Technical University of Munich
Chair of Transportation Systems Engineering

Master's Thesis

Predictive Modelling using Machine Learning Techniques for Railway incident based parameters-
Denmark Railways

Author:
Bhagya Shrithi Grandhi

Matriculation No.:
03686127

Supervised by:
M.Sc. Emmanouil Chaniotakis
Univ.-Prof. Dr. Constantinos Antoniou
Company:
Dr. Felix Laube
Dr. Stephan Thomann

September 11, 2019

Abstract

With the growing concerns for global warming and climate change immense efforts are being taken to reverse the situation. About 15% of the carbon dioxide contribution comes from the transportation sector, making the transportation sector one of the major contributor to global warming. Among the various modes of transport, railways are considered to be the most sustainable among all. But due to its unreliability and lack of punctuality it is increasingly becoming the less preferred mode than the others as they provide more flexibility and control. To make railways more attractive one of the ways is it make them more reliable by making its Traffic Management Systems more robust. Railway Traffic Management Systems(RTMS) ensure the smooth running of the operations. One of the major tasks of the conventional RTMS is to reduce train delays. This becomes a very microscopic goal and leaves out the big picture of the situation on the network. This is the initial motivation to investigate and use ***Total Delay caused by an incident and the Total Duration it lasts***, as the trigger for a TMS to initiate rescheduling rather than delays in trains itself. In this thesis we investigate this notion by trying to create a model for predicting the two mentioned attributes by using the real incident data from the country of Denmark. We explore the possible attributes that could affect total delay and the duration and using these attributes, creation of a predictive model is attempted. Standalone Neural Network models are created for both Total Delay and Duration. Other models XGBOOST, Generalised Linear Models, Linear Models were also created for the comparison with the neural networks and based on the metric results from all the models the best models for both the response variables are chosen. The chosen models are used to create the operational framework. This process could help in changing the way the conventional rescheduling systems work by considering parameters discussed here than the conventional parameters.

Keywords:Total Delay, Duration, Neural Networks, prediction, RTMS

Acknowledgements

This thesis would not have been possible with help from few very important people. Firstly, I would like to thank Prof. Dr.Constantinos Antoniou of the Chair of Transportation Systems Engineering at the Technical University of Munich, accepting the topic on the first hand.

I am especially grateful to Emmanouil Chaniotakis for accepting to supervise me and guide me through out the way. His interesting ideas with the work and the constant motivation helped me to push myself to do the best. When ever I ran into a problem or have the silliest query, he would guide me into the right direction and patiently answer all my questions. He was available whenever needed and it was never difficult to get in contact with him. Without his help I would definitely be floating in the dark.

I would like to thank Dr. Felix Laube and Dr. Stephan Thomann of Emch and Berger AG Bern, for hiring me into the amazing Danish project. Without them it would have been difficult to get the base idea and the required foundation for the thesis. They have constantly supported and mentored me through out the time. I have learnt a lot, professionally and academically, from both of you and would always remember the amazing talks that I had with both of you. Without their constant support during the months the thesis would not have been successfully conducted. It is because of this project I have acquired immense knowledge in railways and their working, which would not have been possible if either of them was not available.

I would also like to thank Michael Rosgaard and Madeleine Jallit for being such good friends and being there for me through my ups and downs all throughout my stay in Copenhagen. I do not think I would have sanely finished the thesis without their support.

Finally, all this would not have been possible with the unconditional support from my family, my father, my mother and my brother who have stood by me in the most lowest points and encouraged me to keep going come what may and reminded me to never lose my motivation. **I would like to dedicate this thesis in loving memory of my father whom I lost during my masters.** All this would not be possible and I would not be here without his undoubted support in the first place. Thank you so much papa you will always be remembered.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

1.1 Introduction

Railways have always been one of most important means of transport for both goods and people. With the increase for transport demand they face a very stiff competition from the other means of transport like the airways and the road ways especially the passenger traffic(Lüthi, 2009). With the increase in affordability of both airways and road-based transport, people have started to prefer them over railways as they have proven to be cheaper, faster (airways) and more flexible (roadways). With the world facing consequences of climate change and global warming railways are being considered as the most sustainable and environmentally friendly means of transport. Governments around the world have started to promote and invest more in the railways. Besides, the competition from other transport modes, the railway industry is facing difficulties with reliability and with meeting customer service requirements (Törnquist, 2006). To strengthen the Railway's position, they must become more competitive, more efficient and deliver safe, fast, reliable, cheap and comfortable services that correspond to customer's expectations (Lüthi, 2009). Many countries have started to de- regularize the state-owned railways, paving way for new organisations to enter the railway sector, making a competitive environment for the stakeholders to create better performance. This makes it increasingly important for the various stakeholders to reduce delays (Nyström, 2008) and improve their performance in meeting the requirements of both passenger and freight customers. The railway service providers should increase their focus scope from not just earnings and profits but also include provision of better services for passengers and freight forwarders with favourable travelling and transporting experiences.

In Europe, the modal split of railway traffic is actually decreasing, from 11% in 2000 to an expected 8% in 2020 concerning freight transport, and from 6% to an expected 5% concerning passenger transport. In comparison, the share of road traffic represents over 40% of freight and over 75% of passengers transport (Corman, 2010). These figures show that the railways are currently lagging behind and losing public trust. In Europe most governments are investing heavily in the railways and searching ways to make them sustainable and attractive mode. Punctuality and reliability of public (rail) transport are vital components of quality of service and passenger satisfaction and the railway organisations are failing to provide it. Delayed trains, missed connections and cancellations are annoying for passengers as they cause unexpected waiting times and add travel time uncertainty into the picture. Even small delays in the trains can lead to missing of connections leading to the accumulation of large waiting times and thus increasing the passenger delay in total.

In the German railway market context, though the 2018 Railway Market Analysis reports from the Bundesnetzagentur(Federal network Agency) report the figures of a stable performance and punctuality since 2015, but it does mention that the numerous cancellations and delays have been reported which are attributed to the train path unavailability, very high capacity utilisation and the construction works (Bundesnetzagentur fr Elektrizitt, 2018). Whereas in the Danish context, even though the network capacity utilisation is not as high as in Germany but the delays are majorly attributed to the signalling system in Denmark. Denmark is in the process of changing the whole signalling system to ERTMS Level2 resulting in delays because of the change and the beyond lifetime old signalling system. The Japanese are world famous for their railway punctuality, they follow the principle of "prevention better than cure". In this context it means, they perform all the maintenance, replacement and all other repair works before even a problem occurs in the network. This ensures that nothing disrupts the trains operations and ensures punctuality. This process is monetarily very expensive and not feasible for every railway network. In most European railways there is a clear distinction between the infrastructure and the operations but in case of Japanese railways there is none and both the parts well integrated giving a timely and better understanding of the working of the railways as a whole(Van de Velde, 2013).

To ensure stable performance and reduce delays, a prior knowledge of the conditions of the network, its operations and preparedness are few important aspects which can be assured by robust Railway Traffic Management Systems which will be discussed in the literature review section. RTMS are huge puzzles which can be solved by bringing together different components which serve different purposes, meanwhile keeping in mind the complete picture. In the conventional systems, upon occurrence of an incident the immediate and prior action of the system is to reduce the delay and to limit the propagation of the delay further into the network. To achieve this usually the tendency of the traffic operators in such situations is to reduce the train delays of impacted trains which becomes a very microscopic goal for the network under incident which on the other hand requires an emphasis on the ways to mitigate the problematic area by proper scheduling, rearranging and managing the operations in the disrupted area which is more of a macroscopic view. The macroscopic view is generally a bit underrated among the research studies in the field of scheduling and planning components of the RTMS. The conventional TMS systems are rigid and the decisions are usually taken by the experienced operators and the dispatchers. This dependency of these systems on the human decisions is an impediment in the efficiency of the TMS in current scenario of the world where every industry is being digitised and automated. With the help of availability of the abundant amounts of data a lot of data analytic techniques are being applied in various industries to make their systems more efficient. Though there has been not much implemented digitisation in railways over the years but a lot of research is being done both with an academic and an organisational background. Digitisation and big data have paved a path to automation and implementation of new ideas, innovations and concepts in the area of RTMS. One such idea as discussed below is the base inspiration for this thesis.

1.2 Project Background

In Europe, most of the passenger railways are typically based on a periodic railway timetable, where train lines are operated with regular intervals throughout a day and consistent transfers are provided at transfer stations between train lines of different type or directions. Typical cycle time is 1 hour for a regional train and could vary anywhere from 5 minutes till 20 minutes or more for commuter railways. This periodic system with large intervals works well in case of diffused origin/destination matrices, train lines are synchronised at stations providing seamless connections(Goverde et al., 2005). These periodic timetable are very easy for the passenger to remember and usually planned more or less conflict free. But the actual problem arises when there is a cancellation or disruption resulting in delays. In such situations the time gap between trains increases resulting in extended passenger waiting time and propagation of delays into the network. This type of problems inspired in researching for a new concept of implementation in TMS which would help in bringing flexibility in timetables. As mentioned above, Denmark is in the process of revamping their complete railway signalling system. This project is under progress and besides the signalling system, they are also building a completely new RTMS. This development of new RTMS has paved a room for experimenting and implementation of new innovative concepts and ideas. One such innovative idea is the use of **Conditions** and Service Intentions which act as the driving force for rescheduling. To define the terms in the context, upon detection of a disruption during operations, Conditions are attributes which encapsulate the type of disruption and its specifics. Conditions store three important information entities regarding the disruption that is the duration of the disruption i.e. how long the disruption might last based on the historical data, the total delay caused by the disruption and the corresponding actions required to be implemented in the section/area of disruption caused. Secondly, Service Intentions are the time window/interval within which a train is supposed to arrive or depart from the station to ensure service expectations on the customer side remain safeguarded. Doing this gives a flexibility in the train schedules which ensure the smooth running of the trains. Service Intentions, not only include specifics of departure and arrivals details but also includes the service and comfort, in simple words, Service Intentions are the encapsulation of customer requirements. Upon occurrence of disruption, based on the new system, the scheduler uses Conditions as the immediate constraint of the deployment of actions on the network to recuperate from the incident and ensure least disruption in the network by choosing the best combination of train runs that could fulfil all or most service intentions. That would result in the least amount of customer dissatisfaction and ensure continuous running of trains so that the passengers and the freight can reach their destinations as close as possible to expectations.

This new mechanism of scheduling is based on how long that particular type of incident is going to last and the total delay caused by the type of problem which becomes the basis for decisions on the scheduling, rearranging and the reassignment of the tracks and the traffic. The scheduler,upon detection of disruption, based on the various attributes available a prediction for the Total Delay by such disruption and its duration is done. These predicted values are immediately applied on the network and the disrupted network area is blocked/avoided from any operations for the predicted Duration time. The predicted Total Delay can immediately give an idea of impact of the disruption that can be caused on the operations and network, which can helpful in reducing the knock on delays. This is an iterative process and the prediction can be done frequently as an when the there is more information feeded into the scheduler. Based on the predicted values the corresponding service intentions are applied to the network arrangement/management of the railway operations at the particular network area are implemented. This way of scheduling ensures the macroscopic view(mentioned above) is not neglected, parallelly ensuring the microscopic elements of the network are running fine. Reducing delays completely in highly impossible in the context of railway operations because of its complexity. Instead proper traffic management by means of rearranging and flexible timetables based on Total Delay caused by disruption and its duration could ensure continuous running of services which eventually becomes more important than delays.

1.3 Scope, Objectives and Research Questions

1.3.1 Scope

This new concept of using Conditions as the trigger for railway traffic management upon occurrence of a incident is the basis for thesis. This concept has many directions of research like in railway operation and planning context, or actual creation of a scheduler based on this idea and many more. But because of personal interest and for the creation of proper foundation for this concept, this paper discusses the application of big data techniques on the data required for creation of these Conditions. In conventional systems usually the traffic operators use their expertise to estimate Total Delay caused by a incident and Duration of incident and this is a one-time prediction and generally not updated upon further availability of information. It does not question the ability of a human being's expertise, but the railway operation systems are so complex that humans decision making ability can penetrate a layer or two in the process of decision making process. Data analytics on the available information can help traffic operators in making much better-informed decisions as it helps bring out the hidden information and relationships. This thesis aims at using data analytics techniques on the available data and attempts at predicting total delay caused by an incident. In the literature many researches have been found which predict the delays but most of them concentrate on predicting the delay in a particular line/area or on to a particular train. Some studies in prediction of delays are done on the basis of type of the problem or cause of the problem, and the prediction is carried using single predictor type like weather or location etc but not both. In short, the researches in prediction of delays usually concentrate on a very small aspect of the problem rather than considering the whole. Also, no literature was found on the total delay prediction by an incident. To address this gap in the literature the thesis attempts at bridging it by creating a predictive model for Total

Delays which considers different attributes that can contribute to the prediction. These attributes are selected on the basis of literature, place from which the base data has been collected and basic knowledge from the field. In the railway jargon different words have been used for different types of incidents on basis of how long they last or their cause of the problem etc. But for the purpose of clarity and to avoid any confusion the word "incidents" is being used for all kind of problems in the railways network addressed in this thesis. For this thesis only passenger traffic related disruptions are considered for analysis. Also, for the thesis R programming language is used extensively for modelling and Python is mostly used for data collection and network analysis.

1.3.2 Objectives

Based on the research gap mentioned above, this thesis aims at bridging this gap by creating predictive models for Total Delays and Duration of Incidents. Through this thesis, an understanding on what are the different types of predictor variables are to be considered in the prediction of the mentioned response variables, how much these predictors affect the response variables? In this thesis due to data and time limitations, the considered variables are based on the literature review, expert opinion and data availability. Secondly, an attempt is made for finding the best model fit using the predictors by trying different combinations of predictors, hyper-parameters and different machine learning techniques. After finding the best model for both the response variables(total delay and the duration of the incident) an attempt is made at making a framework using two models where the predicted duration for an incident is passed to the total delay prediction model and then the predicted values and compared to the original values and then the accuracy of the framework is computed.

1.3.3 Research Questions

Based on the scope and the objectives discussed above, the following research questions are formed which this thesis eventually attempts at answering.

1. How Total Delay and Duration of incidents can be predicted?
2. What are the attributes required for this task?
3. Can neural network be the best model in predicting total delays and duration?

Based on these primary questions the secondary questions are made which would address much detailed aspects of the study.

- Which machine learning technique works the best in this context?
- Are the used data and attributes sufficient for the predictive model or more features are to be considered?
- How much does each attribute effect the response variable?

1.4 Thesis Framework

This report is divided into 7 chapters, the first chapter is the Introduction to the thesis where the topic is introduced and gives brief outline on how the topic was chosen, the need for this research based on which research questions were formed. In the second chapter detailed literature review is provided addressing the various researches in TMS, rescheduling and predictive modelling. Also, addresses the gap in this field. In the third chapter the data collection and the considered variables are explained. In the fourth chapter a descriptive analysis is carried out on the complete data set and all the latent relationships are addressed. In the fifth chapter all the model algorithm used for building different models are explained based on their implementation. The parameters used for each models and the process chosen for the model building is explained . In the sixth chapter all the results from the built models are addressed and discussed. The last chapter is Conclusions where the final learning's are presented. Also, future work and the limitations are also discussed here. Figure 1.1 shows the complete process followed for the thesis.

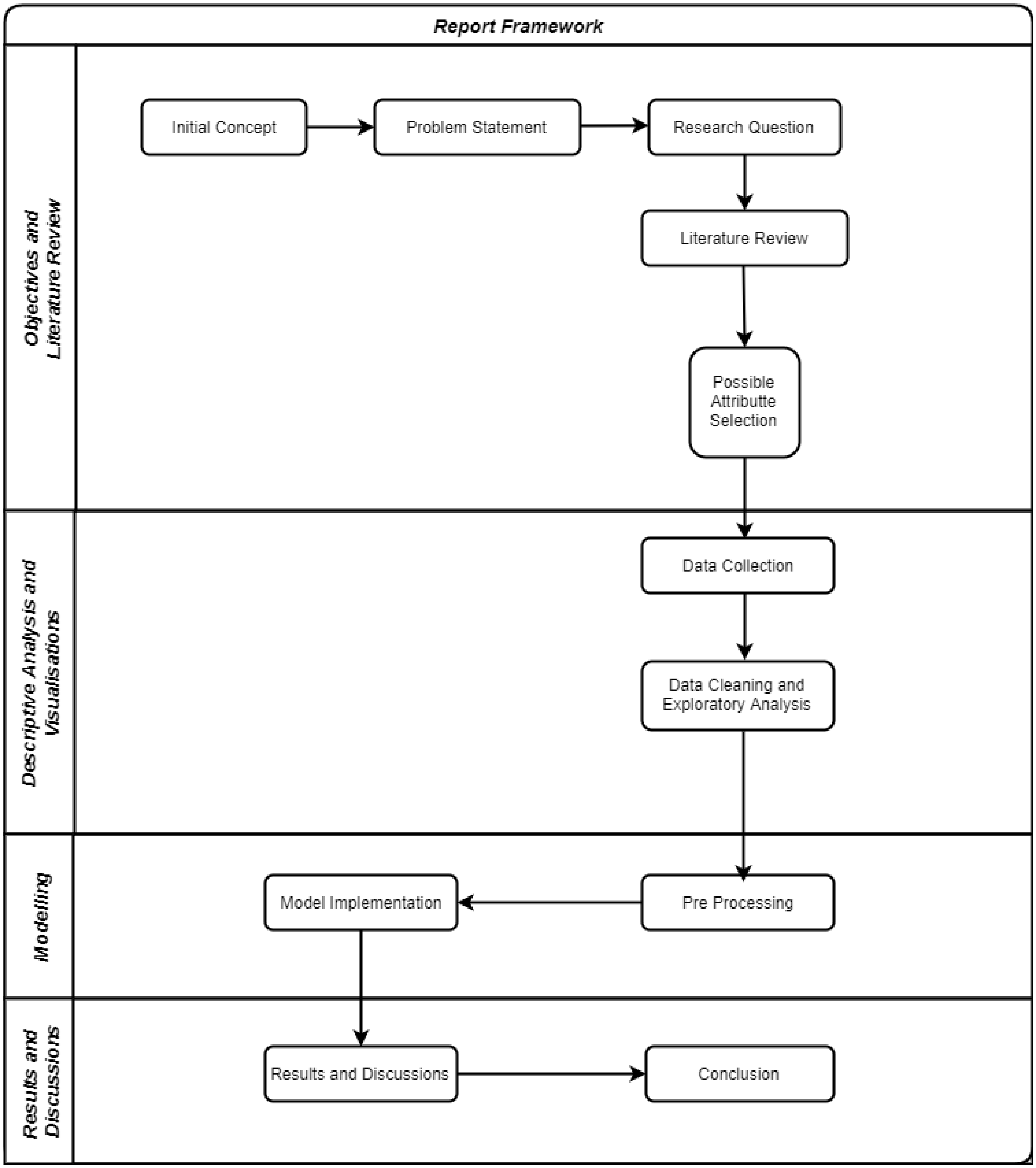


Figure 1.1: Diagnostics Plot Linear Model- Total Delay

Chapter 2

Literature Review

In this section, an overview of state-of-the-art researches in Railway Traffic Management is discussed to understand the workings of the system and the requirements for the operations. The available types of traffic management concepts, their setup, the system requirements and the important variables considered for such research are discussed. The different frameworks and tools created based on the these RTMS types are mentioned. Also, the different components of TMS, and the research model dealing with them are looked into. Since all these mechanism are for reducing poor service quality and improving the railway operations to eventually provide better services for the public, the factors affecting the operations are studied based on which the attributes for the predictive model are selected.Further more, the researches in the field of railway delay and duration prediction is discussed and specifics form these studies are considered for this thesis. Further more, the different machine learning algorithms and their implementation is discussed in the current context. This overview gives an essence of the state of the art in the field. Additionally, there is a peek into the usage of machine learning in this field, discussing which type of models and concepts are being used extensively in which areas.

2.1 Railway Traffic Management Systems

Rail Traffic Management Systems (TMS) are systems designed to support managing the complexity of rail services and rail networks by providing an integrated and holistic view of operational performance. They enable the rail operational staff in making better informed decisions to better balance the competing demands in aspects such as track resource availability, train crew resources etc (Davey, 2012). TMS in general could be said to be the integration of different components intended for different functions. Railway Systems involve various types of functions and a clear distinction is required among them for the purpose of better understanding the tasks involved. Among, the various categorical distinctions found in literature the most simple and generic functional structure was the three level pyramid mentioned in (Lüthi, 2009) which divides the structure into three layers namely Strategic, Tactical and Operational as shown in the figure. These layer's functional purposes vary

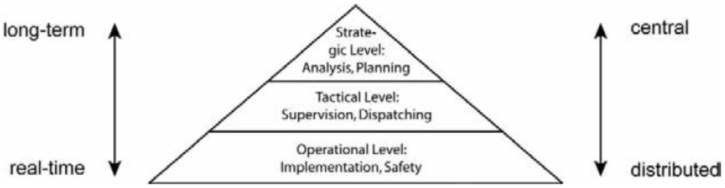


Figure 2.1: Functional Structure for railway operations and planning

with different TMS concepts. Interconnections between the layers also vary from each type of TMS. Few have strictly separated layers and few have crossovers. In (Lüthi, 2009), the Strategic level is associated with Analysis and Planning, the Tactical Level with Supervision and Dispatching, and the Operational Level with Implementation and Safety. In Europe, most of the Railway operating organisations keep the three layers separated from each other with varying amount of crossovers where as in contrast Japanese railways have more integrated system(Van de Velde, 2013). Nowadays, the function of dispatching and operations are strictly separated in the systems but it is foreseen that merging both the layers can result to a much efficient and improved performance of rail operations (Van de Velde, 2013). Whereas Davey in (Davey, 2012)defines the levels differently, namely a Traffic Management Layer which deals with more operational and information aspects of running trains having indirect control of them and Signalling. The Control Layer, as the name suggests has direct control on the trains by maintaining route setting and train operations and the Safety Layer handles the safety aspect of the operations. It can be seen that functionalities remain same but the names and distinctions differ based on the conceptual requirement of the TMS. All these functionalities can be termed under long term tasks and short term/real time tasks as each time differs on it implementation time. It is the real time and the short term tasks that require mostly constant data support.

Basic Definitions

The following are the basic terms generally used in the conventional TMS systems.

- **Block Section**: The tracks are divided into set of blocks to avoid collision of trains. Only one train is allowed to occupy one block at one time. It can also be said as the track occupancy detection section.
- **Operation** :The passage of a train through a block section.
- **Route**: A sequence of operations to be traversed by a train.
- **Running time**: Travelling time required by each operation.
- **Dwell time**The scheduled stopping time of each train at a platform
- **Conflict**: Two or more trains claim the same available block section simultaneously.
- **Deadlock**: A set of trains, where each train claims a block section ahead even though it is not available.

- **Dispatching**: The process of controlling and facilitating train movements.
- **Rescheduling**: The process of changing the railway timetable/plan upon occurrence of an incident.

Dispatching and the rescheduling can be said as the result of disruptions and irregularities in the railway systems.

2.1.1 Overview of RTMS

To give a basic overview of various categories of tasks and their classification following two mind maps have been used. These mind maps show the classification of tasks in the old and new systems, depicting the differences in both the systems. In the Figure 2.2 a mind map for old RTMS is shown with a basic classification of tasks within TMS. It can be seen that the three major tasks of TMS classified as the **Planning**, **Dispatching** and **Asset Management**. In the branch of **Planning** the major task is the collection of all Train Operating Companies(TOC's) requests and then pooling of all the resources required for the requests. After which keeping my mind any future planned activities on the network, the requests are assigned with resources and then these requests are set into the timetable and eventually a complete time table is created.

Simultaneously, on the other hand **Dispatching** takes place which is ensures the running of the trains in the network. Dispatching, according to old systems is majorly based on the operation rules. These rules define the process and ensure safety on the tracks which can be seen in the further branching. One can say that Instruction flows are defined based on the operation rules. Instruction flows is the instructions to be followed once in operation which requires decision making. A major part of these instructions is Signalling as in the old systems signals are the only way driver knows what is to be done next when he/she do not know the current network information.

Asset Management being the third branch takes care of the infrastructure and is basically maintaining it. It consists of all the planned, required and spontaneous maintenance activities on the tracks based on which the rescheduling takes place. In case a maintenance activity is planned well ahead of time then such activities are considered while timetable creation and operations are planned around them. Whereas, in case of an incident on tracks which blocks the network or requires an immediate maintenance activity, triggers the dispatching of activities around the maintenance activity area.

In the case of the new TMS the classification differs as shown in the Figure 2.3 it is branched into 4 tasks. The task mentioned in the Figure 2.2 are rearranged and have grouped with more tasks in the new TMS. The dispatching part in the old TMS is split to **Booking and Negotiations** and **Task Execution**. Signalling became a branch in itself as it is more prioritised and separated in the new system. Asset management mentioned in the old system is replaced by **Conditions Registry** which serves as interface to a self contained asset management. It communicates what asset management has as effect of resources, that is it keeps of all the restrictions related to resources and safety. This can be called as the

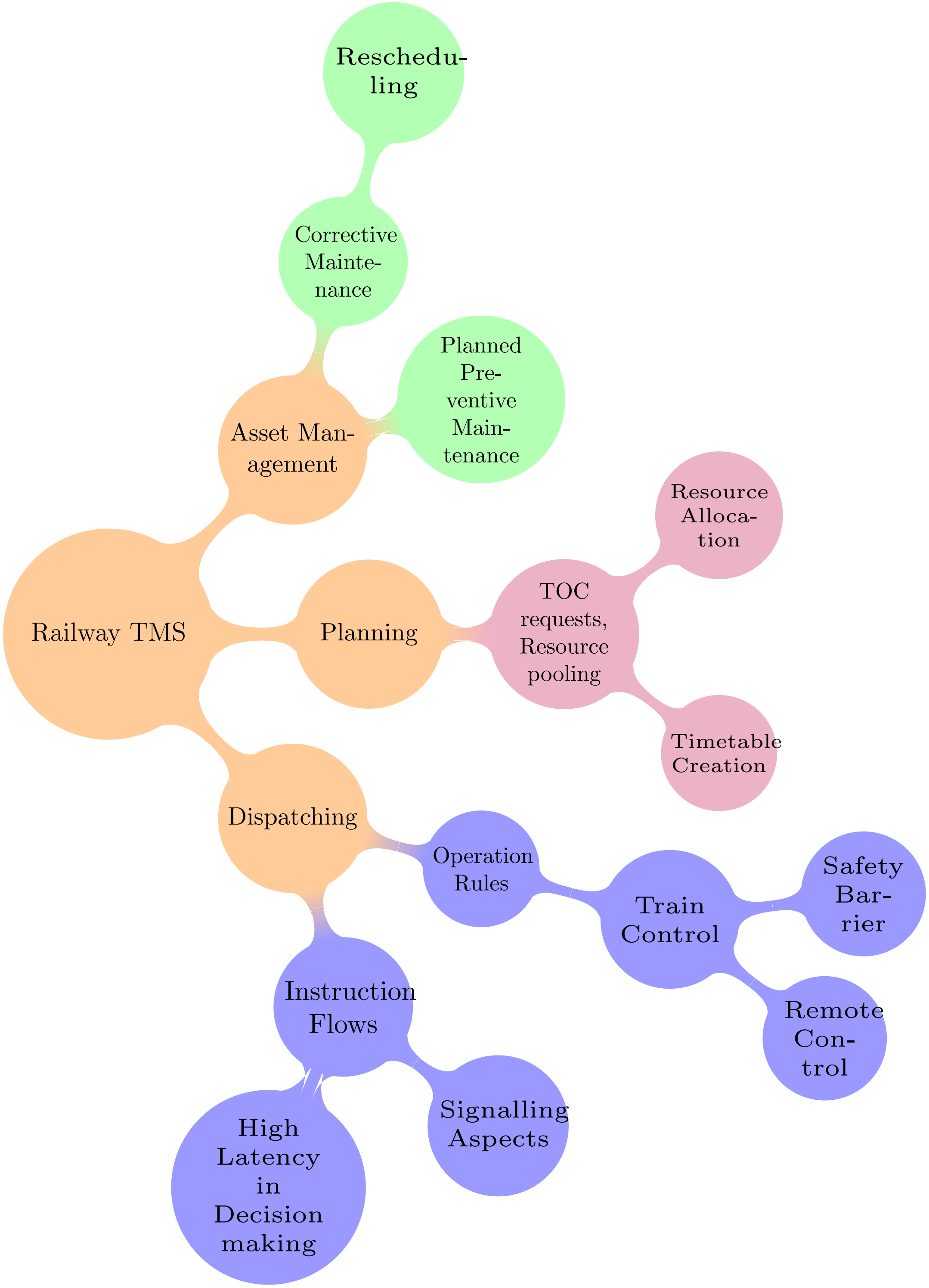


Figure 2.2: Mind map for TMS system

deconstructed operation rules, which is very rigid and strict. It also consists of all possible incidents/failures and the corresponding measures to be taken and applied upon occurrence.

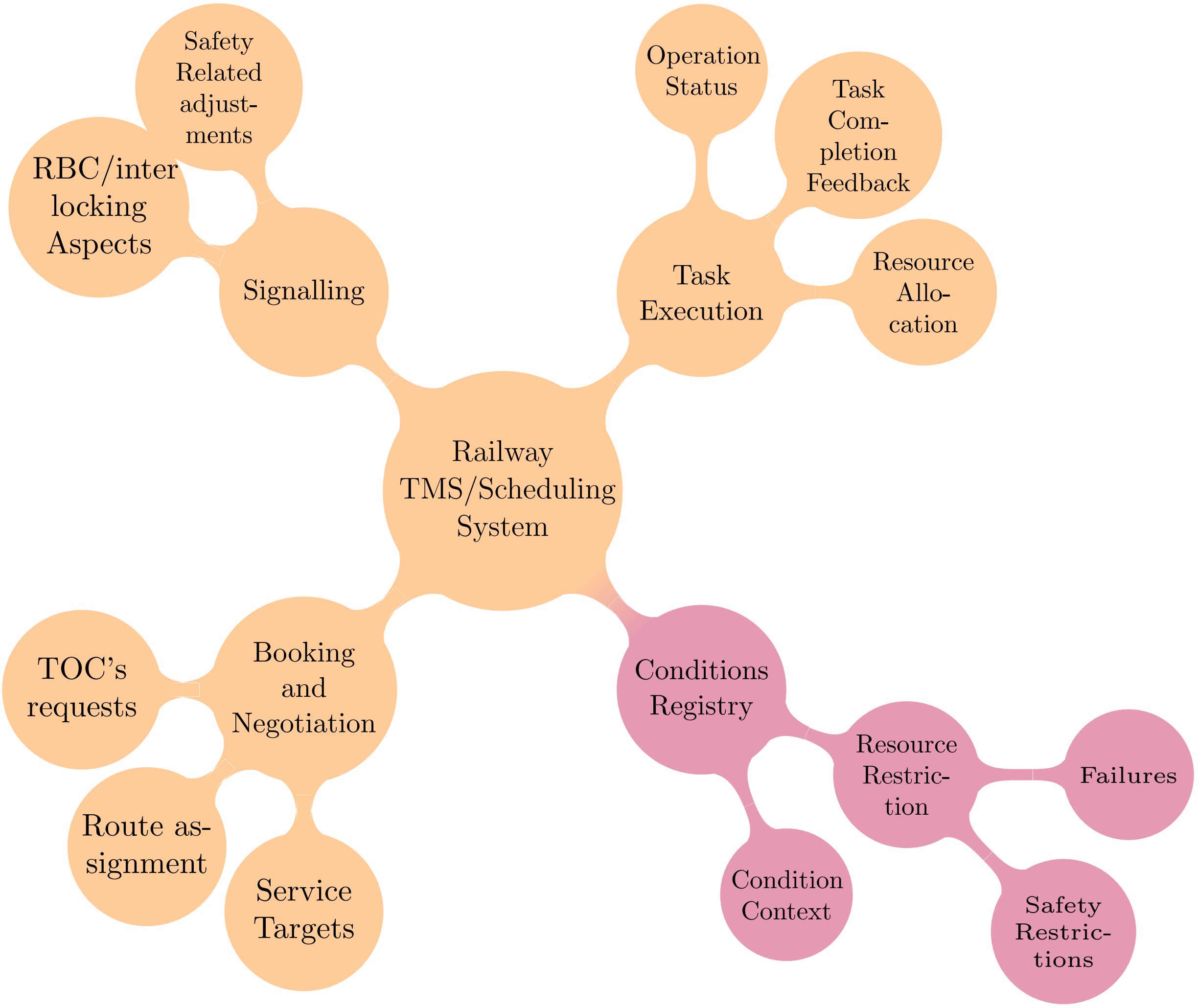


Figure 2.3: Mindmap for new TMS system

Booking and Negotiation takes in all the requests from the TOC's and other stakeholders. Then allocates the service targets to each of the request. Based on which the route assignment for the trains is done. **Signalling** which was earlier under dispatching, coming the third layer, is now a branch in itself which increases the importance of signalling. This includes the RBC and interlocking controls and adjustments required to ensure safe movement of trains.

Task Execution is the branch that could be said the most dynamic and real time branch. It basically has the current status of the operations with each and every stakeholder of the operation(drivers, signal operators, picops and others) giving constant updates on the task at hand based on which decisions regarding resource allocation could be made in real time.

