2022-04

Road Collision Analysis and Prediction Using Machine Learning Approaches

Owjimehr, Omid

UNIVERSITY OF CALGARY

Road Collision Analysis and Prediction Using Machine Learning Approaches

by

Omid Owjimehr

A THESIS
SUBMITTED TO THE FACULTY OF GRADUATE STUDIES
IN PARTIAL FULFILMENT OF THE REQUIREMENTS FOR THE
DEGREE OF MASTER OF SCIENCE

GRADUATE PROGRAM IN ELECTRICAL ENGINEERING

CALGARY, ALBERTA
APRIL, 2022

Abstract

Road travel accounts for most traffic accidents worldwide. Improvements in road safety, education, recent technology advancements, and other environmental factors have decreased the number of collisions in developed nations. Many countries, provincial, and local governments envision the possibility of zero fatalities or serious injuries in the near future. Thus, it is essential to develop road traffic accident prediction models to support such a vision.

On the one hand, classical statistical models have been applied to develop prediction models throughout the literature. These models provide interpretable parameters at the expense of poor generalization when faced with complex and nonlinear relationships. On the other hand, data-driven methods utilizing Machine Learning (ML) approaches have been used recently to deal with the drawbacks of classical models, which showed promising results.

Road accidents result from many factors, including spatial, temporal and external factors. Those factors may influence the occurrence of accidents differently, according to the location and time of accidents. Thus, it is essential to consider the area-specific influential factors while analyzing and developing prediction models. Canada is the second coldest country globally, and its extreme weather has a higher effect on accidents than the other countries, which must be addressed.

This thesis seeks to explore determinants of road collisions, emphasizing Canadian weather. It then compares classical and ML models for collision prediction. Furthermore, it introduces the most influential factors in crashes with respect to Calgary's weather. All study parts are performed on the collisions data in Calgary, Alberta, Canada, between 2017 to 2020. It is shown that all the weather attributes are correlated to collisions. It shows the importance of considering the weather attributes in accident analysis and prediction. Based on the nature of the collisions dataset, which is tabular and heterogeneous, Neural Networks showed higher performances than the other investigated model, with 92% accuracy. The proposed models can be used for policy-making and individual usage in Canadian cities since the effect of all the weather features is already embedded in the models. In order to demonstrate the thesis's applicability, a new speed limit is recommended utilizing the developed models for Deerfoot TR SE. Results showed, for instance, if the speed limit is decreased from 100 to 90 km/h on Deerfoot TR SE, a 5% accident reduction is predicted.

Preface

This thesis is an original work by the author. No part of this thesis has been previously published.

Acknowledgements

I want to express my gratitude and appreciation to my supervisors, Dr. Laleh Behjat and Dr. Merkebe Getachew Demissie, for their excellent guidance and support throughout my studies at the University of Calgary. Their insightful comments in my thesis were the key to my success.

I also would like to express my appreciation to my supervisory committee members Dr. Lina Kattan and Dr. Poornima Jayasinghe for their great mentor ship and support throughout my journey.

A great thanks to the examination committee members, Dr. Saeid Saidi, Dr. Roberto Souza, and Dr. Behrouz Far for giving me time and making valuable comments on my thesis.

It was a great pleasure working with and learning from those awesome people.

I want to express my gratitude to the CREATE - IISC (Collaborative Research and Training Experience - Integrated Infrastructure for Sustainable Cities) for providing the trainees with all the valuable resources and financial support.

Dedicated to

My respectful parents and beloved wife, Mrs. Nazanin Aeenmehr, without whose constant support this thesis paper was not possible. At the same time, my thanks also go to my caring sister, Mrs. Mehri Owjimehr, whose advice really worked for me.

Table of Contents

List of Figures

List of Tables

List of Symbols, Abbreviations, and Nomenclature

Symbol	Definition
ML	Machie Learning
GLM	Generalized Linear Model
APM	Accident Prediction Model
HSM	Highway Safety Manual
SPF	Safety Performance Function
CMF	Crash Modification Factors
DNN	Deep Neural Network
IG	Information Gain
GG	Gini Gain
NN	Neural Network

Chapter 1

Introduction

1.1 Overview

Road traffic accidents cause millions of injuries and deaths every year. Based on the World Health Organization's latest report, approximately 1.35 million deaths and more than 50 million people suffer from non-fatal injuries every year [1]. Therefore, it is crucial to investigate and find the critical factors in accidents in order to reduce the number of accidents. Researchers have tried to understand the impact of these factors on traffic accident occurrence using two different approaches: classical statistical methods [2, 3, 4] and contemporary data-driven methods such as Machine Learning (ML) and data mining approaches [5, 6].

Statistical models define the problem as a mathematical formula by considering the relationship between the model variables. These are mostly complicated models with several dependent and independent variables and predefined assumptions. Therefore, if one of the assumptions is incorrect, the results might differ from the expected results. Another drawback of these models is the insufficient ability to deal with nonlinear relationships.

Researchers started taking advantage of new technologies with the advancements in computers and the introduction of different machine learning techniques. These ML techniques have shown superiority over classical models [7]. Unlike statistical models, data-driven methods try to infer the relationship between the factors that affect an accident's occurrence without considering any predefined underlying relationship among features and develop a model mostly based on optimization techniques. The main advantage of the latter approaches is their generalization capability to model nonlinear relationships, which is often available in transportation systems that statistical models fail to capture. This generalization enables researchers to investigate the effect of external factors on road accidents without the complication of statistical approaches

and without requiring new utility functions. One external factor that has received little attention is the weather features throughout the literature. Utilizing the mentioned ML approaches allows investigating the effect of weather features on road accidents.

This thesis investigates that gap and addresses the importance of the weather features. For doing so, multi-source heterogeneous data, including spatial, temporal, and distinctive weather features, are utilized to conduct an exploratory analysis and introduce the factors with high impact on accidents. Then accident prediction models are developed utilizing the mentioned dataset that can be used for policy-making and individual use in Canadian cities. Furthermore, proposed speed limits using the developed predictive model present an application of the thesis.

1.2 Problem Statement

"Why is weather an important factor for accident prediction?" Weather, with its attributes, plays a significant role in Canadian life. Canada is the second coldest country globally, with an estimated yearly average temperature of -5.35 degrees Celsius[8]. The analysis on the road collisions data shows that the number of road collisions in Calgary in months with an average temperature below zero is 32% more than in months with an average temperature above zero from 2017 to 2020. In addition, the likelihood of accidents when the relative humidity is between 5% to 10% is three times more than when the relative humidity is between 45% to 50%. These numbers indicate the effect of temperature and humidity as weather attributes on accidents in Calgary and show the importance of this analysis.

There are two main components in all data-centred ML-based accident prediction studies: the data and the predictive model. According to each study area's data availability and development conditions, the quality, amount, and data type can be different. Therefore, a single predictive model cannot be proposed for all regions due to differences in the characteristics and data availability of the areas [9]. Consequently, it is paramount of importance to take all the area-specific features of the study area into consideration, since external factors can contribute differently to the occurrence of accidents in different places.

Previous studies have identified several factors for road accidents, including road users' factors [10], vehicle-related factors [11], and road and environmental factors [12] such as road design and road layout, speed limits, traffic flow state, and supportive pedestrian facilities [13].

To the best of the author's knowledge, there is no data-centred ML-based study in the literature that has included all the available weather features in analyzing and predicting road traffic accidents. This study aims first to use all the different available weather features, including temperature, wind speed, visibility, humidity, overall forecast, wind direction, pressure, dew point temperate and other accident characteristics,

to find the best accidents prediction model and then introduce the most influential accident determinants in Calgary, Canada.

1.3 Objectives

Despite the general efforts made by previous studies in different locations, the question is still remained as to how the different area-specific factors related to the environment surrounding the location of the accidents impact accident occurrence in Canadian cities. An empirical study of road traffic accidents in Calgary, Canada, is carried out to expand the research boundary and provide more empirical evidence. This study aims to understand the impacts of factors that contribute to the occurrence of the accident using shallow and deep learning models.

The study makes the following major contributions:

(1) Firstly, multi-source data, including road characteristics, collision data from the Open Calgary data center[14], and historical climate data center on the Government of Canada's website[15], are brought together to enhance the set of datasets available for traffic accident prediction modelling.

(2) Secondly, Exploratory data Analysis focusing on Canadian weather features is conducted to give insight into how different external factors such as weather attributes are correlated to road accidents based on the available data in Calgary.

(3) Thirdly, this study explores a collection of machine learning models to compare and find the most proper model for Accident Prediction Model development. The model outputs can help identify accident hotspots and inform policy on road traffic safety improvement project developments.

As the main contribution of this study, finding the factors that cause or influence accidents can help make the right decisions by policymakers and recommend proper improvements in order to increase the safety of roads in Calgary. Consequently, a safer environment will save lives and money and make the city more livable. Thus, this is a timely study showing the opportunities of multi-sourced datasets and how effectively such datasets can be utilized to predict the occurrence of traffic accidents using novel machine learning models and then find the influential factors in accidents using those models.

1.4 Thesis Structure

This thesis is organized as follows. A brief description of previous research is given in the related works section(Chapter 2). Then, the methodology is described in Chapter 3, followed by data preparation and augmentation. Road traffic accident data analysis and Collision predictive model development are brought

in Chapter 4. Then, an application of the proposed model is presented in Chapter 5. Lastly, results are generated using the proposed methods detailed in the methodology section, are drawn alongside with the conclusion in the last Chapter.

Chapter 2

Literature Review

2.1 Overview

Accident prediction studies are divided into two main categories, studies based on the classical/statistical models and studies based on the novel ML and data mining models. A brief review of classical studies is brought in section 2.2. Section 2.3 reviews the existing literature on accident prediction using novel ML and data mining approaches. A summary of all the features used in the literature is brought in Section 2.4, to show the gap and the primary motivation of this thesis. Section 2.5 goes over the Generalized Linear Model, and finally, Section 2.6 discusses the 4 ML models used in this thesis.

2.2 Classical/Statistical Studies

Classical approaches for Accident Prediction Model (APM) developments, are divided into models based on the Highway Safety Manual (HSM) and individual models based on Generalized Linear Model (GLM). HSM introduces some predictive methods for forecasting the average crash frequency of a site or facility. Those methods used different regression models named Safety Performance Functions (SPF) as base models. SPFs are constructed on variables such as traffic volume and road length. Upon applying HSM's models on a specific area, some Crash Modification Factors (CMF) and calibration factors need to be applied based on the specific requirements of the area [16]. Some studies, such as [17], and [18] used HSM's models and proposed ideas for CMF selection based on the area-specific requirements of the study area in order to adapt those models in local studies.

GLM is another approach widely used for APM developments in the literature, in which each output is assumed to be in the range of an exponential family distribution. Poisson and Binomial Distribution

are mostly used for accidents depending on how the problem is articulated. For instance, Lina Kattan et al. [19] used a Negative Binomial distribution to make up for the over dispersity of the accidents. They utilized several databases, including a crash database, land-use data, and household surveys in Calgary; some variables reflected land use, demographic conditions, travel, and area characteristics. They found that the number of daily trips contributes to increasing the number of accidents. Another founding was the strong correlation between the daily commercial truck trip with the frequency of accidents.

Poul Greibe[20] performed an analysis on accidents in Denmark, assuming that accidents follow a Poisson distribution and are recurrent. He used the model structure as Equation (2.1) where $E(\mu)$ is the expected number of accidents on the road segment, N is the traffic flow, x variables are the geometry of the road, a, p, and β_j are estimated parameters. That model relates the number of accidents to road design and traffic flow on road segments utilizing GLM. He examined all the traffic flow and the road variables in the model one by one and excluded the ones that had no significant influence on the output using the maximum likelihood method. Since that single model could only perform on the road links, he developed four more models for 3 and 4 arms, signalized and non-signalized junctions, adding more complexity. The best result achieved by those models was 60%, based on the goodness-of-fit metric.

$$E(\mu) = aN^p \; exp \sum \beta_j \; X_i j \tag{2.1}$$

Another study carried out by Alfonso Montella[21] used GLM with a negative binomial distribution for accidents for rural motorways that followed Equation (2.2). The model in that study reached 65.8% accuracy predicting crashes, based on the goodness-of-fit metric. Traffic volume, road characteristics, and speed limits are used as contributing factors, and their main finding was that the road design consistency has the most influence on accidents.

$$P\left(Y = y/\Lambda = \lambda\right) = \frac{\lambda^y \times e^{-\lambda}}{y!} \tag{2.2}$$

A special case of GLM that uses binomial distribution and logit function is called Logistic Regression, which is mainly used for binary classification problems. Logistic Regression models are flexible general models that are used for a variety of applications in road safety. Abdel-Aty et al.[22] developed a real-time crash likelihood prediction model using a matched case-control Logistic Regression model. Their model achieved 69% accuracy predicting crashes. Another study carried out by Tao Lu et al.[23] utilized Logistic regression to find the accident hot-spots in Beijing city. They used different factors, such as road type, vehicle type, driver state, and visual conditions, to develop the model and achieved 86.7% accuracy.

Although classical approaches are widely used throughout the literature, the main downside of those

models is the insufficient ability to deal with complex and non-linear relationships. Furthermore, adding more factors such as detailed weather features to the models will add more complexity. Thus, those models are not a good candidate for the analysis and APMs development considering new factors.

2.3 Novel ML Studies

There are two trends in the literature for APM development using ML methods, accident severity prediction and accident occurrence prediction. This thesis is carried out based on the latter, which is the accident occurrence prediction; however, the models used in these two trends are the same, but the analysis and methodologies are different.

The standard ML models used in the literature are Neural Networks (NN)[6, 24, 25, 26, 27], Support Vector Machines (SVM)[5, 28, 29, 28], Decision trees[5, 30], Random Forests[31, 32, 33], Convolutional Neural Networks (CNN)[34, 35, 36], Naive Bayes[37, 38], Long Short-term Memory Recurrent Neural Networks(LSTM-RNN)[39], K-nearest Neighbours (KNN)[40, 41], Deep Neural Networks (DNN)[42, 43], and the hybrid models combining the models mentioned above[30, 44].

Neural Networks have shown promising performance in extracting complex relationships and finding the latent patterns within the feature space; thus, they have been used several times in different traffic accident prediction studies. In a study by Akin et al.[24], a Neural network with standard back-propagation was used to classify the intersection's accidents as fatal, injury, and property damage accidents. They used the hyperbolic tangent function as the activation function and managed to get to 91% accuracy. Murat et al.[45] developed a real-time accident detection model utilizing a feed-forward Neural Network and the sensor data on a highway in Istanbul, Turkey. They also used the traffic data and reached 99% accuracy; however, the false alarm rate was high due to the imbalanced nature of the data. Thus their model just could be used as an accident warning system.

M.A Sahrai et al.[46] compared the performance of a Neural Network with a Logistic Regression model to predict future accident frequency and the risk factors of accidents. They used location, date, road characteristics, traffic volume, summer/winter variables as the feature space. They concluded that the NN model has a high capacity to predict the frequency of accidents on different road sections. Using the proposed model, they predicted that accidents will grow 11.2% annually from 2020 to 2030 in the study area. Deep Neural networks (DNN) are used in a few studies containing more layers in order to find the more complex patterns; for instance, Zekun et al.[42] utilized a DNN-based framework for multi-task severity accident prediction. Their model outputs three different categories for injury severity, death severity and property loss severity from 25 independent input variables. They compared the results of their model with a regular

neural network and a Logistic Regression model. They showed that their model outperformed the other two models in accuracy and AUC scores.

Recurrent neural networks (RNN) are a class of Neural Networks with memory elements to keep track of the short dependencies in a data sequence; however, to find the long-term temporal dependencies in a data sequence, LSTM-RNN is introduced. In addition, if the accident is assumed as a recurrent event in relationship with the previous accidents, the future accidents can be predicted by LSTM-RNN models. Abdel-Aty et al.[39] took advantage of that scenario and developed a real-time APM to predict the crashes using historical and live data from loop-detectors. For the sake of comparison, they developed a Logistic regression model on the same problem to show the effectiveness of their proposed model. The results showed, even though the accuracy of the LSTM-RNN model output is 7% higher than logistic regression, the models are not good enough for deployment purposes.

Another technique used in the literature is Support Vector Machines(SVM), a linear classifier that can be used to find non-linear relationships with modified kernels. SVM models are not perfect classifiers for accident prediction applications but have fewer complexities compared to NN models. A few studies, such as [47, 48], used SVM models and employed tweet data from Tweeter APIs for real-time traffic accident detection. The drawback of these studies is the ability only to detect accidents after the occurrence and not predict the accidents.

Tree-based methods such as Decision Trees, Random Forests, and Gradient boosts have received less attention than NN-based methods and statistical methods. Their capacity for making decisions on structured heterogeneous datasets has been underestimated. In an analysis by Shakil et al.[38], a Decision tree predictive model was developed to predict the casualty severity of accidents in the UK; they also compared the result with a Naive Bayes model and showed the higher performance of the former model.

Random Forests are used primarily in the literature to estimate variable importance [32, 49, 31], by measuring the prediction error while the variables are permuted. RF models have been used rarely for accident predictions; however, in a study by Yassin et al.[33], the authors used an RF model to predict the severity of accidents into serious injury, light injury, and fatal injury using variables such as driver experience, light conditions, driver age, time, and date. The experimental results of their study showed promising performance of the RF model compared to the conventional models. Another study conducted by Kim et al.[43] compared three models, NN, RF, and Gradient Boosting decision trees. They developed those three models to predict the occurrence of accidents in different time intervals in a port in Koria. The models reached 98.2%, 90.4%, and 98.3% accuracy, respectively, for 1 hour time interval, which showed the high capability of those three models.

2.4 Knowledge Gap

As mentioned in the previous sections, all ML-based studies include two main parts, the data and the model. As discussed in section 2.3, different ML methods, including stand-alone models and hybrid models, have been widely used in the literature for APM developments. Furthermore, each model performs differently based on how the problem is articulated and what kinds of data are used for the analysis. This section presents different data and various features used in the literature to show the existing gap in the literature as the primary motivation of this thesis.

Table 2.1 represents all the different features used in the literature for APM developments; road's characteristics [50, 51, 24, 49, 52, 37, 46] are the features that are used the most, then the traffic volume [49, 31, 53, 46] is the second attribute. Following those, speed, average speed and the speed limits [22, 27, 31, 37] are seen repeatedly in the literature. Driver's aspects [50, 54, 33] are the next feature used in the previous studies, followed by vehicle characteristics [24, 27, 33]. Some studies used distinctive features such as GPS data [36], Tweeter data [48, 47], and congestion index [31]. Among all those features, the weather attributes have got little attention; many studies completely ignored the weather [28, 26, 27, 48, 55, 56, 45, 57, 53, 37, 35, 38], and some only considered one factor for the weather, which could be sunny, cloudy, snowy, rainy, Etc [51, 22, 30, 49, 58, 59, 46].

Few studies considered two weather features in their analysis, such as overall weather with visibility or temperature [60, 36]. However, due to the study area's environment, those features have not shown a significant impact on the accident frequency or severity. Moreover, although one study [43] utilized four different weather features in the analysis, including temperature, humidity, wind speed, and precipitation, due to the location of the study, which has a subtropical climate with infrequent events of extraordinarily high or low temperatures, these features did not have a significant impact on the prediction in that study. Thus, the question has still remained how important the weather features are in accidents in Canadian cities with severe winter conditions and how those features can be used to develop APMs.

Table 2.1: Different features used in APM development

Paper	Weather features	Features	Year
[51]	Dry vs wet	Driver characteristics, vehicle characteristics, accident type, road surface condition, and light condition.	2004
[22]	Summer vs Non-summer	Speed, traffic volume, and average occupancy	2004
[50]	Rainy vs sunny	Speed limit, road characteristics, time and date, driver characteristics	2007
[28]	No weather feature	Traffic volume, speed, and time headway	2009
[24]	Overall weather	Time, light and surface conditions, driver and vehicle and road characteristics	2010
[26]	No weather feature	Traffic volume, location, intersection type, traffic control mode	2010
[30]	Overall weather	Speed limit, road description, date, time, light conditions, vehicle and driver characteristics	2010
[27]	No weather feature	Traffic volume, speed, driver and road characteristics, crash type, light condition, accident place	2011
[49]	Clear vs adverse	Traffic, road characteristics, vehicle speed, traffic occupancy	2013
[29]	No weather feature	Speed, occupancy and traffic volume	2013
[60]	Sunny vs non-sunny & Visibility	Accident type, time, date, location, vehicle type, road conditions	2013
[48]	No weather feature	Twitter stream	2015
[47]	No weather feature	Twitter stream	2015
[31]	No weather feature	Traffic volume, peak hour, average speed and congestion index	2015
[55]	No weather feature	Driver characteristics, Accident category, Time , date, Location, Lighting, Accident severity, Road type	2016
[56]	No weather feature	Location, time, severity level	2016
[45]	No weather feature	Velocity, Occupancy, Capacity usage, Weekday/Weekend, Rush Hour	2016
[54]	Overall weather	Date and time, accident type, driver characteristics, speed limit, light condition, road characteristics	2016
[58]	Overall weather	traffic flow, light	2017
[52]	Overall weather	human factors, vehicle factors, road geometric factors, traffic factors	2017
[57]	No weather feature	location, time, longitude, latitude and accident type, traffic	2018
[61]	Moderate and heavy rainfall	Speed, driver behaviour	2018
[62]	Overall weather	Accident features, roadway features, environmental features, vehicle features, casualty features	2019
[53]	No weather feature	Traffic data, road characteristics, crash type, speed limit, date, car characteristics	2019
[33]	Overall weather	Time, Driver characteristics, Type of vehicle, Service year, Location, Road condition, Light condition	2020
[59]	Overall weather	Accident severity, time, day, road surface condition, location, daylight condition, vehicle type, pedestrian characteristics	2020
[37]	No weather feature	Speed limit, road width, types of signs, pavement, road pattern, crossroads	2021
[35]	No weather feature	Driver characteristics, time, accident data	2021
[43]	Temperature & Humidity & Wind speed & Precipitation	Movements of containers, accident data	2021
[36]	Overall weather & Temperature	Accident and road data, taxi GPS data	2021
[42]	Overall weather & visibility	Driver characteristics, time, road condition	2022
[38]	No weather feature	Driver characteristics, road condition, vehicle condition	2022
[46]	Summer vs winter	Date and time, road characteristics, traffic volume	2022

2.5 Generalized Linear Models

Nelder and McCullagh first formulated a Generalized Linear Model (GLM) to model the response variable y_i as a function of explanatory variables $x_i^T\beta$.

There are three components in a GLM:

- A random component that represents the probability distribution of output y_i (it can be Normal, Poisson, Binomial, etc.).
- Systematic components specify the linear combination of the input x_i variables ($\beta_0 + \beta_1 x_1 + \beta_2 x_2 +$ ..., where β_i are the estimated parameters)
- A link function shows the relationship between random and systematic components. The link function specifies the relationship between the expected value of the response variable with the linear combination of the input variables.

For different applications, distinctive probability distributions and unique link functions are used. For instance, if the identity link function is used incorporating the normal distribution, the whole equation will be a linear regression model as shown in Equation (2.3).

$$y_i = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + ... + \beta_n x_n \tag{2.3}$$

2.5.1 Binary Logistic Regression

Binary Logistic Regression is a special case of GLM when a Binomial distribution is assumed for the output y_i, and a logit function as the link function as shown in Equations (2.4), and (2.5), where π_i is the probability of success for the variable y_i, x_i are the explanatory variables, and β_i are estimated parameters.

$$Logit(\pi_i) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + ... + \beta_n x_n \tag{2.4}$$

$$Log(\frac{\pi_i}{1-\pi_i}) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + ... + \beta_n x_n \tag{2.5}$$

Binary Logistic Regression is used for classification applications where the output variable is categorical such as 0 or 1, pass or fail; however, this model suffers from complete separation, which means if two explanatory variables separate the output classes perfectly, the model tends to optimize the weight to infinity and never converges. To deal with the mentioned problem, regularization terms are used while estimating the β_i parameters.

2.6 Machine Learning Models

Machine Learning is a subset of Artificial Intelligence (AI) that deals with algorithms and data to mimic the way humans learn during a training process. ML is divided into three main categories, supervised, unsupervised, and reinforcement learning. Supervised learning uses labelled datasets to classify new samples or predict new outputs. Supervised learning has applications in many domains and includes methods such as Neural Networks, Decision Trees, Naive Bayes, Support Vector Machines, Random Forests, to mention a few. Unsupervised learning is mainly used for clustering unlabelled datasets into subgroups. It is a powerful tool for finding the similarities and differences in data. Dimensionality reduction is another application of unsupervised learning, which removes the variables with high similarities to reduce the complexity in problems with many explanatory variables. Two well-known unsupervised methods are K-means clustering and Principal Components Analysis (PCA). And finally, Reinforcement learning is a sort of behavioural learning that finds a sequence of successful attempts by trial and error for a specific problem.

Machine Learning models have been getting more attention over the last decade due to the advancements in computational power. The main advantage of ML models is the prediction power at the expense of so-called less interpretability than the statistical models, especially when the problem contains many explanatory variables. The following sections explore some considerations and four supervised ML models that are used in this thesis for APM developments.

2.6.1 Model Evaluation

Evaluating the results in any machine learning application is crucial since the results show if the proper methods and procedures have been taken along the model development. Furthermore, the results are the most important part of any study as they answer the questions that initially caused the model development. In this thesis, the prediction of accidents is formulated as a binary classification problem. The binary classification results can be shown in a matrix form called confusion matrix, as shown in Figure 2.1. The confusion matrix shows a comparison between the actual samples and the predicted samples, in other words, a comparison between ground-truth and the model predictions as follows:

- TP or True Positive is the number of positive samples predicted as positive
- FP or False Positive is the number of negative samples predicted as positive
- TN or True Negative is the number of negative samples predicted as negative
- FN or False Negative is the number of positive samples predicted as negative

False Positive and False Negative are the two important metrics of a classification model, which show the number of misclassified samples. Based on the application, either of those might be critical and high values of those can have drastic consequences. For example, a False Negative shows the model failed to predict an accident in an accident prediction application. For instance, assume you are using a mobile application that predicts the occurrence of accidents during your daily commute, and the goal is to avoid roads with probable accidents. The application predicts that the road segment you are taking will not have any accidents, but later, you will be involved in an accident that you did not expect. If this situation is infrequent, it means the model performance is acceptable since the False Negative is low; however, if that situation happens again and again, it means the False Negative is high, and the model performance is not satisfactory.

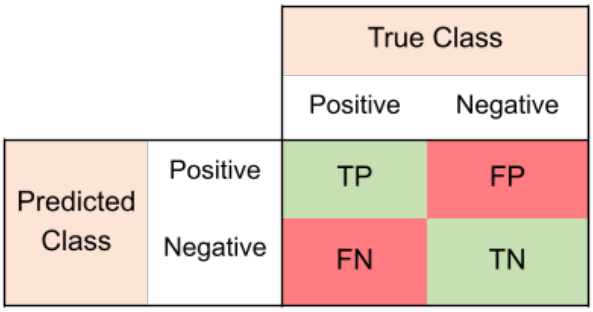


Figure 2.1: Confusion matrix

The following metrics can be derived from the confusion matrix:

Accuracy

Accuracy shows the model's overall performance and is calculated by the number of correct predictions over the number of predictions, as shown in Equation (2.6).

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \tag{2.6}$$

Precision

Precision shows how many percent of the positive predicted samples are truly positive and calculated by Equation (2.7).

$$Precision = \frac{TP}{TP + FP} \tag{2.7}$$

Recall

Recall shows how many percent of the positive samples are correctly predicted and calculated by Equation (2.8).

$$Recall = \frac{TP}{TP + FN} \tag{2.8}$$

As an example, consider a case where a model predicts 100 samples of accidents and non-accident, in which 60 samples are accidents and 40 are non-accidents. After the prediction, the results show that, for example, 50 out of 60 accidents are correctly predicted, and 35 out of 40 non-accident are also correctly predicted. The accuracy shows how many samples are correctly predicted; in this case, 85 out of 100 samples are correctly predicted, which shows the accuracy of the model is 85%. The Precision shows how many percent of accident predicted samples are truly accidents; in this case, 50 accidents are correctly predicted, and five non-accident are predicted as accidents; therefore, the Precision is equal to 50/(50+5), which is 90%. The Recall shows how many percent of accidents are correctly predicted as accidents; in this case, there are 60 accidents, and the model only predicted 50 of those accidents as accidents; therefore, the Recall score is 83%.

F1 Score

F1 Score is the harmonic mean of Precision and Recall representing a combination of those metrics and calculated by Equation (2.12)

$$F_1 Score = 2(\frac{Precision * Recall}{Precision + Recall}) \tag{2.9}$$

ROC Curve

The Receiver Operating Characteristic curve (ROC Curve) plots the true-positive rate against the false-positive rate with considering different cut-off thresholds for the output. Some classification models such as logistic regression and Neural networks produce a probability for the predicted output. By considering a cut-off threshold, that probability is converted to a label; for example, if the cut-off is set to 0.5, the probabilities greater than 0.5 are converted to 1, and smaller than 0.5 are converted to 0. The cut-off threshold can affect the performance of the model. ROC curve not only shows the model's performance but also provides the optimal cut-off value based on the application's requirements.

ROC-AUC

Area Under the ROC Curve (ROC-AUC) shows the performance aggregation of the model with all cut-off values for binary classification and provides a probability between 0 and 1. The greater the ROC-AUC is, the better the model's overall performance is.

2.6.2 Machine Learning Considerations

In addition to the metrics discussed in the previous section, another important concept in ML studies is Generalization. In essence, ML models are trained with historical data to predict future events; more specifically, a classifier is trained on the available data to classify unseen or new data in a classification problem. Those models can suffer from Overfitting and Underfitting. Overfitting refers to when the model has learned the patterns and the noise too well. It somehow memorized the random fluctuations and noise that might not exist in unseen data, which leads to poor performance when dealing with unseen data. Non-parametric and complex models are prone to overfitting, and some pruning and regularization techniques need to be taken into account while training those models.

Underfitting refers to when the model has not learned the patterns well from the data and does not perform well on either the available or unseen data. The actual reason for underfitting is that either the model structure is too simple or the parameters and/or hyper-parameters in the model are not optimized.

2.6.3 Decision Trees

Decision Trees is a non-parametric supervised learning method inspired by a real tree with branches and leaves. It uses an upside-down tree-like decision model, including condition nodes, branches, and decision/leaf nodes, as shown in Figure 2.2. This figure shows a simple decision tree with three conditions that represent three features. The conditions are inferred from the data features, and the first condition is called the root node.

Each condition represents a split, which is determined by Information Gain (IG) or Gini Gain (GG). The split is chosen from a set of possible splits where the IG or GG is maximized in each condition. Two criteria are used to find the maximum IG or GG on each split, Gini impurity for GG and Information Entropy for IG. Gini impurity shows how pure each classified label is, and a Gini impurity of 0 is the best possible impurity, which is calculated by Equation (2.10), where G is the Gini impurity, C is the number of classes, and $p(i)$ is the probability of randomly picking a sample in class i.

$$G = \sum_{i=1}^{C} p(i)(1 - p(i)) \tag{2.10}$$

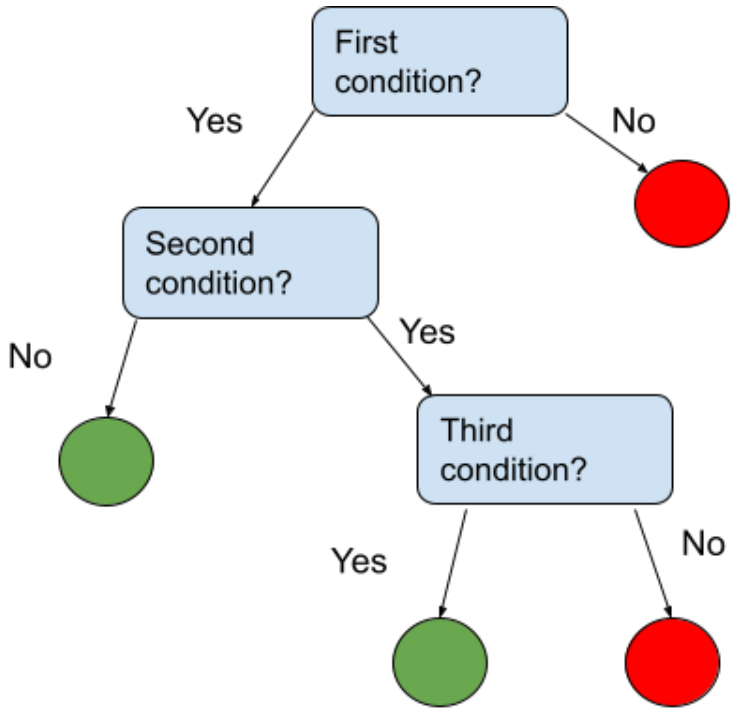


Figure 2.2: Decision Trees Structure

Information Entropy measures the variance in a set of samples or splits and determines the goodness of splits on each condition. Information Entropy is calculated by Equation (2.11) where E is the Entropy, C is the number of classes, and $p(i)$ is the probability of picking a sample in class i. When a split is performed, the impurity of the parent node is compared to the sum of the impurities of the child nodes to evaluate how good that split is, and the best split will maximize the comparison.

$$G = -\sum_{i=1}^{C} p(i) \log_2(p(i)) \tag{2.11}$$

Construction of the decision tree starts from finding the best candidate for the root node utilizing IG or GG, and it goes on to cover all the features in the feature space; however, when the feature space contains numerous features, some stopping criteria are considered to avoid building a very complex tree than can result in overfitting. The stopping criteria can be one of the following conditions:

- The number of conditions is equal to the maximum predefined depth
- No candidate produces the IG or GG greater than the minimum predefined IG or GG
- No candidate has a number of instances greater than the minimum predefined instance per node

Decision trees are simple to understand and visualize, can find the non-linear relationships, can handle both

numerical and categorical variables, and need less data preparation than other ML models. However, this method is susceptible to small changes in the data, which leads to constructing a different tree of decisions, that can be alleviated with some techniques such as bagging and boosting, which will be covered in the coming sections.