This explanation basically gives an outline picture of the complete system. The **Conditions Registry is the focus area for this thesis**marking it with a different colour alone. In recent years various types of TMSs have been researched and through the literature few of the significantly important systems have been discussed here.One the most discussed is the **Real Time** and the **Dynamic System** and there has been a major research in this area. The difference between the two type is that the real time systems are adjusting the operations and the timetables on the spot in case of delays in the system. Whereas the dynamic systems improve the system by providing flexibility to time tables (D'Ariano, 2009). Various models have been created over the years and it has been observed that dynamic dispatching has always been a challenging task (Harrod, 2012). Various studies discuss the idea of implementing real time systems in the context of Railway Dynamic Traffic Management(RDTM). RDTM is discussed in the literature review because, the study carried out for this thesis could be used in the context of RDTM.So having an understanding of such systems could be really useful.

2.1.2 Railway Dynamic Traffic Management(RDTM)

RDTM system dig into the possibility of having flexible time tables for the operations breaking with the rigid railway systems.(D'Ariano, 2008) gives a brief overview of the concept of RDTM and different studies researching it. The basic idea behind it is to keep the train traffic flowing in the bottlenecks by avoiding unnecessary waiting times and this could be achieved by relaxing some of the timetable specifications like arrival/departure, sequencing and train routing. (Schaafsma, 2001) proposes RDTM model and based on this concept further studies were conducted from Technical University of Delft which could not be found on the Internet even after extensive search but have been mentioned in (D'Ariano, 2008). They describe RDTM process to be based in the following points.

1. Strict arrival and departure times are replaced by time windows of minimum and maximum arrival and departure times at each platform and relevant timetable points of the network. This band width accommodates flexibility. In this case the railway managers and dispatchers have a min/max arrival and departures timetable where as the passengers timetables have a maximum arrival time and minimum departure time. Longer travel times could be compensated for improved reliability of trains services as in travel times and connections.

2. The defining of scheduled orders of the trains at overtaking sites or junctions, provisionally or partially in operational timetable and finally determined in the real time. In case of any conflicts, they could be solved in real time by changing the order allowing the reduction of delay propagation.

3. Provision of a set of feasible platform/passing tracks instead of a default set leaving the decision to traffic controller.

Instead of defining every piece of information in the offline timetable, leaving some information to be defined based on the real time data gives better control on the traffic. With the increase of degrees of freedom, more is left for the real time traffic control systems to make actions which requires computerised TMS to support the dispatchers to make informed decisions. In (D'Ariano et al., 2007) using the RDTM procedures and under different degrees of freedom two types of dispatching systems are compared. One which is based on the ARI(Automatische Rijweg Instelling) system used in the Netherlands. Where as the other is a more advanced real time traffic management system based on the ROMA tool which has been already mentioned above. In (D'Ariano, 2009) implements the ROMA application in context of RDTM in 2 complicated and densely used areas.

ROMA(Railway traffic Optimisation by Means of Alternative graphs), to automatically recover disturbances. ROMA is able to automatically control traffic, evaluating the detailed effects of train reordering (D'Ariano et al., 2007) and local rerouting (D'Ariano, 2008) actions, while taking into account minimum distance head ways between consecutive trains and the corresponding variability of train dynamics. It is based on the blocking time theory for recognition of conflicts in case of disturbances and a general discrete optimisation model based on alternative graph theory the evaluation of the train re timing, reordering and rerouting of trains. In (D'Ariano, 2009) ROMA was extended to short term traffic prediction in dispatching are which was under strong disturbance through which an attempt on evaluation of affects of the rescheduling for a certain time period. This tool has been integrated in various other studies too. To reduce the the computational time an improvement was implemented where in from a pool pf different instances in a dispatching area, using a tabu search for smaller instances (Corman et al., 2010). This software has been extensively used in the creation of Dynamic Traffic management platforms.

2.1.3 Real Time Traffic Management and Rescheduling

Many Real Time Traffic Management frameworks have also been created over the years and few have them been mentioned below. It can be seen that Machine Learning has been used extensively. (Mazzarello and Ottaviani, 2007) mentions about a new real time TMS system called as COMBINE which an improvised version called as COMBINE 2. This system addresses the problem of real time traffic regulation and optimisation in railway networks equipped with different signalling systems. This paper concentrates on core modules dealing with local traffic optimisation and control, which were named as the **Conflict Detection and Resolution(CDR)** which was developed based on the idea of alternative graphs and Speed Profile Generator(SPG) which was based on prediction of the position and speed of trains using dynamic programming technique. Alternative graph method is also used in (Sam, 2018) from which a formulation of the Mixed Integer Linear Programming is used to model decisions. The alternative graph method has been implemented in both macroscopic and microscopic simulations. These frameworks have just reached the pilot tests but never tested in the real time scenarios. (Quaglietta et al., 2016) also discusses a real time traffic management based concept of framework which works on the basis of the **closed control loop concept**. The idea is to forecast a traffic plan tackling any possible conflicts based on the network and the current traffic scenario on the network. A simulation platform namely HERMES was used for simulating perturbed scenarios in different railway networks across Europe in combination with the re-planning tools ROMA and RECIFE. The proposed framework integrates algorithms for traffic state prediction and monitoring, conflict detection and resolution and automatic route setting. (Lüthi, 2009) discusses an integrated real time planning system using feedback control loops. This framework divides the system into two parts the inner and the outer feedback control loops. The inner loop takes care of the train driving and infrastructure and the outer loop supervises the re-planning, train traffic and infrastructure state. A new concept of production plans is introduced which is a detail description of the task and its start and end times assigned to each resource involved in the production.The new production plan is developed based on optimisation objectives in case of deviations Prediction of the future state based on the reasons of primary delay is an important aspect of this system. Integrated systems ensures precise running of trains and improve the performance of the system. A working example of this framework has been simulated for a test-bed in Switzerland where only the inner loop is explicitly modelled while a manual procedure is used in place of optimal re-planning algorithms in the outer loop(Quaglietta et al., 2016).

There are many other models used for solving the rescheduling and the re planning problem. One of the popular models is based on the Heuristic measures which could found in (Gholami and Törnquist, 2018) and (Carey, 1999). The former uses a combination of heuristic measures with a Mixed integer linear programming and the later concentrates more on the knock-on-delays and rely majorly on the probability of the delays.

These TMS frameworks are the big picture but to make informed decisions these systems require relevant accurate information. This is where predictive models come in the picture. Predictive models could be used for many other purposes but the most relevant for this thesis are the prediction of delays and duration's. Researches related to them have been discussed below.

2.2 Researches on Predictive Models

In current systems enormous amounts of data is generated from the operations, however, in most organisations this data remains unused as the data amount is gigantic and doing analysis requires a lot of effort. With the advent of big data analytics in every possible industry, they are trying to improve their performance using these data. Even in railways there is definitely logging of the delay and incident data in the systems(if there is no such system in place then one should be created for collecting valuable information). Using these disruption, deviations and delays data one could gather very valuable information from these systems. Recently,many researchers have started working towards applying big data analytics on the railway data. There are many studies that have used Data Analytics techniques for complex railway planning problems Machine learning techniques help solve complex problems which are difficult for a human to solve. Machine learning is a data analytics technique where the machine learns from earlier experiences. Use of these Machine learning(ML) techniques in delay prediction is extensive. Identifying the delays, their causes and inter dependencies have been a very important task required for better performing rescheduling system. To make the rescheduling problem much easier another area of research was found in the prediction of the delays and duration of Incidents. An overview of the most relevant studies has been provided in this section with most important takeaways from the papers are discussed.

2.2.1 Delay Prediction Models

According to the Machine Learning technique categories, this thesis deals with a supervised learning problems where total delay caused by an incident is calculated. (Robert and Kim, 2018) also discusses a similar supervised learning problem. The study predicts the train delays using different machine learning methods like **neural networks, decision tree with and without Ada Boost**. One major variable used in the study was **weather**.The setup here is in the Swedish railways The weather data and the scheduled departure data are combined for training the models. The accuracy metric used for comparison was the **average error, i.e. the difference between predicted and actual delay**. Among the different considered models, Neural Network give out the best result with the least average delay. In another similar study discussed in (Wang and Zhang, 2019), they also use **weather** as one of the important predictor for predicting train delays at each station. This paper also uses ML technique of Gradient Boosting Regression Trees Model for prediction of train delays. The results were not discussed in detail and the access was denied but it does mention that this model provides proof that these models can help predict delays accurately.

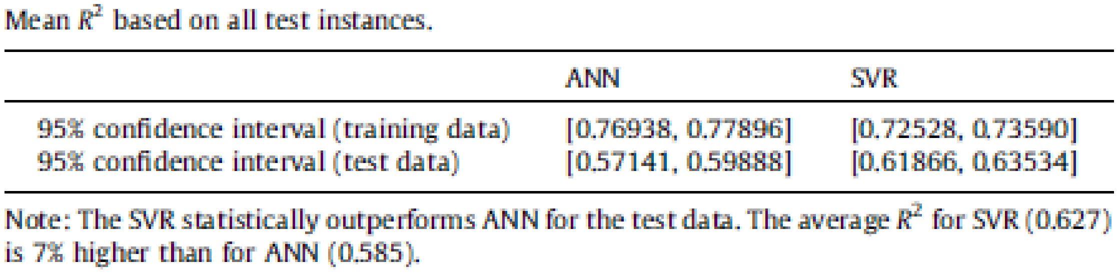
Another, study that also uses neural networks for predicting Railway passenger train delay (Yaghini et al., 2013) in the setup for Iranian railways. This paper is discussed a bit in detail here as the it goes into detail about the neural networks. It proposes a neural network which is then compared with a decision tree and Multinomial Logistic Regression. Different type of model architectures namely **Quick, Dynamic and multiple** with different sets of input variables(normalised real number, binary coding and binary set encoding inputs for the categorical variables) were used for variation of the inputs for the model to find the best model fit. The model is based on the Feed Forward MultiLayer Preceptron algorithm. The dynamic and the multiple methods piques the interest for this thesis, in dynamic method, topology of the network changes during training, with units added to improve performance until the network achieves a desired accuracy. There are two stages to dynamic training: finding the topology and training the final network. Whereas, in multiple method, multiple networks are trained parallelly, and the network with highest accuracy is considered as the final model. Firstly, a several single layer networks were created on the basis of the given neurons and sequence of neurons to be tested. Next, for each single layered network, a set of two layer network is created where in the neurons are again changed in a sequential manner for the second layer and a corresponding network is created. In this study to simplify the original data, a concept of **Data Discretisation** is used to reduce the number of values for a given continuous attribute by converting them into intervals. These intervals are then used to replace the actual values. Using this technique, the data is divided into 10 different intervals and accordingly each intervals gives a one output unit for the neural network which converts the problem into a classification problem. For comparison of the models, the metrics used are accuracy and training time. Nothing more is mentioned about the accuracy, as in the type of metric or how the accuracy is calculated. The values of the accuracy are in percentages and that of training times are in seconds. Also, aggregation and dis aggregation methods were used for improving results. From the results it was shown that the dis aggregated neural network model improved the results with different types of variations mentioned above gave a great accuracy ranging from 90% to 93% with most of the best results coming from a

multiple structure method.

In a slightly different approach of modelling is discussed in (Corman and Kecman, 2018) which introduces a Stochastic train delay prediction model which is based on Bayesian networks. The model basically predicts the probability distribution P of random variable describing an event. This model supports the real time updation of the probability distributions real time. This method represents the complex stochastic inference between random variables. The time dependent random variable is described based on the train delay over time and space represented stochastically. Also, it is further extended to model interdependence between trains sharing same infrastructure or the have passenger transfers. The data set is from a busy corridor in Sweden. Results use different metrics for selecting the predictors and for the models. For the selecting the predictors, a basic linear regression is carried out on the response and the explanatory variables where and based on the R^2values form the these linear models the correlations between the response and the predictor is studied. Whereas **Mean Absolute Error (MAE)**is used to define the performance of the model and for comparison between different models.

Another Bayesian network(BN) model for predicting delays is presented in (Lessan et al., 2019). In this model the three different BN schemes namely, heuristic hill-climbing, primitive linear and hybrid structure, are investigated using a data set from a high speed line. The data rationalised the dependency graph of the developed structures and to avoid over fitting the structure is trained with k-fold cross validation. Eventually a comparison is carried out between the hybrid BN model and the real world bench mark data. This study suggests that BN models can be an efficient tool for capturing interaction effect of train delays. The most interesting takeaway from this paper was the performance evaluation the first basis was the comparison between the predicted and the actual values and then using scatter plots and distributions plots at each station to understand the how good the predictions are in comparison with the original. Next was to quantify the deviation of the prediction from the observed, for which the metrics RMSE(Root Mean Square Error), MAE and ME(mean error) were used. The closer the values to 0 the better are the predictions. For this study the RMSE for predictions was less than 2 minutes which was considered as large ans suggested the existence of outliers. It was found that the prediction match was only 56% because of which the discretisation of the data was considered to improve the results which was discussed also in the paper (Yaghini et al., 2013). After converting the dataset into intervals, the metrics used for evaluation were **Accuracy, Sensitivity, Specificity and Kappa statistic**. An accuracy of 80% is achieved after discretisation. It mentions **True Positive Rate and Recall(Sensitivity)** that measures the proportion of cases that are correctly identified(60%) and **True Negative Rate(Specificity)** represents how effectively the models avoid wrong predictions. **Kappa Statistic**is also used which uses the confusion matrix. It shows the overall agreement between the observed accuracy against the expected accuracy and considered to be less misleading than accuracy as metric. Kappa value of higher than 30% are considered as acceptable.

Another very impressive study with many important takeaways was (Markovic et al., 2015). This study does not predict delays directly but analyses the train delays by capturing relations between train arrival delays and various characteristics of a railway system. Train arrival delays cause due to irregularities in internal sources with in the system. The is study was done from railway data set of the Serbian Railways. The data set includes passenger trains travelling through the Rakovica Station in Belgrade. They build a model based on the Support Vector Regression(SVR) for train delay analysis and then it is compared with an Artificial Neural Network(ANN) model. Among the various variables considered for the study, one of them that was found to be relevant to this thesis was **Headway**. This lead to the consideration of headway as an attribute for the study. The ANN is a generic Feed Forward model trained on the back propagation algorithm with single layer of 50 hidden neurons, using the tangential hyperbolic transfer function. A statistical comparison is conducted to analyse performance of the models using R^2(coefficient of determination) as the metric for comparison. From the Figure 2.4 it is evident that ANN outperformed the SVR model on the training data but in case of the test data the results are otherwise. Also, it can also be seen that there is a big difference between the train and the test R^2 indicating that there is chance of probable over fitting to noise in the case of ANN. To test the difference in mean R^2 a **two way Analysis of Variance(ANOVA)** test is conducted by providing a 95% confidence interval(CI) for mean R^2. This CI reiterates conclusions from ANOVA and allows to rank the models. On the principle of over fitting the SVR model is considered better performing than ANN.



Mean R^2 based on all test instances.

	ANN	SVR
95% confidence interval (training data)	[0.76938, 0.77896]	[0.72528, 0.73590]
95% confidence interval (test data)	[0.57141, 0.59888]	[0.61866, 0.63534]

Note: The SVR statistically outperforms ANN for the test data. The average R^2 for SVR (0.627) is 7% higher than for ANN (0.585).

Figure 2.4: R^2 values for SVR and ANN models for the train and the test models

From the extensive search for literature many researches use ANN for train delays as see in the above mentioned researches and few more (Malavasi and Ricci, 2001), (Peters et al., 2005), (Pongnumkul et al., 2014), (Yaghini et al., 2013), (Markovic et al., 2015), (Oneto et al., 2017) have generally outperformed than the multiple regression models which can found in these researches (Seriani et al., 2016), (Goverde et al., 2005), (Gorman, 2009), (Flier et al., 2009). This leads to implementation of Artificial Neural Network as the main model for this thesis.

2.2.2 Duration Prediction Models

Whilst the section above discusses about the research in the field of delay prediction, on the contrary, this section discusses about the papers in the area of Duration Prediction. Since the studies in the context of railways are way fewer, studies from other fields are also discussed for the better understanding the scenario. In (Zilko et al., 2016) discuses about the lengths of the railway disruptions. In general shorter incidents are defined as a disturbance, which could be overcome by a simple timetable adjustment while the longer incidents are referred as the disruptions which may require more adjustments than just timetable like the crew and rolling stock etc. But for analysis purposes they consider both of them as disruptions and attempt to determine the length of a disruption. In practice, they found a three step prediction process employed where in P1 is the first prediction of disruption length based on the average of the past similar incidents. P2 is the mechanics prediction which is based on their diagnosis of the situation. Any update is after P2 is noted and the final prediction informed to the control center is called as the P3 based on which they know when to resume the operations.To tackle this uncertainty in prediction of disruption length they created a dependence model between the disruption and the influencing factors. Akaike Information Criterion (AIC) is used as the basis for comparing different structures of models which was introduced in (Akaike, 1974) which is defined as

$$AIC = 2k - 2ln(L)$$

where k represents the number of parameters and L represents the likelihood. Various other algorithms have been used to calculate the goodness of fit for the models. Using Copula Bayesian method they created a model which upon occurrence of adisruption, realisation of the influencing factors conditionalise the model and results in a disruption length distribution that is specific to a specific situation, through which the disruption length is deduced. Other studies have been found which are also based on Bayesian modelling (Lessan et al., 2019)which are predicting train delays. Other Bayesian concept based models for prediction of incident duration can be found in (Boyles et al., 2007),(Pettet et al., 2018) , (Li and Cheng, 2010) and many other such researches.

Not much researches was found that implemented ANN for duration prediction, on the contrary little regression based research was found in predicting railway incident duration. One of them is part of an actual implemented TMS which is employed by the organisation SNCF(Société Nationale des Chemins de Fer francais) which is Frances national state owned railway company. Their TMS system **EXCALIBUR** provides historical incident database which includes all the details about the incidents. The variables related to the incidents are classified into three types: Static, Dynamic characteristics and consequence information where static gives information about the time, location, resource and type of incident etc., dynamic characteristics give information about consequence on traffic, type of operations, source of in formations etc., consequence information gives total perturbation and total incident duration of the incident(Chandesris, 2006). Based on the data distinction two models were created a static and a dynamic model. The former model is the calculation of the mean duration of the past incidents and the prediction is improved based on the variables using a generalised linear model.The later model was based on the regression tree model and was a much more elaborate model with both static and dynamic characteristics as input.This is one of the models in working with real time traffic. No details about the metrics to measure the model accuracy were mentioned.

In regression models one of the most famous models is Partial Least Squares Regression(PLSR) and (Wang et al., 2013) uses this method for prediction of incident duration but in the context of Roadways. The paper classifies incidents into different types and then eventually compares them with a model that considers all the models. The performance evaluation for the models was done based on comparing the **MAE(Mean Absolute Error)**. The prediction accuracy ranges from 71% to 88% for different models and eventually states that with that PLSR can be used for incident duration prediction.

In a completely different field of implementation that is for incident duration prediction is the roadways(Pereira et al., 2013). It piques the interest for this study because of the use of text for gathering information can be very helpful in railways too as there are times when very important information cannot be entered in a numeric or a model readable format. Most of the data captured in transport system has text fields than a constrained value fields which contains invaluable information related to the incidents. For this purpose they introduce a topic modelling technique as a tool for extracting the information and develop a machine learning model for that integrates the textual and the non textual features. Latent Dirichlet allocation (LDA) is a generalisation of pLSI developed by (M. Blei et al., 2003), that allows documents to be mixtures of topics. This approach could be really helpful in applying to current railway systems where most of the data entries are manually done and have textual entries. Various models like SVR, ANN, linear regression, Decision trees and Radial Basic functions models were used. For assessment of performance **Correlation Coefficient** and Mean Average Error (MAE) are used. This study is worth noting because of its unique implementation idea. The result showed that modelling with textual feature outperformed the other models by decreasing the error up to 28%. There is a lot of research in this area of interest, even if it is in the context of road ways. But the basic concept can always be extended and applied to other fields and areas. (Li et al., 2018) provides a complete overview of the researches and, the models, algorithms, methods and the concepts implemented in them. The data sets and the data sources are mentioned. An overview of performance evaluation metrics, model combinations and structures, outlier prediction, combining different data sources and model improving techniques are discussed. This (Li et al., 2018) study can be used to give an initial overview and idea on how to start with research in this field.

2.3 Types of Delays and Cause Relationships

To carry out efficient rescheduling of trains in case of a disturbance the understanding of the reason and causes for a disturbance can be vital information for rescheduling, contain consequences and enhance the service performance of the system. Having this information can give better support for the dispatcher to make more informed decisions and the consequent reactions could be efficiently handled. Rescheduling is the next step when a disturbance occurs and it is the task of the scheduler system and dispatcher to ensure the new schedule doesn't have any new conflicts or carry forward delays from before. Railway companies are extensively working and researching ways to reduce perturbations and detect them early and deploy preventive measures. To understand and act towards these delays a deeper understanding is required on the correlations and causalities leading up to the delays. Though this area is not the focus for this thesis it indeed provides a good grasp of the system and it complements this thesis. Also, Machine Learning has been extensively used in this area of focus too.

Incidents, disruptions and deviations cause delays and sometimes cancellation of the services in the railways, making them an unreliable option for passengers. Disruptions effect the operations adversely causing delays which according to the experts have been distinguished into two types names the **primary delays** and the **secondary delays** also called as the **knock-on-delays** (Kono et al., 2016). A knock-on delay which are usually caused by route conflicts, prolonged alighting and boarding times of passengers, and other exogenous delays to railway operations, may significantly affect global punctuality, as the delay to one train may propagate to other trains (Hwang and Liu, 2010). Secondary delays are propagated by a primary delay, and one trains delay may propagate to other trains and result in delay chaining. On the other hand, the primary delays are the immediate result of a perturbation (Lee et al., 2016).

(Lee et al., 2016) discusses a model for discovering the root cause of a knock-on-delay and consequently adjusting the timetable accordingly. The interactions between the trains classified into **'meeting','overtaking'** and **'waiting'**. The main factors leading to knock-on-delays as timetable itself, train behaviours at the station which include waiting, meeting and overtaking maneuvers, railway system capacity and prolonged passenger boarding. Based on this idea and classification a decision tree based is built to identify the cause. Since delays usually happen in a sequence there is a latent pattern, which upon identification can be helpful information. (Cerreto et al., 2018) attempts at recognising delay patterns from large data sets using the data mining technique of **K clustering** on heavy traffic line in the north of Copenhagen. Mining such patterns could lead to the actual cause. The clusters are formed based on the common factors of the observations which would be a basis for **delay pattern**. After the cluster, the centroid for the corresponding cluster is found and then the new centroids are found with the mean of the centroids and repeating it until convergence making it an iterative process. The proposed method is claims to be flexible, simple, unbiased and automatic which can serve as a **feedback/learning** loop in the process of planning train timetables. (D Student and Schittenhelm, 2009) mentions about the lack of feedback/learning loops in the Railway Net Europe international timetabling process and how automatic pattern recognition algorithms could help in filling this gaps. Patterns in delays have been considered as a very important source for finding the causes of delays and inter-dependencies among themselves. (Cule et al., 2011) uses frequent pattern mining methods in the form of **episodes**.To find the patterns in delays themselves and establish relations between delays rather than associating them with external patterns. The algorithm implementation was done on the Belgian Railways data set Infrabel. **Association rules**, a data mining technique has been found to be used extensively in various focus areas and one of them being in delay pattern mining. (Kono et al., 2016) which applies the concept of Association Rules on historical data for identifying the cause of delays. They consider different types of primary delays as items and find the relationship between these items. This concept is used to identify the primary delays which will usually cause the secondary delays.

Another very different approach for identifying the dependencies is in (Conte, 2008) which implements a stochastic graphical modelling approach called the **Tri-Graph**. Delay propagation is distinguished into three ways, one being propagation along the same train and, other being propagation from one to another train due to connections and the third being the propagation from one train to another due to limited capacity. These capacities are converted into capacity constraints of rescheduling problem using linear regression and then a optimisation problem is solved. They consider the stations as nodes and the track lines and activities in between the stations as edges. It also suggests breaking down the activities into smaller sub activities to have a better network analysis and understanding.The constraints are distinguished as **implicit** and **explicit** dependencies in which the former because of limited resource between activities and called as a resource constraint. Where the later is because of the precedence relationships in activities from organisation or technical requirements.Therefore it is called as the temporal constraint. Implicit constraints are discussed in the context of a periodic timetable.

2.4 Influencing Factors

In the various studies mentioned above, it can be seen that in most of them one or more influencing factors other than the usual railway attributes like headway, time, schedules etc are used. Two major influencing factors noticed from the literature above was weather and the railway network properties. Many studies mentioned above are designed around large networks, others concentrated on dense network regions or complex junctions or busy stations. Topological and locational aspects of the region have been considered in their studies.

2.4.1 Network Centrality

To consider these aspects, many studies have used centrality measures in their studies, but in completely different areas of focus. (Barthélemy, 2004) uses betweenness centrality(BC) for large complex networks. In a large complex network, it is important to know the most important nodes and how it they effect the whole network upon removal of such nodes, in short the importance of the nodes is to be known. A good measure of the centrality of a node has thus to incorporate a more global information such as its role played in the existence of paths between any two given nodes in the network thus leading to selection of betweenness centrality(Barthélemy, 2004).This just uses the BC for knowing the important nodes in the network. Network Centrality(NC) has also been used in the context of public transport(Metros)(Derrible, 2012). This used NC and specifically BC as measure to evaluate 28 metro network systems in the world and check for global trends in the area and then a detailed analysis was done for important stations in these metro systems around the world. Not going in the detail of the study, the results were astonishing as there were similar patterns in betweenness centrality all over these 28 systems and it was concluded that considering centrality into account in the planning process can be extremely valuable for better distribution of the flows of passengers. Similarly, a study in Switzerland also discusses about the network analysis of road and railways networks in Switzerland. From this study it was deduced that the Closeness and Betweeness Centrality prove to be very important in the network analysis of complex transport networks(Erath et al., 2009). This lead to selection of Centrality measures in the current study.

2.4.2 Weather

From the literature it was evident that weather is considered a major influencing factor for railway delays, as extreme weathers lead to various track related problems. Many studies like (Robert and Kim, 2018), (Wang and Zhang, 2019), (Hansen et al., 2010),(Xia et al., 2013) and many other studies have considered weather as the major attribute to carry out the prediction. Weather, temperature, wind and precipitation, effect railway operations adversely, an outline was found in (Xia et al., 2013). Warping of tracks due to uneven thermal expansion in the summer or build-up of snow and ice on the tracks in the winter can lead to decreased speeds and derailment. Extreme cold causes brittle tracks and track separation.Railway locomotives are high-profile vehicles, so high crosswinds may influence their stability. Another issue is that extreme weather may affect the presence of personnel. Standard linear regression models are used to estimate the effect of weather conditions. Levels of wind gust, precipitation and snow, the difference between the maximum and minimum temperature within a day and a measure of the relevant temperature are included in the study and are found to have a very positive correlation with the railway infrastructure disruptions. One unique observation was that the number of cancellations is not related to precipitation. Weather of course changes from region to region. And in terms of some extreme weather Scandinavian countries can be considered as places that faces some extreme weather conditions. (Zakeri and Olsson, 2018) address the extent to which weather factors such as temperature, snow, precipitation and wind influence the punctuality of trains on the Norwegian railways. Using **Pearsons Coeffienct of Correlation** and **regression analysis**, it was found that using the correlation coefficient punctuality and the independent variable, sum of snow depth, were found to have highly significant relationship(r-0.44). For the year 2008, 2014 and 2015 Norway experienced weak winters which was shown in the results too as most of the factors had less significant values. From the regression analysis it was also deduced that the snow depth was the variable that best explained the variations in the punctuality with a R^2 value of 0.51, 0.42 and 0.43 for the years 2010, 2012 and 2013, respectively. From these deductions supported the consideration of weather as another important variable for the study.

2.5 Conclusions from the Literature Review

It has been observed from the literature that understanding the delays has been considered very important as it can lead to better preventive measures and reduce the effect of the cause. In general it has been observed that conventional systems are largely reactive and not preventive which has been actually confirmed in the research. The many new types of TMSs being discussed and mentioned in the literature review attempt to make a systems which is preventive but on the basis of delays. It perceived that the reducing of delays could be the way to cater for the needs of the customers but the argument formed here is that concentrating just on the reduction of delays in a specific line reduces the focus on the overall situation. It is observed that rescheduling is driven based on the idea of reducing delays rather than finding/ arranging services which would ensure the running and provision of services. This doesn't mean that delays are not important and should not be considered in scheduling. Rather, it should be one of the triggers for rescheduling and not the only parameter. Provision of the perceived services could be a way towards making the railway attractive for the people. This could be achieved through a flexible timetable approach and making an integrated system rather than a separated system similar to the conventional system currently in operation in many parts of Europe and world. This can be seen being discussed in much research mentioned in the literature but with a delay as a driving force. Not much literature was found which would research in the area of considering the service intentions as a system driver.

Also, much research has been found on predicting the delays, relations with knock on delays and their occurrence based on the various factors like the weather, the resource conditions, the type of the delay being studied etc. For the conventional systems, applying this methodology could be helpful in curbing the delays in a particular service but loose the hold on the overall propagation of delays even though there are algorithms to address this problem. Rather, a new approach of Conditions is studied here where in the duration of the disruption becomes a driving force on occurrence of disruptions. Also, various models have been found which predict the duration of an incident and are based on variety of concepts. However, many of these models remain just as an information provider for the traffic operators rather than system integrated with them. Arranging continuous flow of services around the effected area for the calculated duration could be a priority.

Another observation was that delay causality, predictions and interrelations found in the papers, are restricted to very few cases related conflicts from timetables and resulting waiting and crossing problems. An association of delays with the root cause is attempted but the classifications for them were found to be very generic resulting lack of details about various other types of disruptions and delays other than the one mentioned in the classification itself. Additionally, many of the mentioned studies were not tested with real time scenarios.

In the tables below a complete overview of the most relevant research and the portion of it that is discussed here are mentioned with their corresponding details related to the study. In Table 2.1 an overview of the delay, duration and inter-dependency models is given and the Table 2.2 shows the overview of RTMS frameworks and the rescheduling algorithms discussed in the literature review.

Table 2.1: **Prediction and Dependency Models**

Area	Research Citations	Goal	Tools/Methodology
Identifying Root Causes& Delay dependencies	(Lee et al., 2016)	Root Cause detection & timetable adjustment	Decision Tree
	(Cerreto et al., 2018)	Recognising Delay Patterns	K-means Clustering
	(D Student and Schittenhelm, 2009)	Finding Delay Causes &inter-dependencies	Automatic Pattern Recognition
	(Cule et al., 2011)	Relations and Patterns in Delays	Frequent Pattern Mining
	(Kono et al., 2016)	Identifying delay causes	Association Rules
	(Flier et al., 2009)	Resource conflicts & blocking dependencies	(Rob et al., 1998),Synchronization
	(Conte, 2008)	Identifying the Dependencies	Tri-Graph
Predicting delays and their duration	(Zilko et al., 2016)	Predict Length of Disruption Length	Copula Bayesian Concept
	(Pereira et al., 2013)	Prediction by text analysis	topic modelling technique &Machine Learning
	(Chandesris, 2006)	Predict Incident Duration	Regression Tree Model Generalised Linear Model
	(Wang and Zhang, 2019)	Predicting Train Delays	Machine Learning, Data Fusion
	(Robert and Kim, 2018)	Delay Prediction Based on Weather Data	Decision Tree, Adaboost, Neural Networks
	(Yaghini et al., 2013)	Delay Prediction	Neural Network
	(Markovic et al., 2015)	Analyse Train Delays	Support Vector Regression, Artificial Neural Network
	(Corman and Kecman, 2018)	Train Delay Prediciton	Bayesian Networks

Table 2.2: **Complete Overview of RTMS and Rescheduling Algorithms**

Area	Research Citations	Goal	Tools/Methodology
Dynamic Systems	(Schaafsma, 2001)	RDTM Criteria	Feedback loops/Control cycle
	(D'Ariano, 2008)	Real time system with RDTM criteria- ROMA	Alternative Graph
Real-time Traffic Management	(Mazzarello and Ottaviani, 2007)	Local traffic Optimization and Control	Alternative graphs & Dynamic Programming
	(Sam, 2018)	Possible alternative Routes	Alternative Graph & MILP
	(Quaglietta et al., 2016)	RTM Framework based on SOA	ROMA, RECIFE, HERMES railML
	(D'Ariano et al., 2007)	Application of ROMA with RDTM	ROMA, Route Optimization Dutch ARI system
	(D'Ariano, 2009)	Application of ROMA on complicated and densely used areas	ROMA
	(Corman et al., 2010)	Improve computational time	Tabu Search
	(Kersbergen et al., 2013)	New Rescheduling Method	Switching Max Plus Linear Method(SMPL) & MILP
	(van den Boom et al., 2011)	RDTM traffic control Model	Switching Max Plus Linear Method(SMPL) & MILP & Model Predictive Control(MPC) Permutation algorithm
	(van den Boom et al., 2012)	Rescheduling in Large Scale Network	Max Plus Linear
	(Kersbergen et al., 2016)	Railway Management For entire network	MPC , Mixed Integer Quadratic Programming
	(Caimi et al., 2012)	Dispatching Assistant, Alternative Routing Possibilities around large station	MPC , Binary Linear Optimization Model
	(Gholami and Törnquist, 2018)	Resheduling and Replanning	Heauristics Measures, MILP
	(Carey, 1999)	Resheduling and Replanning	Heauristics Measures Delay Probabilities

Chapter 3

Data Collection and Variable Description

This chapter discusses all the variables considered in the analysis their types and their source for collection. Firstly, a complete description of the base data is provided, including explanation for each variable. In the next section the topological setup of the Denmark Railway network is discussed in addition to the basic network analysis with regards to the centrality of the network. Then the collection of temperature data is discussed, ending with a table giving an overview of all the variables.