2.6.4 Random Forests

Random Forests (RF) is an ensemble supervised learning method that uses the concept of the wisdom of crowds. An RF classifier includes many individual Decision Trees, and the majority of the votes from those decision trees make the final decision. The individual trees in an RF model are constructed with the least correlation to compensate for the error in other trees. Figure 2.3 illustrates the structure of an RF classifier.

Since the correlation between individual trees is the most critical element of an RF, some procedures need to be taken into consideration. As mentioned in the previous section, Decision Trees are sensitive to the small changes in the data; therefore, by randomly picking different features out of the feature space for each tree, distinctive trees will be grown with low structure correlation; this technique is called bagging. The bagging technique helps create different trees with different error rates, and averaging the outcome of all the trees will result in higher accuracy and lower overfitting.

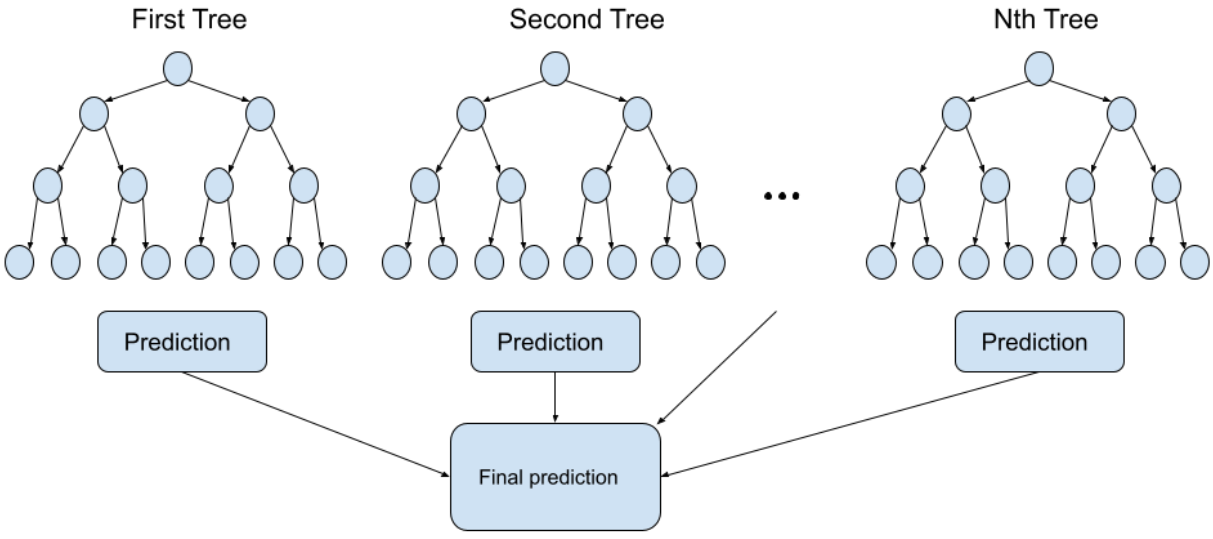


Figure 2.3: Random Forests Structure

2.6.5 Gradient Boost

Gradient Boost (GB) is an ensemble supervised learning method and a powerful technique for building enhanced predictive models. The main idea is to have a weak hypothesis or weak learner as the base model, keeping it intact and working on the shortcomings to build another weak learner and add it to the base model for better prediction. Figure 2.4 illustrates a simple GB classifier to demonstrate how a GB classifier is constructed. The top left rectangle represents the very first weak learner that classifies the given points into two categories, green circles and red triangles. This weak learner can be a simple decision tree that does not perform well as three green circles have fallen into the wrong category. GB uses those three green circles from the outcome of the first weak learner and builds another weak learner shown as the top middle rectangle in Figure 2.4 to compensate for those three wrong points utilizing the Gradient Descent method. Although the new weak learner correctly classifies those three green circles, it now misclassifies two red triangles into the wrong category. Again another weak learner is constructed based on the shortcoming of the previous weak learner. Eventually, the constructed weak learners are weighted and aggregated to build the final predictive model.

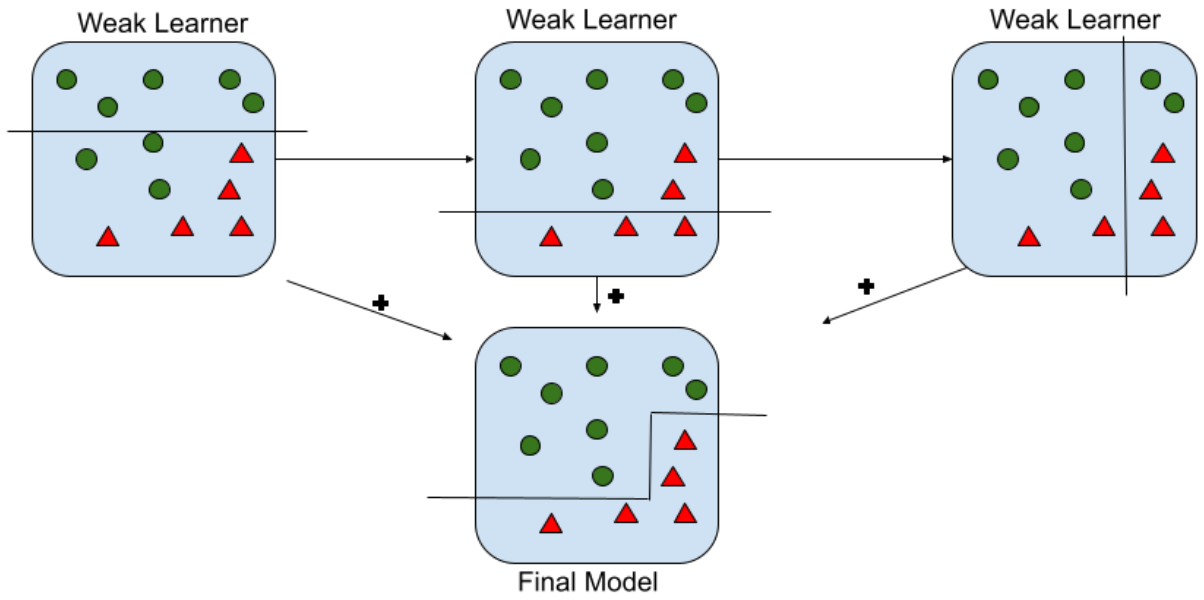


Figure 2.4: Gradient Boost structure

Gradient Descent is a well-known technique in optimization problems, which is used abundantly in machine learning applications, for example, to optimize the coefficients of a regression problem or optimize the weights in a Neural network. Since decision trees are a non-parametric algorithm, to utilize Gradient Descent, trees are parameterized to be modified by Gradient Descent, which is called functional Gradient

Descent.

To avoid overfitting, some constraints and penalization factors are considered while training the model, such as tree constraints, shrinkage, random sampling, and regularization factors. All the rules and constraints in a Decision tree classifier can also be applied in GB; however, having more weak learners will lower overfitting compared to having fewer stronger learners.

2.6.6 Neural Networks

The neural network (NN) is a supervised machine learning model that resembles humans' brains. A NN contains interconnected processing units called neurons, similar to neurons in a brain. Figure 2.5 illustrates the functionality of a neuron that applies a function to a set of inputs and creates an output.

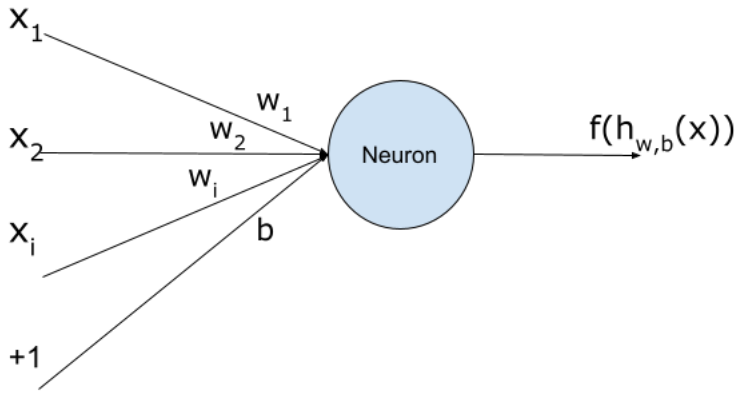


Figure 2.5: Structure of a Neuron

Based on the structure of a neuron shown in Figure 2.5, x_i are the inputs, w_i are the assigned weights to the inputs to adjust the effect of each input on output, b is a constant that adds bias to the output, in case all the inputs are zero, there will be an activation in the neuron. Function f is the activation function and can be in form of step, tanh, Sigmoid, or ReLU function, and $h_{w,b}(x)$ is a linear combination of inputs as shown in equation (2.12).

$$h_{w,b}(x) = b + \sum_{k=1}^{i} w_k x_k \tag{2.12}$$

A neural network structures the neurons in different layers: input, output, and hidden layers, as shown in Figure 2.6. The input layer includes all the attributes in the feature space; for example, in this thesis, road characteristics and weather features are applied through the input layer. The output layer represents the output of the whole model; for example, in a binary classification problem, the output is a probability

between 0 and 1 that can be converted to a label using a cut-off threshold. The hidden layers are known as black boxes, which do most of the computations. The number of hidden layers can be different based on the complexity of the problem. If there is only one hidden layer, it is called a feed-forward neural network; however, if the number of hidden layers exceeds one layer, by convention, it is called a deep neural network.

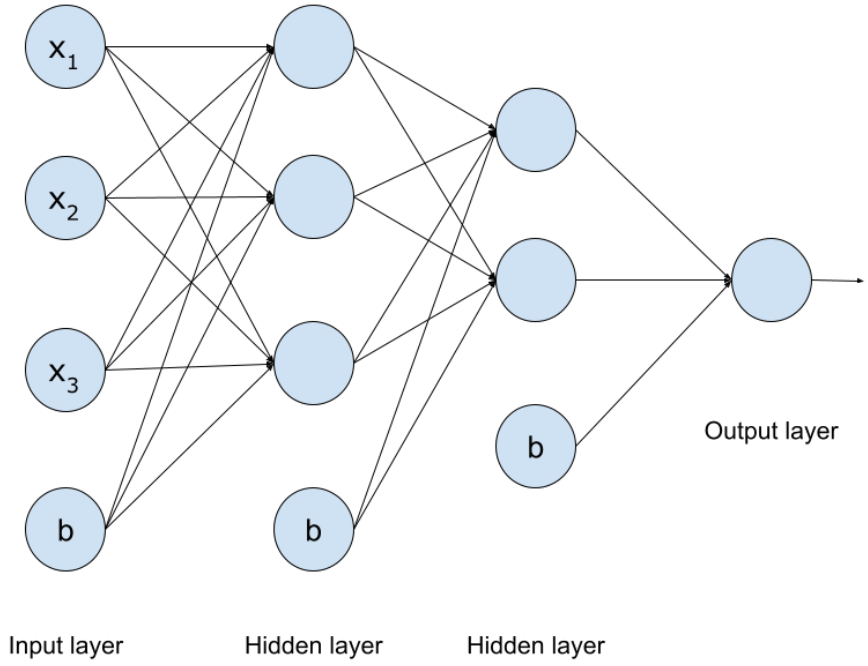


Figure 2.6: Neural network structure

The neural network is trained to find the optimal weights and biases using the back-propagation technique during a training process. At first, arbitrary values are assigned to the weights and biases in each neuron, and then the training data are applied to the input layer. The neurons in the hidden layers generate an output based on the input data and arbitrary parameters and give it to the output layer. The output layer receives the result from the hidden layer, compares it with the actual outputs, and generates an error using a loss function. The generated error goes back to the hidden layers to adjust the neurons' parameters to decrease the error in the next iteration. This procedure is repeated until reaching a minimum desired error.

Chapter 3

Methodology and Data Preparation

3.1 Data Preparation

The City of Calgary, which had a population of roughly 1.5 million in 2020, was selected as the study area for this research. By continuously building new communities, the city of Calgary is growing fast; therefore, it is essential to provide timely recommendations to policymakers for the development of the city, considering the safety aspect of the roads.

In any data-centred study, the fundamental part of the study is the dataset; therefore, accurate data preparation is critical. An accident in this study is defined as a collision between two or more vehicles on a road segment. In addition, predicting an accident refers to estimating the probability of collision occurrence at a specific road segment and time using relevant explanatory variables.

This thesis utilizes different data preparation techniques: filtering, sorting, cleaning, combining, dealing with missing values, scaling, transforming, encoding, normalizing, undersampling and oversampling to have the most precise data. Since the extracted data from the sources mentioned earlier contains too many missing values and noise, different techniques are applied to deal with this issue, such as interpolation, padding, mean insertion, imputation, and manual insertion. The congregated dataset comprises a group of 15 features, including three categorical features encoded in numbers (month, day, hour), three categorical features encoded in strings (road class, road segment, and weather), and seven numerical features (temperature, wind speed, visibility, relative humidity, pressure, wind direction, dew point temperature, road speed and road traffic).

All the weather features are measured hourly in Calgary's International airport station, which is accessible on the Government of Canada's website[63]. The temperature is measured with degrees Celsius; wind speed

is km/h; Visibility is measured by Km; humidity is measured based on percentage above 0% and below 100%. Road traffic shows the annual average daily traffic on each road segment. The daily traffic volume is used in this analysis because the hourly traffic volume for the study area is not available. In the absence of accurate data, using the average data for the model development is common. In other words, using average traffic volume data is better than not using any traffic volume data. Road speed represents the posted speed limit. Road class represents the category of roads containing arterial street, skeletal road, collector, urban boulevard, industrial arterial, parkway, neighbourhood Blvd, and residential street. Time features contain the month, day, and hour. The weather includes clear, cloudy, snow, rain, mostly cloudy, and so on.

3.2 Data Augmentation

As a rule of thumb, real-world datasets encompass mostly normal samples with a minority number of interesting samples, leading to an imbalanced dataset[64]. The accident prediction problem in this study is defined as a binary classification problem that needs two categories: positive accident events and negative accident events. After the first preprocessing stage, the accident dataset is a set of accident samples that all show the positive events of accidents and are considered interesting samples. This dataset is completely imbalanced without even one sample showing a non-accident event; therefore, negative accident events must be generated.

Accidents are assumed to happen in 1 hour time slot, thus, the general approach to generating negative samples is to consider all the time slots in which accidents have not happened from 2017 to 2020. There are approximately 14,835 samples of accidents in 100 different road sections during the given four years in the dataset. If all the one hour time slots without accidents over the 100 places are considered non-accident events, there will be approximately 3.5 million non-accident events. In this case, the proportion of negative samples will be way more than positive samples. The dataset will be highly imbalanced with a ratio of 236:1, which means there are roughly 236 negative samples in the dataset for each positive sample.

In many applications, such as intrusion detection and disease detection, and this thesis, correctly predicting the minority class is essential with highly imbalanced datasets. For instance, the dataset may contain 99% negative samples compared to 1% positive ones; then, the accuracy could get to 99% without even predicting one positive sample [65]; in this case, the model is highly biased, leading to biased predictions and misleading accuracies.

One approach to deal with this issue is resampling, which includes oversampling the minority group or undersampling the majority group. In this study, both undersampling on the majority class and oversampling on the minority class are examined, and a hybrid approach is taken to deal with the class imbalancity of the

dataset.

In the first step, undersampling on the majority class, the negative samples, is performed utilizing six common methods and the results are shown in Table 3.1. Random Undersampler picks a portion of samples in the majority class randomly using a random function[66]. Since the random function selects samples differently using different random states, undersampling is performed ten times with different random states, and the average of the metrics is used.

Table 3.1: Undersampling the Majority Class

Method	Accuracy %	Recall %	Precision %	F1 Score
Random Undersampler	66	65	65	0.65
NearMiss version 1	61	62	61	0.61
NearMiss version 3	65	66	64	0.65
AllKNN	80	82	81	0.82
EditedNearestNeighbours	76	77	76	0.76
RepeatedEditedNearestNeighbours	82	85	83	0.84

The next method is NearMiss, which uses the nearest neighbours algorithm to pick the most distinctive samples based on the average distance between samples[67]. The next method is Edited Nearest Neighbours, which uses the nearest neighbours algorithm to pick the samples and then removes samples that do not agree enough with neighbours[68]. Repeated Edited Nearest Neighbours works the same as the Edited Nearest Neighbours method, but the undersampling is performed several times to perform better[69]. Moreover, the last method is All kNN which is similar to the Repeated Edited Nearest Neighbours but increases the number of neighbours at each iteration[70]. Since different undersampling methods have different results based on the nature of the data, a Classification model using a Neural network is used to evaluate examined approaches. Table 3.1 shows the result of the undersampling experiment with four essential metrics: Accuracy, Recall, Precision, and F1 Score. Based on the results, the most appropriate undersampling method for the accident dataset is the Repeated Edited Nearest Neighbours method which shows the highest performance in all metrics.