3.1 Base data

The base data used for the analysis is taken from the disruption logging system for the Denmark railways called as the RDS. The data entry in to this system is manual entry and the details are filled in by traffic operators or drivers. For the purpose of this thesis 18 months of incident data was collected, incidents data from September 2017 to February 2019. Data set contains incidents all over the country network and it also includes data related to the Copenhagen commuter lines.

After collecting the 18 months of incident data, since both the duration and the delay were to be considered for the study incident entries consisting of both delay and duration information were considered. For collecting the total delay information, firstly all the entries with same incident ID were pooled together. These entries with same incident ID belonged to the represented various impacts caused by the same incident and the delay caused for that particular impact. Summing up all delay of these similar Incident ID entries gives the total delay caused by a particular incident. To calculate the duration of the same incident, each of them had the variable Start Time and End Time. These start and end times remain same for each entry with same Incident ID. The difference between the start and the end time would give the total duration the incident lasted. Apart from the duration and the delay data, there was other specific information about the incident available from the system. From that information the most relevant data columns relevant for the analysis were selected. The selected attributes are Type Code, Severity, Hour of start, Month of occurrence, Day of the week, Area of the incident occurrence and the Track Type. The variables Hour depicts the hour during which the incident occurred. Month(**Categorical Variable**) is the month in which the incident occurred and Day of the Week(**Categorical Variable**) is the day of the week did the incident occurred. All the mentioned three variables belong to the Start time of the incident. Whereas the attributes Type Code, Severity, Track Type and Area of the incident occurrence are factor variables with various levels. These levels have been assigned to numeric values to avoid data reading problems in the analysis. Each level for each variable is explained below. Also, how these variables have been considered into the data is explained.

Type Codes

Type Code is the variable which has the information regarding the type of incident. It gives an idea about the cause of the incident on the basis of which the incidents have be classified. The notation used to represent the types is by using numbers between 100 to 900. All the three digit numbers wherein the first digit(hundredth place) defines which category of the incident and the next two digits i.e. the tens and the one's place point to more detailed type of error in the same category. The different categories are defined below

- 1XX - Errors in Planning and Dispatching
- 2XX - Error/Mistakes by Drivers
- 3XX - Incidents in Freight Trains
- 4XX - Material Errors in Engines and wagons
- 5XX - Trains with inbuilt Engines
- 6XX - Passenger or train guard related incidents
- 7XX - Signalling and Interlocking errors
- 8XX - Train Operating Companies
- 9XX - Accidents due to weather or unexpected external influences

For a better understanding here is an example a code with number 720 means that It comes under the category of signalling and interlocking errors and the 20 denotes incidents caused by infrastructure problems related to signalling like track switch, crossings etc. For this thesis we only used the initial classification of hundredth digit that is 1 to 9. We are not using further classification as they go into more technical aspects of the railways operations which not the scope of the thesis.It is **categorical variable**.

Severity

Severity is the variable which describes how severe the incident was. This variable was also taken from the system and is divided into 4 levels from 1 to 4 each describing the kind of severity therefore it is a **categorical variable**.

- 1 - Incidents causing major impact, requires manual assistance for solving the problem
- 2 - Incidents causing impact less than the level one, also requires manual assitance
- 3 - Incidents causing much less impact but can be fixed without manual assistance
- 4 - Incidents that don't require manual assistance at all and the impact is negligible.

This notation is system specified and not created for the purpose of the thesis. These levels are used by operators and the drivers. Based on the personal conversation with the traffic operators and the driver it was deduced that there is not specific criteria for choosing a specific level for incident rather they choose based on personal experience and assessment for the incident and some types based on the type of the problems too.

Area

In this dataset each incident is attributed to a nearest station rather than the actual coordinates of the incident. Since there are hundreds of stations all around the country, it is very difficult to consider each and every station into the analysis. For simplifying, to consider the locational aspects of the incident the whole network is divided into different sections based on the type of line, the railway operator and the regional distinctions. These areas have been attributed to numbers for easy data analysis making it a **categorical variable**. These codes have been assigned by a railway expert and not by the author. Below is the complete overview of the areas and the description of the

- 1 - EDL-E : A line with medium traffic, will be operated in ETCS starting from December 2019 and the first line in East Denmark with commercial train services and ETCS Trains running from Roskilde to Næstved and back, no other patterns except from very few freight trains.
- 2 - ØSJS : Private line South of Køge, Separated from other lines with its own traffic pattern, but connections in Køge to trains going to Copenhagen.
- 3 - SL-M : A few private lines, merged per area, where also some of the lines meet. This set comprises of mainly smaller stations in the middle of Zealand, geographically compact and separated from rural areas in the North and South of Zealand
- 4 - SL-S : Private lines in the South of Zealand, sparsely populated, not a lot of traffic here.
- 5 - Westcoast : Another set of stations, similar to Ti, but trains usually do not run from Westcoast to Ti, they only meet in certain nodes.

- 6 - NJ : Nordjyske is an operator that took over traffic North of Aalborg completely. The line North of Aalborg to Frederikshavn is ETCS equipped and runs commercially with ETCS
- 7 - S-Bane : Commuter lines in the Capital Region- Copenhagen
- 8 - ML-1 : Mainline, West of Glostrup. The line connects the important nodes Ringsted and Roskilde.
- 9 - Kb : Kalundborg line. Traffic between Kalundborg and Ringsted.
- 10 - Kh : Major regional line in Copenhagen
- 11 - Pa : Line to Padborg. Important for international traffic to Germany.
- 12 - Fyn : Line going from the east to west of the island of Fyn
- 13 - Kystbane : Regional line between Helsingør and Copenhagen
- 14 - Svg : Line going from the Odense to South of Fyn, Svendborg
- 15 - Fa : Line from Fredericia(immediately after Fyn island) to Århus
- 18 - JL-Mid-1 : Middle of Jetland, merging a few smaller lines with small amount of local traffic. Very rural areas.
- 17 - JL-Mid-2 : Similar to JL-Mid-1
- 16 - JL-Mid-3 : A bit similar to JL-Mid-1. Includes important node Herning
- 19 - Es : Two lines From Esbjerg splitting at Bramming
- 20 - Ab : South of Aalborg, Commuter trains and express trains connecting the towns South of Aalborg down to Aarhus.
- 21 - Ti : Similar to SL-S, but in Jetland, West coast.
- 22 - Others : International Lines
- 23 - SL-N : A number of private lines in the North of Sealand

For better understanding of the areas the map in Figure 3.3 and the next section could be used as reference which mentions all the major important topological aspects of the railway network and the country itself.

Track Type

Danish railway traffic is not the largest when compared to other countries. Most of the lines in the country you find are either single or double track lines. In Denmark, usually all the lines span out from few major important points which here are considered as junctions. So, to consider this element of the network which also encapsulates part of the traffic, the variable Track Type was introduced which is a **categorical variable**. It has three levels as mentioned below

- 0 - Junction
- 1 - Single Track
- 2 - Double Track

Train Intervals/ Frequency of Trains

Train Intervals is basically the frequency of the trains. This variable was considered for the analysis to include the schedule aspects of the operations.Because of the lack of schedule data for the operations if was not possible to include the exact schedules of the trains but to include the traffic aspects into the analysis, frequency of the trains was the best representation. This could give an idea on how much trans will get affected upon occurence of an incident i.e. lower the frequency, lower the chance of propagation, higher the frequency higher the chance of delay propagation.

To include this variable into the analysis, the approach used here is that instead of taking frequency of each line in the network, the intervals are considered on the basis of the Area Codes mentioned above. An average of the frequencies in the lines belong to the area mentioned above is considered and the stations belonging to that particular area is assigned with the average frequency of the area. The frequency information is provided by a expert traffic operator.

3.2 Network Centrality and Topological Setup

3.2.1 Railway Topology

Denmark topology is divided into five major parts that are Northern Jylland(Nordjylland), Central Jylland (Midtjylland), Sjælland(Zealand), South Denmark(Syddanmark) and Capital Region(Hovedstaden) which have been depicted in the Figure 3.1 The 4 four largest cities in the country are Copenhagen, Aarhus, Odense and Aalborg. All these 4 major cities are in the four major regions of the country. The main lines passes through all the these four cities which can been seen in the Figure 3.3 Copenhagen is the largest city, with the largest population in the country, with commuter lines spread all over the city and the suburbs around.

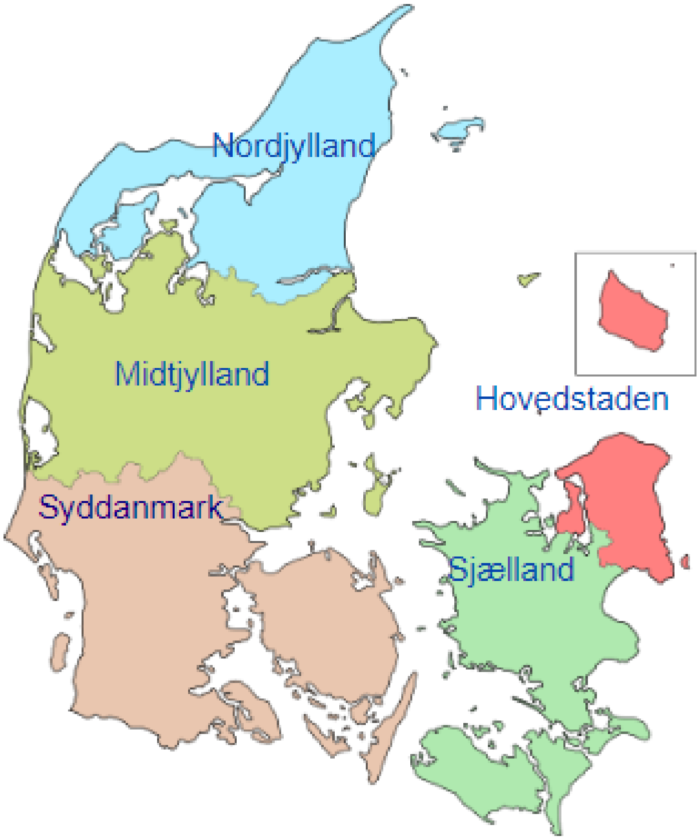


Figure 3.1: Regional Division of Country Denmark

That region has dense railway traffic with high ridership. Train from Sweden and Germany travel to and fro Denamrk from the Øresund bridge in towards Sweden and 2 connections to Germany. One of the connections being ferry from Rødby Denmark to Puttgarden. The second connection is overland via Padborg, Germany. In the ferry connections the diesel trains go into the ferry and are transported across via ferry whereas over the land, the line is electrified. Most of the lines are not electrified in the country and are in the process of electrification. The islands of Fyn and Sjlland are connected by the famous Great Belt bridge which became a very important line for traffic flow from east to west and vise versa. In terms of traffic capacity, Denmark has seen an increase in the traffic over the years(see Figure 3.2) but the infrastructure has remained pretty much the same. This situation can lead to very quick propagation of the delays into the network which can effect both the train operations and the passengers. Denmark has a more homogeneous operation of the railway network than other countries (e.g.Germany and France), which allows the high capacity utilisation. Homogeneous operations meaning that there is no high-speed operation and only a limited amount of freight transport(Landex, 2008).

From the Figure 3.3 it can be seen that the apart from the state run lines, the lines in the white in represent the county and the privately run lines. The infrastructure is owned by

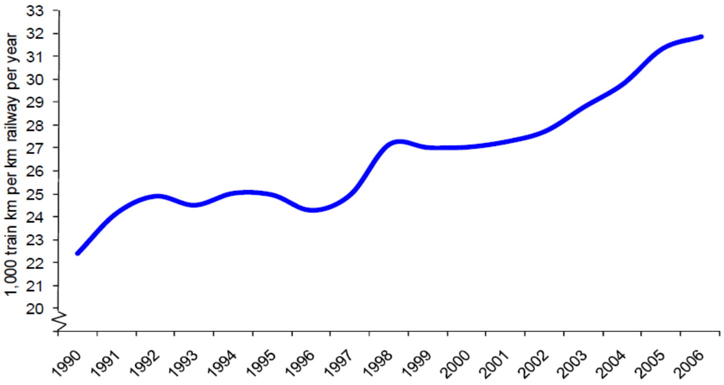


Figure 3.2: Denmark Infrastructure utilisation development. (Landex, 2008)

BaneDanmark which is state owned company. Danske Statsbaner(DSB) is the largest state owned train operating company(TOC). There are other private TOCs also operating in some parts of the network. Few of the other operators are Arriva, MJBA Midtjyske Jernbaner, NJ Nordjyske Jernbaner, Statens Järnvägar and Deutsche Bahn Cargo.

3.2.2 Network Centrality

The other locational variables mentioned above consider the type of lines, the operator and other attributes but it does not actually consider the actual location(coordinates) of the incident in the analysis. Upon research the concept of Network Centrality which is a part of Network Analysis in Graph Theory was found to be a perfect fit for considering actual locational influence in the predictive model. Railway networks are heterogeneous networks made of various elements which have different importance levels in the network. Node and edges are the different elements of which a base network is represented in. The concept of centrality is the measure of importance of node in heterogeneous network in comparison with other nodes. This importance can be quantified by the network centrality, as we expected that the most central nodes are the most influential ones and can propagate their information content easier than other nodes (Rodrigues, 2019). There are various metrics to measure the centrality which can be used to characterise complex networks.

Among the various metrics present, Degree, Closeness and the Betweeness Centrality are used as the metrics for considering the centrality measures in the system. To measure the mentioned metrics firstly the network data was to be collected. Both Python and R have packages which are used to do network analysis on network data. For this thesis Python is used for the network analysis. Two packages NetworkX and OSMNX are available in Python which can be used for network analysis. OSMNX (Boeing, 2017)package also can be used for getting network data from Open Street Maps. By passing the name of the place and its coordinates, the package helps in getting the node and edge data for the railway network from Open Street Maps(OSM). The data from the OSM is yard based and not station based. To find the station locations in the network graph, from the base data the coordinates of the stations were available. To bring these station points on to the network graph the get_nearest_node function provided by the OSMNX is used to find the nearest node to the

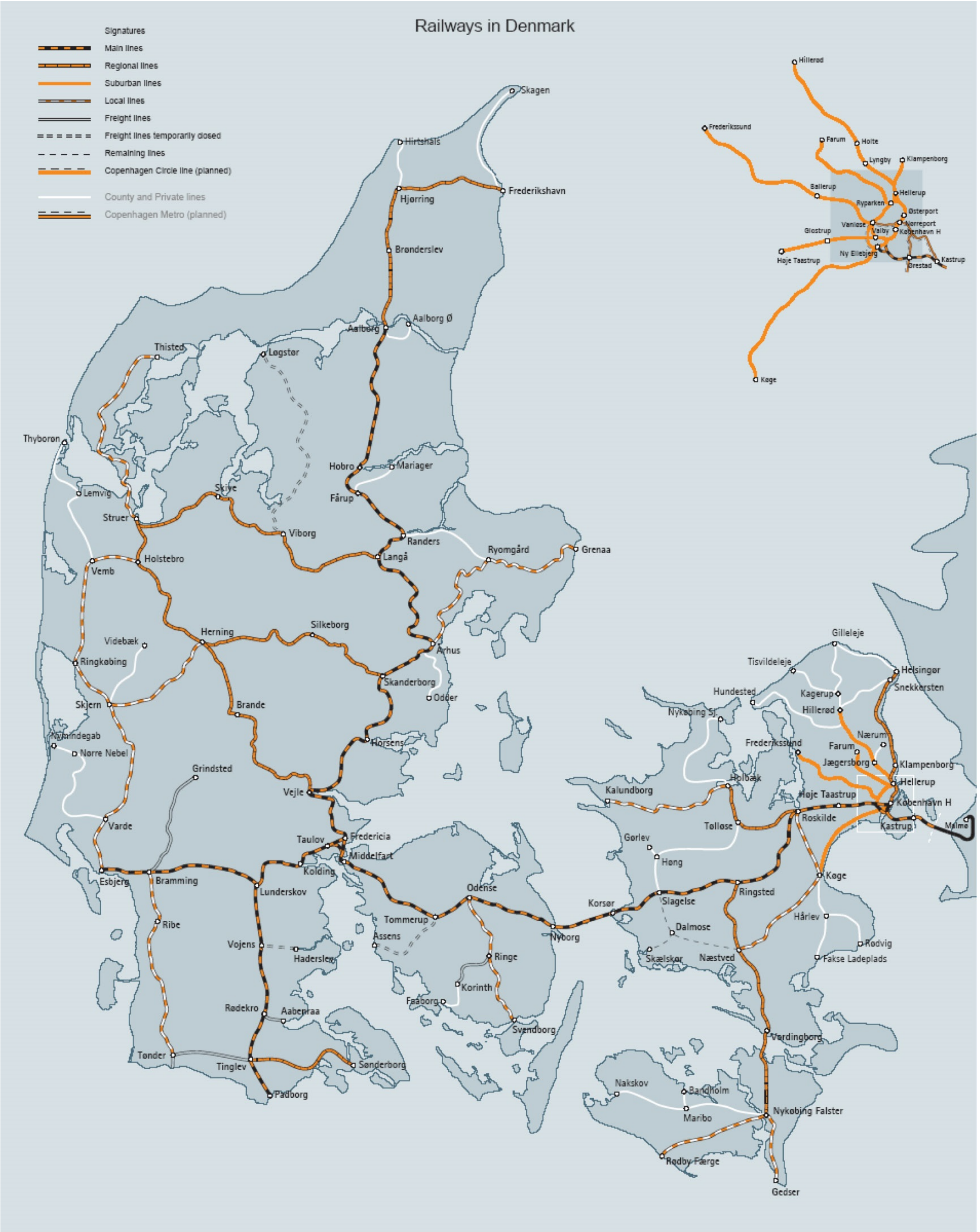


Figure 3.3: Railway network of Denmark with all the major stations and types of lines

station coordinates. The method options available are either euclidean or haversine.

The next step of the analysis of the network is the calculation of the different metrics for the network analysis. The OSMNX package also enables in measuring the various metrics available. Based on the details available in the paper (Boeing, 2017) following mathematical

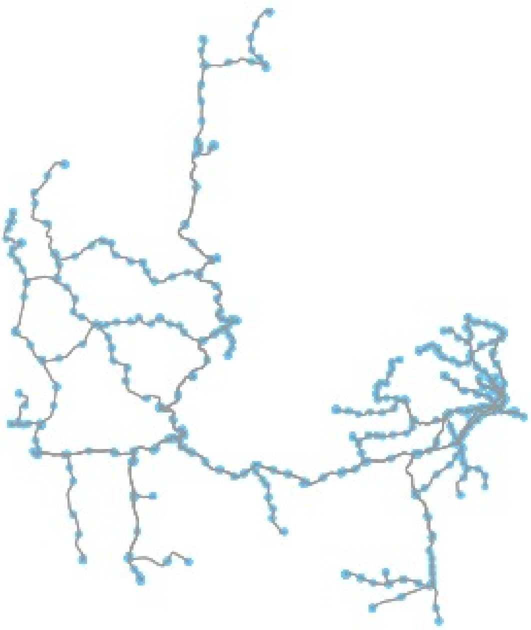


Figure 3.4: Line graph of Railway network of Denmark based on shape file from OSM

formulaes where used for the calculation of the all the mentioned metrics which are Degree, Betweenness and the Closeness Centrality. The terms are explained below with the method of calculation mentioned.

Degree Centrality(DC)

Degree Centrality is the considered the most simple centrality measure which can be defined as the number of connections or nodes attached to a particular node. The network structure is considered as a network matrix A_{ij} where i and j are the nodes and if $A_{ij} = 1$ then it can be said that there is a connection between the nodes i and j otherwise A_{ij} is not equal to 1 denoting no connection between i and j. The Degree Centrality of node i can be calculated using the following formula i.e the sum of the elements of row i of A (Rodrigues, 2019).

$$k_i = \sum_{j=1}^{N} A_{ij}$$

where N is the total number of nodes in the complete network and k_i.

Closeness Centrality(CC)

This is another measure of centrality where it considers the shortest path. The distance between nodes i and j is given by the number of edges in the shortest path connecting them. Closeness Centrality uses the idea that central node is close to all other nodes in the network in terms of distance. To define Closeness Centrality (Rodrigues, 2019).

$$C_i = \frac{N}{\sum_{j=1,j\neq i}^{N} d_{ij}}$$

, $d_{ij}$is the length of the shortest path between i and j, and N is the number of nodes in the network. Lower the value of CC the more central is the node and Higher the CC value more the isolated is the node.

Betweeness Centrality(BC)

In Betweenness Centrality the base idea is that the if flow particles in considered the network then centrality can be defined in terms of the load. It can be said that more the central the node is the more it receives the largest number of particles in a defined time interval. Assuming that these particles move following the shortest distances, the load in a node i is given by the total number of shortest paths passing through i. Since we can have more than one shortest path between a pair of nodes a and b, it is more suitable to define the load in node i as the fraction of shortest paths connecting each pair of nodes (a,b),a,b = 1, ...,N, that includes i.

$$B_i = \sum_{(a,b)} \frac{\eta H(a,i,b)}{\eta H(a,b)}$$

, where $\eta H(a, i, b)$ is the number of shortest paths connecting vertices a and b that pass through vertex i and $\eta H(a, b))$ is the total number of shortest paths between a and b.

The above mentioned 3 centrality measures are considered as the variables for the predictive model. Nearest node near the location of incident can be used for calculating the centrality measures which can be helpful in understanding how the location of the incident can affect the network and delay propagation into the network. Higher the value of BC higher more important is the node and can be said the more particles flow from the node. After calculating the centrality measures for all the incidents from the data-set the following network graph plot depicting the Closeness Centrality of the network was created which clearly shows the most central and the most isolated nodes in the network. In the figure 3.5 shows the brightest nodes are the nodes with highest CC values pointing the most central nodes whereas the darkest nodes are the nodes most isolated with least CC values. From the Figure 3.5 it can be seen that the nodes in the Copenhagen region are brightest and the nodes connecting the two regions of Sjælland(west), Jylland(east) and Fyn. Also, you can see the norther most nodes in Jylland are dark indicating the farthest and most isolated nodes. This intensity plot is very similar very much with the railway traffic in the region indicating more traffic in the region with the brighter nodes and lesser in the region of darker nodes.

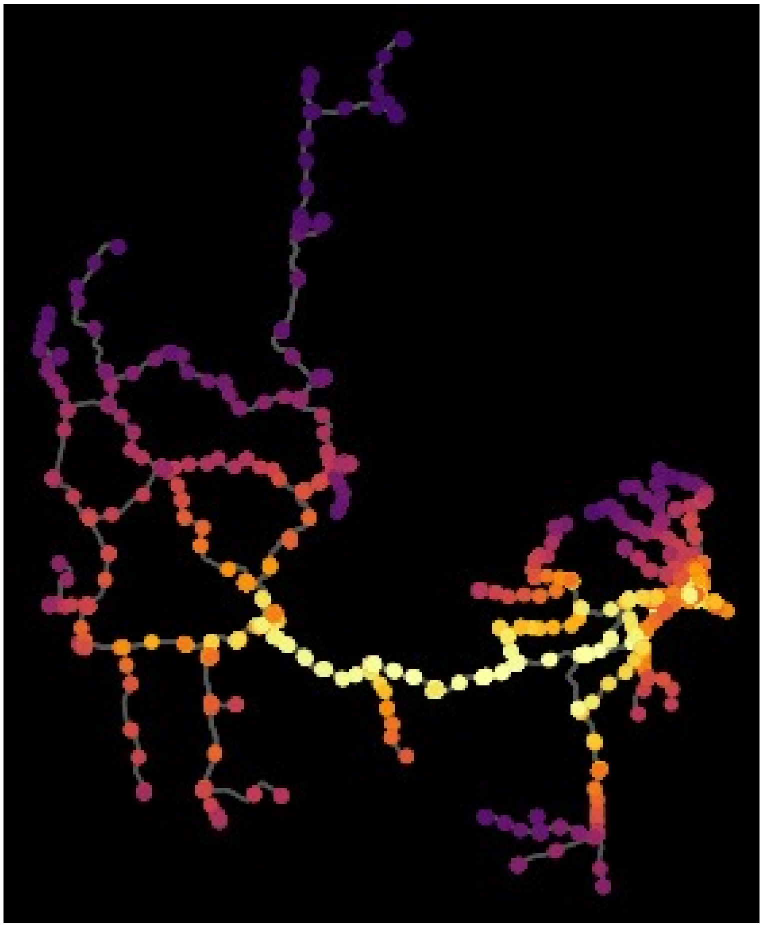


Figure 3.5: Line graph showing the the intensity of closeness centrality in the whole railway network

3.3 Temperature

From the literature review it can be seen that the weather plays an important role in railway incidents. The worse the weather more the railway incidents. To consider the weather as a variable in the analysis, temperature was chosen as the variable. Due to lack of wind data for the country of Denmark, it was not possible to use wind pressure for the analysis. For the collection of the Temperature data the DarkSky API was used. This API has many wrapper libraries for various programming languages and for this thesis darksky package for R is used for collecting the temperature data. By passing the locational coordinates and the date and time for the which the information is required, the corresponding hourly weather data can be collected for that particular day. The table below gives complete overview of the variables considered in the analysis

Table 3.1: **Complete Overview of Variables used**

No.	**Variable**	**Type**	**Levels/values**	**Units**
1	Total Delay(TDH)	Continuous	Numeric	Hours
2	DURATION	Continuous	Numeric	Hours
3	No. of Trains Impacted	Continuous	Numeric	Count
4	Headway Intervals	Continuous 2015	Numeric	Hours
5	Track Type	Factor	3 levels	SingleTrack(1), DoubleTrack(2), Junction(0)
6	Hour	Factor	24 levels	Hours
7	Day of the week	Factor	7 levels	7 days of the week
8	Month	Factor	12 levels	12months of the year
9	Severity	Factor	4 levels	1to 4-1 being most severe
10	Temperature	Continuous	Numeric	Degrees Celsius
11	Betweenness Centrality	Continuous	Numeric	Number
12	Closeness Centrality	Continuous	Numeric	Number
13	Degree Centrality	Continuous	Numeric	Number
14	Type Codes	Factor	9 Levels	1 to 9 each depicting type of problem
15	Area Codes	Factor	23 Levels	1 to 22 each depicting region

Chapter 4

Descriptive Data Analysis

In this chapter, a basic exploratory analysis is carried out on the variables discussed in the previous chapter. It can be helpful in finding any patterns or some important detail hidden in the details. Also, the pre-processing steps for preparing the data for the modelling is discussed. (Dasu and Johnson, 2003) is a book for exploratory data mining, it mentioned that 60%-80% of the data analysis time is spent on the data cleaning and preparation.

4.1 Exploratory Analysis

In the previous chapter the data collection and the variables parts of the base data were addressed. From the data collected initial count of incidents were around 20,000 of where few of them had the duration data and few of them had the delay data. Eventually, only incidents with both the data were considered for the analysis. Which turned out to be around 10,000 entries. The initial range of duration values ranged from 0 to 10,000 hours and for the total delay the values range from 0 to around 800 hours. In parallel, basic linear regressions were run on the whole data too and the results, R^2 specifically signified that the dataset needed improvements. Since the source of the base data was from manual entry it can be said that these huge values which amounts to over months can be ignored because these values might have arisen from the delayed closure of the incidents or maintenance works that actually required months in time. To select the most relevant incident entries, the first step was to filter entries that had duration values of less than 250 hours which is around 10 days. Because of the data limitations a wide range/variety(based on the duration)of incidents are considered. The initial filtration is done on the basis of duration which was based on the initial idea that more the duration, more will be the corresponding delay. This notion is discussed in detail in the next section.

4.1.1 Total Delay and Duration relationship

To better understand the distribution of the total delays and durations, a deeper analysis is required. In the Figure 4.1, it can be seen that the relationship between the total delay and duration are is in the shape of an ”L”. From the shape one thing can be said that duration and total delay don’t necessarily need to have a linear relation ship which is evident from the graph. Higher duration values can have lesser total delays because in few cases the corrective measures are applied immediately and the operations are arranged accordingly in short time resulting less accumulation of delays. On the contrary, cases with very high duration have more total delay which can be explained by an example. If a technical problem occurs at critical junction over which there is a heavy traffic of trains, then even minor incidents can cause to accumulate delays as the delay can propagate fast to the trains in queue.

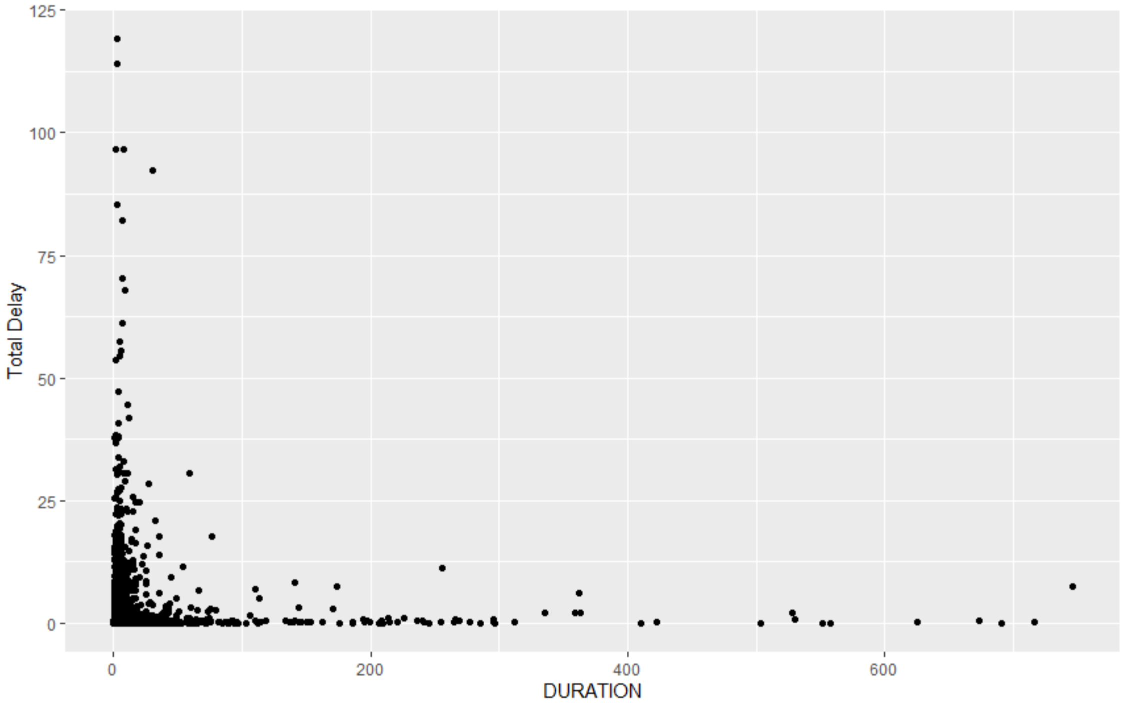
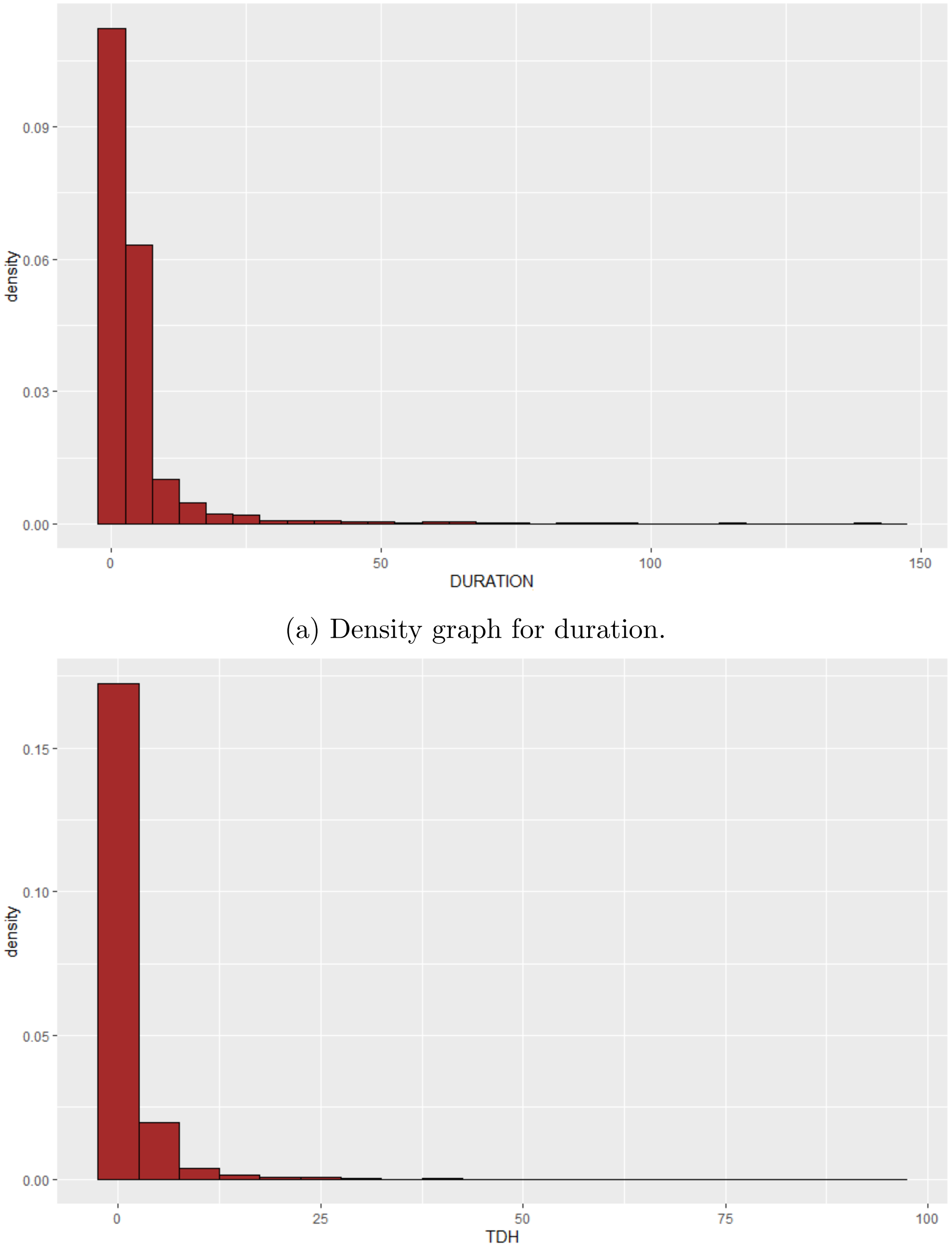


Figure 4.1: Scatter plot between the Total delay and Duration

From the Figure 4.2 are shown above and it can be seen that both have similar kind of distributions. From the data set it was found that about 93% of the incident entries were having delay values of less than 5 hours, about 4% have a value between 5 and 10 hours and only one percent of entries have a delay value greater than 20 hours. In case of Duration data about 80% of the incidents have values less than 5 hours,and about 11% of the incidents have values of in between 5-10 hours, 3.3% of values between 10-15 hours and about 7% values greater than 20 hours. The consolidated table can be found in the appendix. The ”L” shaped plot seen in Figure 4.1 usually indicate the presence of a heavy tailed distribution. This can be checked by using the power law function which is a functional relationship between two variables where change in one results in a relative proportional change in the other. To check if the data follows the a power Law distribution, using R package powerLaw was applied on the data. By using the goodness of fit measure, it can be said that whether power law distribution fits the data or not. If p value is greater that the significance interval of 10% that is $p > 0.1$ or more lenient condition $p > 0.1$ can be said to follow the power



(a) Density graph for duration.