In the second step, oversampling on the minority group, the positive samples, is performed. Even though there are several oversampling approaches in the literature, such as Random Oversampling[71], SMOTE[72], ADASYN[73], KMeans SMOTE[74], SVM SMOTE[75], and Borderline SMOTE[75], according to the categorical features in the accidents dataset, Only Random Oversampling and a subset of SMOTE called SMOTE-NC is performed and the results are brought in Table 3.2. The rest of the techniques can perform oversampling on a dataset including only numerical features.

One of the limitations of oversampling in this thesis is that oversampling is performed by multiplicating the existing positive samples, not generating new positive samples. This multiplication might result in

contamination or leakage in the test set, which needs to be taken care of when the model's performance is evaluated. In order to avoid having contamination or leakage in the dataset, a small proportion of the data as the test set is held aside without involving in the oversampling process for the model performance evaluation. In future works, the possibility of using techniques such as Generative Adversarial Networks (GAN) can be examined to generate new samples of accidents to have better predictions.

Table 3.2: Oversampling the Majority Class

Method	Accuracy %	Recall %	Precision %	F1 Score
Random Oversampler	80	87	77	0.82
SMOTE-NC	72	73	72	0.72

Based on the results in Table 3.2, Random Oversampler performed better in all metrics. Thus, a hybrid approach including Repeated Edited Nearest Neighbours Undersampler and Random Oversampler are utilized to deal with extreme class imbalancity of the accidents dataset in this thesis.

As the result of over/under-sampling and other preprocessing steps mentioned earlier, the final dataset encompasses 2 million balanced samples, one million generated synthetic negative samples, and one million positive samples that are generated by random multiplications of the existing positive samples. The final dataset has 15 explanatory variables, including month, day, hour, road class, road segment, road speed limit, traffic volume, weather, temperature, wind speed, visibility, relative humidity, pressure, wind direction, and dew point temperature. Table 3.3 shows all the explanatory variables with their variable types, measuring units and nominal values in the dataset.

Table 3.3: Explanatory variables

Feature	Variable type	Nominal value and measuring unit
Month	Categorical/Discrete/String	January, February,...,December
Day	Categorical/Discrete/Integer	1,2,...,31
Hour	Categorical/Discrete/Integer	0,1,...,23
Road class	Categorical/Discrete/String	Skeletal Road, Arterial Street,Residential Street, Urban Boulevard, Industrial Arterial, Neighbourhood Boulevard, Parkway, Collector, Primary Collector
Road segment	Categorical/Discrete/String	Deerfoot TR SE, Glenmore TR NE, 16 Av NE,... (100 road segments)
Speed limit	Numerical/Discrete/Integer	30, 40, 50, 60, 70, 80, 100 (km/h)
Traffic volume	Numerical/Continuous/Integer	from 5000 to 110,800 (Annual Average Daily traffic)(Vehicle/day)
Weather	Categorical/Discrete/string	Cloudy, Clear, Snow, Rain, Fog, Smoke, Ice Crystals, Haze, Thunderstorms, Drizzle
Temperature	Numerical/Continuous/Float	From -32.2 to +36.4 (°C)
Wind speed	Numerical/Continuous/Integer	From 0 to 69 (km/h)
Visibility	Numerical/Continuous/Float	From below 0.1 to 80.5 (km)
Relative humidity	Numerical/Continuous/Integer	Greater than 0% and below 100%
Standard pressure	Numerical/Continuous/Float	From 86 to 91 (kPa)
Wind direction	Numerical/Continuous/Float	From 0.0 to 36.0 (10s deg)
Dew point	Numerical/Continuous/Float	From -35.8 to 18.7 (°C)

3.3 Feature Analysis

In this section, an analysis is performed on the feature space to find the feature importance and feature correlation on the generated dataset. In first section, a correlation measure is performed to find the relationship between each feature and accidents. The Mutual Information(MI)[76] technique is utilized to deal with this dataset that contains both numerical and categorical features. The mutual score shows the dependency of accidents with individual features using reduction in entropy. A high mutual score means there is some shared information and patterns between associated features and accidents, and a low or zero score means no common information.

The green bars in Figure 3.1 show the score of different features using the MI method. Based on this result, seven attributes including road class, speed, road segment, traffic volume, visibility, the hour of the day, and weather have more common information with accidents than other attributes. Among those features, visibility and weather belong to the weather features that shows weather features are related to accidents, based on MI score.

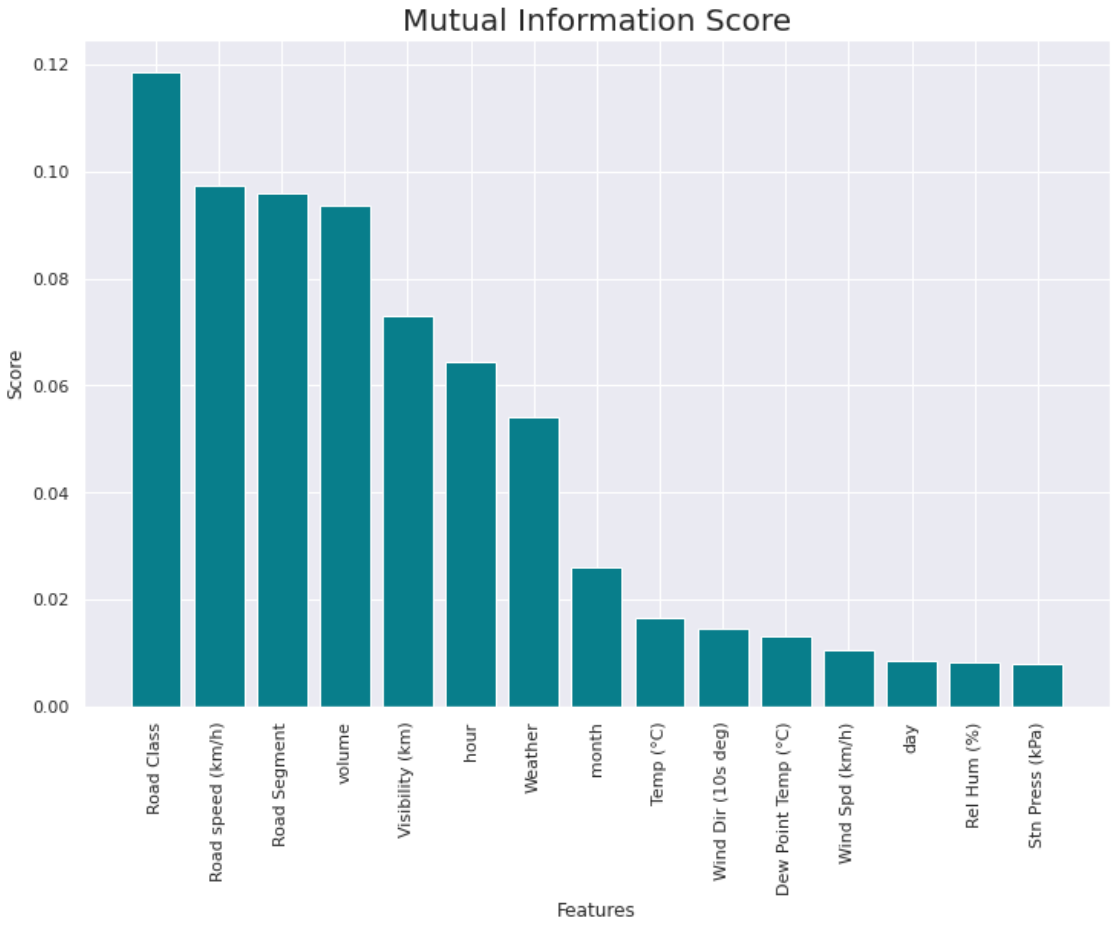


Figure 3.1: Mutual Information Score

Following the feature correlation, the box-plot method is used to visually investigate the result of the above feature importance based on the data distribution on some categorical features. As shown in Figure 3.2, the distribution of hours in accident and non-accident events is more distinctive than days; that is why hours are more important than days in terms of accident occurrence, and it has a higher score in MI.

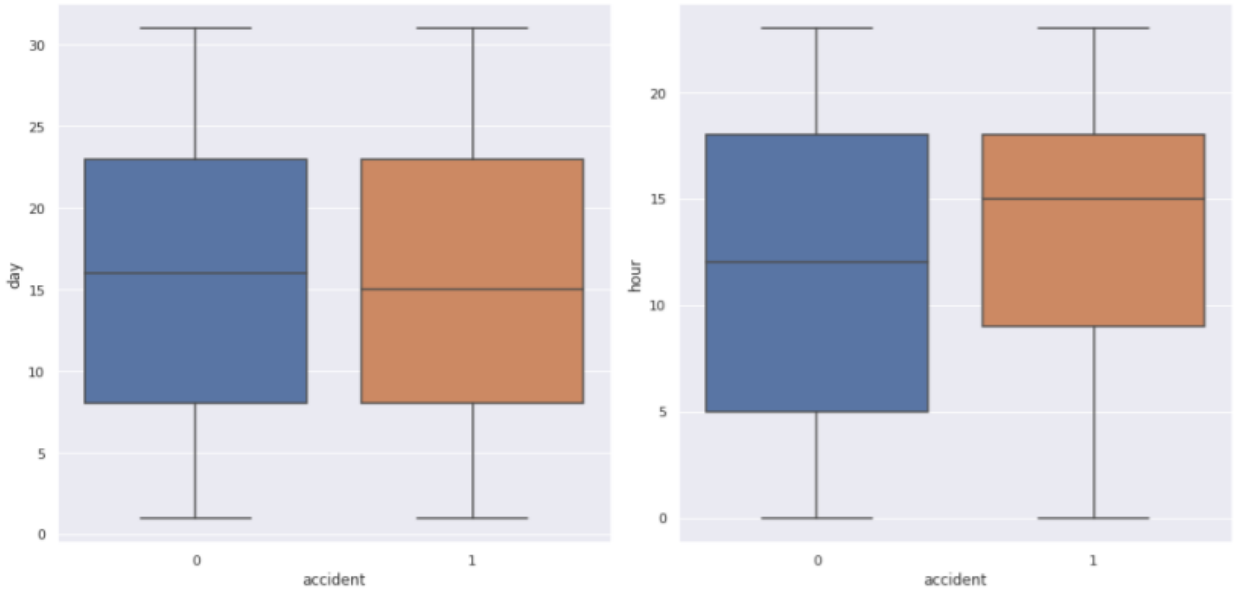


Figure 3.2: Distribution of day and hour in the dataset using Box-plot, blue is the distribution of non-accident and orange is the distribution of accidents, left figure is the distribution with respect to day and right figure is the distribution with respect to hour.

The correlation matrix is another tool that shows the similarities between pairs of features. A correlation matrix is computed on all the features and brought in Figure 3.3. The matrix diameter has the highest score and shows the correlation of each feature with itself which is always 1. Based on this matrix, temperature and dew point temperature are interconnected and carry similar information. After that, road speed and traffic volume are similar with a lower degree. Following those, road segment and road speed also have some similarities, and then road segment and traffic volume. Temperature also has some similarities with the visibility.

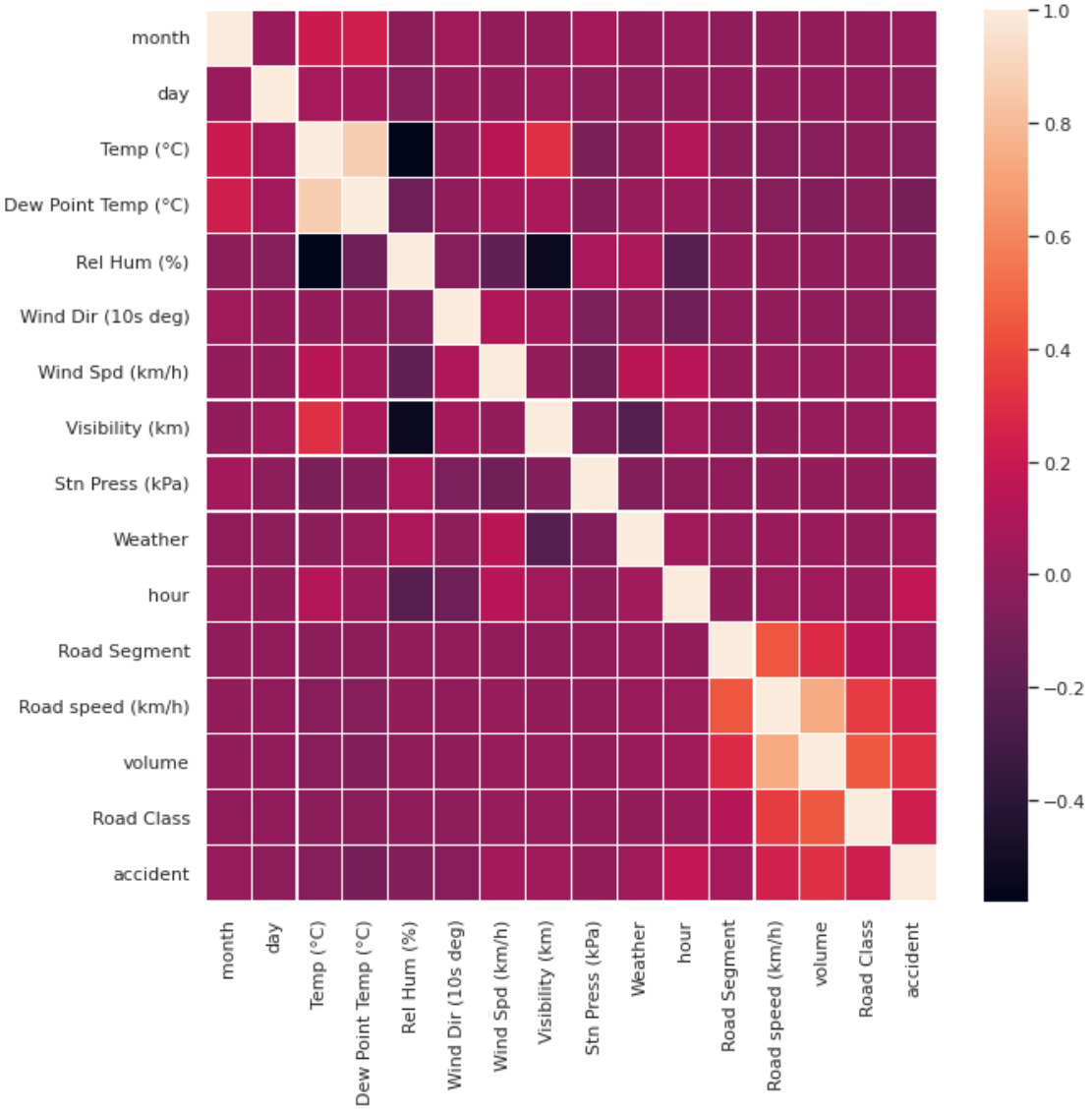


Figure 3.3: Correlation Matrix that shows the correlation between each variable with other variables in the feature space, light red is more correlation and dark red is less correlation.

Thus far, the importance of all different attributes in the dataset has been investigated; some features such as road class, speed, road segment, and some weather attributes are more important than others, but the question is how important the other features are and can they be removed from the dataset without harming the whole dataset since they have less information to provide. Principal Component Analysis (PCA)[77] is a well-known method for dimensionality reduction in problems with too many variables to remove variables that are too similar to others and have no new information but impose more computational expenses. PCA transforms the data into principal components based on the number of features and shows the amount of variance or sparsity using various components.

PCA has been used to evaluate the effect of feature exclusion, and the results are brought in Figure 3.4. The green curve in Figure 3.4 illustrates the variance of using the various number of components, in other words, the importance of using various features. For instance, using only two components has 40% of all variance, and using eight components, has roughly 80% of all variance. Based on this curve, at least 14 out of 15 features must be used to have all the variance within the dataset. The two features that carry almost the same variance could be temperature and dew point temperature, according to the correlation matrix in Figure 3.3. Even though it seems removing one of these two features will not harm the dataset, none of them is removed since only one extra feature will not increase too much computation expenses in this case. In conclusion, all the 15 provided features will be utilized for model development.