(b) Density graph for total delay

Figure 4.2: Density graphs for total delay and duration

law distribution (Clauset et al., 2009). Three distributions were applied discrete, continuous and log normal and the obtained values for p and α were 0.01, 0.95, 0.27 and 2.28, 2.14, 2.28 respectively. From the p values it can be said the null hypothesis that, power law distribution fit exists, could not be rejected. Resulting in the pointing of the power law distribution in the data(the results and the graphs are shared in the section appendix). From this it could be inferred that a small change in either total delay or duration could lead to a significant change in the other.

4.1.2 Other Variables

Based on the Area Codes, discussed in the previous chapter,from the Figure 4.3 it can be seen that area 11 and the area 7 have the highest number of incidents which correspond to the areas near the border with Germany in the area of Padborg and the S Bane(commuter lines) region in the capital city of Copenhagen. The reason can be said as the change in country at the border requires change in systems. And as mentioned earlier that the highest traffic with ridership can be seen in the same area. Since the frequencies between the trains is small delay propagation can be really high. Also, it can be seen that the majority of the incidents in all the areas belong to category 1(most severe). From the dataset, it was found that incidents of Severity type 1 amounted to up to 86% indicating that majority of the incidents are delay causing type incidents.

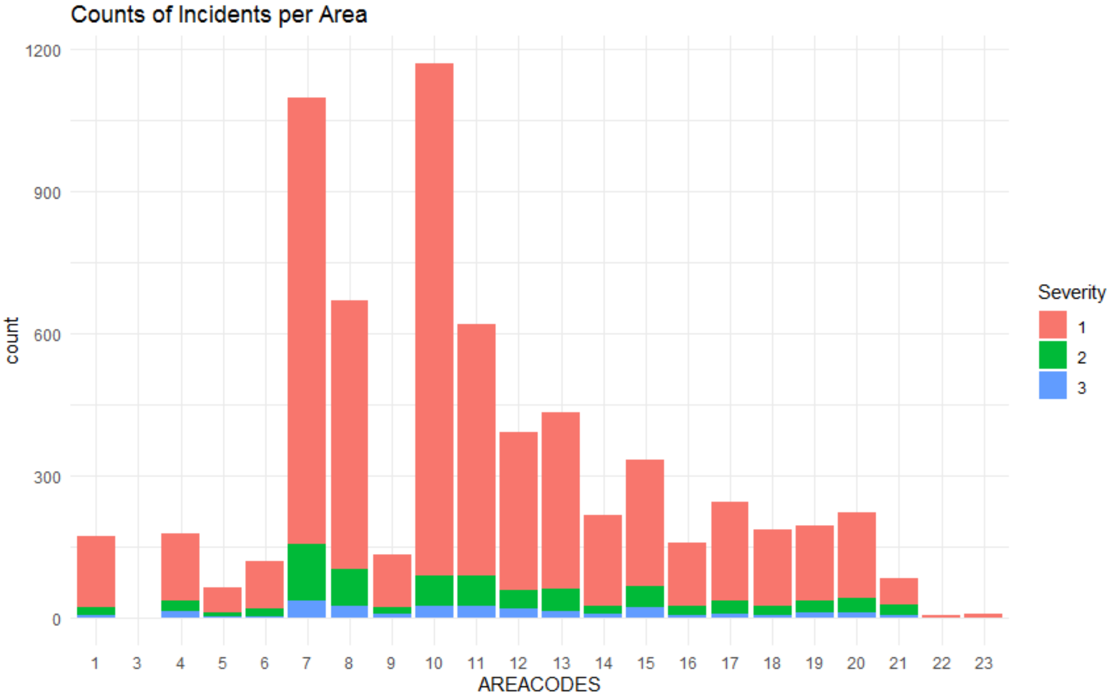


Figure 4.3: Number of incidents per AreaCode of different Severity Type

There was another classification of TypeCode in the dataset. When classified on the basis of the TypeCode it was found that almost Majority of the incidents belonged to the category of 7 and 9 those are the Incidents related to Interlocking/signalling related problems and External influences like weather, respectively. Indicating the ongoing signalling related works have been creating a lot of problems in the network. Additionally, as discussed in the literature, it can be seen that weather does has an immense influence on the operational performance. It also can be seen from the Figure 4.6 that these weather related incidents(adding to the external forces) to be equally distributed over all the months.

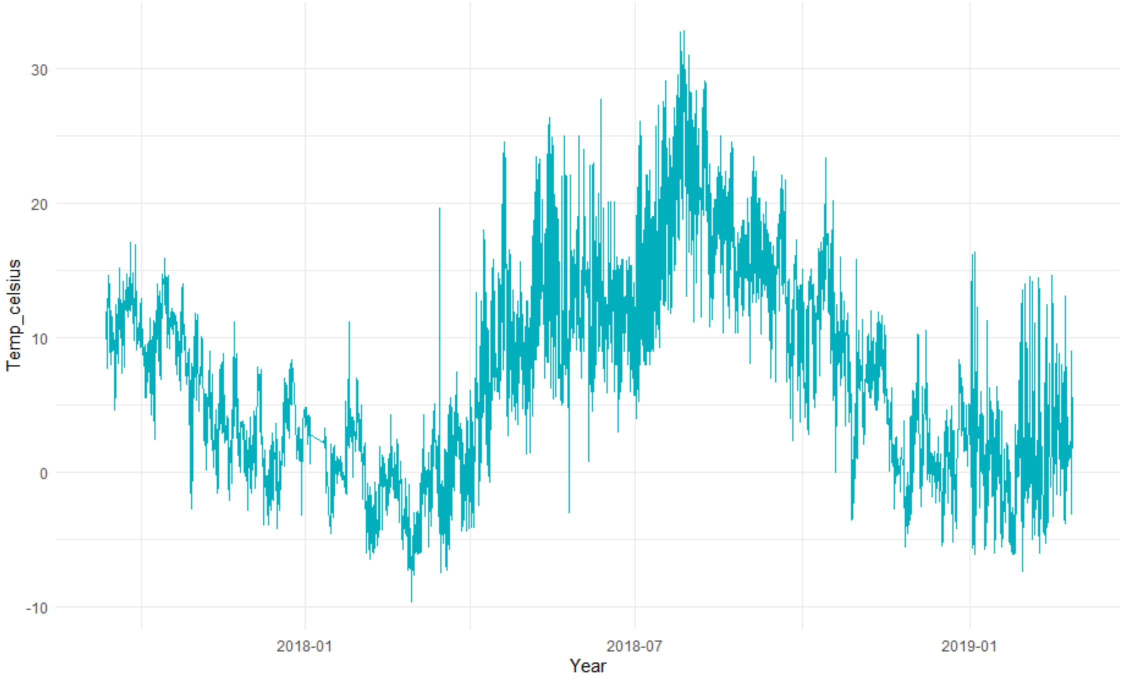


Figure 4.4: Temperature vs months

The line graph above in Figure 4.4 shows that the variation of the temperature with the months. It can be seen over the whole time for the 18 months of data the temperature falls below zero during the winter months and the temperature values rise during the summer months. Around 47% of the incident occured during the 5 months of winter considered from October to March. Though the sum of incidents in the other 7 months was found higher than the winter months. But from the Figure 4.14 it can be seen that the increase in the incidents during the winter months. Though not all the incidents are caused because of the weather conditions but the increase could be attributed to it.

As mentioned in the previous chapter, the track system is divided into 3 types that is node, single track and double track. The double track lines are usually the main lines which connect the major stations. It can be expected that the incident count and the total delay can be high in the areas with double track because the frequency could be high in double track as each line could be dedicated to one direction of travel. From the dataset it was found that the around 62% of the incidents where caused in the double track sections. 17% of the incidents belongs to the nodes/junctions. But that is considerably high as the number of major junctions around the country amount to just 15 and the percentage of total delay caused was 21%, which is higher than the that of the percentage of total delay for caused by single track which is at 17%. This shows that junctions have more influence on the operations in case of an incident. Average duration and the total delay for all the three types of track sections was found to be very close to each other but in both cases the average total delay and duration are higher for junctions than single track section. This clearly shows that the junction has more impact on the operations in case of an incident. (Sogin et al., 2013) mentions that double track and single track behave differently and the reason for delay in

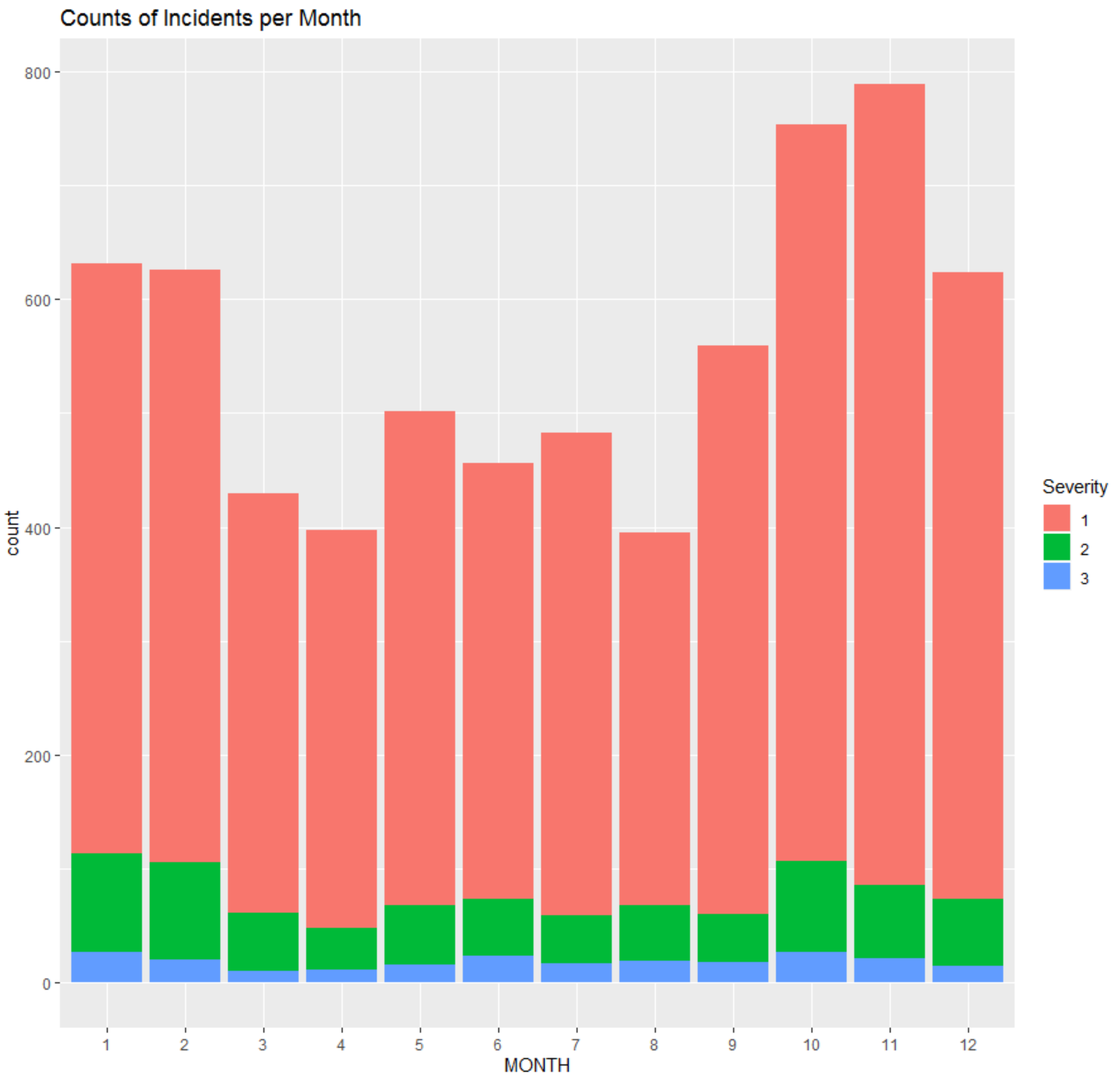


Figure 4.5: Incident count per month and classified based on Severity

both the scenarios is different. In case of double track, the major reason for delay apart from incidents is the difference in speed between the train types. Whereas in case of single tracks the delays are usually because of the meeting and overtaking maneuvers. The Figure 4.7 shows that distribution of incidents based on the track type in each area. It is evident that area 7, 8, 10 and 11 have the highest incidents and every area has highest two track incidents.

Talking about track sections and type another important variable that influence operations is the frequency. The Figure 4.9 below shows the number of incidents that happened based on the frequency. Also, the share of each track type is shown to understand the effect of it. It shows that highest incidents are seen when the frequency is 5 minutes, deducing high total delay. This can be explained because the propagation is very easy as the frequency between two trains is so small that it lead to propagation thus accumulating a high total

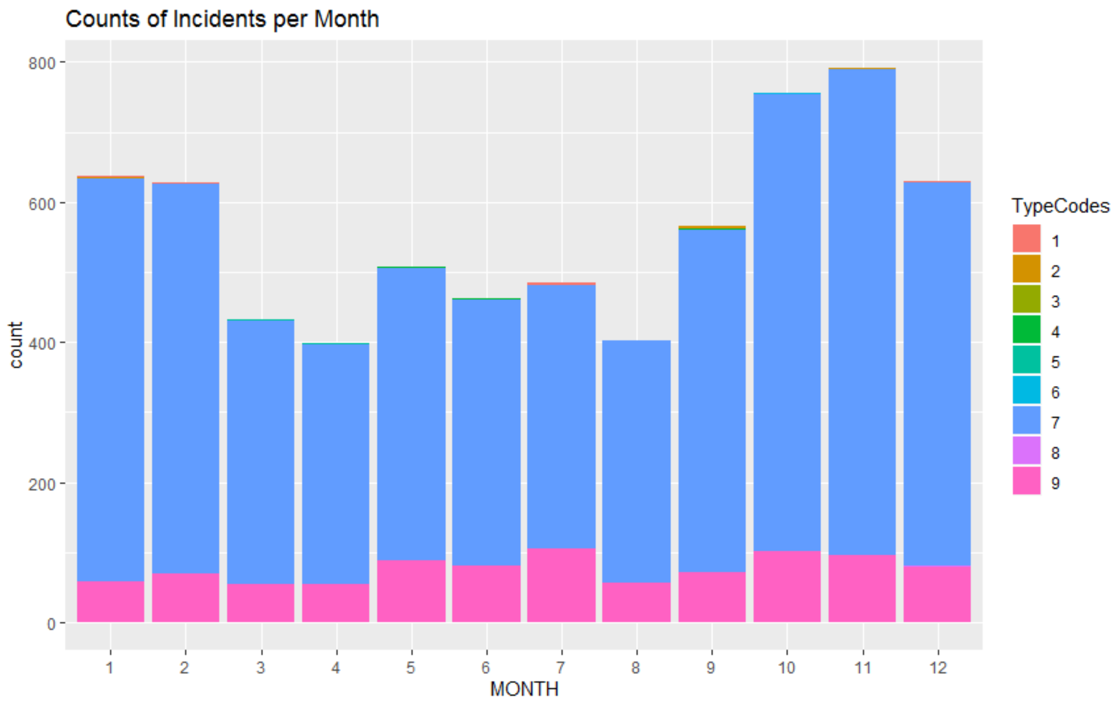


Figure 4.6: Incident count per month and classified based on type

delay. Though the incident duration might be small the total delays could be higher as the propagation into the network is fast because of a small frequency. Another very important thing to note is the significance of the junctions and the single tracks based on the high and low frequencies. It can be seen that with lower frequencies, junctions and double tracks have more significance. High average delays were found for values less than 15 minutes with highest average for 10 minutes indicating more delay, even though the incidents are less. On the other hand, when the frequencies are long, incidents on the single track are more frequent. In a single track section if an incident occurs then the track is closed down and all the trains that need to use the track have to be reroute which adds into their travel time making them deviate from their timetable, accumulating delays,and increase in the total delay for that particular incident. In the Figure 4.7(b), it can be seen that the duration values are fairly uniformly spread for the all the intervals.

Now since the frequencies are discussed, lower the frequencies values the more traffic on the network. This makes the Number of Impacted trains an important variable to be investigated. In the Figure 4.10a, it is clearly evident that the relation between the duration and the number of impacted trains is linear. It can be seen that the increase in total delay results in increased impacted trains or rather, the more trains are impacted the larger the delay. But it can also be seen that there are few data point where the total delay is significantly higher compared to the number of delays. This can be due to the fact that sometimes few trains are directly affected by the incident causing them to stop for long time or completely stop the service and shift the people into other trains. Such problems can cause few trains to accumulate a lot of delay leading to increase in the total delay and with few impacted

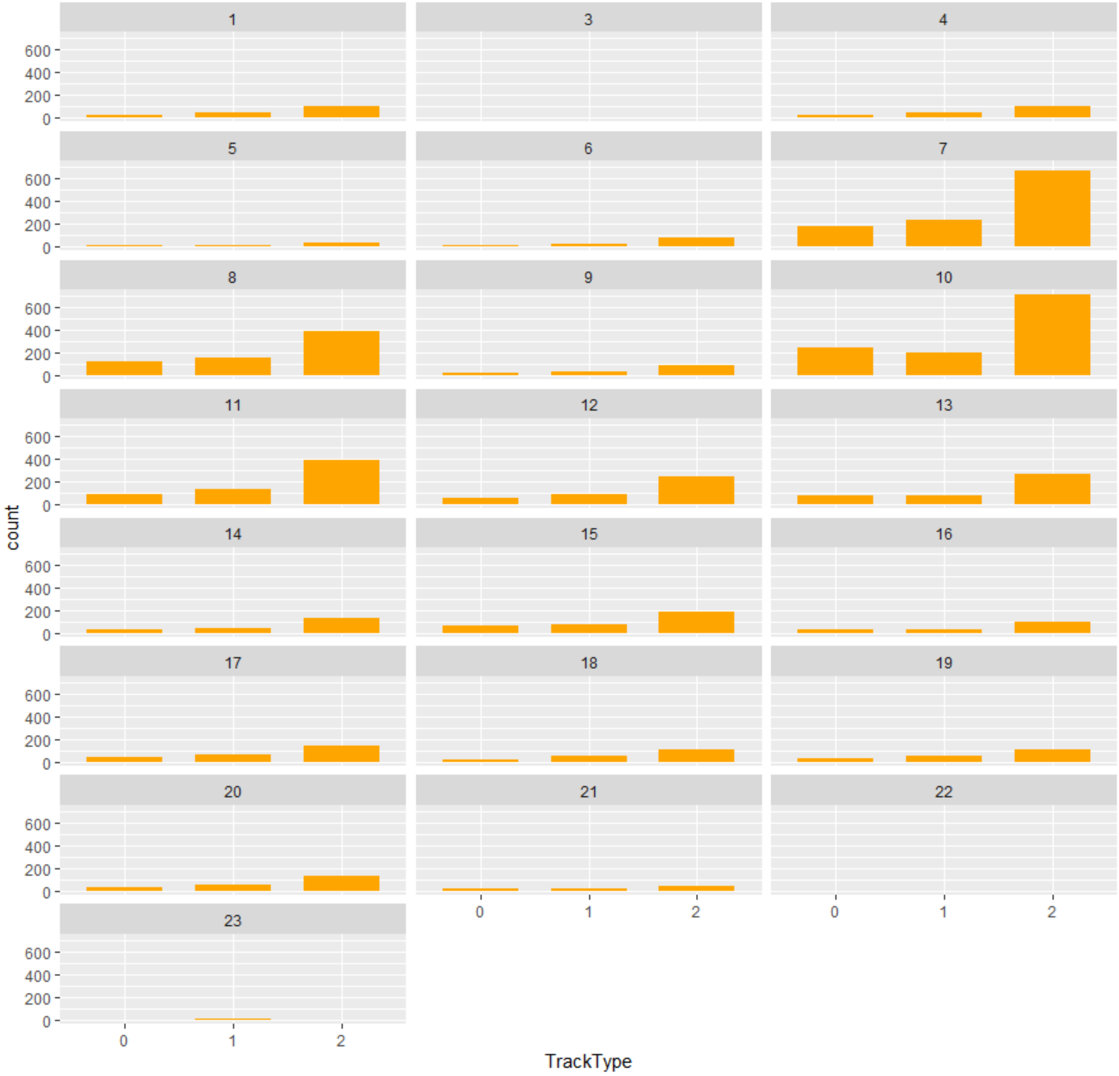
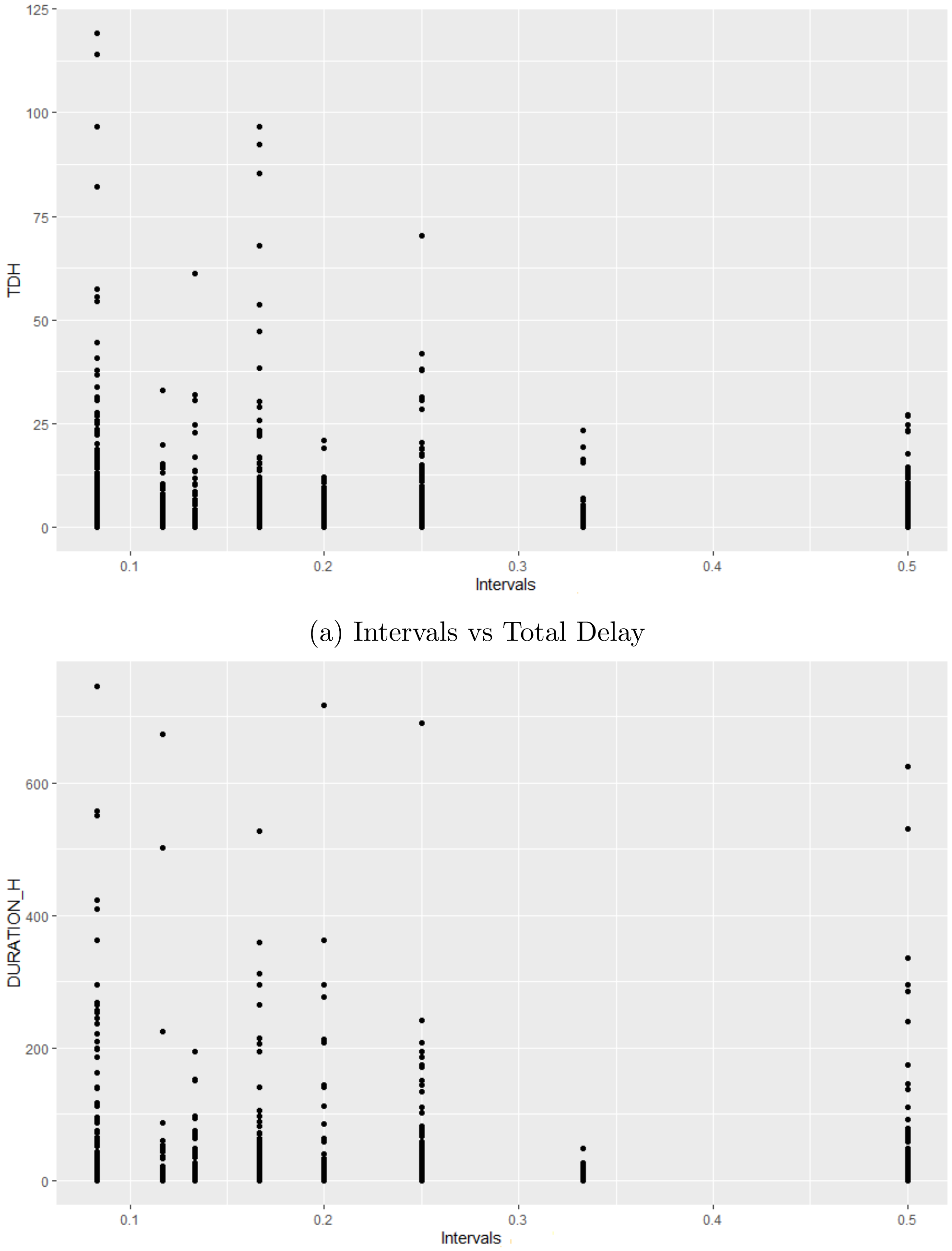


Figure 4.7: Number of incidents on each Track Type per each Areacode

trains. On the other hand, there are very few points where the impacted trains are many but the total delay is little, indicating that the incidents can cause more impact on trains but can cause only a small delay to each of them. In terms of duration, a similar ”L” shaped plot can be seen in the Figure 4.10b where there is a very strong effect on both the variables duration and the number of impacted trains.

From the Figure 4.10b it is clearly evident that the longer the duration less the impacted trains whereas on the other hand the more the impacted trains the less the duration, but there are some uniform values to the lie at the vertex of the ”L” curve. One can clearly not say that with a longer duration of incident more trains are impacted. Additionally, though this paper does not mention the role of the maintenance management team, it indeed is an important aspect to consider when trying to understand these relationships. In the Figure 4.11 which shows the relation between the impacted trains and the Intervals, it can be seen



(a) Intervals vs Total Delay

(b) Intervals vs Duration

Figure 4.8: Duration and Total Delay vs Intervals

that there is more noise in the area of lower Intervals. This proves the statement made above in relation with the intervals that smaller the frequencies more chances more trains get effected by the incidents.

Another very important variable considered were the centrality measures. For the betweenness centrality, the higher the values the more important are the nodes. Whereas, in case of

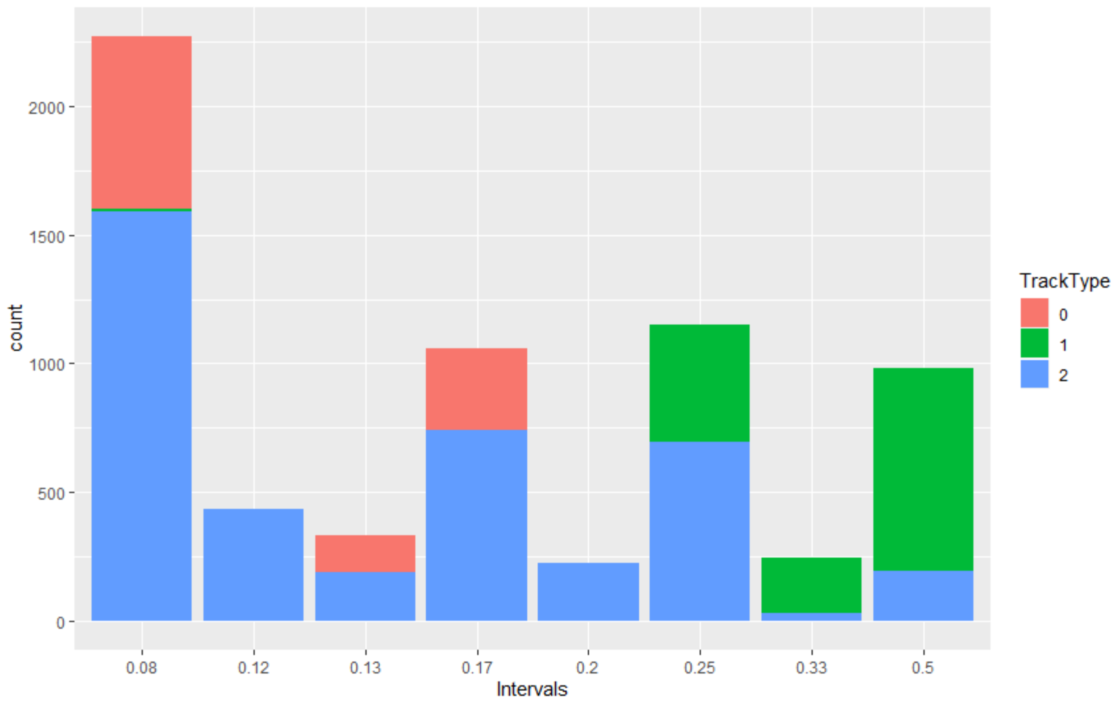
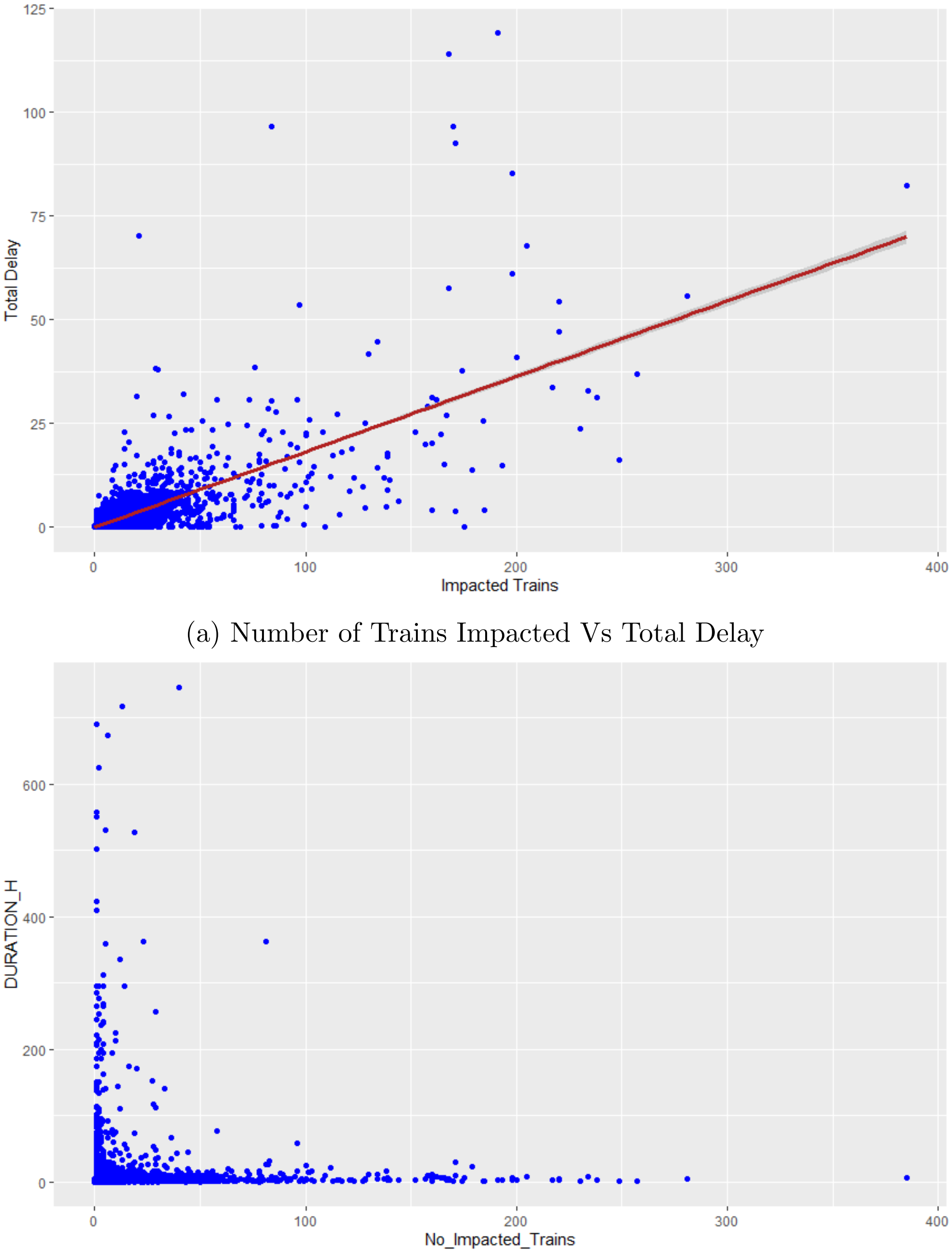


Figure 4.9: Number of incidents based on the time intervals

closeness centrality, the higher the value less central are the nodes. From the Figure 4.13a can be seen that the higher the betweenness centrality, the longer the duration and the delay values indicating the more the important values get affected more causing more delay. In case of closeness centrality the smaller the value the longer the duration and the total delays are more in the central areas which is self explanatory because in a central location the delay spread quickly, adding up to a larger number. In case of Degree Centrality, Figure 4.12 shows its relation with Betweenness Centrality and Total Delay. It can be seen that medium values of the degree centrality have the majority of the incidents pointing that most of the nodes have other nodes attached to it pointing to the fact that major incidents occur at nodes with degree of medium value. The month element of the date time variable has been already investigated. The other elements are the weekday and the hour and a incident count plot for the both the variables are shown in Figure 4.14. It can be seen that the number of incidents are the least during the weekend that is on Saturday and Sunday. The percentage of incidents on the weekends are around 7-8 % and which 6-10% less than the other days. Though the average total delays are uniform among all the days but the average duration was not uniform as it ranged from 12% to 21% during the weekdays but during the weekends it reduced to 8-9%. Also, the number of incidents increase during the start of the working day at 6.00 in the morning, continues until 18.00 in the evening which is evident in the Figure 4.14a.