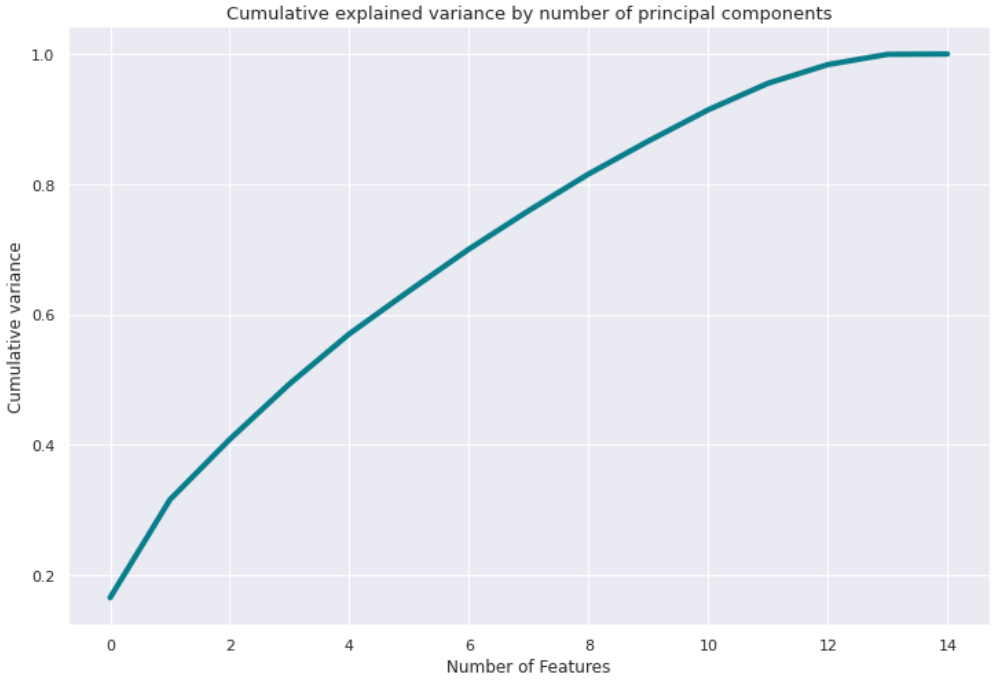


Figure 3.4: Principal Component Analysis that shows the percent of variance considering different number of features.

3.4 Models

This Section explores the development of logistic regression and four ML models, including Decision tree, Random Forest, Gradient Boosting, and Neural Network. Since the aggregated dataset includes numerical and categorical explanatory variables, some preprocessing techniques have been applied in the previous stages to prepare the data for model development. Data Cleaning, Data Transformation and Normalization, Data Reduction, Category Encoding, and data resampling have been employed in the preprocessing stage. Then, the balanced dataset including 2 million samples of accidents and non-accident are used to train the models.

Figure 3.5 illustrates the entire pipeline used for model development. Road characteristics, historical collision data, and comprehensive weather data are aggregated into a CSV file after preprocessing unit. After cleaning, the data is divided into two parts; 80% of the data as the training set for training the models, and 20% of the data as the test set to test the actual performance of the trained models.

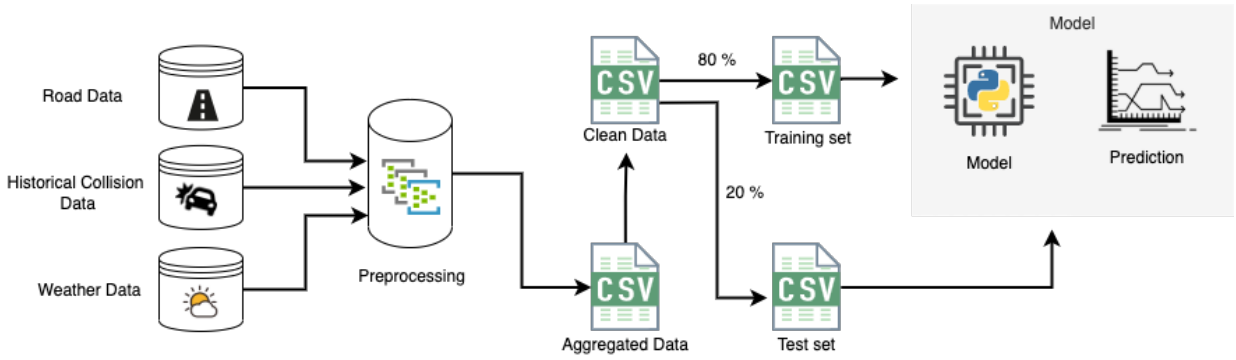


Figure 3.5: Model pipeline flowchart

3.4.1 Logistic Regression

Logistic regression is a special case of GLM, which is a linear model for classification that uses a logistic function. Python programming language is used utilizing the Scikit-learn[78] framework to implement Logistic regression in this thesis. As mentioned in the previous sections, the basic Logistic Regression suffers from complete separation; and a regularization term is added to the cost function to alleviate that effect by penalizing big values of weights. The cost function in Scikit-learn library is in the form of Equations (3.1), (3.2), or (3.3), which provide three regularization terms called l_1, l_2, and Elastic-Net, respectively.

$$min_{w,c}\frac{1}{2}w^T w + C\sum_{i=1}^{n} log(exp(-y_i(X_i^T w + c)) + 1) \tag{3.1}$$

$$min_{w,c}||w||_1 + C\sum_{i=1}^{n} log(exp(-y_i(X_i^T w + c)) + 1) \tag{3.2}$$

$$min_{w,c}\frac{1-\rho}{2}w^T w + ||w||_1 + C\sum_{i=1}^{n} log(exp(-y_i(X_i^T w + c)) + 1) \tag{3.3}$$

where X_i is the feature space, y_i is the labels vector, w is a matrix of coefficients, c are the estimated parameters, and ρ is a constant to adjust the regularization strength.

Scikit-Learn is equipped with several hyper-parameters in the Logistic Regression library as shown in Table 3.4. 'penalty' specifies which of the l_1, l_2, or Elastic-Net is used in the cost function; 'tol' is the

stopping criteria of the solver; 'C' is a constant parameter in the regularization terms; 'random_state' sets the random seed used by the Python programming to run random function; 'solver' specifies the optimization algorithm to minimize the cost function; 'max_iter' sets the maximum number of iterations for the solver; 'multi_class' specifies if it is a binary or multi class classification problem.

Table 3.4: Scikit-Learn Logistic Regression Hyper-parameter

Parameter	Value	Default
penalty	'l1', 'l2', 'elasticnet', 'none'	'l2'
tol	a float number	1e-4
C	a float number	1.0
random_state	int, random state instance	None
solver	'newton-cg', 'lbfgs', 'liblinear', 'sag', 'saga'	'lbfgs'
max_iter	an integer number	100
multi_class	'auto', 'ovr', 'multinomial'	'auto'

3.4.2 Decision Tree

In this thesis, the decision Tree library from the Scikit-Learn framework is used to implement a decision tree classifier for accident prediction. Scikit-Learn library uses the CART technique (Classification and Regression Trees) to construct the tree[78]. The CART technique minimizes the cost function $G(Q_m, \theta)$, as shown in Equation(3.4), at each node to find the best split criteria; where m is the number of node, Q_m is the data in node m, θ is the split candidate with feature j and threshold t_m, N_m is the number of samples in a Q_m data, and $H(Q_m, \theta)$ can be either Gini Impurity or Information Entropy as discussed in the literature review section.

$$G(Q_m, \theta) = \frac{N_m^{left}}{N_m} H(Q_m^{left}(\theta)) + \frac{N_m^{right}}{N_m} H(Q_m^{right}(\theta)) \tag{3.4}$$

Decision trees are prone to overfitting by creating complex trees; however, using some constraints will help avoid overfitting. Those constraints can be applied through the predefined parameters in the Scikit-Learn framework, brought in Table 3.5; where 'criterion' is the function to measure the split quality that can be either Gini Impurity or Information entropy; 'splitter' specifies the strategy to choose the split at each node; 'random_state' used to run random function; 'max_depth' shows the maximum length of the tree; 'min_sample_split' is the minimum number of samples required to split an internal node; 'min_sample_leaf' is the minimum number of samples required to be at a child node; 'max_features' is the number of features to consider when looking for the best split; 'max_leaf_nodes' specifies the maximum number of leaf nodes; 'min_impurity_decrease' set the threshold for impurity while splitting, 'ccp_alpha' is the complexity parameter to prune the tree.

Table 3.5: Scikit-Learn Decision Tree Hyper-parameter

Parameter	Value	Default
criterion	“gini”, “entropy”	”gini”
splitter	“best”, “random”	“best”
random_state	int, random state instance	None
max_depth	int	None
min_samples_split	int or float	2
min_samples_leaf	int or float	1
min_weight_fraction_leaf	float	0.0
max_features	int, float or “auto”, “sqrt”, “log2”	None
max_leaf_nodes	int	None
min_impurity_decrease	float	0.0
class_weight	dict, list of dict or “balanced”	None
ccp_alpha	non-negative float	0.0

A critical aspect of every ML model is the generalization ability that shows how a trained model performs on the new data. Lack of generalization in ML models is known as overfitting, which means the model memorizes the training set and performs poorly on the test set. The generalization degree is measured by comparing the accuracy of the model on training and test sets, and based on the nature of the problem and the requirements of the application, a difference threshold can be considered. For instance, for applications such as demand forecasting or product recommendation, the difference can be up to 5% or even more, if the overall accuracy is high; however, for other applications such as disease detection or accident prediction, since the prediction error can have drastic consequences, the difference threshold is considered less than 1%.

In order to control the overfitting in a Decision tree classifier, the hyper-parameters must be tuned, which means all the possible combinations of hyper-parameters are set, and the performance of the model on training and validation set is compared; the combination that has the highest accuracy and its difference between training accuracy and validation accuracy is less than or equal to a threshold is picked as the best model. The number of combinations follows the Factorial function, which means the more the number of hyper-parameters is, the more time-consuming and computationally expensive the tuning is.

3.4.3 Random Forest

Random Forest is constructed with several Decision Trees using different sub samples of the features. This technique is used to control overfitting and improve the accuracy of the Decision Tree model. Scikit-Learn[78] framework is used in this thesis to develop Random Forest model. Random Forest applies all the hyper-parameters mentioned in the Decision tree section with some additional hyper-parameters regarding the number of trees and the method of sampling as shown in Table 3.6; where 'n_estimator' specifies the number of trees in the forest, 'bootstrap' determines if all the samples are used for each tree or a proportion of the

samples, 'n_jobs' is the number of trees that the model can train at a time for constructing the whole forest, and 'max_samples' specifies the number of samples used for growing trees if the parameter 'bootstrap' is True. All the parameters must be fine-tuned to have the highest accuracy and lowest overfitting. In order to do so, a grid search algorithm is utilized, and the complexity of the algorithm and the time it takes to find the optimal hyper-parameters follow the factorial function.

Table 3.6: Scikit-Learn Random Forest Hyper-parameter

Parameter	Value	Default
n_estimator	int	100
bootstrap	bool	True
n_jobs	int	None
max_samples	int or float	None

3.4.4 Gradient Boost

To implement the Gradient Boost model, Catboost[79] framework is used in this thesis. This framework grows a prediction model F^T as an ensemble of weak learners f^t as shown in Equation (3.5).

$$F^T = \sum_{t=1}^{T} f^t \tag{3.5}$$

The weak learners in this problem are Decision trees that are built sequentially, and each tree is created to approximate negative gradients g_i of the loss function. The g_i negative gradient is denoted by Equation (3.6), where a_i denote $f(x_i)$ and x_i are the inputs.

$$g_i = -\frac{\partial l(a, y_i)}{\partial a}|_{a=F^{T-1}(x_i)} \tag{3.6}$$

3.4.5 Neural Network

The Keras deep learning API on top of the TensorFlow [80] framework are utilized to develop the deep Neural network classifier in this thesis. Unlike the investigated models so far, hyper-parameter tuning in a Neural Network incorporates layer structure tuning and hyper-parameters tuning. Layer structure tuning includes determining the number of hidden layers and the number of neurons in each layer, and hyper-parameter tuning is finding the best hyper-parameters such as activation functions, optimizer, learning rate and epochs. Since there are numerous choices for the number of layers, neurons, and other hyper-parameters, finding the best ones is cumbersome.

There are two common beliefs on the number of layers and neurons among researchers. The first belief is

that every function can be approximated using only one layer, and all the relationships between explanatory variables and the target variable can be found in many practical problems[81, 82, 83, 84]; The second belief recommends setting the number of neurons in the hidden layer as follows[85]:

- The number of neurons should be between the size of the input and the output
- The number of neurons should be 2/3 the size of the input plus output
- The number of neurons should be less than twice the size of input

An essential element of the neurons is the activation function that links the weighted sum of the input from the previous layer to the next layer. Dense hidden layers and output layers utilize activation functions but use different functions based on the functionality.

The common activation functions used in the hidden layers are as follows:

- Rectified Linear Activation (ReLU)
- Logistic (Sigmoid)
- Hyperbolic Tangent (Tanh)

The ReLU function outputs the same input if it is positive; otherwise, it outputs zero. The sigmoid function applies a logistic function, and the output can be either 1 or 0. The Tanh function is similar to the Sigmoid function, but the output ranges between 1 and -1. Both Sigmoid and Tanh functions suffer from gradient vanishing problems if the number of neurons and layers are high; thus, the ReLU activation function is used for all the hidden layers in this thesis. They also showed better performance during the hyper-parameter tuning.

The common activation functions used in the output layer are as follows:

- Linear function
- Logistic (Sigmoid) function
- Softmax function

The linear activation function transforms the weighted sum of inputs from hidden layers to the output and is used for regression problems. The Sigmoid function is a good candidate for the binary classification problem since it outputs 1 or 0, representing two predicted labels; finally, the Softmax function outputs a vector of values with the sum of 1, which can be used for multi class classification applications. A Sigmoid function is used for the output layer in the proposed Neural Network model.

Another important element of a Neural network is the Optimizer used to tune the weights assigned to each neuron to minimize the cost function. The Optimizers are divided into two families based on the learning rate choices; gradient descent optimizers and adaptive optimizers. In the first family of optimizers, the learning rate to converge to an optimal point should be adjusted manually; however, it is adjusted automatically during the learning process in the second family of the optimizers. The gradient descent optimizers such as SGD(Stochastic Gradient Descent) are prone to get stuck in local minima and very sensitive to the learning rate adjustments. Adaptive optimizers such as Adam and RMSprop automatically adjust the learning rate based on specific algorithms that compensate for the drawbacks of the first family of optimizers. The Adam optimizer was found to have the best performance on the accident data after the hyper-parameter tuning.

Neural Networks use optimizers to find the optimal values of weights through a recurrent process. After each iteration or epoch, the error for that epoch should be calculated to be reflected throughout the network by the back-propagation algorithm. There are some loss functions to estimate the error, which are selected based on the nature of the prediction model; loss functions such as Binary Cross-Entropy, Hinge Loss, and Squared Hinge Loss are suitable for binary classification problems. The last two are used in models that output in the range of 1 and -1, and the Binary Cross-Entropy is the best candidate for the proposed models.

Chapter 4

Accident Analysis and model development

4.1 Exploratory Analysis

In this section, an analysis is carried out on collision data considering different features. In Figure 4.1, a heat-map of collisions in Calgary is shown during December 2019; most of the collisions occurred in and around downtown, along the Deerfoot Trail, Glenmore Trail, Crowchild Trail, and Macleod Trail. These locations are considered collision hot-spots.

The collision dataset encompasses approximately 14,835 samples of actual accidents in Calgary from 2017 to 2020. In Figure 4.2, the red bars show the cumulative number of accidents amongst different years for the given time. The highest number of accidents can be observed in 2018, $4,458$ accidents that had increased by roughly 24% from $3,603$ accidents in 2017. A 30% accident reduction happened in 2019 compared to 2018 from 4,458 to 3,119 accidents. The effect of the COVID-19 pandemic on road safety is complicated. One study conducted by Vanlaar et al.[86] analyzed the self-reported driving behaviours and showed a 5.5% increase in speeding and 4.2% increase in distracted driving during the COVID-19 pandemic in Canada. The preferred method of travelling has also changed during the COVID-19 pandemic; A report by the Traffic Injury Research Foundation (TIRF)[87] showed that although a 69.9% increase in using personal vehicles happened, active transportation such as walking and cycling have significantly increased between 120% to 150%. Furthermore, in 2020 there are 3,655 accidents; roughly equal to the accidents in 2017.