(a) Number of Trains Impacted Vs Total Delay

(b) Number of Trains Impacted Vs Duration

Figure 4.10: Number of Impacted trains and relationship with other variables

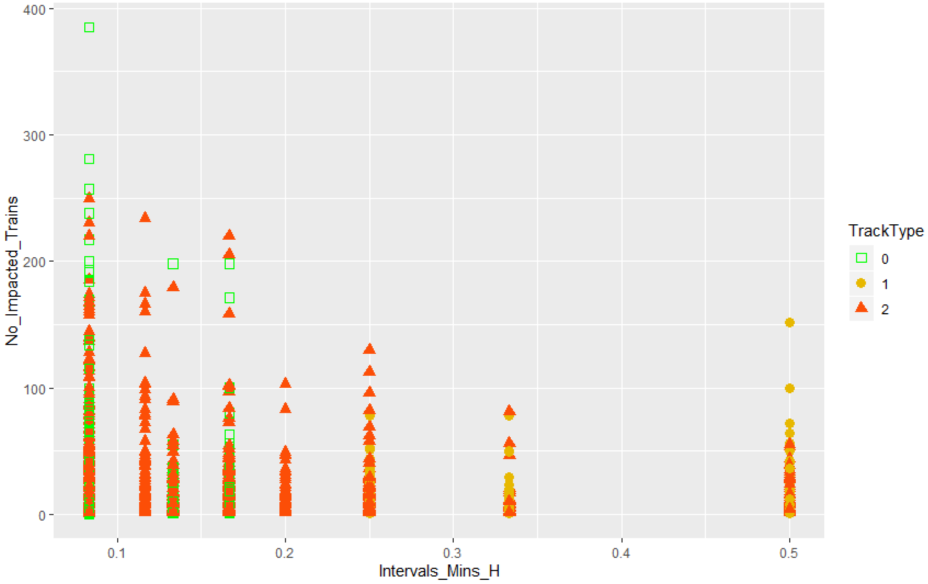


Figure 4.11: Frequencies Vs Number of trains Impacted

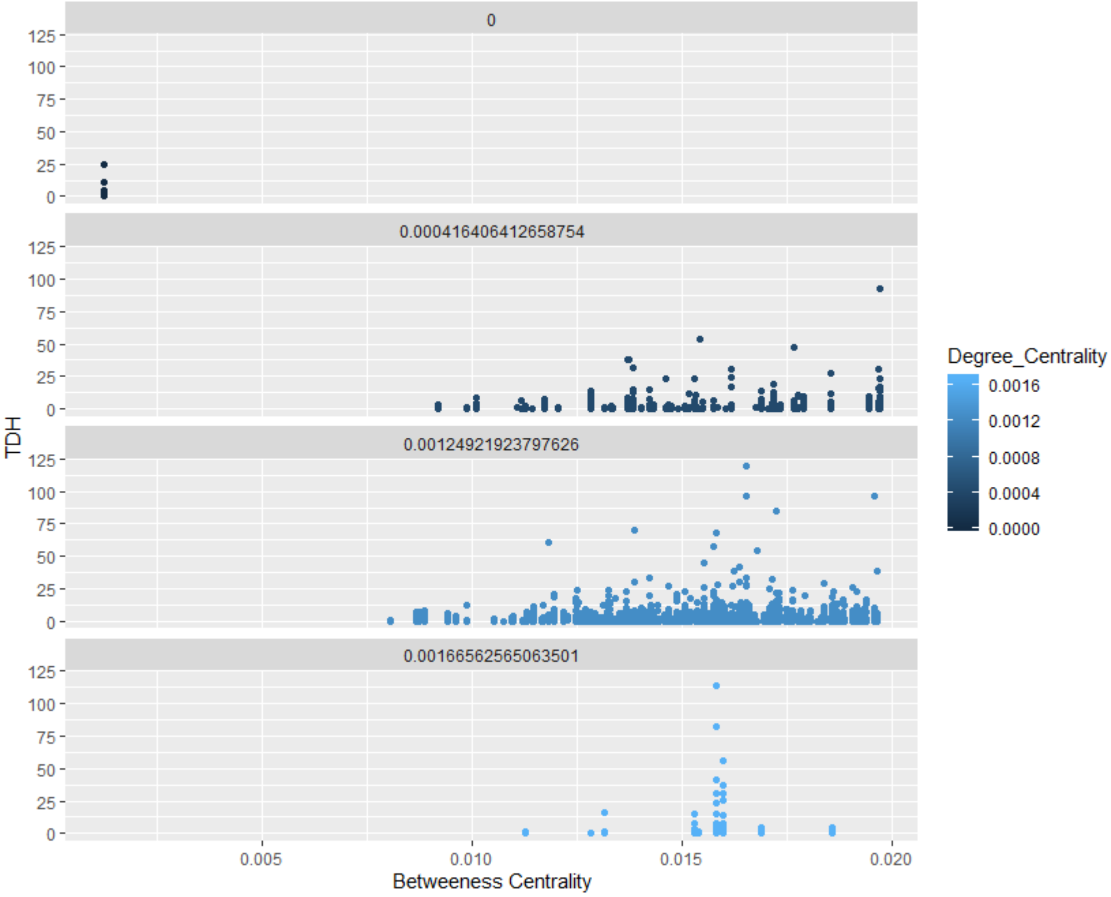
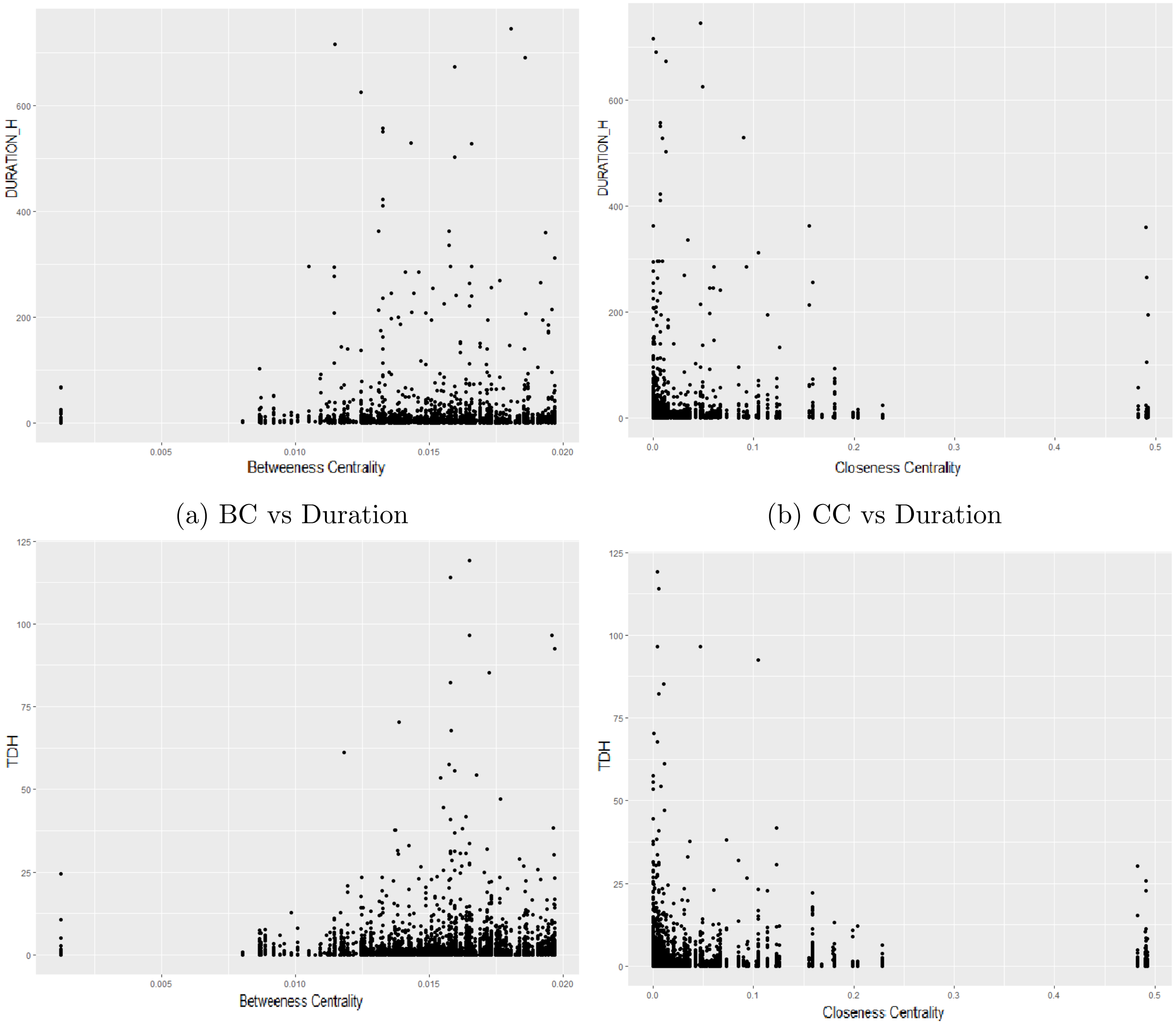


Figure 4.12: Degree Centrality in relation with Closeness Centrality and total delay



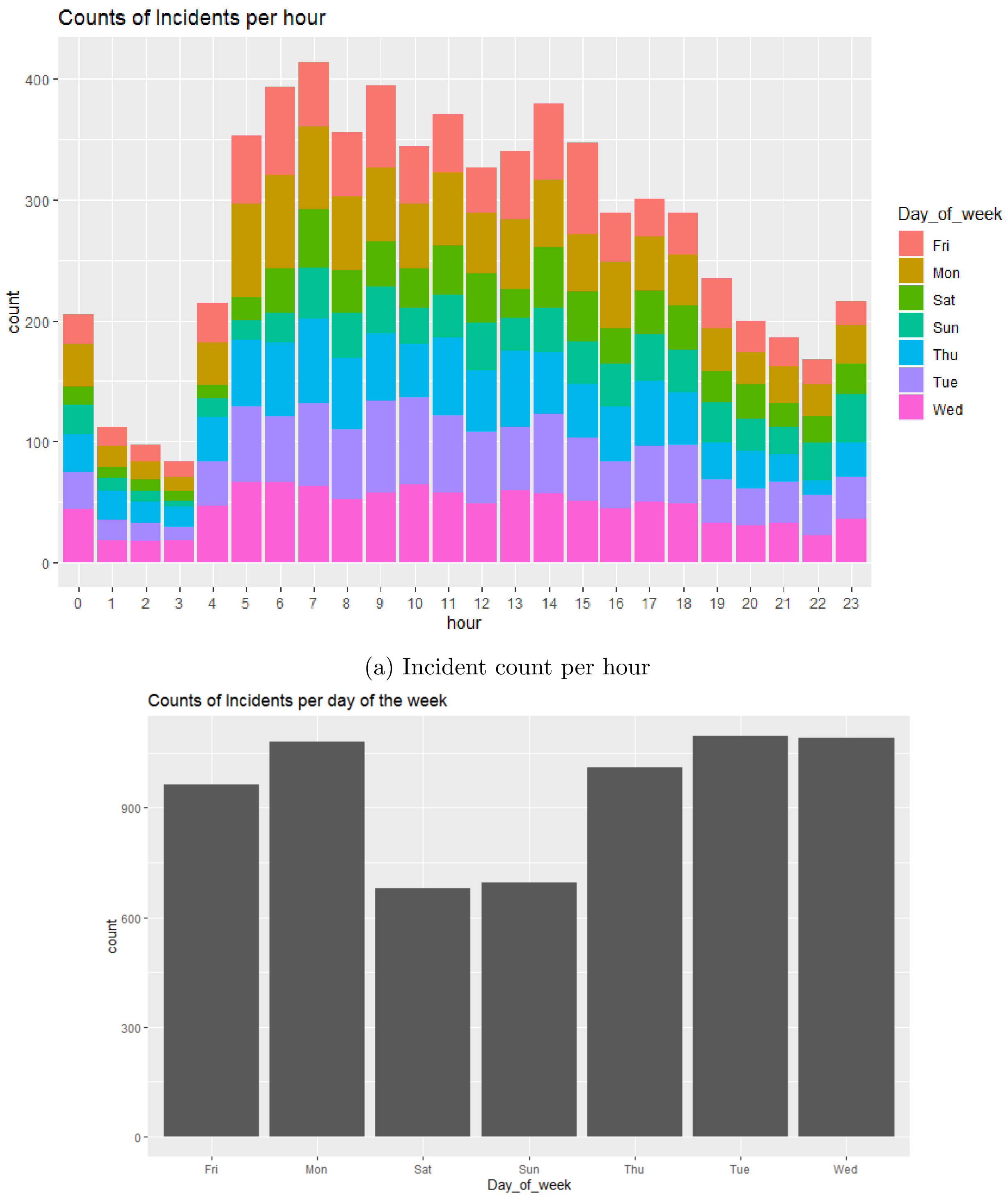
(a) BC vs Duration

(b) CC vs Duration

(c) BC vs Total Delay

(d) CC vs Total Delay

Figure 4.13: Betweenness and Closeness Centrality w.r.t Duration and Total Delay



(a) Incident count per hour

(b) Incident count per day of the week

Figure 4.14: Incidents per based on hour and the day of the week

4.2 Pre-Processing

To Prepare the data for the modelling, preprocessing is one of the very important steps because real time data is prone to lot of anomalies, missing information and inconsistencies. Data preprocessing aims at improving the quality of raw data and, consequently, the quality of mining results, and preparing it for further analysis(Jambhorkar and Jondhale, 2015). Preprocessing prepares the data by removing noise, removing missing data and bringing out the hidden trends. It also refers to the addition, deletion, or transformation of training dataset (Kuhn and Johnson, 2013). In the start of the chapter, it was mentioned that the values with very high duration values were removed and not considered for the analysis as they lasted for over months. It led to a dataset with 6693 entries where the maximum total delay value is 120 hours and the max duration value of 746 hours. From the descriptive analysis above it can be seen that duration and total delay are related to various other variables. The wide range of values for both duration and total delay, outlier removal using box plots would not help because there are too many values. So, considering the whole dataset with 6693 entries one of the first test to carry out was to check for collinearity. The presence of collinearity can pose problems in the regression context, since it can be difficult to separate out the individual effects of collinear variables on the response. (James et al., 2013). Since, the dataset consists of both categorical and continuous numerical variables using a **ggcorrplot** would not help as the it only takes the numerical variables. So, firstly the correlation between the numeric values is checked which can be seen in the Figure 4.15b and it can be seen that Intervals and Closeness Centrality have a negetive relationship whereas Betweenness Centrality and Closeness Centrality have the highest(but not enough to remove) correlation values.

Correlation

For the categorical and numeric values together to find the correlation, **GoodmanKruskal** is used to do this task which enables calculating the correlation between both categorical and numeric values. From the Figure 4.15a it can be seen that there are many variables which seem to have absolutely perfect correlation which also raises questions regarding the removal of such variables actually help the data. So, a further crosscheck for multicollinearity was done by computing the ***Variance Inflation Factor(VIF)***. The smallest possible value for VIF is 1, which indicates the complete absence of collinearity. As a rule of thumb, a VIF value that exceeds 5 or 10 indicates a problematic amount of collinearity. Using the ANOVA function in R the table. In the Figure 4.15c it can be seen that none of the variables have a VIF value greater than 5 indicating less or no collinearity between variables. So, none of the variables were removed from the dataset.

Outliers

Coming to problem of outliers using the box plot method did not seem like a viable option as the numerous points lying outside the quartile regions. So other methods were searched and the outlier diagnostics with Cooks distance was found to be a good method to do the outlier detection(Türkan et al., 2012). Cooks distance can be defined as the measure of distance between the estimates of the regression coefficients with the i-th observation β and without the i-th observation β_{-i}. D_i can be defined below

$$D_i = \frac{(\beta - \beta_{-i})^T (X^T X)(\beta - \beta_{-i})}{(\sigma)^2 p}$$

Cook suggests that Di be compared to a central F distribution with p and np degrees of freedom. This gives however too high cutoff values. Practically, a cutoff value of $\frac{4}{n-p}$ seems more reasonable and an observation is considered as an influential observation/outlier when D_i exceeds the cut-off point of $\frac{4}{n-p}$ (Türkan et al., 2012). Using this three different linear models were run to find the outliers. First, a linear model was total delay against all the variables, second a linear model with duration against all the variables and third a linear model with both duration and total delay against all the variables. Upon investigating the influential values from the three linear models the influential values from the third model were most relevant outliers. From the figure it can be seen that only the most extreme values are shown in the Figure4.16. All the values seemed relevant entries with all aspects when crosschecked manually. One point to note is that, during modelling every possible change in parameters are done and tested for the best model. Models with data, with and without outliers are tested to find the best results.

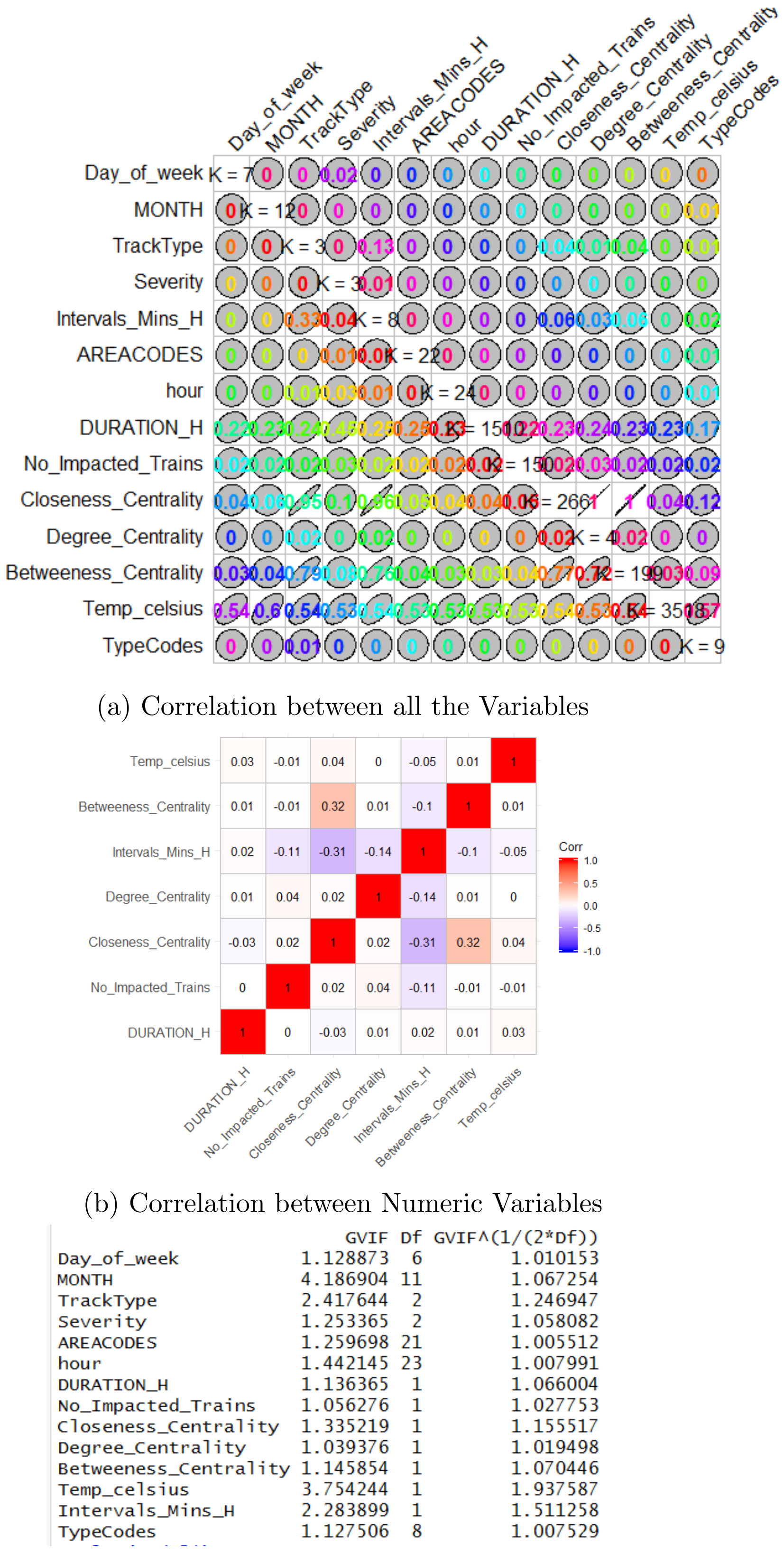
Data Transformation

Normalisation is one of the methods in data transformation. It is a very important aspect in predictive modelling because transforming the data into a standard format improves the training process, making it more accurate and faster (Kuhn and Johnson, 2013). There are many different methods for normalisation of data and one the most famous method and the one used for this thesis is ***Min-Max Normalization***. The following formula is used for normalisation

$$z_i = \frac{x_i - min(x)}{max(x) - min(x)}$$

Here, x_i is the data element,*min(x)* is the minimum of all data values, and *max(x)* is the maximum of all data values. This method transforms all the scores into a common range of [0, 1]. But the normalisation process can be used only for the numeric values in the dataset. So, the approach followed for this dataset is, that numeric and the categorical values are separated and then numeric values are normalised using min- max method. For the categorical values the variables are transformed into factors first. There are various methods for normalising the categorical values, in this thesis the function **DummyVars** provided by the ***Caret*** package in R. The method of **One Hot Encoding** was used to normalise the categorical values. The function **DummyVars** changes levels of the attribute into a variable with binary values(0,1) where entry with a value of 1 can be said to have that level as true and where 0 would be false. After the normalisation the total dataframe size was 6693 rows and 88 columns. Dimension Reduction is another method in data transformation. There are many methods for reducing the dimensions, however the one chosen for this thesis was **near zero variance**. In many cases, the data generating mechanism can create predictors that only have a single unique value (i.e. a zerovariance predictor).On the other hand, predictors might have only a handful of unique values that occur with very low frequencies. The concern here is that these predictors may become zerovariance predictors when the data are split into crossvalidation/bootstrap subsamples or that a few samples may have an undue influence on the model(Kuhn, 2012a). These **nearzerovariance** predictors may need to be identified and eliminated prior to modelling. This method was selected because upon normalisation the levels of categorical variables become attributes and few variables have very less variance. So, after creating the full data frame, using **nearZeroVar** function the variables with less variance are subsetted. Once the subset is created the variables are investigated and then decided whether to be removed or not. This method was applied during modelling before the data was applied and then investigated, whether upon removal of these variables the results of the model improve or not and accordingly decided to apply this on the data for better results of the model.

The next chapter discusses about the approach for the model building and the methods used.



(a) Correlation between all the Variables

(b) Correlation between Numeric Variables

```
                          GVIF  Df  GVIF^(1/(2*Df))
Day_of_week           1.128873   6         1.010153
MONTH                 4.186904  11         1.067254
TrackType             2.417644   2         1.246947
Severity              1.253365   2         1.058082
AREACODES             1.259698  21         1.005512
hour                  1.442145  23         1.007991
DURATION_H            1.136365   1         1.066004
No_Impacted_Trains    1.056276   1         1.027753
Closeness_Centrality  1.335219   1         1.155517
Degree_Centrality     1.039376   1         1.019498
Betweeness_Centrality 1.145854   1         1.070446
Temp_celsius          3.754244   1         1.937587
Intervals_Mins_H      2.283899   1         1.511258
TypeCodes             1.127506   8         1.007529
```

(c) VIF values for all the variables

Figure 4.15: Correlation between variables

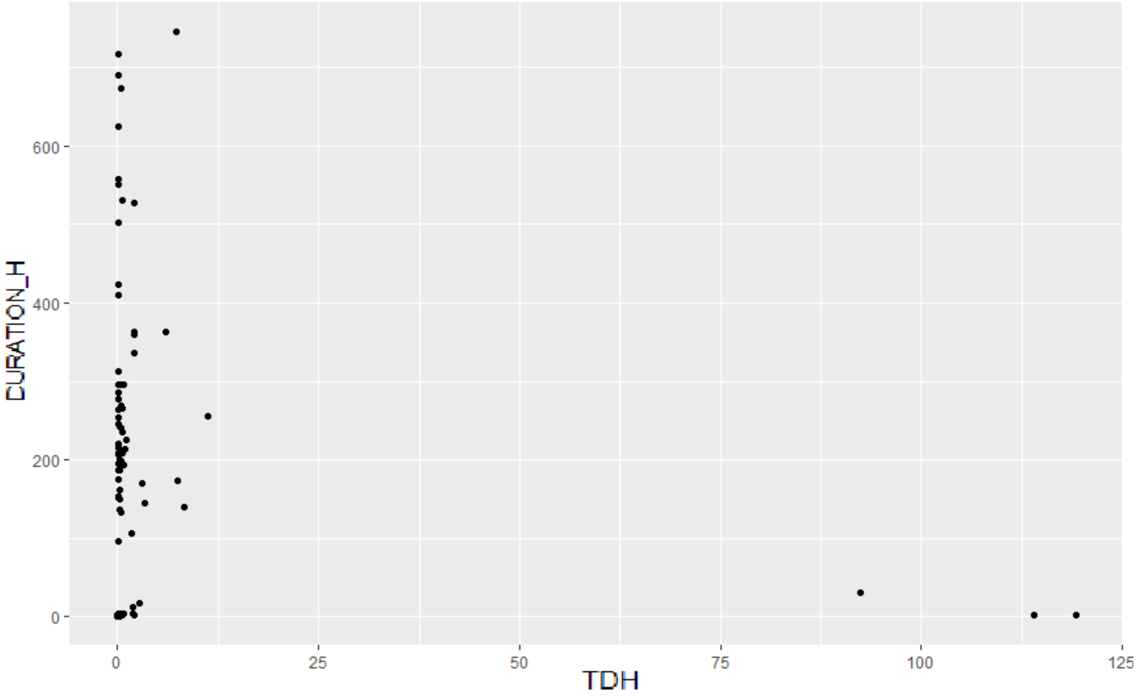


Figure 4.16: Duration VS Total Delay for outlier entries

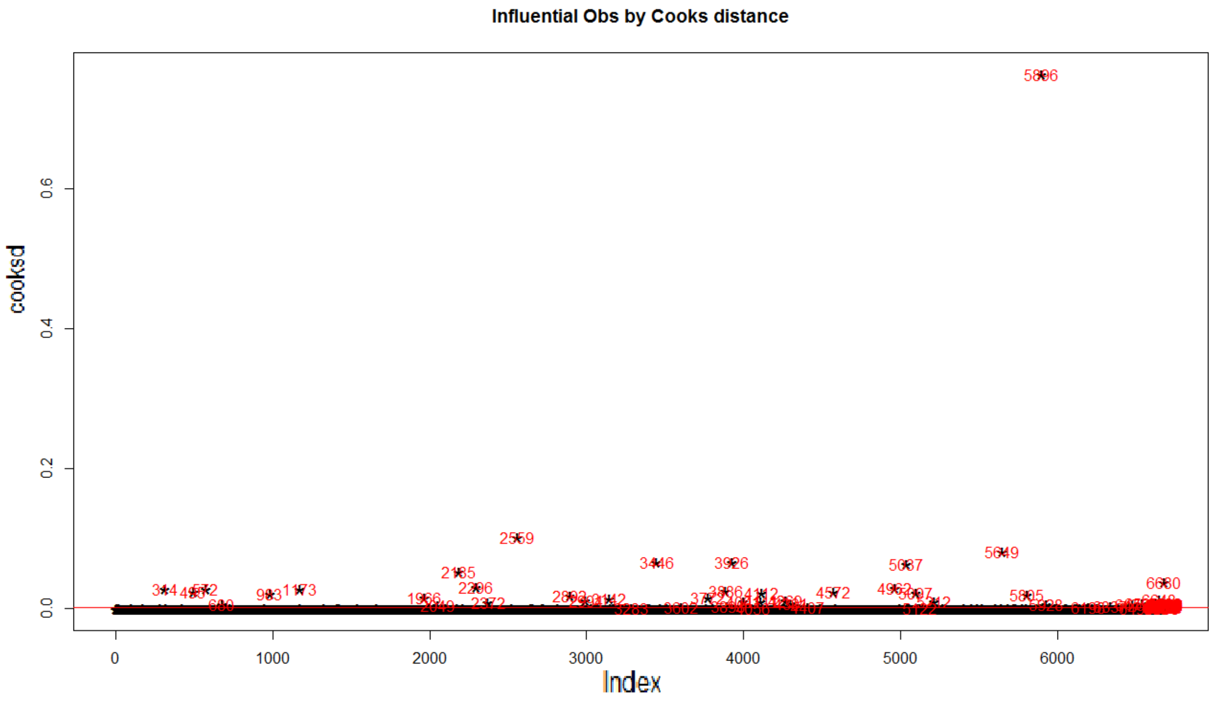


Figure 4.17: Influential Observations by Cooks Distance

Chapter 5

Models for Total Delay and Duration

This chapter starts with an introduction to the concept of Neural Networks and introduces all its parameters and specifications. After which the approach followed for the model building is discussed. Also, the other models that have been considered will also be discussed. The reason for choosing Neural Networks, as mentioned in the literature review, was found to be the most popular algorithm in the field and found to give pretty good.

5.1 Introduction to Neural Networks

Artificial neural networks also called Neural Networks(NN) are popular machine learning techniques that simulate the mechanism of learning in biological organisms(Aggarwal, 2018). Significant amounts of data is being processed by the human brain which is carried out by neurons in the brain. Similar to a biological neuron, ANN's define the *neuron* as central processing unit. These neurons perform a simple task of activating upon receiving a signal that exceeds the activation threshold. The main function of the of the NN is the computation of the output of all the neurons. A basic neural network constitutes of **Input layers, Hidden layers, output layers, neurons, weights, bias and activation functions**.The hidden input and output layers are made up of neurons and the hidden neurons can be single or multiple layers. The output from one layer is used as the input for the next hidden layer. Such networks are called the *feed forward* networks. Weights play a very important role in converting this input to output. Weights scale the inputs connecting the neurons and defining the computational function of the cognitive process. The learning process takes place in the selection and change of the weights that link the neurons. Training data is fed to the network which contains the inputs units and the output is reached as the process teaches the network model. Mathematically it can be represented as shown below

$$Y_k = \phi(W.X) = \phi(\sum_{j=1}^{d} w_j x_j)$$

Y_k is the output layer, ϕ is the activation function, x is input variables and w is weight assigned to each input variables.

Neural Networks can be used for both supervised an unsupervised learning and for classification problems and regression problems. For this thesis we use have a supervised regression problem in hand.

5.1.1 Parameters

Neural Networks can consists of single or multiple computational layers which Parameters are the components that change and help fine tune the model. The weights denoted as $w_1, w_2, w_3 ...$ contains features that are multiplied with the input and added at the output. Bias is added as the weight of an edge, it can also be said as the intercept added to linear equation $output = \sum weight * inputs + bias$. The Figure *5.1shows the difference in model with an without bias.

An activation function is applied to this output, which alters the model. Without this neural network acts like any other linear model.The output is always a post activation and pre. This activation is what adds the non linearity into the model. In the book (Aggarwal, 2018) are summarised the principal activation functions for Neural Networks. v below is the argument of the Neural Network.

$$\phi = sign(v), \quad \text{Sign Function}$$

$$\phi = \frac{1}{1+e^-1}, \quad \text{Sigmoid Function}$$

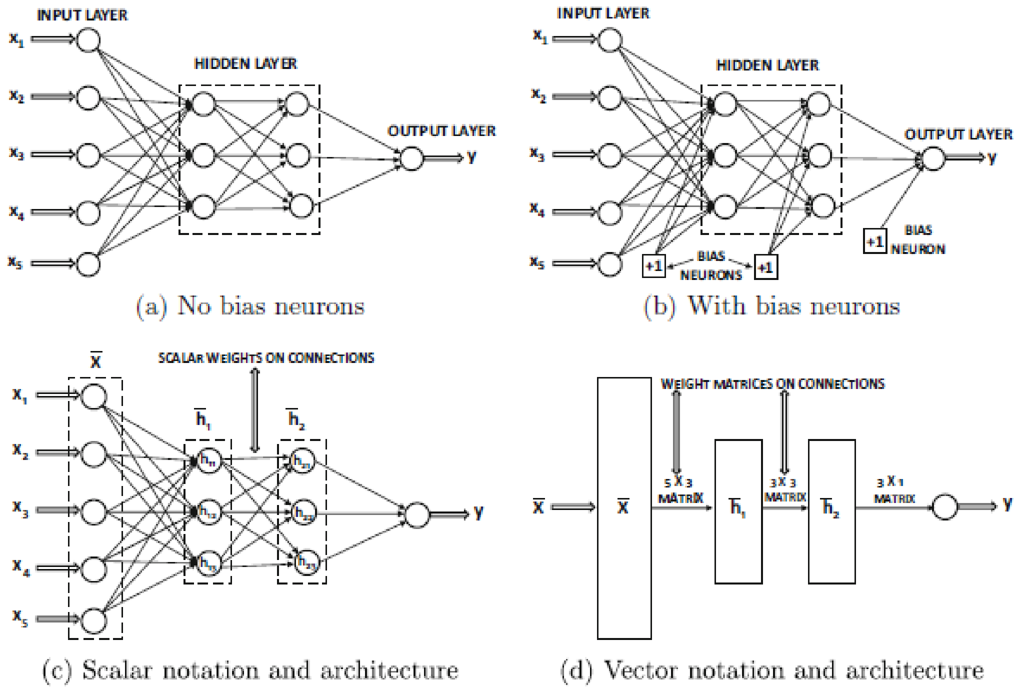
$$\phi = \frac{e^{2v}-1}{e^{2v}+1}, \quad \text{Tanh Function}$$

$$\phi = max(min[v,1]-1), \quad \text{Hard tanh}$$

$$\phi(v) = v, \quad \text{Linear}$$

$$\phi = max(v,0), \quad \text{Rectified Linear Unit(ReLu)}$$

Linear function is the most basic function and provides no non linearity to the model and used generally when the output is real value. While, the *sign activation* is used to map binary outputs at prediction time and prevents the use of loss function in training. The most widely used activation function is the *sigmoid function*. Here the outputs are generally binary (0,1), which is helpful in performing computations whose output are interpreted in probabilities. It is also helpful in constructing loss functions derived from the maximum-likelihood models. On the other hand *tanh* is related to sigmoid function $tanh(v) = 2.sigmoid(2v) - 1$. The tanh function is more preferable than sigmoid when the outputs are desired to both negative and positive. In recent year, ReLu has gained more popularity which is a piece wise linear activation function. The ReLU and hard tanh activation functions have largely replaced the sigmoid and soft tanh activation functions in modern neural networks because of the ease in training multilayered neural networks with these activation functions.The use of nonlinear activation functions is the key to increasing the power of multiple layers(Aggarwal, 2018).



(a) No bias neurons

(b) With bias neurons

(c) Scalar notation and architecture

(d) Vector notation and architecture

Figure 5.1: Multilayer Preceptron model with and without bias

These outputs from one layer are passed forward as input, in the forward direction, to the neurons in the next computational layer. The default architecture of feed-forward networks assumes that all nodes in one layer are connected to those of the other. Therefore the neural network architecture could be completely defined once the number of neurons and the number of layers are defined. Now the model has to be trained which is and the most commonly used algorithm is ***backpropagation algorithm***. This training process is relatively easy in case of single layer NN because error (or loss function) can be computed as a direct function of the weights, which allows easy gradient computation. Where *loss function* is the function which helps optimise the learning of the model. The loss function has to be minimised to attain more accuracy. In case of a Multilayer model is a bit tricky as the loss function in this case is a complex composition of weights from the previous layers. It consists of two phases the forward and the backward phase which are explained below. *Forward Phase* is where the fed inputs for a training data cascade into the layers using current wights and give the output which are compared to the training data and the derivative of the loss function w.r.t the output. This derivative of loss is not to be computed with respect to the weights in all layers in backward phase. The total errors at the output is calculated using the following formula. $E = 0.5(t - y)^2$ where $E =$ the squared error, t is the target output for training sample and y is the actual output *Backward Phase*'s goal is to learn the gradient of the loss function and update the weights accordingly. These gradients are learnt backward that is from the output layer to the input layer, this phase is so called as the backward phase. The illustration below depicts the backward propagation

Based on the above explanation and the lecture notes for the course Statistical

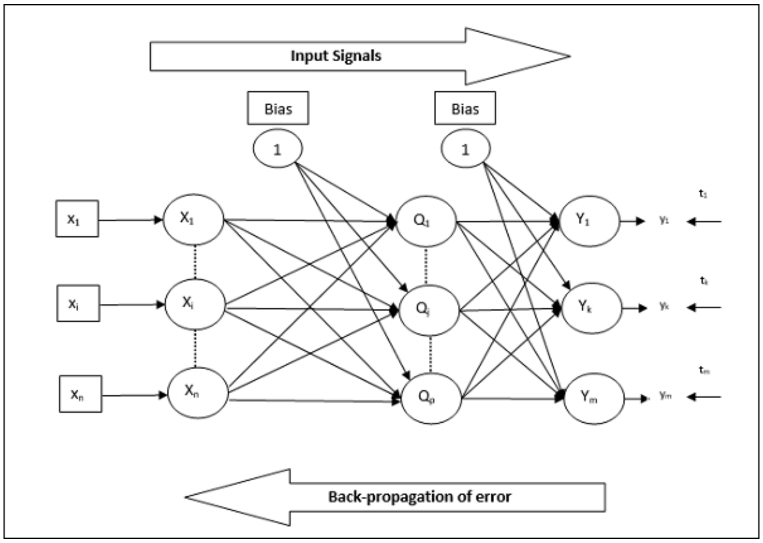


Figure 5.2: Back propagation in Neural Network

Learning and Data Analytics For Transportation Systems in the summer semester of 2019, the neural network could be summed up into the following steps

- Send the input X_i
- Calculate the output Y_k
- Given the correct output O_k calculate the error
- Adjust the w_{ij}
- Adjust the w_{jk}

5.2 Specifications for the Model Building

Since this dataset and the work was done was the first time in such context, it was important to focus on other methods too and not just focus on the Neural Network. Just to review again the aim of the prediction models is three fold.

1. Prediction of the total delay caused by an incident
2. Prediction of the duration of the incident
3. Using the predicted duration as an input to the total delay prediction model and check the output.

Just to note that the whole modelling process was carried out using R and packages provided by R.

Linear Regression Model

To start with the predictive modelling, first step was to understand the behaviour of the data, so as to accomplish this, a simple multiple linear regression model was used. The linear regression was used to check the various properties in the data, its linearity, homoscedasticity and significant predictors. The linear regression is applied on both the duration and the total delay. The main takeaways of the linear models was the does the data have non linearity, significant predictors and is linear model sufficient for the prediction. A log model is also tested.

5.2.1 Specification for the Neural Networks

For building the models the package used is ***CARET*** (Kuhn, 2008). Caret stands for ***Classification and Regression Training,*** contains numerous tools for developing predictive models using the rich set of models available in R. It focuses on simplifying the training process, provides various packages for variable importance, preprocessing, model visualisation etc. It enables in running multiple models parallelly and ensemble models. Because of its flexibility, caret is used for the thesis. The basic rule of thumb for the selectiopn of number of neurons in the hidden layers are the following:

1. Number of hidden neurons should be between the size of the input layer and the size of the output layer.

2. Number of hidden neurons should be 2/3 the size of the input layer, plus the size of the output layer.

3. Number of hidden neurons should be less than twice the size of the input layer.

But to start with the number of testing neurons in this case, it was chosen such that the number of neurons in the first layer were less than number of neurons in the second layer and the passed number of neurons were less than the total number of inputs in each layer. The best model is chosen on the basis of the R^2 value.

Coming to the neural networks, the final package used for the model building was RSNNS that provides a convenient interface for the Stuttgart Neural Network Simulator(SNNS) with caret. Both **nnet** and **neuarlnet** packages provided by **R** were tried but nnet allows only one layer NN and neuralnet, though allows two layers building a dynamic method model was not possible. Additionally computation time was higher for it when compare to others. So, in Caret **MLPML** model is used which uses RSNNS package for the model computation. RNSSN allows multiple layers and much more flexible training process. (Yaghini et al., 2013) mentions about the dynamic and multiple structures for the implementation neural networks. For this thesis a mixed method of both dynamic and multiple is used where in the number of neurons in the layers for testing are given by the user and then based on the increment values given for each layer every combination is tried out with corresponding increments. After finalising the package specifications the process for the model building is to be discussed. Firstly, the whole process of finding the best model is iterative. The components changed for each iteration were the variables, hyper parameters, data with and without outliers, number of neurons, increment number of neurons, layers, removing/not removing the near zero variables(nzv) and the training control options. This applies to both the duration and the total delay models. For every run one of the components mentioned above is changed to check if the results are improving or not. Another very important change of the parameters was the inputs itself. Apart from the nzv attributes, based on the linear regression variable significance results removal and addition of less significant variables were also tested.

Most of the changing parameters have been discussed but hyper-parameters and train control are very important aspects not discussed in detail. One of the most important hyper parameter is the **learning rate**. Learning Rate is the parameter that determines how quick the model learns and adapts also called as the **shrinkage**. The backpropagation keeps changing the weights until there is greatest reduction in errors by an amount known as the learning rate. The learning rate has been changed from 0.0001 to 0.1 and the lower the learning rate slower the model learns but the results are much reliable. Higher the learning values faster the models learns but the results could be less reliable and the results could change drastically when predictions are carried out. For the backpropagation algorithm there are various algorithm available for implementation in RSNNS for example Standard BackPropagation, BackPropagation with momentum, Back Propagation through time, Quickprop, Resilient Backpropagation, Backpercolation and few more. But for this model we use the Standard Back Propagation for the purpose. This backpropagtion algorithm requires something to assign weights to the neurons, the initFunc of type "Randomize_Weights" is used to initialise the weights to the network and it is usually assigned with a range of values for this model the values use were -0.3 to 0.3. The Std_Backpropagation uses two parameters called the gradient descent also called as the learning rate and d_{max} which is the maximum difference between the target and the output that is tolerated. The d_{max} value was set to 0, implying the threshold for the tolerating error is 0.

The other important parameter to be considered are the activation function. In the package RSNNS there are many different activation functions available of which themost relevant for feedforward networks are**Act_logistic**(used for classification) and **Act_Identity**(for regression). The **Act_Identity** function is used here as this is a regression problem. It also has the function called **linOut** used for defining if the problem is a regression problem or a classification problem. Most of the Neural Network packages are suitable for both regression and classfication problems. In this case, since the output is a continous quantitative number the linOut function is to be considered TRUE. The default value of the function being FALSE. Another function called Maxit which is defined as the number of iterations the i.e, the number of training epochs to perform. This value was set to 200.

For evaluating the models and compare the results the metric used is the R^2 **(Coefficient of Determination)**. In the literature review it was found that few studies have used the same R^2 as metric for evaluation. In this case R^2 was chosen for easy comparison between models. Also, usually R^2 values greater than 0.5 can be considered as good models and the aim for the models could be to reach a R^2 value greater than **0.5**. RMSE is generally mostly used for evaluation but since the lack of benchmark for a RMSE value the evaluating of the models is considered a bit difficult. R^2 can be defined using the following formula:-

$$R^2 = 1 - \frac{\sum_i (Y_i - y_i)^2}{\sum_i (Y_i - y_{avg})^2}$$

where $Y_i - y_i$ is the unexplained variance and $Y_i - y_{avg}$ is total variance to be explained and Y_i is the observed value, y_i is the predicted value and y_{avg} is the average value of the observed values.

Data Splitting For predictive modelling the data had to be split into two parts that are the **training** and the **test** datasets. The split is done of the whole data set with a ratio of 70 to 85 % for training set and 30 to 15% for testing set. **The split used for this thesis was 75% - 25% for training and testing respectively.**

Resampling is an important step in predictive modelling. It is helpful in validating the stability of the machine learning. To get an assurance that the built model got all the patterns from the data correct and it is not picking up too much of noise, other words it maintains low on bias and variance. The re-sampling method used here is the ***k- Fold Cross Validation*** where the sample is divided into k sets roughly of equal sizes. A model is fit using all the samples except first subset. The held out sample are used for prediction by the model and to estimate the performance measures. This process repeated for k over the whole training data set. The k re sampled estimates of performance are summarised (usually with the mean and standard error) and used to understand the relationship between the tuning parameter(s) and model utility(Kuhn and Johnson, 2013).For this thesis ***5-Fold Cross Validation*** was used which was define in the **Train Control** function of the caret package.

5.3 Gradient Boosting Machines

Apart from the neural network model, other models were also developed to show a comparison and to study the behaviour of the various models with the data. **Gradient Boosting Methods**(GBM) was selected as one of the algorithms for model building because of its high flexibility and its customisability with any particular data driven task. These characteristics of the algorithm are from its performance improving ability.Gradient Boosting Machines defined in the (Friedman, 2001) and (Friedman et al., 2000) is one of the many types in the family of the **boosting techniques**. The base principle of the boosting techniques is the addition of new models to a ensemble sequentially, where in an ensemble approach is combining of different weak simple models to do get a strong ensemble prediction(Natekin and Knoll, 2013). The GBM models are a gradient descent based formulations. The principle for the GBM's is to construct new base learners to correlate it to the negative gradient of the loss function, associated with the whole ensemble (Natekin and Knoll, 2013). The are wide variety of loss functions up for use but the most widely used loss function is the squared-error loss function which would result in a consecutive error fitting. There is no specific rule for the selection of the loss function for the model and usu sally it depends on the user's choice. The main working of the GBM's can be classified into three parts

1. **Base Learning Models** : Boosting framework starts off with a base learning model which is usually considered to be a weak model. Among the various models, the most commonly used model is the decision trees. Additionally, Linear models and Smooth Models are also used as the base learner models.

2. **Training Weak Models** : The error rate of a weak model is slightly better than random guessing. A sequential model builds a simple weak model to slightly improves the remaining errors. Combining weak models has benefits as it avoids over-fitting, Slow learning making it to perform well and improved speed as the weak models are easy to build.

3. **Sequential training with respect to errors** : This is the most important part, as the boosted trees are grown/learned each new tree uses the information from the previous tree. It can be represented using the following equations where x represents the features and y represents the response

    - Fitting the base learner model to the data – $F_1(x) = y$
    - Fitting next model for the residual of the previous – $h_1(x) = y - F_1(x)$
    - Adding the new model to the main algorithm – $F_2(x) = F_1(x) + h_1(x)$
    - Fitting next model for the residual of the $F_2(x)$ – $h_2(x) = y - F_2(x)$
    - Adding the new model to the main algorithm – $F_3(x) = F_2(x) + h_2(x)$
    - Continuing the process further until the process is stopped by actions like cross validation.

    The generalised GBM be represented using the formula below which is a simple stage wise additive model of b individual trees.

    $$f(x) = \sum_{b=1}^{B} f^b(x)$$

4. **Gradient Descent**: It is generic optimisation algorithm used for finding the optimal solution for wide variety of problems. Most algorithms focus on minimising the residuals and the error. The main idea of the gradient descent is to tweak the parameters iteratively in order to minimise the loss function for a given set of parameters (θ and takes the direction of the descending gradient and once the gradient reaches zero.

There different types of loss functions used for GBM machines based on the type of response variable i.e. whether it is a categorical, continuous or other types. An important parameter by in gradient descent is the size with loss function are the steps called as the **learning rate**. After discussing the workings of the algorithms the tuning parameters and the model implementation is discussed here. To get the best results and tune the model the following parameters play a major role

- Number of trees : Total number of trees to fit.

- Depth of trees : The number of splits in each tree; control the complexity of the ensemble. Usually its value is greater than 1 and less than 10

- Learning Rate : Also called as the **shrinkage**, controls the rate at which the model learns. Smaller values avoid over-fitting but increases time.

- Sub-sampling : Fraction of the available training observations. A values less than 1 mean implementing a Stochastic Gradient Descent. It helps reduce over-fitting.

Apart from the above mentioned parameters there are few more parameters that can be used based on the model requirements.

There are various packages use available for implementing GBM's in various software's, among all one of the most famous one's is the XGBOOST(Extreme Gradient Boosting) package by (Chen and Guestrin, 2016). This package is very famous for its faster processing time than other available packages. For the implementation of this package initial data preparation is required for the model. As used in the neural network the categorical data is converted into numerical variables using one-hot encoding using the dummyVars function from the caret package. In general practice in case of decision tree implementations, usually scaling/normalisation is not implemented on the data. Also, xgboost takes the data in the format of matrices where in the **xgb.DMatrix** function is used to convert the response variable data and the predictor data into to separate input values. The parameters used to change for the best results were the number of iterations that is the **nrounds** the variation was done from 50 and increasing it by 50 ach time. The next parameter was the depth of each tree whose values were varied from 1 to 10. The learning rate was also varied for various values from 0 to 1. The regularisation value where held constant, gamma value to 0. Also the cross validation was held for 5. The final model is shared in the results section. The model was run for both the response variables the Total Delay and Duration of the incident. The training and test data partition was set at 75-25%.

5.4 Generalised Linear Regression Models

For the thesis different models were implemented and linear regression was one of them. In the output it was found that the results from the output was not normally distributed. This prompted in the implementation of the Generalised Linear Regression also called as the Generalised Linear Models(GLM). The main idea of the Generalised linear regression is to unify different models into one generalised framework. It brings together various distributions which includes normal, binomial,Poisson and gamma distributions among others. These distributions are also called as the exponential dispersion model(EDM). Instead of explaining the whole mechanism of the algorithm the model implementation part is only discussed here (Dunn and Smyth, 2018). The continous EDM are the normal and gamma distributions where as the discrete EDM's are Poisson, binomial and negative binomial distributions. The package used here for the implementation of the algorithm is **Glmnet** in R. Glmnet is a package that fits a generalised linear model via penalised maximum likelihood. This package fit linear, logistic and multinomial, Poisson, and Cox regression models. The regularisation path is computed for the lasso or elasticity penalty at a grid of values for regularisation parameterλ (Hastie and Qian, 2014). Glmnet solves the following

$$min_{\beta_0,\beta}\frac{1}{N}\sum_{i=1}^{N} w_i l(y_i\beta_o + \beta^T x_i) + \lambda[(1-\alpha)||\beta||_2^2/2 + \alpha||\beta||_1]$$

where the λ are the grid values covering an entire range. Here $l(y, n)$ is the negative log-likelihood contribution of observation i; e.g. for the Gaussian case it is $\frac{1}{2}(y-n)^2$. The elastic-net penalty mixes these two; if predictors are correlated in groups, an $\alpha = 0.5$ tends to select the groups in or out together.The tuning parameter λ controls the overall strength of the penalty (Hastie and Qian, 2014). This package algorithm uses cyclical coordinate descent which successively optimises the objective function over each parameter with others fixed, and cycles repeatedly until convergence (Hastie and Qian, 2014). The main parameters for glmnet are the alpha value which lies between 0 and 1 and lambda values ranges from 0 to infinity. For the implementation of the model caret package was used.

5.5 Operational Framework

Until now the concentration was more on both the response variables separately and the models created were standalone and not interconnected. In the duration model the total delay was not considered as a parameter whereas for the prediction of the total delays, duration was one of the parameters, that too an important parameter. This lead to the idea of understanding the impact of duration on the total delay. Apart from the feature importance calculations(discussed in the next section) there was need for understanding of the actual behaviour between them. This lead to the creation of the operational framework where a bridge between the two best models for each response variable is created and the impact of one on the other is calculated. With both the models in hand an operational framework was created. The operational framework would help in creating response surface where in the change in behaviour of one response variable affects the other. Also, from the feature importance results(discussed in the next chapter) of the total delay it was evident that the duration was a significant variable in the prediction of the total delay which prompted in the creation of the operational framework. As mentioned before that the for the duration model total delay was not used as a parameter, so the main idea of the operational framework using the parameters in hand, duration of the incident is predicted first using the best neural network model. Using this predicted values of duration **as input** for the finalised model of the total delay, the total delay is predicted eventually and the difference in the metric values are observed. Randomly selected 1000 entries from the main data are used to for the estimation of the operational framework. The figure 5.3 below shows the flow chart for the operational system. By doing this a possible response surface could be created where a relation between the duration and the total delay can be established and their behaviour can be studied.

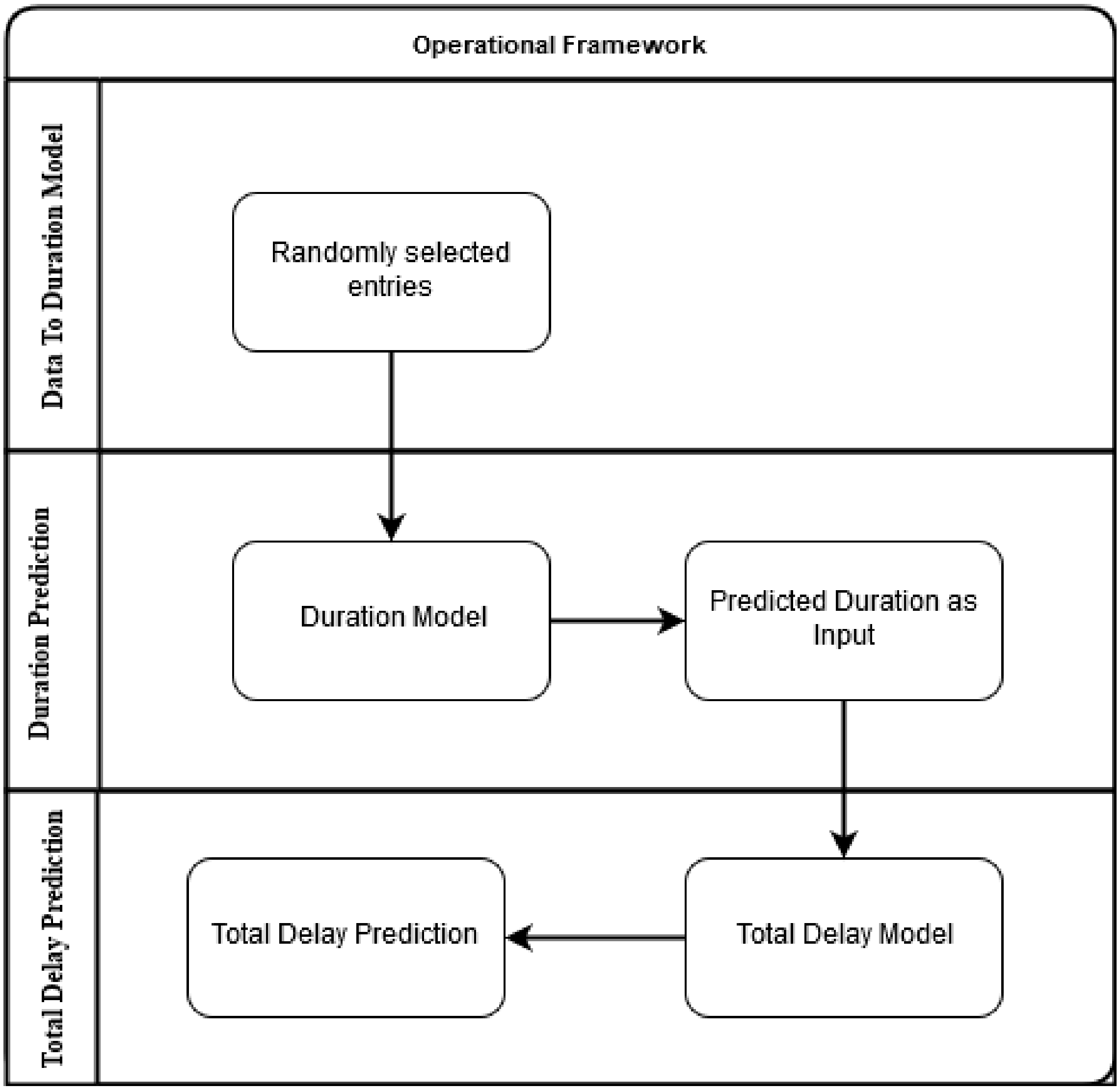


Figure 5.3: Process diagram for the operational framework

Chapter 6

Results

This chapter summarises all the results from the model discussed in the previous chapter. To start with the first model discussion is for the Linear Regression Model.

6.1 Linear Regression

As mentioned in the chapter in the Linear Regression section. Both multiple linear regression model were applied on both direct response variable and on the log value of both the response variables duration and the total delay.

Total Delay

For the general understanding of the data first a linear model was run on the data with response variable as the total delay and the best model was found with an R^2 of **0.5423** which could be considered as preferably an okay model. In the Figure 6.1 regression do not stand proven as the Normal Q-Q plot is not linear with tails in the start and the end. In the Scale-Location graph, the regression line is not a horizontal and the points are concentrated at one end of the line. This indicates heteroscedasticity in the data-indicating application of non-linear method. Further up the line the points are spread. Also in the residuals Vs fitted graph it can be seen that the variance of the residual is very non uniform and the values are concentrated at the starting though the red line is horizontal. This prompted for investigating a log model with and the results can be seen 6.2 the models doesn't improve at all rather worsens even more with an R^2 value of **0.3482**. Indicating that neither log model helps in improving the linear model. The significant variables are noted **(Day of the week, Duration, Month, TrackType, Severity, Areacodes, Hour, Impacted Trains, Closeness Centrality, Degree Centrality, Betweeness Centrality, Temperature in celsius)**

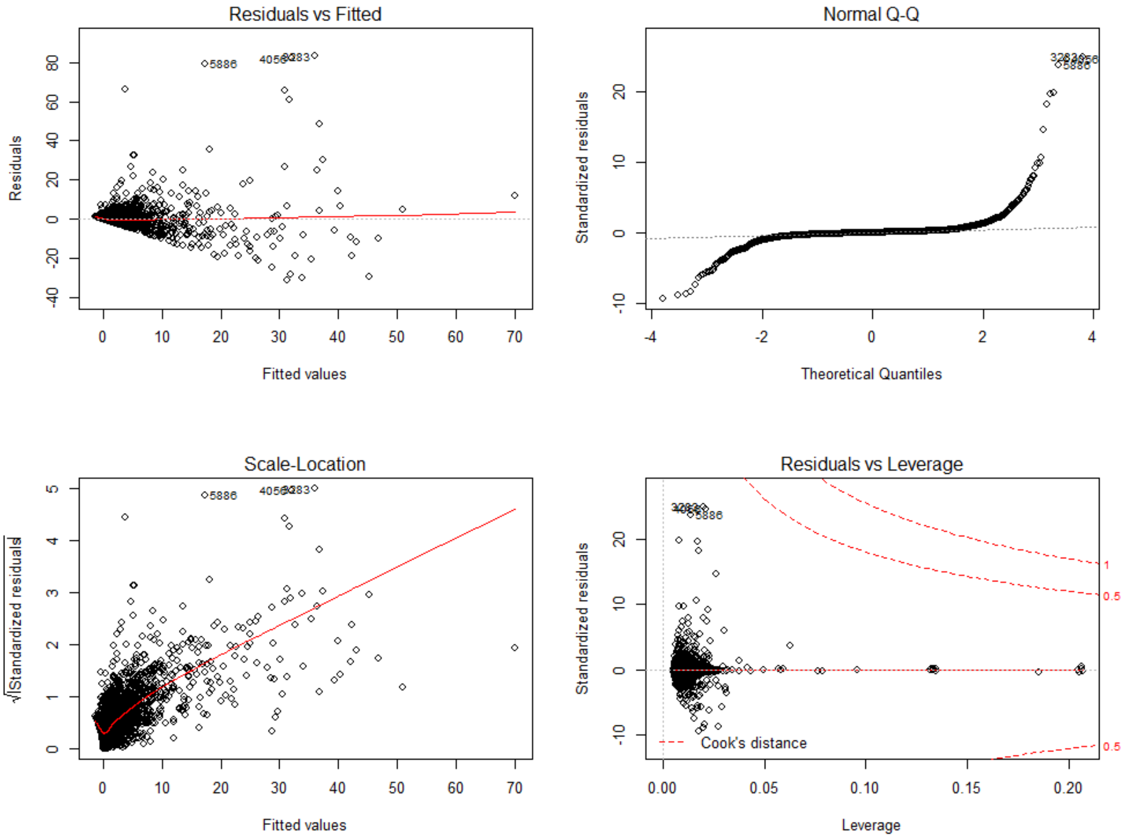


Figure 6.1: Diagnostics Plot Linear Model- Total Delay

Table 6.1: **R Square values for linear models Total Delay and Duration**

No.	**Model Type**	**Response Variable**	**Train R^2**	**Test R^2**
1	Log Model	Total Delay	0.3482	0.3089
2	Normal model	Total Delay	0.5298	0.6097
3	Log Model	Duration	0.3032	0.2918
4	Normal model	Duration	0.1292	0.1143

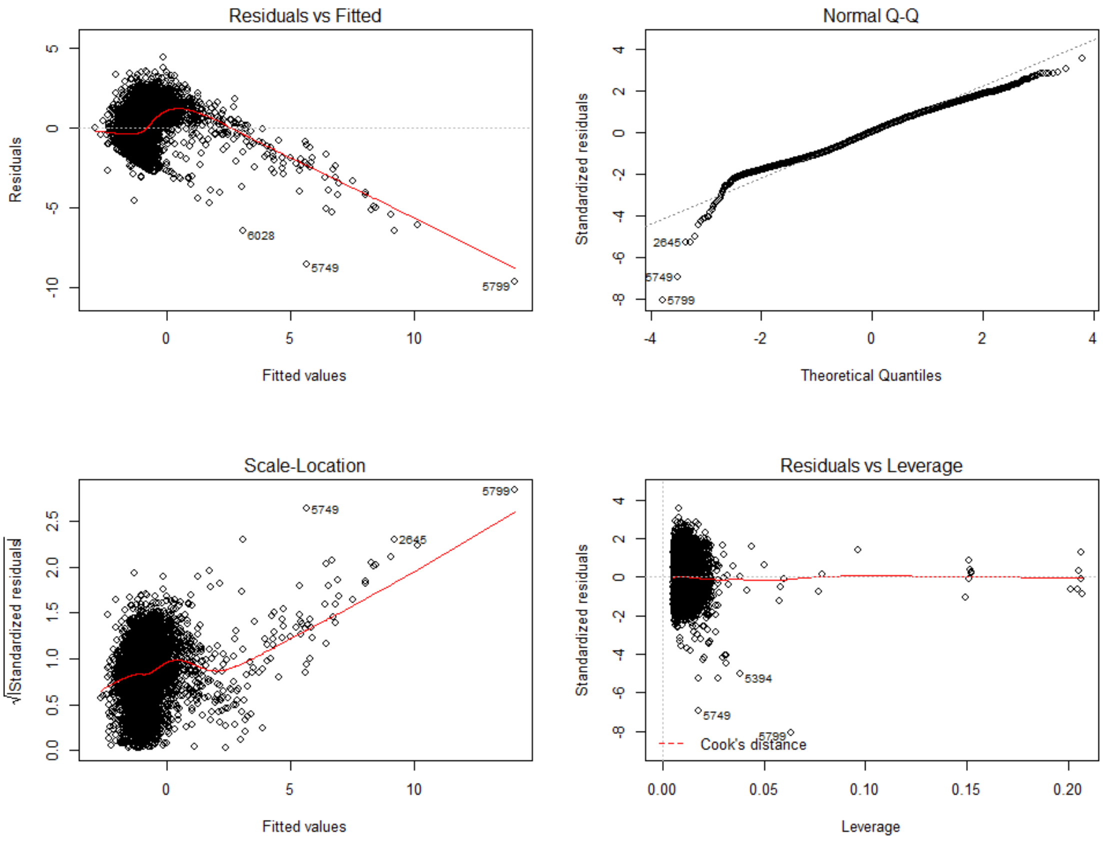


Figure 6.2: Diagnostics Plot Log Model-Total Delay

Duration

Similar behaviour as seen in linear model for total delay is seen in the linear model where with response variable is Duration. From the Figure 6.3 and 6.4 it can be the duration models behave worse than the delay models. With R^2 less than the models from the delay and the diagnostic plots behaving even more weirdly. From the Figure6.3 it can be seen that the that there two parts of concentration in both residuals vs fitted and scale location graphs. Though the Normal Q-Q graph seems to improve in the log model but all other perform really poorly. Both the assumptions homoscedascity and the normal distribution of the residuals are negated and it signifies the clear presence of nonlinearity in the data. Among other things the significant variables are noted **(Day of the week, Month, Severity, Areacodes, Hour, Impacted Trains, Closeness Centrality, Betweeness Centrality, Temperature in celsius)**

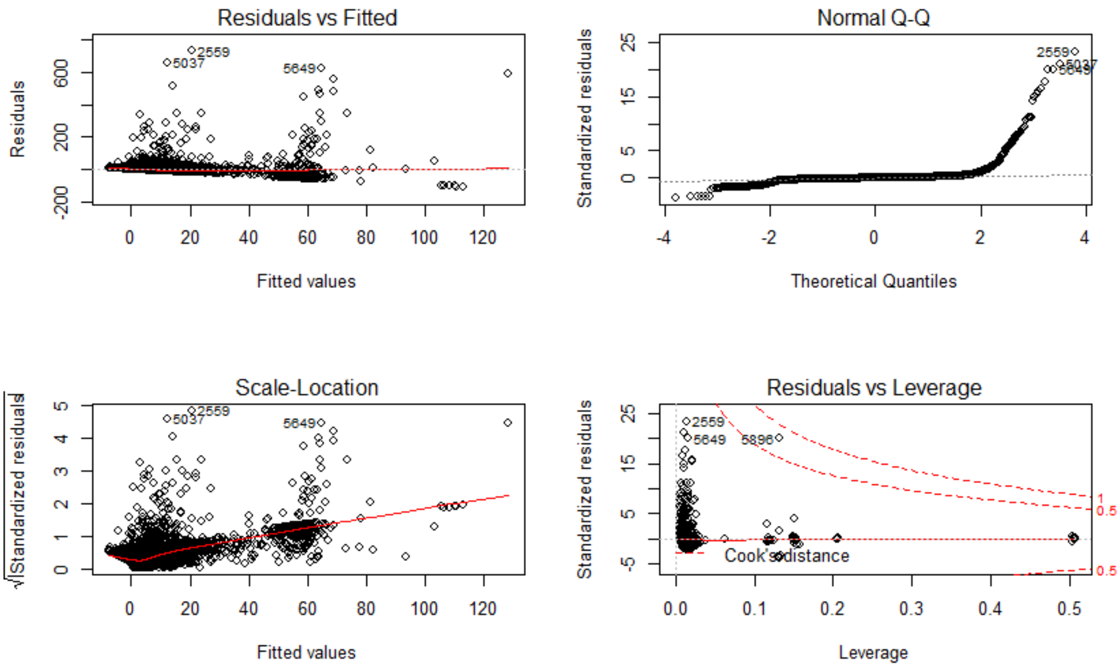


Figure 6.3: Diagnostics Plot Linear Model-Duration

This lead to the implementation of the Neural Network which is explained in the next section.