Since the distribution of the accidents is non-uniform amongst different years, it can be inferred that the year as a feature can be influential in the occurrence and consequently prediction of the accidents; hence,

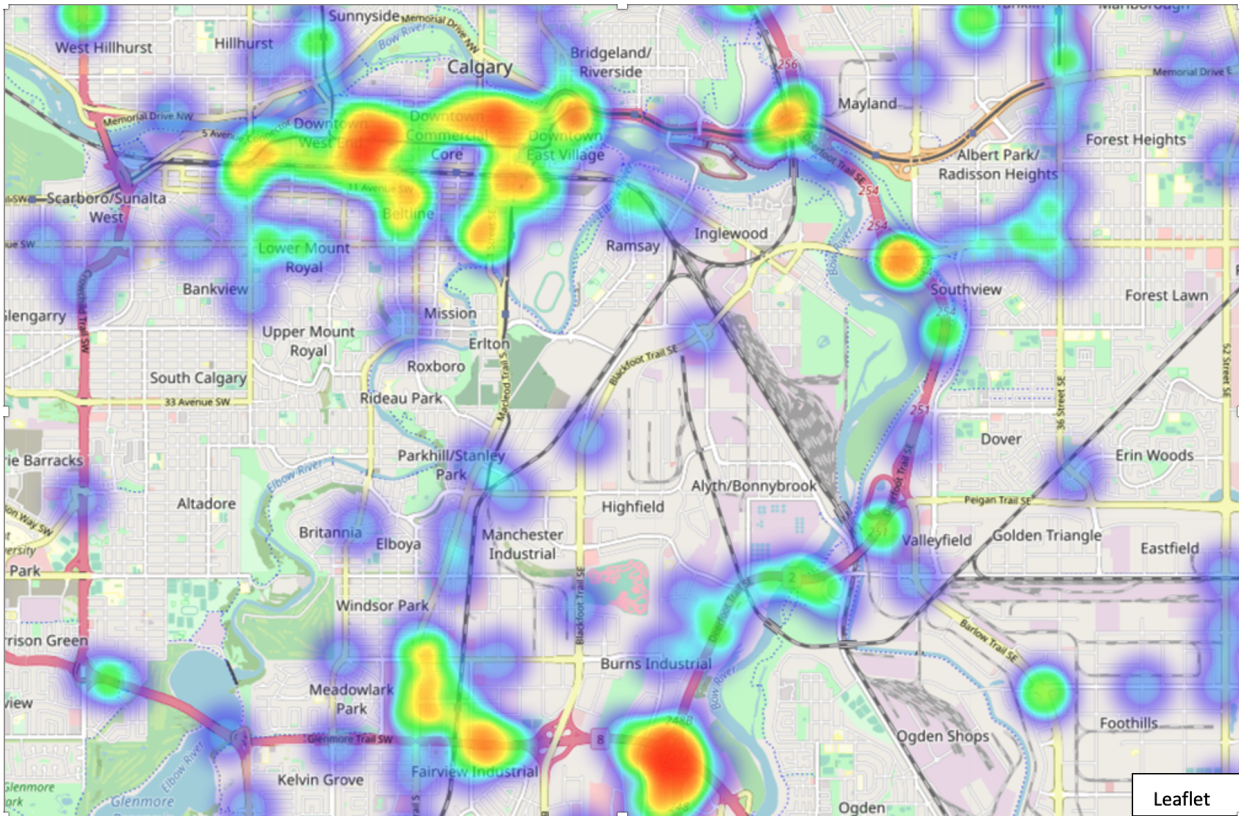


Figure 4.1: Calgary's December 2019 collision heat-map

the year is ignored in the model development as a feature since the model can be used to predict future accidents; there is no sense to use the year in the model development stage.

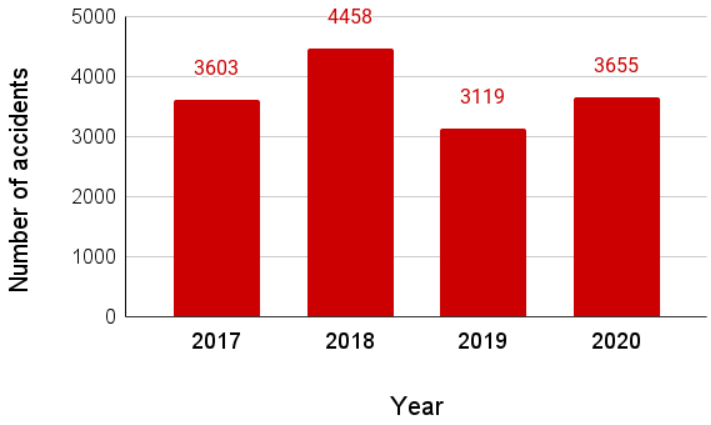


Figure 4.2: Distribution of accidents in different years

Following exploration of the dataset reveals the trends and patterns in distinctive weather features and the relationship between accidents and the road and the traffic characteristics.

4.1.1 Weather

Red bars in Figure 4.3 illustrate the distribution of accidents in different weather conditions. As shown, 0.74% of the hours with ice crystals or ice fog weather there are accidents in very low temperatures, which is roughly two times more than when it is cloudy or clear. This ice fog is composed of particles of ice that occur at very low temperatures, mostly below -30 degrees Celsius. That shows that driving on the road with ice crystals, which is too slippery, might have two times more probability of being involved in accidents than the clear weather.

The second most influential weather condition on accidents is the snowy weather, which has 0.62% accidents proportion and roughly 1.5 times more accidents than clear or cloudy weather conditions. The next two conditions are the haze and the smoke that cause low visibility and increase the perception time, decision time, and reaction time of the drivers on roads and contribute to accidents in Calgary. Figure 4.3 illustrates a non-linear relationship between the weather types and the rate of accidents and shows how harsh weather conditions such as ice crystals in very low temperatures and snow are correlated to the number of accidents and should be considered in APM developments in locations with extreme weather conditions such as Calgary.

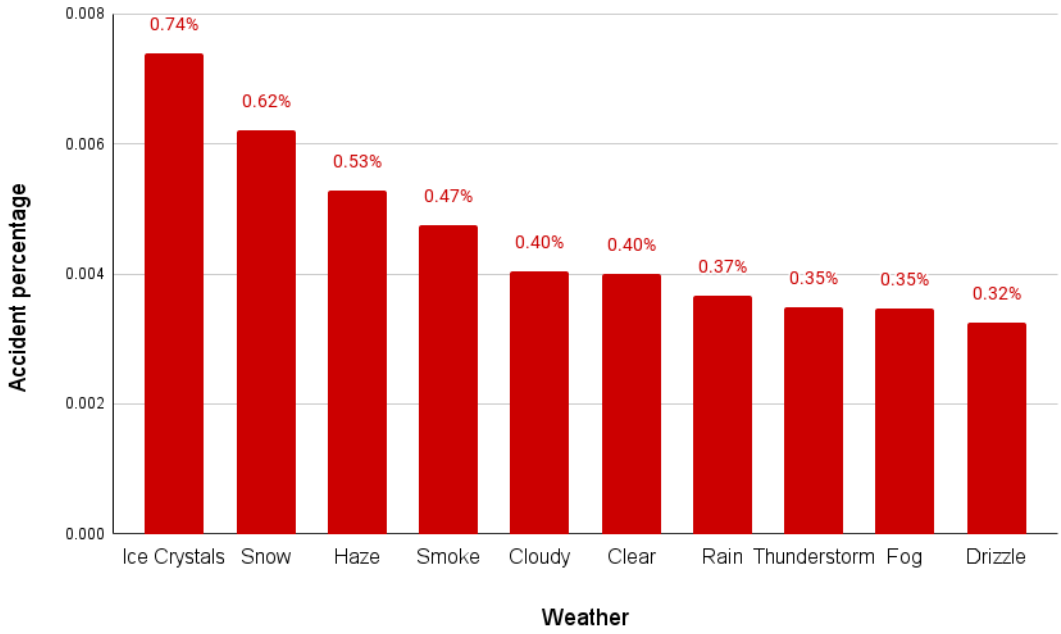
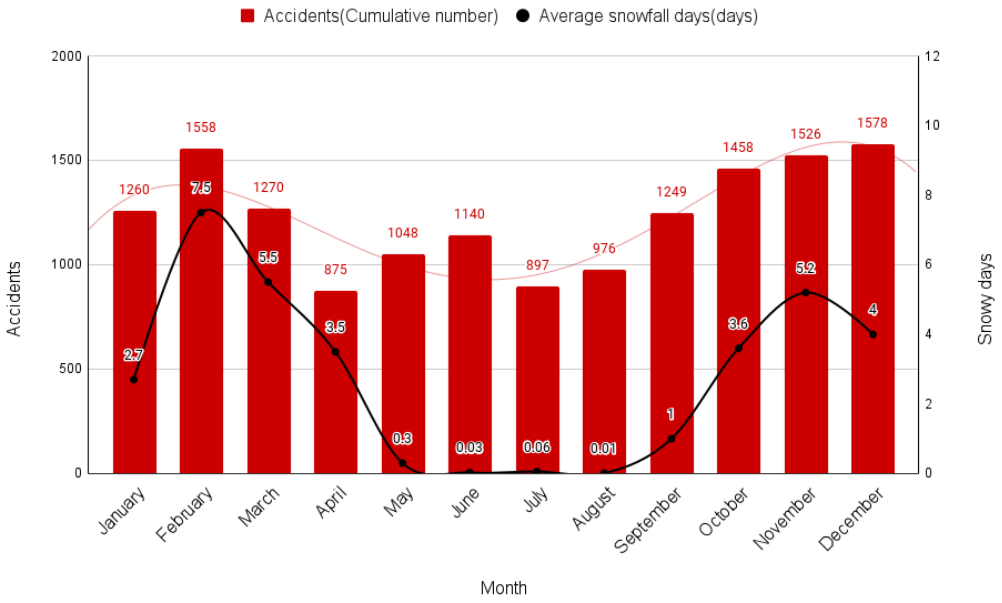


Figure 4.3: Distribution of accidents in different weather conditions

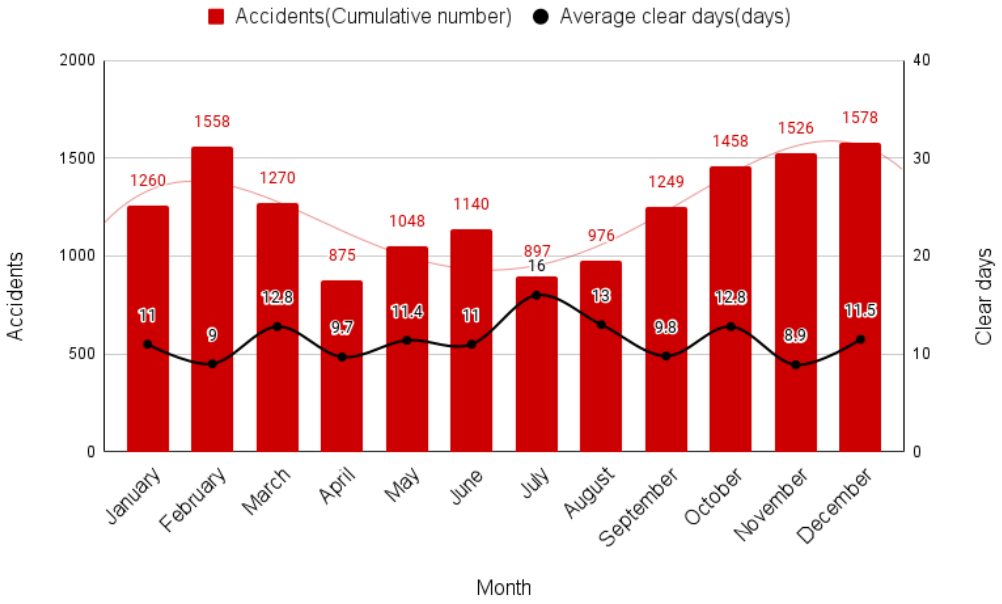
To support the results of weather type analysis, a monthly effect of snowy and clear days are conducted as follows: In Figure 4.4a, the black curve shows the trend of average snowfall days per month between

2017 to 2020, and the red bars are the cumulative numbers of accidents during the same period of time. The collisions have correlation with the average snowy days. There are the highest number of accidents in February, March, November, and December, the cold months with the most significant snowy days. For example, there are 1,558 car accidents over the given four years in February with 7.5 snowy days on average this month. On the other side, there are only 897 and 976 car accidents in July and August, consecutively, with almost zero snowy days.

In Figure 4.4b, the black curve shows the average clear days per month from 2017 to 2020, and the red bars are the cumulative number of monthly accidents. The average number of clear days fluctuates during different months, but it does not show a distinctive relationship with the number of accidents. For instance, although there is the same number of clear days in May and December, the number of accidents is significantly different, 1,048 in May and 1,578 in December; December has roughly 1.5 times more accidents. Furthermore, even though there is somehow the same number of accidents in April and July, the number of clear days is different; July has roughly 1.5 times more clear days than April.



(a) Monthly accidents with monthly average snowy days



(b) Monthly accidents with monthly average clear days

Figure 4.4: The effect of snowy and clear days on accidents

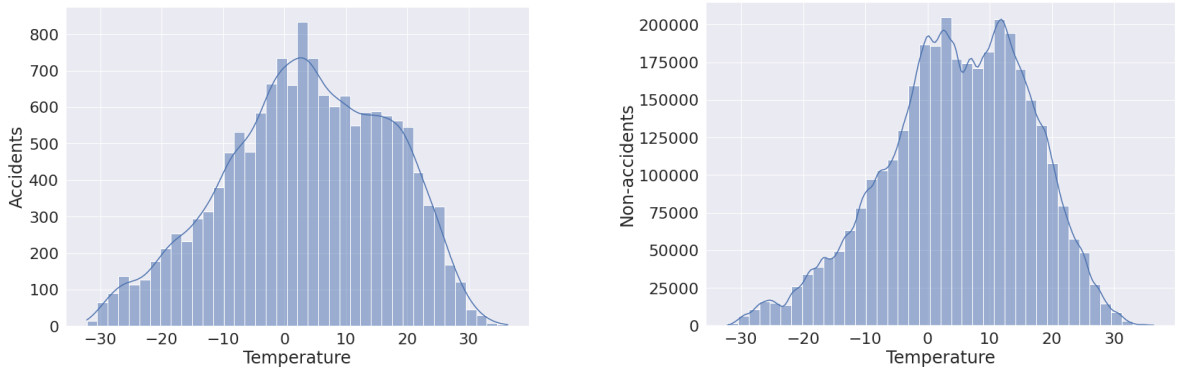
4.1.2 Temperature

Another influential factor in collisions is temperature; the temperature indicates how hot it is regardless of how it feels or affects other environmental factors. Below zero temperatures cause frozen roads and surfaces,

and consequently, more accidents happen due to a lack of control and skidding. Figures 4.5a and 4.5b show the distribution of accidents and non-accident events with respect to different temperatures below -30 to above +30 degrees Celsius.

By comparing the slope of the left sides of the two histograms, the slope of the non-accidents curve is steeper than the accidents curve, which shows that lowering the temperature correlates with the increase in the proportion of accident events than non-accident events.

Another insight can be derived from comparing the peaks of the two histograms. The non-accidents histogram has two peaks, one between 0 and +5 degrees and another one between +10 and +15 degrees, whiles the accidents histogram has only one peak between 0 and +5 degrees Celsius; it indicates that even though the temperature between +10 to +15 is common in Calgary, this temperature has less effect on the accident compared to a temperature between 0 and +5. That difference is significant which is correlated to roughly 300 more accidents when the temperature is between 0 to +5 degrees Celsius.



(a) Histogram of accident

(b) Histogram of non-accident events

Figure 4.5: Distribution of accident and non-accident events with respect to the temperature

Relative accident distribution can be calculated utilizing the accident and non-accident histograms that can picture more insights into the influence of temperature on accidents, as shown in Figure 4.6. Based on the red bars in Figure 4.6, when the temperature is below -30 degrees Celsius, almost 1% of the time, there is at least one accident in Calgary, 2.5 times more than when it is between -5 to +5, and three times more than when it is between +5 to +15. The colder weathers exponentially affect the rate of accidents and based on this dataset, the safest temperature for driving is between +5 to +15, with the lowest rate of accidents, which is 0.33%. Furthermore, driving in temperature between -30 to -25 has 2.5 times, between -25 to -20 has two times, and between -20 to -15 has 1.7 times more likelihood of involving in a road accident compared to the safest temperature. The right side of the distribution of accidents in Figure 4.6 shows in the temperature above +15 degrees Celsius, the rate of accidents increases somehow linearly. The temperature analysis showed that temperature significantly correlated with the occurrence of accidents and should be

utilized for APM development in Calgary.

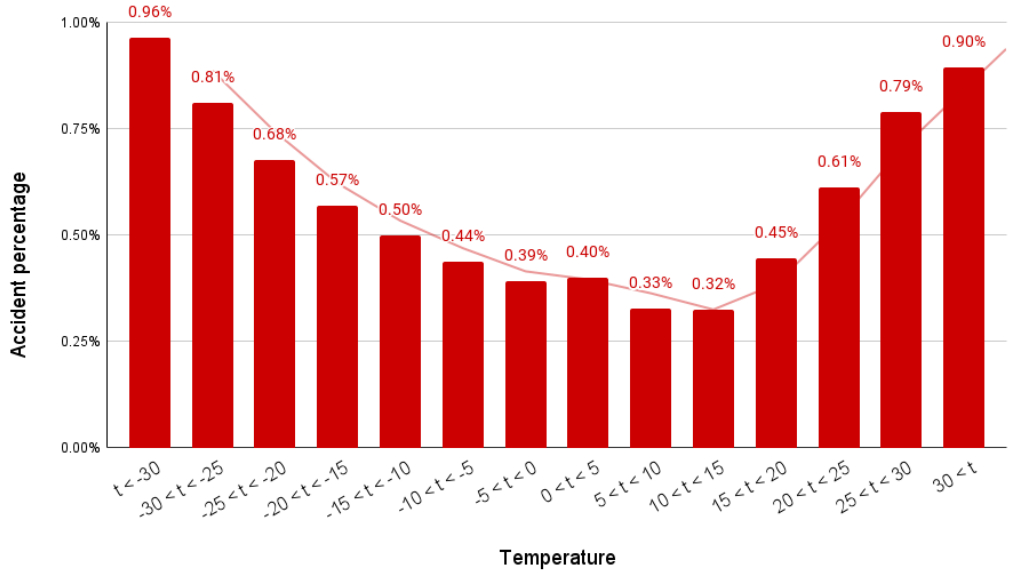


Figure 4.6: Distribution of relative accidents in different temperature ranges

Since the temperature changes is significantly correlated with the rate of accidents based on analysis shown in Figure 4.6, a monthly temperature effect analysis on accidents is carried out to give a more understandable outlook.