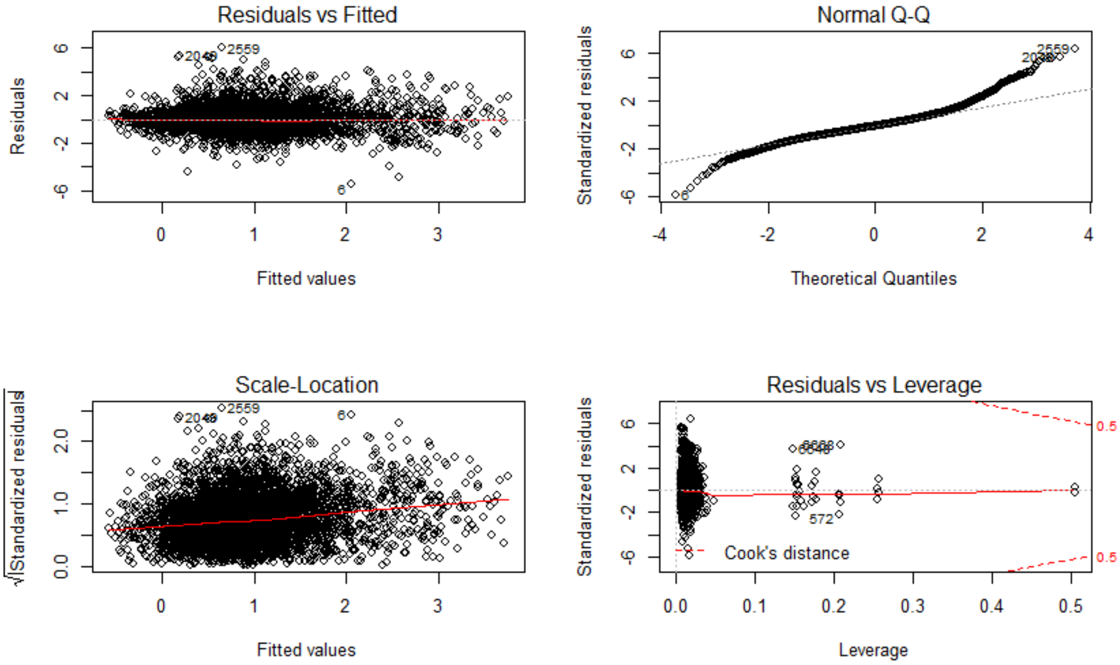


Figure 6.4: Diagnostics Plot LOG Model-Duration

6.2 Neural Networks

As mentioned in the section 5.2.1 about the iterative approach for finding the best model. The modelling done in the neural networks is three fold where in the first model is for the prediction of the Total Delay.

6.2.1 Total Delay

For this model as mentioned before iterative method was adopted with change of parameters for every model run and the results were checked accordingly. From the table 6.2 it can be seen few of the best model runs have been shown. Variables, learning rates, layers and step increase in layers were changed. In some model runs the variables removed were based on the linear model, some of the variables removed based on the results from the near zero variance. Also, with the data, few model runs included the influential points and in few they were removed. Learning rate is one more important aspect changed here. This was the testing phase to find the best parameters for the model. From the table it can be seen the best R^2 value is obtained at the learning rate of 0.05 with variables that were removed based on the linear regression. Also, full data was used i.e, influential points were included and the near zero variables were also included. All these models were run on the basis of the dynamic and multiple method. The model number 11 in table is considered because on any further iterations the R^2 value did not increase at all. 0.597 was highest that could be reached. A double layered model was chosen because the R^2 results from the single layered results were ranging from **6.91316e-02** to **6.462937e-05** with the **RMSE** values reaching a value of **5.256169**.

Best Model

Based on the results from the table 6.2 the best model was considered to be the one with **28 neurons in the first layer and 19 neurons in the second layer**. To further discuss the results of this selected model following results were obtained. From the Figure 6.5 it can be seen that the with the change of neurons in layer 1 and 2 how the R^2 value is changing.

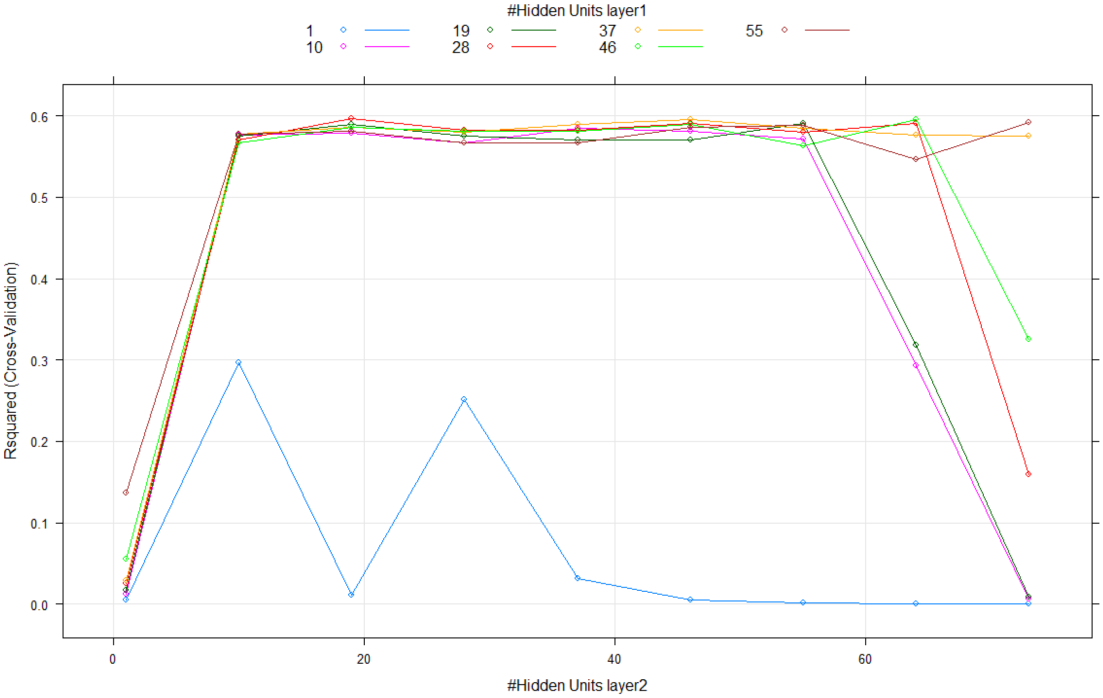


Figure 6.5: RSquare values Vs the neurons

Upon prediction (using the predict function in caret)of the values on the test data the R^2 of predicted values was **0.60856** which is slightly more than the that of the train R^2 value. This could indicate a slight over-fitting of the model. From the plot of the prediction and the actual data for test results can be seen in the Figure 6.6a though there is not a proper linearity but there are not many stray point. The line is a 45 degree line. Overestimation can be seen in the values near to zero and one in the very large values.

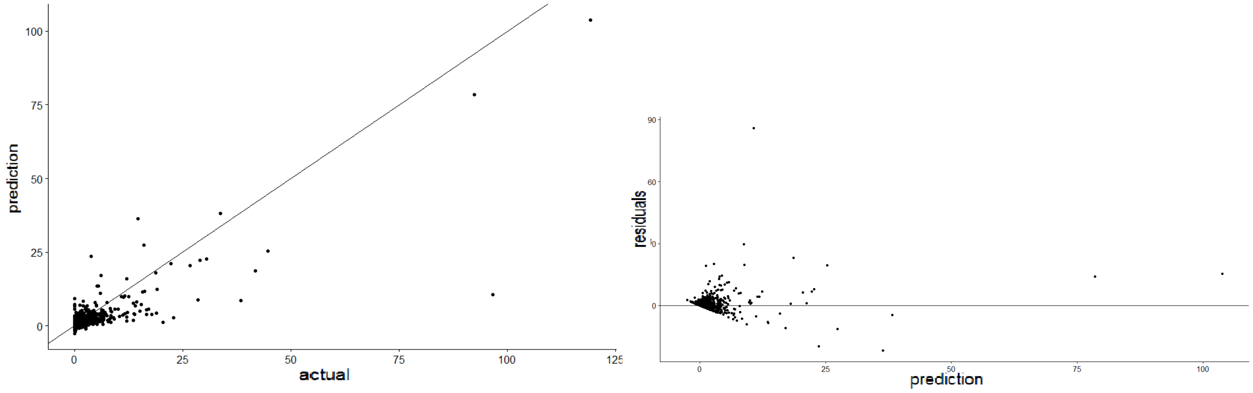
From the prediction vs residual plots in Figure 6.6b can be seen that the most of the residual values are not too high expect for few prediction values which could indicate the presence of possible outliers. There is a pattern of a cone seen at the near zero values. The variable considered for the model were. DURATION, Number of Impacted Trains, Intervals, Closeness Centrality, Betweeness Centrality, Degree Centrality, Temperature, Month, Track Type, Day of the week, Area Codes, Hour and Severity.

Table 6.2: **Best Tune R Square values(training) for Total Delay Neural Network models**

No.	**Layer1**	**Layer2**	**Learning Rate**	**RMSE**	**Train R^2**	**MAE**	**Parameter Changed**	**Input**
1	41	56	0.001	0.06178	0.45394	0.03232	Starting	87(all)
2	5	40	0.01	0.05742	0.5035296	0.02567	Input variables learning rate	64
3	3	45	0.1	0.05637	0.51907	0.02458	learning rate	64)
4	6	16	0.01	0.05789	0.50240	0.02517	learning rate Input Variables (linear)	63
5	11	26	0.01	0.02961	0.52303	0.01163	Influential pts removed Input Variables(-nzv)	52
6	9	30	0.01	0.03229	0.57458	0.01156	Influential pts removed Input Variables(-nzv) (linear)	42
7	16	9	0.01	0.03254	0.57119	0.01350	Influential pts removed Input Variables (linear)	64
8	19	43	0.1	0.032002	0.57624	0.01188	Influential pts removed Input Variables(-nzv) (linear) learning rate	41
9	13	7	0.05	0.02642	0.57626	0.01051	Influential pts removed Input Variables (linear-1) learning rate	78
10	55	19	0.05	0.02563	0.59144	0.00989	full training data Input Variables (linear-1)	78
11	**28**	**19**	**0.05**	**0.02840**	**0.59698**	**0.01013**	**full training data** **Input Variables** **(linear-1)**	**78**

Table 6.3: **R Square values for best Neural Network model for Total Delay**

No.	**Layer-1**	**Layer-2**	**Train R^2**	**Test R^2**
1	28	19	0.59698	0.60856

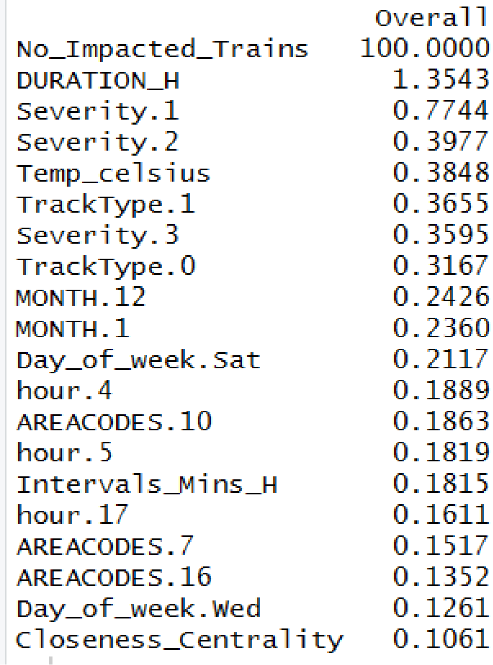


(a) Plot of actual vs test values on the test set

(b) Plot of fitted vs residual values for test set

Figure 6.6: Final Neural Network plots

From the Figure 6.7 the feature importance can be deduced by knowing which feature is impacting the response variable the most. The function used here is from caret package. It uses the **garsons** method to calculate variable importance(Kuhn, 2012b). Only the top 20 variables are shown. It can be seen that the Number of trains has the most impact on the total delay which was inferred from the descriptive analysis. All other variables have very less impact on the response variable of total delay. The next impacting variable that can be seen is the duration. There is a huge difference between the value of impacted trains and duration. This proves that duration is a important variable. The plot is attached in the appendix. Figure 6.8 is the plot of the final neural network model for the total delay.The picture very blurred because of the lack of the resolution power.



```
                          Overall
No_Impacted_Trains     100.0000
DURATION_H               1.3543
Severity.1               0.7744
Severity.2               0.3977
Temp_celsius             0.3848
TrackType.1              0.3655
Severity.3               0.3595
TrackType.0              0.3167
MONTH.12                 0.2426
MONTH.1                  0.2360
Day_of_week.Sat          0.2117
hour.4                   0.1889
AREACODES.10             0.1863
hour.5                   0.1819
Intervals_Mins_H         0.1815
hour.17                  0.1611
AREACODES.7              0.1517
AREACODES.16             0.1352
Day_of_week.Wed          0.1261
Closeness_Centrality     0.1061
```

Figure 6.7: Variable Importance

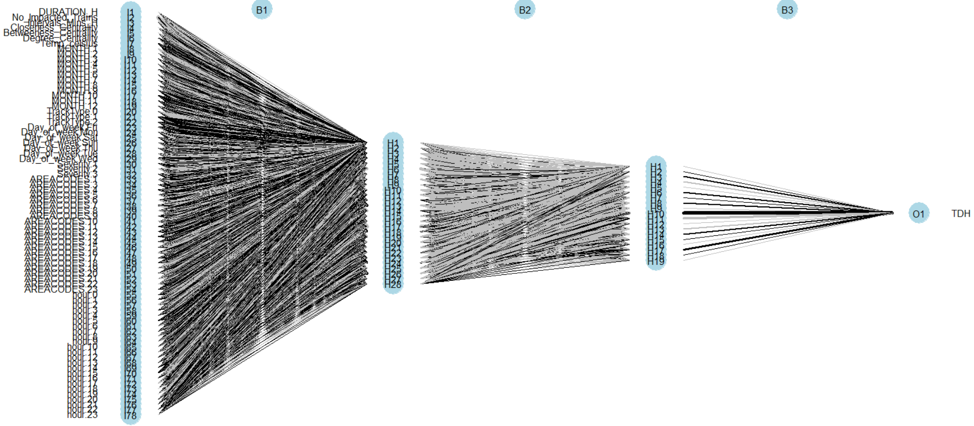


Figure 6.8: Neural Network for the Total Delay

6.2.2 Duration

In case of the Duration model, the total number of variable's initially considered are the 87 variables. One less than the those used in the Total delay model. Here the total delay values are not used for the prediction. From the linear regression it can be seen that the duration models were not performing really well. For the various model runs the same procedure used in the total delay model were used. Change of hyper-parameters, change of variables based on the linear regression and near zero variables. Also model runs with and without outliers were run. The table summarises the results of the model runs. Many model run other than the mentioned below were done. The entry 3 in the table 6.4 has the highest value of R^2 value of **0.15008** which is not a good value to start with, **consisting of 27 neurons in the first layer, 5 in the second and 5 in third**. So, it is very clearly evident that this not the best model for evaluation but the best for only this condition. Figure 6.9 show the variation of the RSquare value upon change in neurons for layer 1, layer 2 and layer3. This is a very messy plot but shows how the change in so many neurons could not help in improving the model. This model is still considered for further analysis.R^2 values from the table clearly indicates that the values are really poor. Because of poor R^2 values the number of layers was increase form 2 to 3 and then tested and the 3 layer model gave the best results. Upon prediction on the test data it can be seen from the table 6.5 that there is slight increase in the R^2 value when the values are predicted for test data set. From the graphs in the Figure 6.10a it is can be seen that points are not linear along the 45 degrees line with points above the underestimated and the points below the line are overestimated. The spread of the points is mostly along the y axis indicating major overestimation in the values near zero. From the residual plot Figure6.10b it is seen that the there is downward slope with very high residual values.

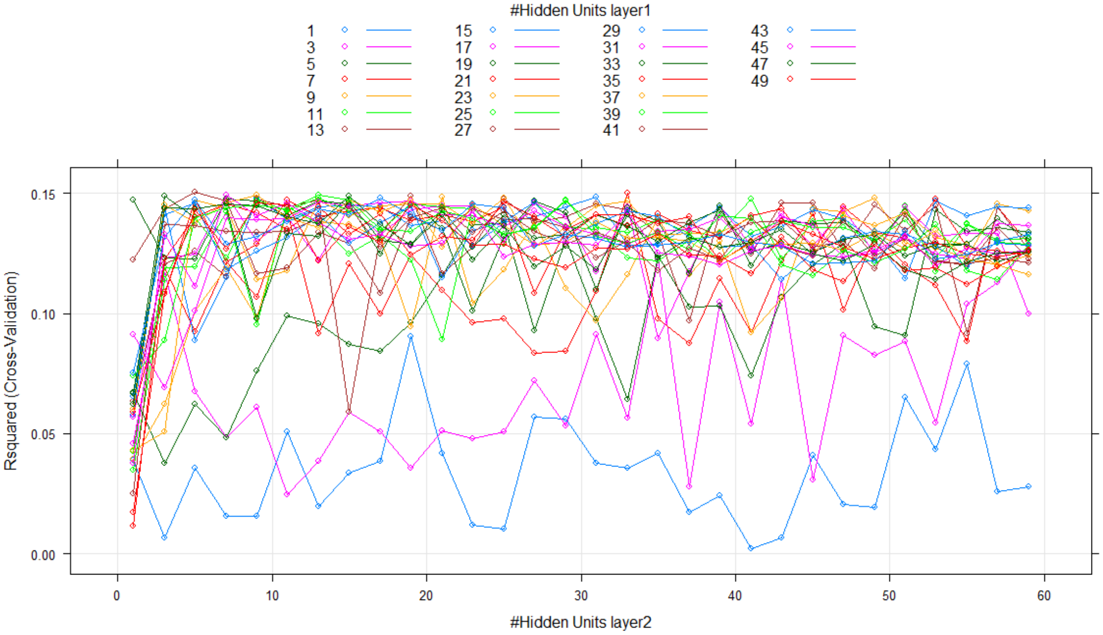


Figure 6.9: RSquare values Vs the neurons for duration model

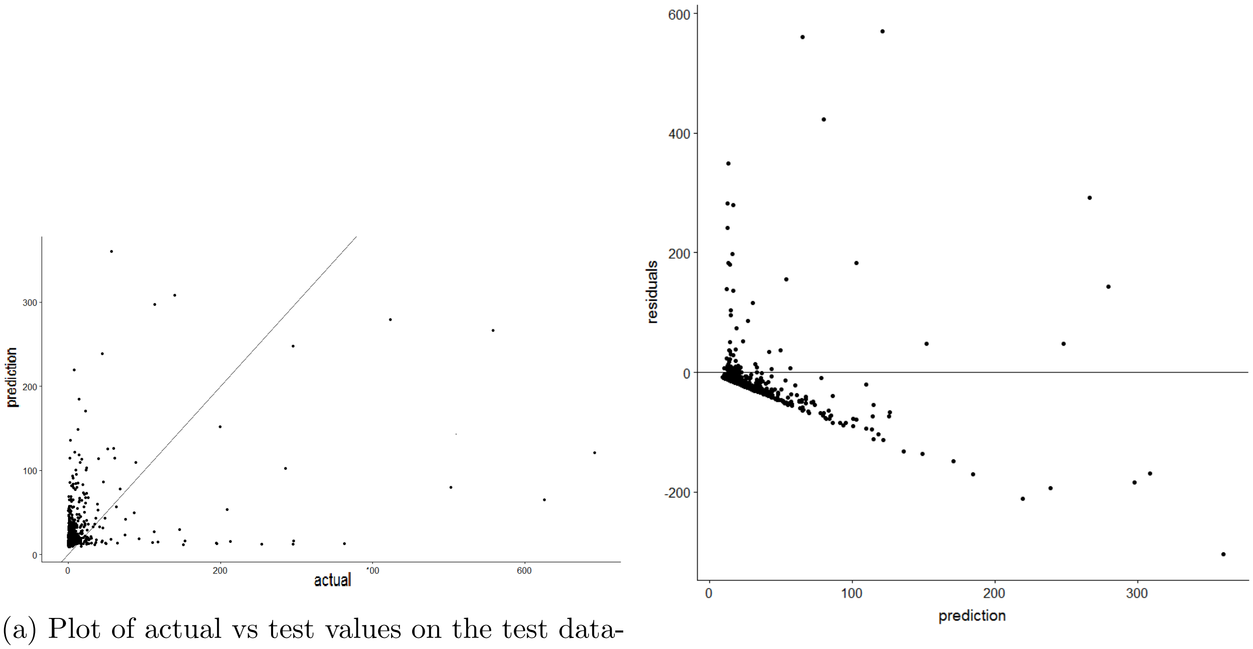
Eventually the Variable importance for the initial 20 variables can be seen that show that the most Severity plays the most important role in defining duration. Which is a bit self explanatory, because more the severe incident more is its duration. The next attribute is the AreaCode of 10 which is the major regional lines in the Copenhagen area. The third

Table 6.4: **Best Tune R Square values for Duration Neural Network models**

No.	**Layer1**	**Layer2**	**Layer3**	**Learning Rate**	**RMSE**	R^2	**MAE**	**Parameter Changed**	**Input**
1	21	6	0	0.1	0.06447	0.13541	0.028960		
2	21	26	0	0.01	0.06558	0.13639	0.02987	learning rate Inlfuential pts removed variabels(-linear)	63
3	**27**	**5**	**5**	**0.05**	**0.06214**	**0.15008**	**0.02641**	**nzv variables removed** **learning rate**	**50**
4	11	31	0	0.1	0.06471	0.14513	0.02945	learning rate variables removed(-linear)	77
5	1	51	0	0.25	0.0656	0.1364	0.02987	learning rate	77
6	11	31	0	0.1	0.06471	0.14512885	0.02945741	learning rate	77
7	21	21	0	0.05	0.06468	0.14552	0.02731	learning rate	77

Table 6.5: **R Square value For best Neural Network Model for Total Delay**

No.	**Layer-1**	**Layer-2**	**Layer-3**	**Train R^2**	**Test R^2**
1	27	5	5	0.15007	0.17210

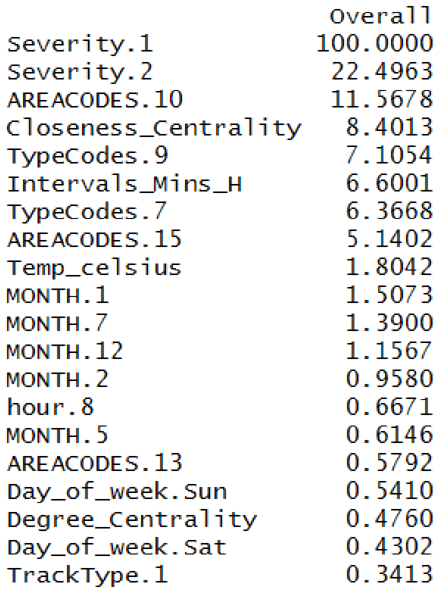


(a) Plot of actual vs test values on the test data-duration

(b) Plot of fitted vs residual values for test set

Figure 6.10: Final Neural Network plots

most important variable is Closeness Centrality, more central location an incident occurs the duration can be affected accordingly, less central node more the duration and more central the node lesser the duration. The importance values are much impact than the one's in the total delay model. The neural network for the duration prediction is illustrated in the Figure 6.12. The variables used in this model were **Number of Impacted Trains, Intervals, Closeness Centrality, Betweeness Centrality, Degree Centrality, Temperature, Month, Track Type, Day of the week, Are Codes and Hour**



```
                          Overall
Severity.1              100.0000
Severity.2               22.4963
AREACODES.10             11.5678
Closeness_Centrality      8.4013
TypeCodes.9               7.1054
Intervals_Mins_H          6.6001
TypeCodes.7               6.3668
AREACODES.15              5.1402
Temp_celsius              1.8042
MONTH.1                   1.5073
MONTH.7                   1.3900
MONTH.12                  1.1567
MONTH.2                   0.9580
hour.8                    0.6671
MONTH.5                   0.6146
AREACODES.13              0.5792
Day_of_week.Sun           0.5410
Degree_Centrality         0.4760
Day_of_week.Sat           0.4302
TrackType.1               0.3413
```

Figure 6.11: Variable Importance for the Duration Model

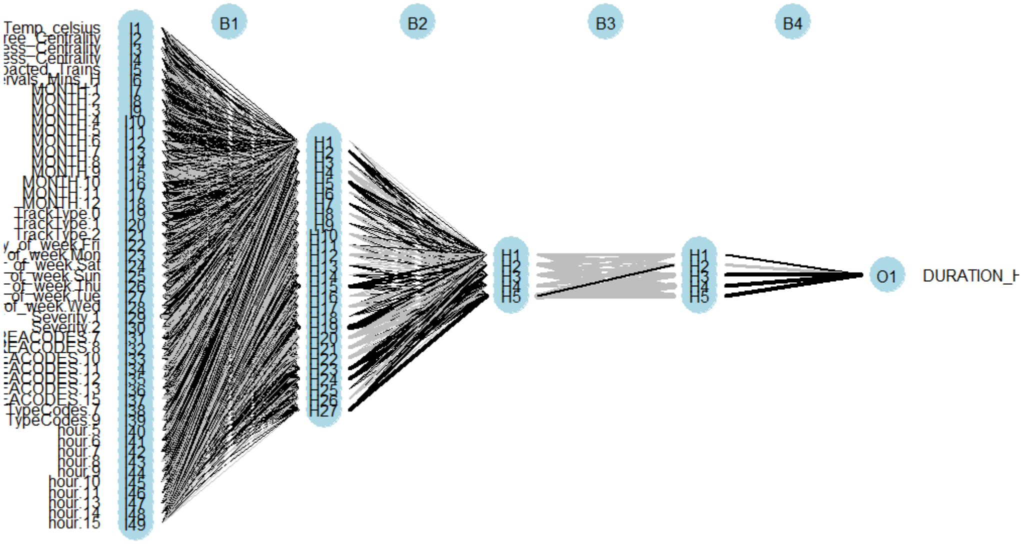


Figure 6.12: Neural Network for the Duration

6.3 Other Models for Total Delay and Duration

6.3.1 Extreme Gradient Boosting Machines

As discussed in the previous chapter the extreme gradient machines was also implemented. Both the total delay and duration prediction models were created. The first model discussed here is the **total delay model**. With the variation of the various parameters mention in the section the following tune was found to be the best among all. The initial parameters used for achieving the best model are, the maximum depth used were 3,6,9. The learning rate and the gamma value is held constant at 0.3 and 0 respectively. The best model achieved was at the iteration of 100 with a maximum depth of 3 and a learning rate of 0.3. Upon estimating the model the highest value of R^2 of value **0.59915**. This model was achieved while considering all the variables and full dataset with the outliers from the cooks distance analysis. Using the model a prediction is carried on the test data and the R^2 value is **0.44787**. It can be seen that there huge difference between the estimated and the predicted R^2 prompting to make the deduction that there is a possible over-fitting in the model estimation part.

From the figure 6.13 it can be seen that with tree depth of 3 highest R^2 is reached and the there is a variation of values upon increase in iterations. Also it can be seen that with increase in boosting iterations the R^2 values worsen indicating less number of boosting iterations are required.

From the Figure 6.14 it can be seen that there are few overestimated values whereas which are on the left side of the graph and few underestimated values on the right. Though the spread of values is not completely linear it can be but they do follow the 45 degree line indicating a linear tendency.

Coming to the feature importance it can be seen from the Figure 6.15, it is clearly evident that the similar to the previous models Number of impacted trains is the most important variable among all, after which closeness centrality and the duration of the incident are next important features. Duration was also an important feature in the other models indicating it does effect the total delay caused by an incident prompting in creating an operational framework for the the two response variables.

Table 6.6: **R Square values for best Boosting model for Total Delay and Duration**

Model Type	**Iteration**	**Learning rate**	**Train R^2**	**Test R^2**
Total Delay	100	0.3	0.59915	0.44786
Duration	40	0.2	0.11262	0.0581

On the other hand the same procedure was carried out for the duration prediction model. The results from the models for the duration model were really worse with R^2 going as low

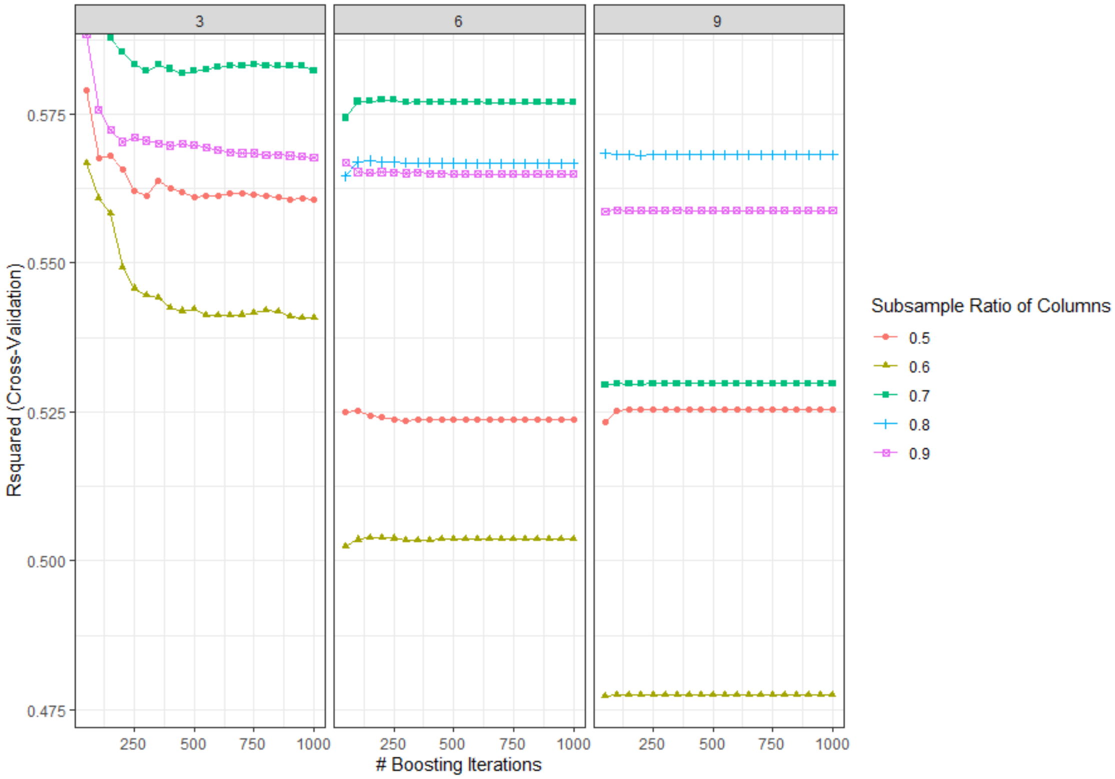


Figure 6.13: Diagnostics Plot Linear Model-Duration

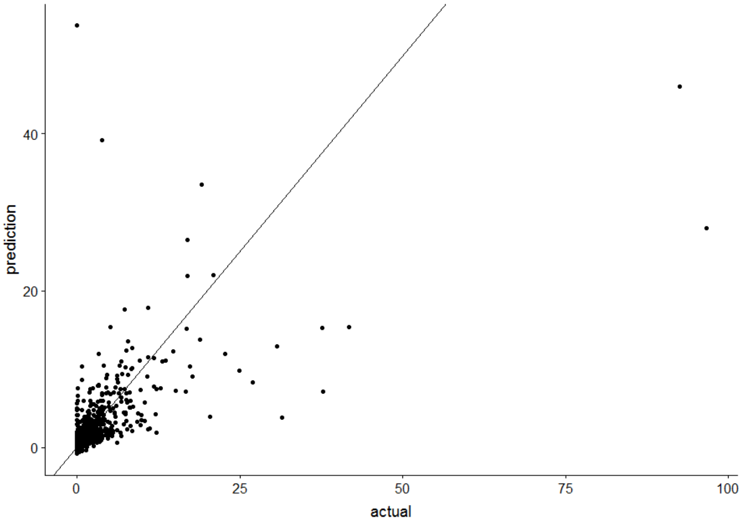


Figure 6.14: Prediction Vs Actual scatter plot for Boosting model on test set

as 0.01 indicating very poor model fit. This poor performance was also seen in the Neural networks model for the duration but the xgboost performance is worse resulting in a very bad prediction. Also, to for cross checking the variables input for the model were changes based on the importance of the variables, also, the removal of outliers is also carried out to check the performance of the model. The highest performance value of the metric is considered as

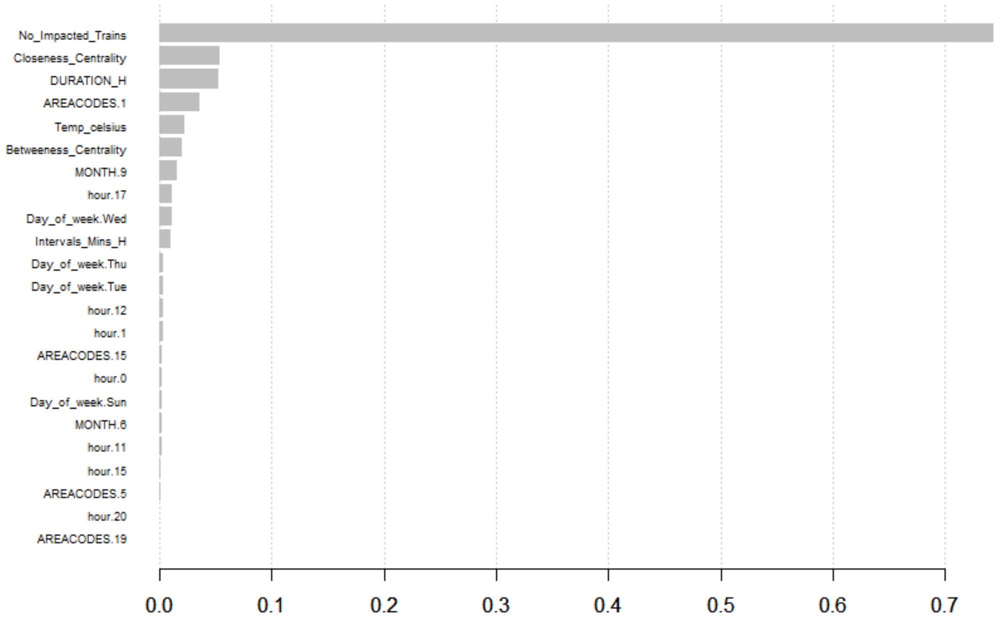


Figure 6.15: Feature Importance Plot for the Total Delay model

the best model. Table 6.6 shows the estimation and the prediction values which indicates that the duration model was over-fitted as there is huge difference between the test and the train R^2 values indicating that this model is not the best way for duration prediction.