The black curve in Figure 4.7 shows the pattern of changes in monthly temperature in Calgary throughout 2017 to 2020, and the red bars illustrate the cumulative number of accidents in different months over the same given time. The temperature is mostly below zero in the months, with a high number of collisions in January, February, March, November, and December. The temperature has a reverse relationship with the accidents. This effect is valid in all months except May, June, and July, which shows a vague relationship.

It seems that in months with cold and freezing temperatures, the number of accidents is significantly higher; In months with moderate and warm temperatures, the accidents do not show to have correlation with temperature, which complicates the relationship between temperature and accidents.

Another insight that can be derived from Figure 4.7 is that even though the number of accidents in November and December is very close to the number of accidents in February, there is a big difference in the average temperature in those months.

4.1.3 Relative Humidity

The relative humidity is another attribute of the weather condition; it shows the relative water vapour in the air, and higher temperatures can hold more water vapour. Red bars in Figure 3.12 represent the accident

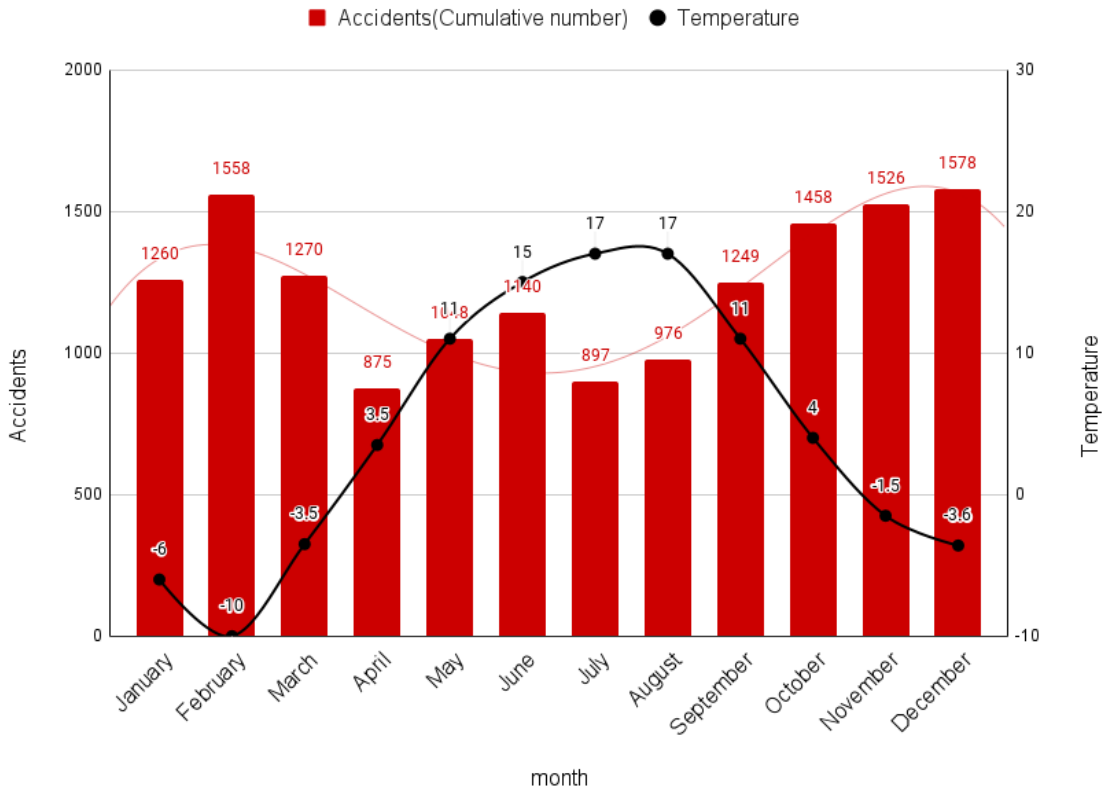


Figure 4.7: Distribution of accidents and temperature in different months

rate in different humidity ranges above zero and below 100% in Calgary in the given time.

The rate of accidents remains constant with minor fluctuations from 95% to 40% relative humidity, but below 40% humidity is correlated to the rate of accidents. The lower the relative humidity, the higher the rate of accidents but non-linearly. For instance, the rate of accidents in the humidity between 5% to 10% is 89%, which is three times more than the rate of accidents in the humidity between 45% to 50%. Taking the average of the accident rates above and below 40% humidity, shows that the rate of accidents when the humidity is below 40% is twice the rate of accidents when the humidity is above 40%.

The relative humidity analysis pictures the importance of this variable in APM development since it has been shown to have a huge correlation with the probability of accident occurrence, especially in low temperatures as low relative humidity comes in low temperatures.

4.1.4 Atmospheric Pressure

Atmospheric pressure or air pressure is another attribute that always comes with the weather, which is examined in this thesis to find how much that pressure is correlated to road accidents. Figure 4.10 illustrates

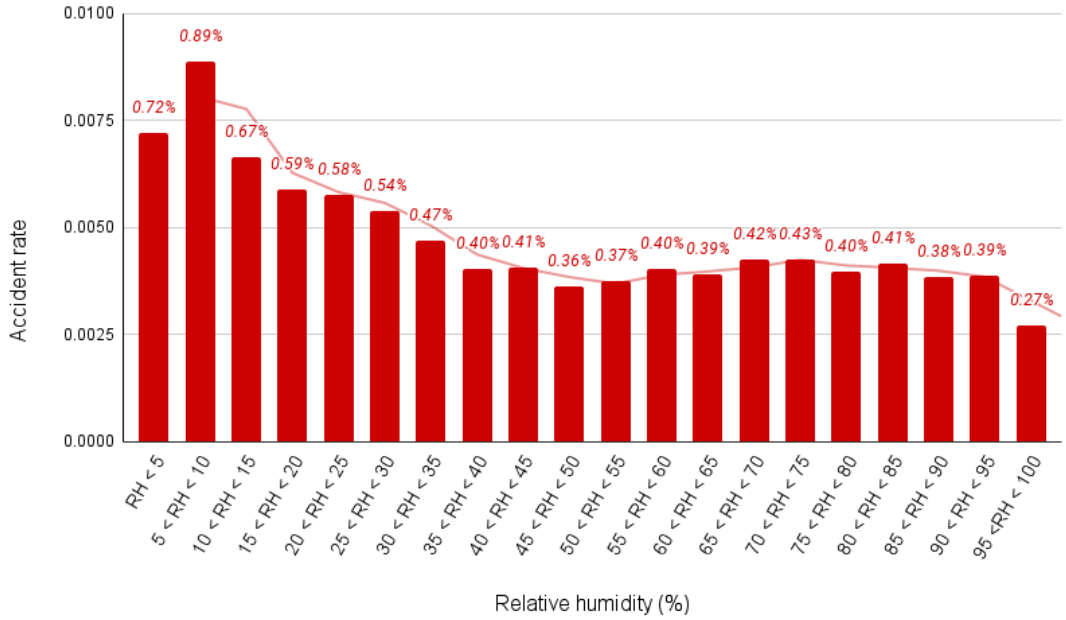
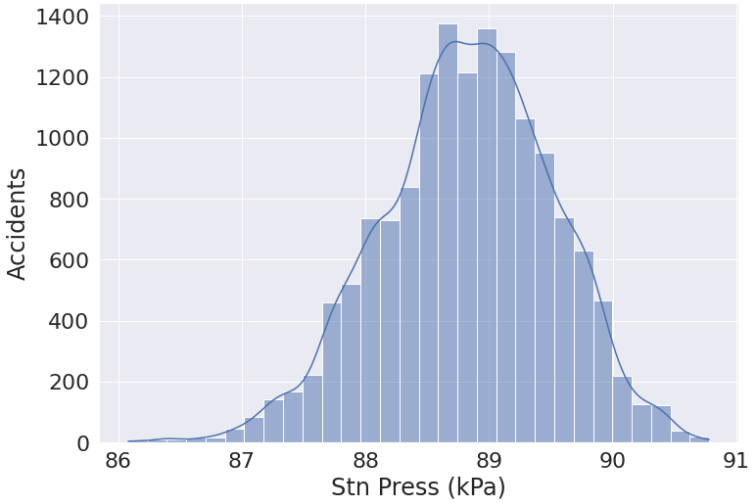
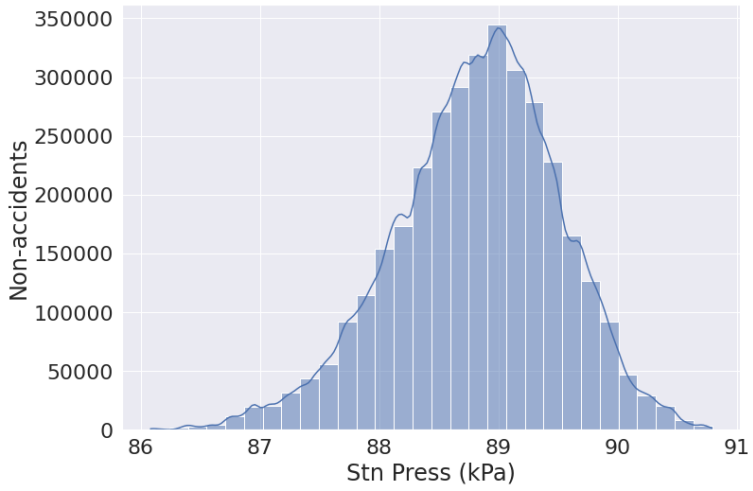


Figure 4.8: Distribution of relative accidents in different relative humidity percentages

the distribution of accident and non-accident samples in the accident dataset with regards to air pressure. The bottom histogram displays the distribution of non-accident time slots, which is slightly left-skewed; the center of the data is at 89 kPs, while the mean of the data is 88.8 kPs. The top histogram depicts the distribution of accidents which is more symmetrical than the bottom one; the center of the data is 88.7 kPs, while the mean of the data is 88.8 kPs. These two distributions are slightly different, which shows accidents are more likely to happen in some air pressures than others. For instance, on the left tail of the histogram, which shows lower pressure, accidents tend to happen more than the right tail of the histogram. Therefore, air pressure is correlated to the frequency of accidents, which will be used as a feature in APM development.



(a) Histogram of accident



(b) Histogram of non-accident events

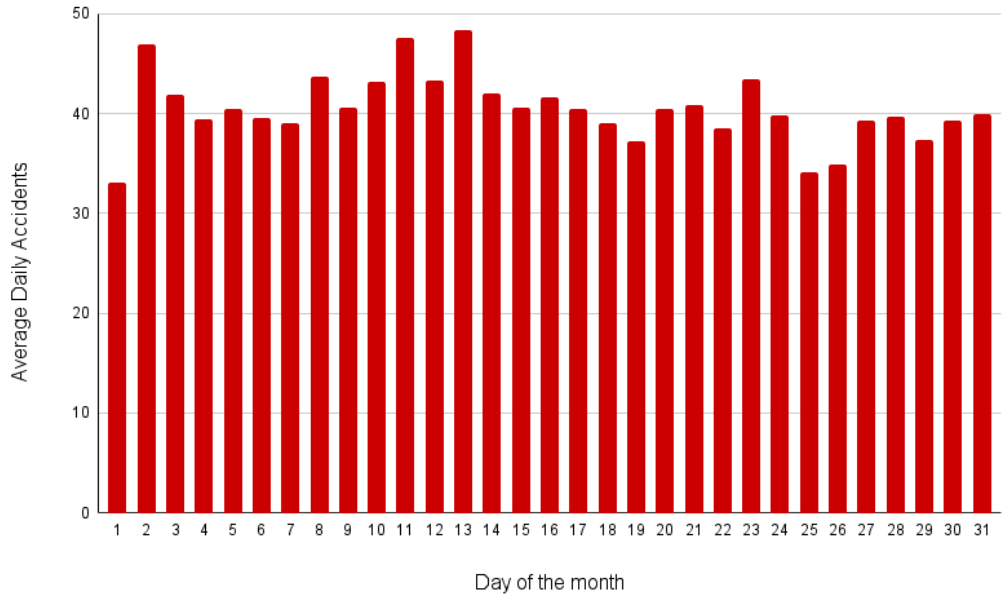
Figure 4.9: Distribution of accident and non-accident events with respect to the air pressure

4.1.5 Day of the Month and Week

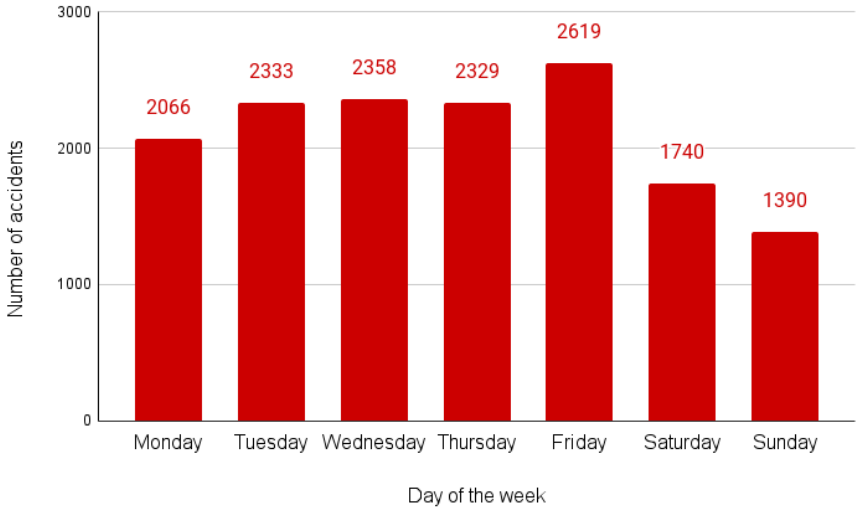
Whether the day of the month is an important factor in accidents is ambiguous, at least by looking at the distribution of accidents with regards to the day of the month shown in Figure 4.10a. The average number of accidents fluctuates during consecutive days throughout the month. It seems that travelling in the first half days of the month might slightly increase the probability of accidents compared to the second half days because the average number of accidents in the first half is 6% more than the second half.

Although the day of the month has not been shown to have a high correlation with accidents, the day of the week has. The red bars in Figure 4.10b show that the number of accidents on the weekdays is 33% more than the number of accidents on the weekends. Furthermore, Most accidents happen on Fridays and the

least on Sundays. In conclusion, the day as a determinant can be highly correlated to accidents and must be used in APM development.



(a) Distribution of accidents with regards to day of month



(b) Distribution of accidents with regards to day of week

Figure 4.10: Distribution of accident with regards to month and week

4.1.6 Hour of the day

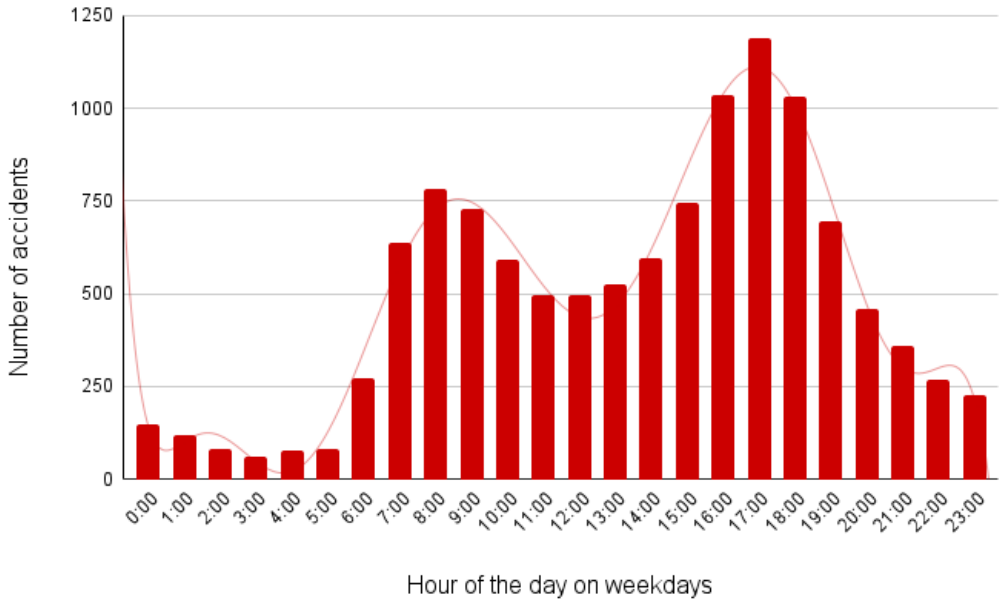
As seen in the previous section, weekdays and weekends have different correlation with the accidents because the kind of travel is different in nature on those days; on the weekdays, travels are more routine and mostly

work travels versus on the weekends, travels are not routine, and mostly leisure travels. This difference results in having two distinctive accident patterns on weekdays and weekends.