6.3.2 Generalised Linear Regression Models

GLM algorithm is applied for the total delay model. As mentioned in the previous chapter the lambda and the alpha values are used to variate and check for different models. The Figure 6.16 shows the metric values for the models when the two parameter values are changed. The best model was found to have a R^2 **0.59126** and for this model all the variables were considered and without removal of the influential points. The λ and α values for the best model were found to be at 0.31586 and 0.2 respectively. But the estimation on the test set was found to be R^2 value of **0.46885**. This gap indicates a over fitting in the model, also the there is an over estimation and underestimation of the values can be seen in the Figure 6.17.

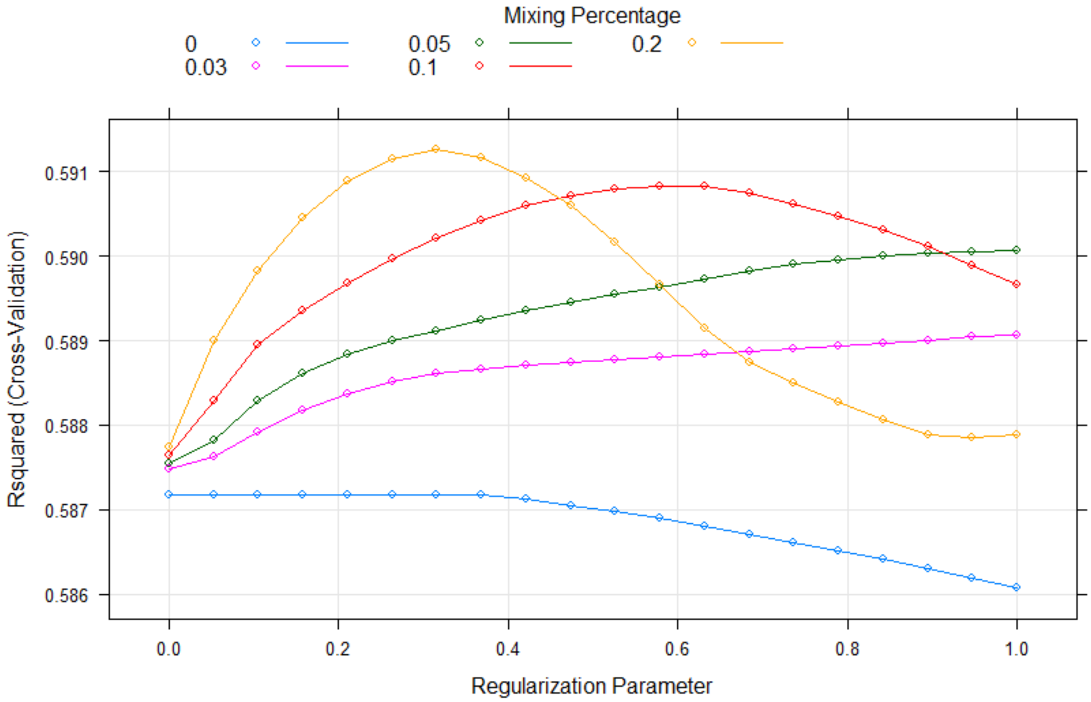


Figure 6.16: GLM Model results with parameter variation

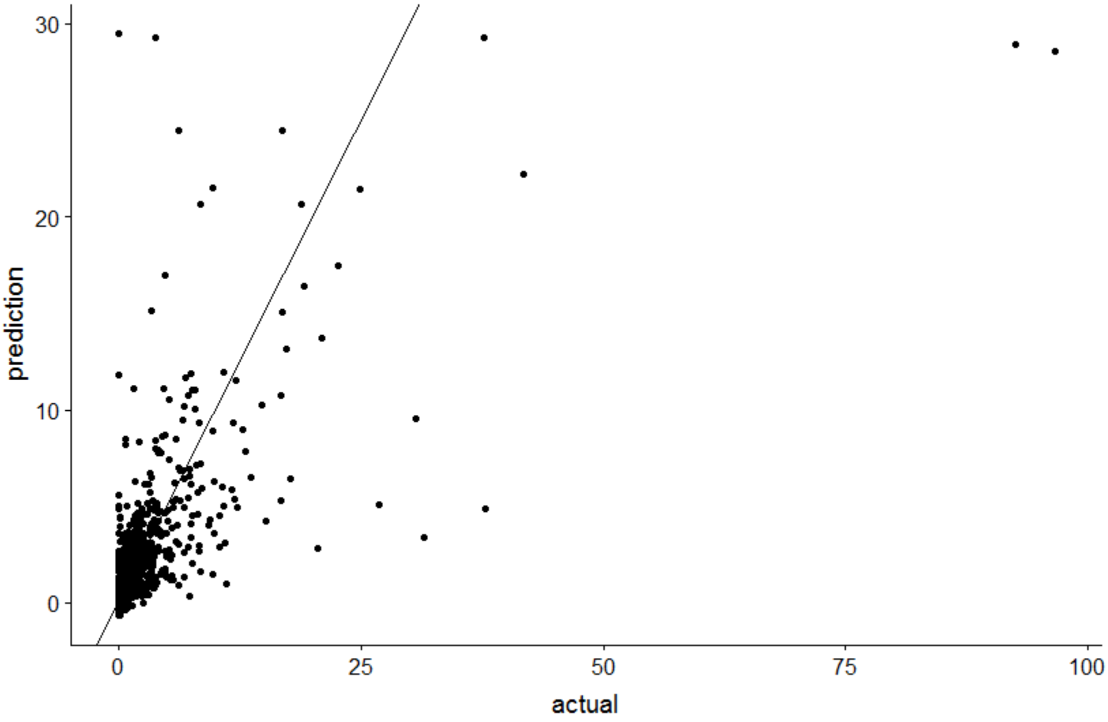


Figure 6.17: Actual vs Predicted values for the glmnet model on test set For Total Delay

Upon estimation of the duration model, the results, similar to the boosting model were found to be very poor. The table 6.7 shows the R^2

Table 6.7: **R Square values for best GLM model for Total Delay and Duration**

Model Type	λ	α	**Train** R^2	**Test** R^2
Total Delay	0.31585	0.2	0.59126	0.46885
Duration	1.05272	0.5	0.13318	0.06070

It can be seen that the R^2 value of 0.13 states that the model is poor but is slightly better than the Boosting model for duration. Different models were tried with different variations and the best model obtained was the one which gave the results mentioned in the table 6.7. The best result was obtained when the all the parameters with all the data values including the outliers. The eventual conclusions are discussed in the next chapter.

6.4 Operational Framework

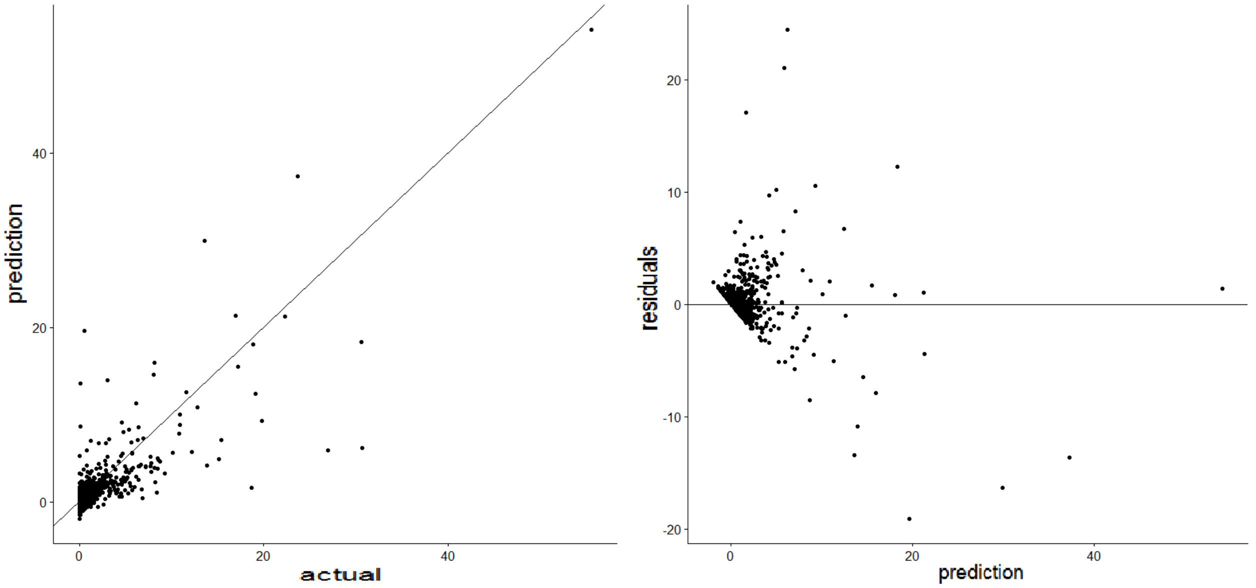
In the chapter 5 the operational framework is discussed which provides the basis for the results here. From the models mentioned above the most reliable models seemed to be that of the Neural Networks with less over-fitting than the models based on other algorithms. The principle of the framework is that the best neural network model discussed above is used to predict the duration for the data of 1000 rows that is randomly selected from the main data. Next the newly predicted duration is passed as the duration values to the best neural network model for the delay prediction. Using these newly predicted duration values the total delay is predicted. This can also be called as the **sensitivity analysis**. Here also **two methods** were applied to see the difference on how the framework itself if behaving.

1. For the first framework, the data of 1000 rows was taken from the main data pool(6693*88). The influential points(outliers) obtained from cooks distance analysis discussed in the Chapter 5 were **not** removed from the main pool data before 1000 row selection.

2. For the second framework, the data of 1000 rows was taken from the data pool(6616*88),i.e. the influential points(outliers) **were removed** from the main pool data before 1000 row selection The results from both the frameworks are shown in the table 6.8. In the table the Actual Test R^2 values are the values obtained on the prediction on the test data using the standalone models discussed in the previous section. Whereas the FM_Test R^2 are the values obtained using the framework. The R^2 value obtained from th framework is just for reference. The first framework was given priority because both the standalone models got the best results when the outliers were not removed, so it makes right sense to choose the 1st framework as the framework in hand. The improvement seen in the results was for the duration prediction negligible, whereas on total, the framework improved the R^2 for the total delay by 14.02% which is significant a improvement. The corresponding actual vs prediction plots were plotted only to see that there is an improvement in the linear distribution, along the 45 degree line, for both the predicted and actual values. Another observation made was the in both framework upon iterating over different sets of 1000 rows, the trend observed was with increase of duration model R^2 the overall framework R^2 decreased and the vise versa.

Table 6.8: **R Square values For Framework(FM)(with(w) and without(wO) Outliers)**

FM Type	**Variable**	**Actual Test R^2**	**FM_Test R^2**
FM_w	Duration	0.15008	0.17163
FM_w	Total Delay	0.59698	0.68071
FM_wO	Duration	0.15008	0.19944
FM_wO	Total Delay	0.59698	0.71995



(a) Actual vs Predicted values from the framework

(b) Predicted vs Residual values from the framework

Chapter 7

Conclusions

Model Comparison

From the Chapter 6, consolidation of all the results obtained from the models are represented in the tables 7.1 and 7.2 for both the main response variables. The best performing model is provided by the Neural Networks with comparatively less over-fitting which is can be deduced from the gap between the R^2 values of the train and the test (in other words the estimated and the predicted respectively). Both the XGBOOST and the

Table 7.1: **Consolidated Total Delay Models Results**

No.	**Model Type**	**Train** R^2	**Test** R^2
1	Neural Network	0.59698	0.60856
2	XGBOOST	0.59915	0.44786
3	GLMNET	0.59126	0.46885
4	Linear(Normal)	0.5298	0.6097
5	Log Linear	0.3482	0.3089

Table 7.2: **Consolidated Duration Models Results**

No.	**Model Type**	**Train** R^2	**Test** R^2
1	Neural Network	0.15007	0.17210
2	XGBOOST	0.11262	0.0581
3	GLMNET	0.13318	0.06070
4	Linear(Normal)	0.1292	0.1143
5	Log Linear	0.3032	0.2518

GLMNET models give good values for the total delay while estimating the models, but when the prediction is carried out on test values the metric value seem to reduce than the estimated indicating over-fitting in the model. This can be seen in the linear and the log linear models too, but the effect is comparatively less than the other two mentioned earlier. It can be clearly deduced that the best model for the total delay is Neural Network which was expected to be from the starting.

In case of the models for duration consistently poor values can be seen which indicates that there is definitely a lack of more data specific to other non included attributes. It was found that the Log linear model was the best model for the duration with clearly the largest value of R^2 among all the models with slight over fitting but from its diagnostic plots it can be seen that the data distribution is not linear indicating probable non linear interaction in the data. Also, linear models could be too simple for the data like this and may sometimes lead to less information capture. This lead to the selection of Neural Networks as the next best model and for the operational framework. From the results in can be safely said that Neural Networks, for now, are the best model for this study and probably for further research too. It clearly has more advantages over other models because of its ability to consider more latent information for prediction. Though they are complex and tend to over-fit it, in this case it was found otherwise and the other models were found to over-fit even more making.

Attributes

It consists of various component among which the two main components are **Total Delay** caused by an incident and the **Duration** of an incident. The main objective was to create a predictive model for total delay and duration of an incident and check its behaviour w.r.t the various attributes.The first part of the thesis was to find out the appropriate attributes to be used for predicting the response variables in hand. Apart from the various railway and time related attributes the two other features used were **Weather** and **Centrality Measure**. Neural Networks were chosen as the main models for the purpose as they were found, from the literature review, to be the most popular in the field. It was observed, from the results of the two standalone models created for the corresponding target attributes, that both weather and centrality measures influenced both the response variables where in duration was more influenced by both of them than the total delay. It is also observed from the results of both the standalone models that the total delay even though has less influence over all other variables other than number of impacted trains, still performed well. Winter months saw a slight increase in incidents(as discussed in the descriptive analysis) which supports the idea of weather inclusion in the attributes. Type of incidents ,i.e. the type codes were not used in both models because the majority of the incidents belonged to the only two types which are Signalling and interlocking problems and the problems caused by the weather and other external influences which can be explained by the ongoing signalling system change to ERTMS level2. Because of this immense bias it was not considered in the modelling. From the feature importance results for the neural network model of total delay it can be seen that the **Number of Impacted Trains** has the most impact making it the most important variable of them all. In case of the total duration model the duration was next best attribute which become of the reason to make the operational framework. On the other hand, among the centrality measures considered the most important measure was the closeness centrality and the temperature was among the top 5 attributes hence proving that weather does have effect on the total delay. Severity 1 and 2 where also among the most important attributes but it can be said that the most of the incidents belonged to these two categories on the contrary it is a important feature on its own. In case of the duration model it can be seen that the **severity** is the most important attribute of them all. Indicating, with more the severe the incident more would be the duration of the incident. Also, closeness centrality is one of the top 5 most important indicating the centrality does have influence on the duration making it an important attribute too in the analysis.

Operational Framework

From the results of the operation framework it can be seen that the there is a slight decrease in the duration prediction R^2 but using the predicted duration values as input for the total delay model, there is a increase in the total delay model performance. Though the results may sound strange but the a negative relation could be noticed in the framework. Which means the increase in duration model performance could bring down the performance of the operational framework. Additionally, one point to be taken note of is the very poor values of the duration model which could make the total delay model behave differently. When the duration model performance increases by 14.36% the performance of the operation framework improves by 14.02%. Whereas, an increased performance of 32% for the duration model causes a 20.60% increase. Though, the performance of framework increases the values itself reduces with increase in the performance of the duration.

To sum it up this thesis investigates the important variables total delay caused by an incident and the duration of the incident. Predictive models using various machine learning algorithms are created for the two variables mentioned and the models are checked for their performance. Also, various parameters have been selected for the analysis and it was found that most of the selected parameters do have a significant impact on the outputs. But more attributes could be researched for better results. Neural Networks are indeed on the best Machine Learning algorithm available for this purpose. Through this major policy recommendations can carried out for the major railway organisations as it leaps into the lesser used properties of the railways systems.

Limitations and Future Work

The main limitation is the lack of surety regarding the correctness of the data as it is manually entered data from the source. This can bring in the aspect of human errors into the data effecting the overall performance. Also, the exact locations of the incidents could help improve the study even more, i.e. the improved incident detection could help in creating more realistic models. Also, data from a larger network or a network with more traffic could also be helpful in this study as the number of scenarios involved also change making the models better better at learning as there are more patterns to learn from. In terms of improving the models itself data discritisation discussed in literature review in

(Yaghini et al., 2013) could be applied to the data for improved results and it can be applied based on the area sections considered. This seems like a good approach as the categorical parameters are high and the analysis could be carried based on each parameter level in a single model.

Since this is just a initial idea for application it can be further improved using much more deep machine learning methods. This work can also be tested by using it in the context of railway scheduling and the system performance can be tested accordingly. The future work directions after this work are immense and it is upon the user interest to use it the respective direction.

Bibliography

Aggarwal, C. C.
2018. *An Introduction to Neural Networks*, Pp. 1–52. Cham: Springer International Publishing.

Akaike, H.
1974. A new look at the statistical model identification. *IEEE Transactions on Automatic Control*, 19(6):716–723.

Barthélemy, M.
2004. Betweenness centrality in large complex networks. *The European Physical Journal B*, 38(2):163–168.

Boeing, G.
2017. Osmnx: A python package to work with graph-theoretic openstreetmap street networks. *Journal of Open Source Software.*

Boyles, S., D. Fajardo, S. Travis Waller, and A. Professor
2007. Naive bayesian classifier for incident duration prediction. *Transport Research Board.*

Bundesnetzagentur fr Elektrizitt, Gas, T. P. u. E.
2018. Railway market analysis, germany 2018. Technical report, Bundesnetzagentur fr Elektrizitt, Gas, Telekommunikation, Post und Eisenbahnen.

Caimi, G., M. Fuchsberger, M. Laumanns, and M. Luethi
2012. A model predictive control approach for discrete-time rescheduling in complex central railway station areas. *Computers & Operations Research*, 39:25782593.

Carey, M.
1999. Ex ante heuristic measures of schedule reliability. *Transportation Research Part B: Methodological*, 33(7):473–494. Cited By :75.

Cerreto, F., B. F. Nielsen, O. Nielsen, and S. S. Harrod
2018. Application of data clustering to railway delay pattern recognition. *Journal of Advanced Transportation*, 2018:1–18.

Chandesris, M.
2006. Dynamic and real-time prediction of duration of incident. *Railway Research, The Global Reference for Rail Innovation, SNCF.*

Chen, T. and C. Guestrin
2016. Xgboost: A scalable tree boosting system. In *Proceedings of the 22Nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '16, Pp. 785–794, New York, NY, USA. ACM.

Clauset, A., C. R. Shalizi, and M. E. J. Newman
2009. Power-law distributions in empirical data. *SIAM Rev.*, 51(4):661–703.

Conte, C.
2008. *Identifying dependencies among delays*. PhD thesis, George- August- Univeristy, Goettingen, http://hdl.handle.net/11858/00-1735-0000-000D-F124-2.

Corman, F.
2010. *Real-time Railway Traffic Management: dispatching in complex, large and busy railway networks*. PhD thesis, Technical University Delft, Delft, Netherlands.

Corman, F., A. DAriano, D. Pacciarelli, and M. Pranzo
2010. A tabu search algorithm for rerouting trains during rail operations. *Transportation Research Part B: Methodological*, 44(1):175 – 192.

Corman, F. and P. Kecman
2018. Stochastic prediction of train delays in real-time using bayesian networks. *Transportation Research Part C: Emerging Technologies*, 95:599–615. Cited By :2.

Cule, B., B. Goethals, S. Tassenoy, and S. Verboven
2011. Mining train delays. In *Advances in Intelligent Data Analysis X*, Pp. 113–124.

D Student, B. and B. Schittenhelm
2009. Railway timetabling based on systematic follow-up on realized railway operations. *Annual Danish Transport Conference*.

D'Ariano, A.
2008. *Improving real-time train dispatching: models, algorithms and applications*. PhD thesis, Technical University of Delft, https://repository.tudelft.nl/islandora/object/uuid:178b886e-d6c8-4d39-be5d-03d9fa3a680f.

D'Ariano, A.
2009. Innovative decision support system for railway traffic control. *Intelligent Transportation Systems Magazine, IEEE*, 1:8 – 16.

D'Ariano, A., D. Pacciarelli, M. Pranzo, and I. Robin Hemelrijk
2007. Evaluating the performance of railway dynamic traffic management. In *Urban transport and the environment : an international perspective*, volume 11.

Dasu, T. and T. Johnson
2003. *Exploratory Data Mining and Data Cleaning*, Wiley Series in Probability and Statistics. Wiley.

Davey, E.
2012. Rail traffic management systems (tms). In *IET Professional Development Course on Railway Signalling and Control Systems*, Pp. 126–143.

Derrible, S.
2012. Network centrality of metro systems. *Plus One*, 7(7).

Dunn, P. K. and G. K. Smyth
2018. *Generalized linear models with examples in R*. Springer.

Erath, A., M. Löchl, and K. W. Axhausen
2009. Graph-theoretical analysis of the swiss road and railway networks over time. *Networks and Spatial Economics*, 9(3):379–400.

Flier, H., R. Gelashvili, T. Graffagnino, and M. Nunkesser
2009. *Mining railway delay dependencies in large-scale real-world delay data*, volume 5868 LNCS of *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*. Cited By :23.

Friedman, J., T. Hastie, R. Tibshirani, et al.
2000. Additive logistic regression: a statistical view of boosting (with discussion and a rejoinder by the authors). *The annals of statistics*, 28(2):337–407.

Friedman, J. H.
2001. Greedy function approximation: a gradient boosting machine. *Annals of statistics*, Pp. 1189–1232.

Gholami, O. and J. Törnquist
2018. A heuristic approach to solving the train traffic re-scheduling problem in real time. *Algorithms*, 11(4).

Gorman, M. F.
2009. Statistical estimation of railroad congestion delay. *Transportation Research Part E: Logistics and Transportation Review*, 45(3):446 – 456.

Goverde, R., I. Netherlands Research School for Transport, and Logistics
2005. *Punctuality of Railway Operations and Timetable Stability Analysis*, TRAIL thesis series. Netherlands TRAIL Research School.

Hansen, I. A., R. M. P. Goverde, and D. J. van der Meer
2010. Online train delay recognition and running time prediction. In *13th International IEEE Conference on Intelligent Transportation Systems*, Pp. 1783–1788.

Harrod, S.
2012. A tutorial on fundamental model structures for railway timetable optimization. *Surveys in Operations Research and Management Science*, 17:85–96.

Hastie, T. and J. Qian
2014. Glmnet vignette. *Retrieve from http://www. web. stanford. edu/˜ hastie/Papers/Glmnet_Vignette. pdf. Accessed September*, 20:2016.

Hwang, C.-C. and J.-R. Liu
2010. A simulation model for estimating knock-on delay of taiwan regional railway. *Journal of the Eastern Asia Society for Transportation Studies*, 8.

Jambhorkar, S. S. and M. V. S. Jondhale
2015. *Data Mining Technique: Fundamental Concept and Statistical Analysis.* Horizon Books (A Division of Ignited Minds Edutech P Ltd).

James, G., D. Witten, T. Hastie, and R. Tibshirani
2013. *An introduction to statistical learning*, volume 112. Springer.

Kersbergen, B., T. van den Boom, and B. De Schutter
2013. On implicit versus explicit max-plus modeling for the rescheduling of trains. In *Proceedings of the 5th International Seminar on Railway Operations Modelling and Analysis (RailCopenhagen)*, Copenhagen, Denmark.

Kersbergen, B., T. van den Boom, and B. De Schutter
2016. Distributed model predictive control for railway traffic management. *Transportation Research Part C: Emerging Technologies*, 68:462–489. Cited By :13.

Kono, A., H. Yakubi, and N. Tomii
2016. Identifying the cause of delays in urban railways using datamining technique. *Asian conference on Railway infrastructure and Transportation*, 2016.

Kuhn, M.
2008. Building predictive models in r using the caret package. *Journal of Statistical Software, Articles*, 28(5):1–26.

Kuhn, M.
2012a. Data sets and miscellaneous functions in the caret package. *Journal of Statistical Software, Articles.*

Kuhn, M.
2012b. Variable importance using the caret package. *Journal of Statistical Software.*

Kuhn, M. and K. Johnson
2013. *Applied predictive modeling*, volume 26. Springer.

Landex, A.
2008. *Methods to estimate railway capacity and passenger delays.* PhD thesis, Denmark Technical University.

Lee, W.-H., L.-H. Yen, and C.-M. Chou
2016. A delay root cause discovery and timetable adjustment model for enhancing the punctuality of railway services. *Transportation Research Part C: Emerging Technologies*, 73.

Lessan, J., L. Fu, and C. Wen
2019. A hybrid bayesian network model for predicting delays in train operations. *Computers and Industrial Engineering*, 127:1214–1222. Cited By :2.

Li, D. . and L. Cheng
2010. Incident duration prediction based on bayesian network. *Journal of Beijing Institute of Technology (English Edition)*, 19(SUPPL. 2):119–123.

Li, R., F. C. Pereira, and M. E. Ben-Akiva
2018. Overview of traffic incident duration analysis and prediction. *European Transport Research Review*, 10(2). Cited By :2.

Lüthi, M.
2009. *Improving the Efficiency of Heavily Used Railway Networks Through Integrated Real-time Rescheduling*, Institut f"ur Verkehrsplanung und Transportsysteme. ETH, IVT.

M. Blei, D., A. Y. Ng, and M. Jordan
2003. Latent dirichlet allocation. *Journal of Machine Learning Research*, 3:993–1022.

Malavasi, G. and S. Ricci
2001. Simulation of stochastic elements in railway systems using self-learning processes. *European Journal of Operational Research*, 131:262–272.

Markovic, N., S. Milinkovic, K. S. Tikhonov, and P. Schonfeld
2015. Analyzing passenger train arrival delays with support vector regression. *Transportation Research Part C: Emerging Technologies*, 56.

Mazzarello, M. and E. Ottaviani
2007. A traffic management system for real-time traffic optimisation in railways. *Transportation Research Part B: Methodological*, 41:246–274.

Natekin, A. and A. Knoll
2013. Gradient boosting machines, a tutorial. *Frontiers in Neurorobotics*, 7:21.

Nyström, B.
2008. *Aspects of improving punctuality: from data to decision in railway maintenance.* PhD thesis, Lule University of Technology. www.ipd.bth.se/jtr/PhD.pdf.

Oneto, L., E. Fumeo, G. Clerico, R. Canepa, F. Papa, C. Dambra, N. Mazzino, and D. Anguita
2017. Train delay prediction systems: A big data analytics perspective. *Big Data Research*, 11:54–64.

Pereira, F., F. Rodrigues, and M. Ben-Akiva
2013. Text analysis in incident duration prediction. *Transportation Research Part C: Emerging Technologies*, 37:177192.

Peters, J., B. Emig, M. Jung, and S. Schmidt
2005. Prediction of delays in public transportation using neural networks. In *International Conference on Computational Intelligence for Modelling, Control and Automation and International Conference on Intelligent Agents, Web Technologies and Internet Commerce (CIMCA-IAWTIC'06)*, volume 2, Pp. 92–97.

Pettet, G., S. Nannapaneni, B. Stadnick, A. Dubey, and G. Biswas
2018. Incident analysis and prediction using clustering and bayesian network. In *2017 IEEE SmartWorld Ubiquitous Intelligence and Computing, Advanced and Trusted Computed, Scalable Computing and Communications, Cloud and Big Data Computing, Internet of People and Smart City Innovation, SmartWorld*, Pp. 1–8.

Pongnumkul, S., T. Pechprasarn, N. Kunaseth, and K. Chaipah
2014. Improving arrival time prediction of thailand's passenger trains using historical travel times. In *2014 11th International Joint Conference on Computer Science and Software Engineering (JCSSE)*, Pp. 307–312.

Quaglietta, E., P. Pellegrini, R. Goverde, T. Albrecht, B. Jaekel, G. Marlire, J. Rodriguez, T. Dollevoet, B. Ambrogio, D. Carcasole, M. Giaroli, and G. Nicholson
2016. The on-time real-time railway traffic management framework: A proof-of-concept using a scalable standardised data communication architecture. *Transportation Research Part C: Emerging Technologies*, 63:23–50.

Rob, D., M. P Goverde, and R. Goverde
1998. Optimal scheduling of connections in railway systems. *WCTR*, 4.

Robert, N. and H. Kim
2018. Predictions of train delays using machine learning. *EXAMENSARBETE INOM DATATEKNIK, KTH Stockholm.*

Rodrigues, F. A.
2019. *Network Centrality: An Introduction.* Cham: Springer International Publishing.

Sam, Marcella', D. A. P. D.
2018. New alternative graph models and methods for the real-time railway traffic management problem. *CASPT.*

Schaafsma, A.
2001. *Dynamisch Railverkeersmanagement besturingsconcept voor railverkeer op basis van het Lagenmodel Verkeer en Vervoer.* PhD thesis, Technical University of Delft.

Seriani, S., T. Fujiyama, and G. De Ana Rodriguez
2016. Boarding and alighting matrix on behaviour and interaction at the platform train interface. *RRUKA Annual Conference 2016.*

Sogin, S., Y.-C. Lai, T. Dick, and C. Barkan
2013. Comparison of capacity of single- and double-track rail lines. *Transportation Research Record: Journal of the Transportation Research Board*, 2374:111–118.

Törnquist, J.
2006. *Railway traffic disturbance management*. PhD thesis, Blekinge Institute of Technology. www.ipd.bth.se/jtr/PhD.pdf.

Türkan, S., M. Candan, . Etin, and O. Toktam
2012. Outlier detection by regression diagnostics based on robust parameter estimates. *Hacettepe Journal of Mathematics and Statistics*, 41.

Van de Velde, D. M.
2013. Learning from the japanese railways: Experience in the netherlands. *Policy and Society*, 32(2):143–161. Cited By :4.

van den Boom, T., B. Kersbergen, and B. De Schutter
2012. Structured modeling, analysis, and control of complex railway operations. In *Proceedings of the IEEE Conference on Decision and Control*, Pp. 7366–7371.

van den Boom, T. J., N. Weiss, W. Leune, R. M. Goverde, and B. D. Schutter
2011. A permutation-based algorithm to optimally reschedule trains in a railway traffic network. *IFAC Proceedings Volumes*, 44(1):9537 – 9542. 18th IFAC World Congress.

Wang, P. and Q.-p. Zhang
2019. Train delay analysis and prediction based on big data fusion. *Transportation Safety and Environment*.

Wang, X., S. Chen, and W. Zheng
2013. Traffic incident duration prediction based on partial least squares regression. *Procedia - Social and Behavioral Sciences*, 96:425 – 432. Intelligent and Integrated Sustainable Multimodal Transportation Systems Proceedings from the 13th COTA International Conference of Transportation Professionals (CICTP2013).

Xia, Y., J. N. V. Ommeren, P. Rietveld, and W. Verhagen
2013. Railway infrastructure disturbances and train operator performance: The role of weather. *Transportation Research Part D: Transport and Environment*, 18:97 – 102.

Yaghini, M., M. M. Khoshraftar, and M. Seyedabadi
2013. Railway passenger train delay prediction via neural network model. *Journal of Advanced Transportation*, 47(3):355–368. Cited By :22.

Zakeri, G. and N. Olsson
2018. Investigating the effect of weather on punctuality of norwegian railways: a case study of the nordland line. *Journal of Modern Transportation*, 26.

Zilko, A. A., D. Kurowicka, and R. M. P. Goverde
2016. Modeling railway disruption lengths with copula bayesian networks. *Transportation Research Part C: Emerging Technologies*, 68:350–368. Cited By :19.

Declaration

I hereby confirm that this Master's thesis is my own work and I have documented all sources and materials used. This thesis has not been previously submitted elsewhere for purposes of assessment.

Munich, September 11th, 2019

Bhagya Shrithi Grandhi