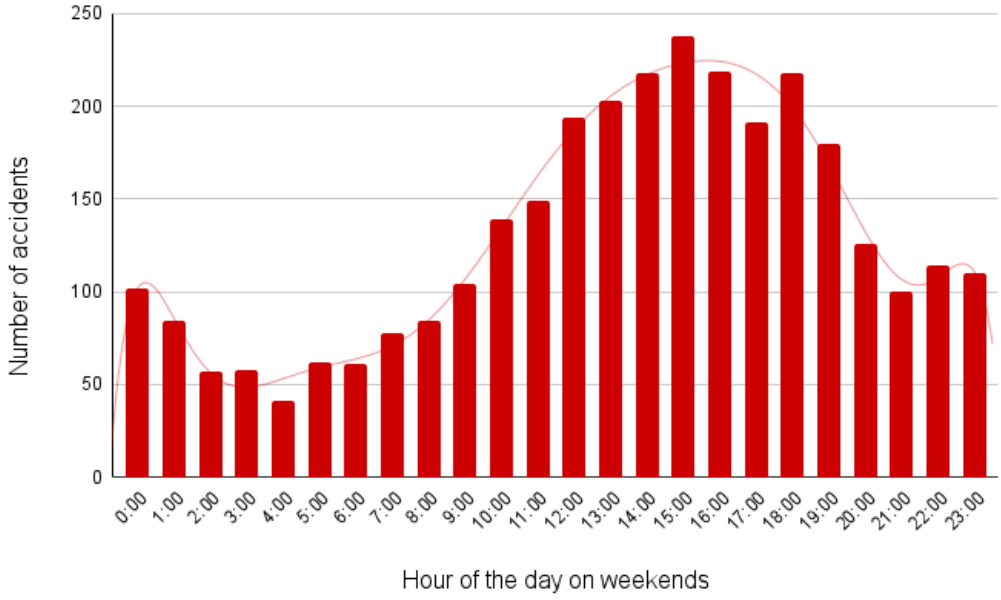
The weekday accident pattern follows the rush hours in Calgary, shown by the red bars in Figure 4.11a. Rush hour is a critical time for every road commuter, not only for the time wasted in traffic but also for the chance of being involved in a road accident. As shown in Figure 4.11a, there are two peaks for accidents on the weekdays; the biggest one with the highest number of accidents happens on the evening rush hour, from 4 pm to 6 pm, and the next one with fewer accidents from 7 am to 9 am, which is the morning rush hour.

The rate of accidents on weekdays varies a lot from 62 accidents at 3 am to 1,190 accidents at 5 pm, which shows that the time of the day on weekdays is significantly correlated with the accidents, and the reason behind that is traffic volume at nights is very lower than days.

The accidents pattern on the weekends, though, is different. Since there is no morning rush hour, the morning has the lowest rate of accidents, as shown in Figure 4.11b; On the other hand, a broader range of time in the evening, from 12 pm to 7 pm, contains more increased accidents. The overall number of accidents is way less than weekdays, and since the pattern is more straightforward, the correlation of weekend hours and accidents is more minor than weekday hours.



(a) Distribution of accidents with regards to the hour of the day on weekdays



(b) Distribution of accidents with regards to the hour of the day on weekend

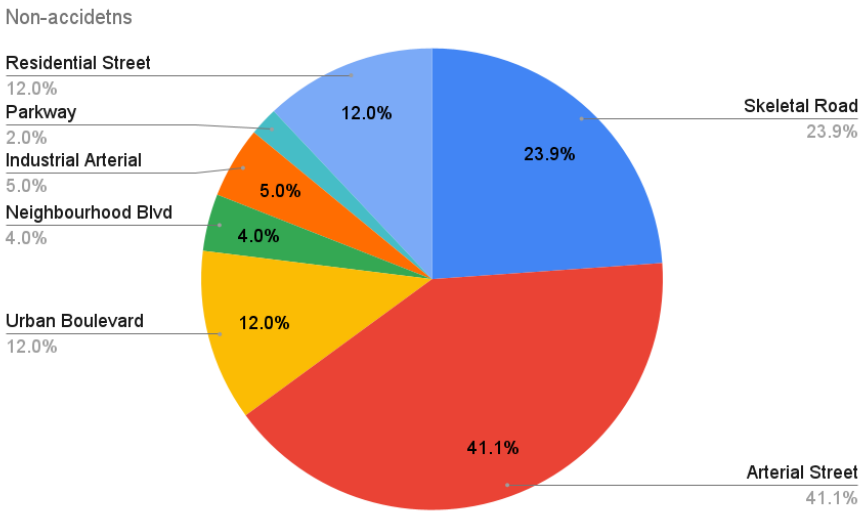
Figure 4.11: Distribution of accident with regards to hours of the day

4.1.7 Road Class

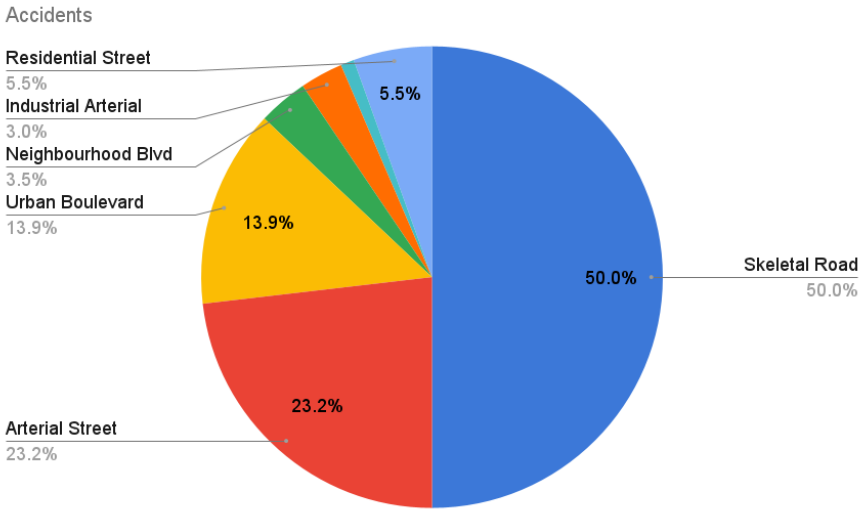
Roads in Calgary are grouped into different types, based on their service. Seven road classes are defined based on traffic density, mobility, access, land-use, and safety as follows[88]:

- Skeletal Roads, known as freeways or expressways, provide the movement of vehicles over long distances and carry over 30,000 vehicles per day.
- Arterial Street provides connections between communities and carries between 10,000 to 30,000 vehicles per day. This kind of road makes up much of the network in Calgary.
- Industrial Arterial is a priority for the movement of heavy trucks but is also used by other vehicles. The speed is typically lower on this type of road.
- Urban Boulevard is primarily a local and regional street focusing on walking and cycling, while a high volume of vehicles is expected.
- Parkway is mainly located near natural parks, waterways, and public institutions.
- Neighbourhood Boulevard is similar to Urban Boulevard but smaller.
- Residential Street is mainly a connection street within and between residential areas.

The pie-chart in Figure 4.12a illustrates the proportion of different road classes within the dataset that shows 41.1% of the roads in Calgary are Arterial streets and 23.9% Skeletal roads. These two roads make up for the majority of the roads in Calgary. Following those two, Urban boulevard and Residential streets with 12% each, are the third types of roads in Calgary. The remaining classes account for the remaining 9% of the road types.



(a) Proportion of road classes in Calgary



(b) Distribution of accidents with regards to the road class

Figure 4.12: Accidents versus Road Class

The pie-chart in Figure 4.12b illustrates the proportion of accidents with respect to different road classes. Skeletal roads account for 50% of the accidents, while only 23% of the roads are Skeletal roads; Furthermore, even though 41.1% of the roads in Calgary are Arterial streets, only 23.2% of the accidents happen on this kind of roads. Residential streets are another type of road that has correlation with the accidents as 12% of the roads are in the Residential street class, and only 5.5% of the accidents happen on these roads. The comparison between the two pie-charts in Figure 4.12 proves that the road type plays a role in accidents as most accidents happen on Skeletal roads.

4.1.8 Road Segment

For the accident analysis in this thesis, roads are broken down into segments as the smaller unit of roads. Some road segments encounter more accidents based on the imposed traffic, the segment's characteristics, and other unknown factors; those segments are called accidents hot-spots. Figure 4.13 shows road segments in Calgary with more than 200 road collisions from 2017 to 2020. As shown by the red bars, Deerfoot TR SE with 966 accidents has by far the highest number of accidents, and following that, Deerfoot TR NE with 805 accidents is in the second place. Following Deerfoot TR, a high number of accidents have happened on Glenmore TR SW, 16 AV NE, and Glenmore TR SE, respectively. There is a decrease in accidents on the following road segments with a low slope from Memorial DR NW with 426 accidents to Bow TR SW with 207 accidents.

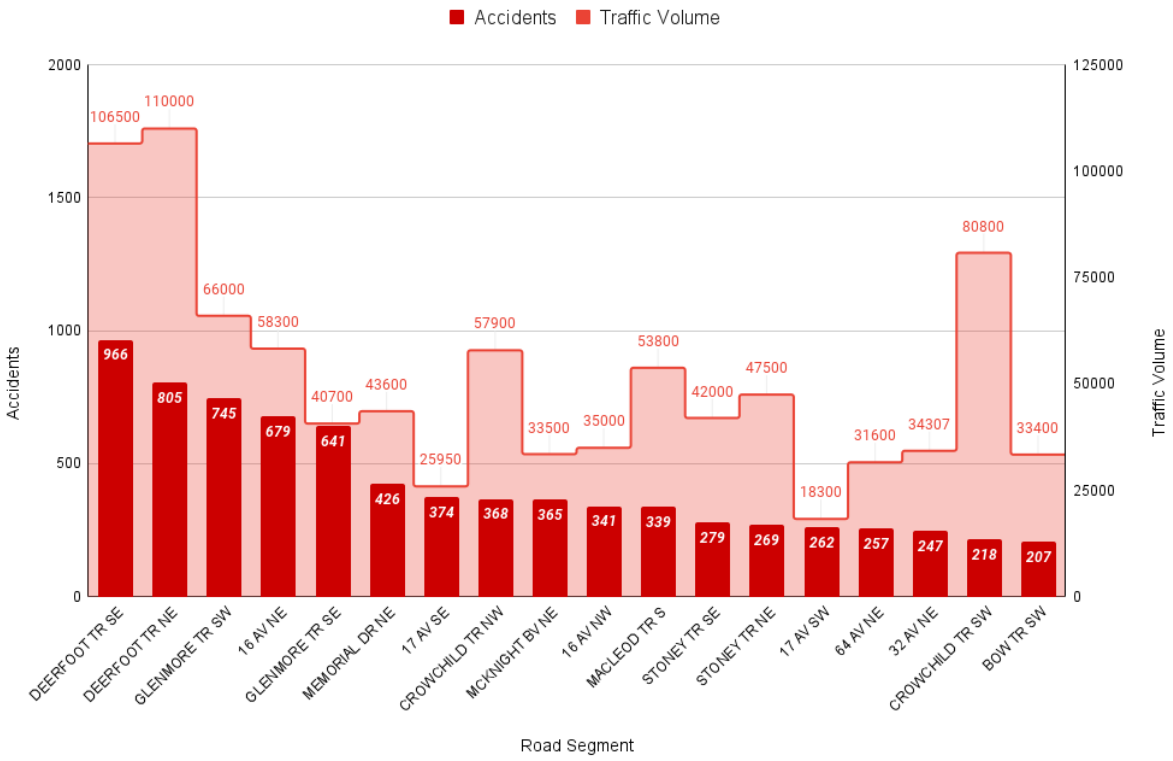


Figure 4.13: Distribution of accidents in different road segments

The light red stepped area in Figure 4.13 illustrates the average daily traffic volume on the road segments with more than 200 accidents in the dataset. The first five road segments with the highest accidents somehow follow the pattern of traffic volume. The higher the average daily traffic volume, the higher the number of accidents; for instance, Deerfoot TR SE and NW, which have the highest traffic volume, also have the

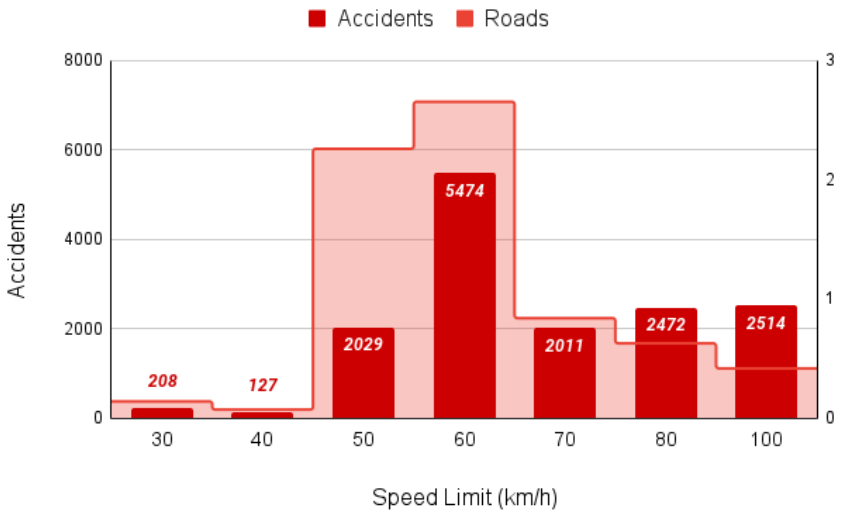
highest number of accidents; moreover, Glenmore TR SW, which has higher accidents compared to 16 AV NE, has greater daily traffic volume. After those five road segments, it seems the rest of the segments do not follow the traffic volume pattern, as Crowchild TR SW, with 80,800 vehicles per day, has only 218 accidents compared to 262 accidents in 17 AV SW with 18,300 vehicles per day.

4.1.9 Speed Limit

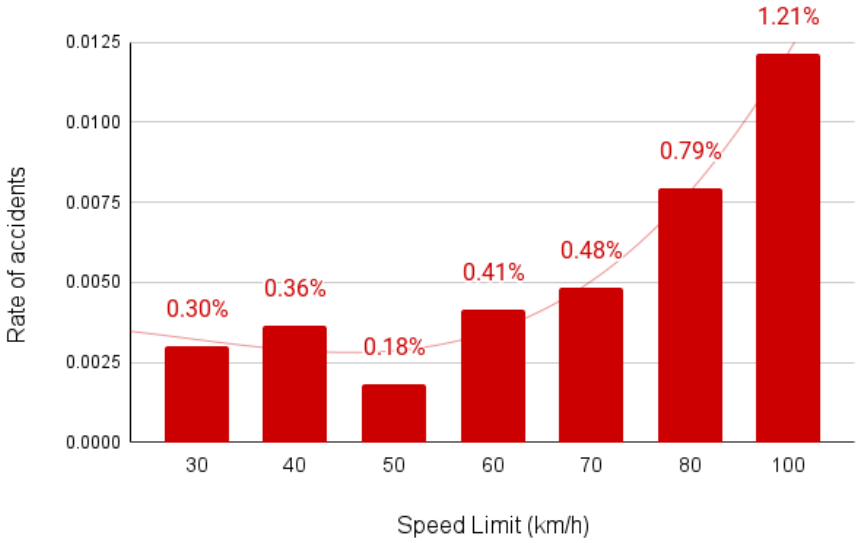
The speed limit is another factor that can change the speed of the traffic flow and the density of the cars on the roads accordingly. The speed limit variability plays an essential role in the occurrence of accidents as well. The red bars in Figure 4.14a represent the cumulative number of accidents that have happened on the roads with specific speed limits from 30 to 100 km/h in Calgary from 2017 to 2020. The highest number of accidents happened in 60 km/h, which is 5,474; following that, 100 and 80 km/h have 2,514 and 2,472 accidents, respectively. The light red stepped area shows the percentage of the roads with the associated speed limits; for instance, the road segments with a speed limit of 60 km/h are 6.4 times more than roads with 100 km/h; and also, the roads with a posted speed limit of 50 km/h are 3.6 times more than roads with 80 km/h.

Although the red bars in Figure 4.14a show that the highest number of accidents happened at 60 km/h, it cannot represent the actual effect of 60 km/h on accidents compared to the other speed limits because a majority of the road segments have a speed limit of 60 km/h, indeed. Thus, an accident rate analysis regarding the speed limits variability was carried out, as depicted in Figure 4.14b. The red bars in Figure 4.14b show the rate of accidents according to different speed limits; the speed limit of 100 km/h has been shown to have the highest correlation with accidents, as 1.21% of all the times, there was at least one accident on roads with 100 km/h posted speed limit. The second important speed limit is 80 km/h since, in 0.79% of the time, there were accidents on roads with the mentioned speed limit.

Above 40km/h speed limits have been shown to have an approximate direct exponential relationship with the number of accidents, as shown by the red curve in Figure 4.14b. However, the rate of accidents is affected by other factors more than the speed limit at and below 40 km/h, as the accident rate at 40 and 30 km/h are 0.36% and 0.30%, respectively, which are not along with the other speed limits.



(a) Distribution of accidents with regards to speed limit



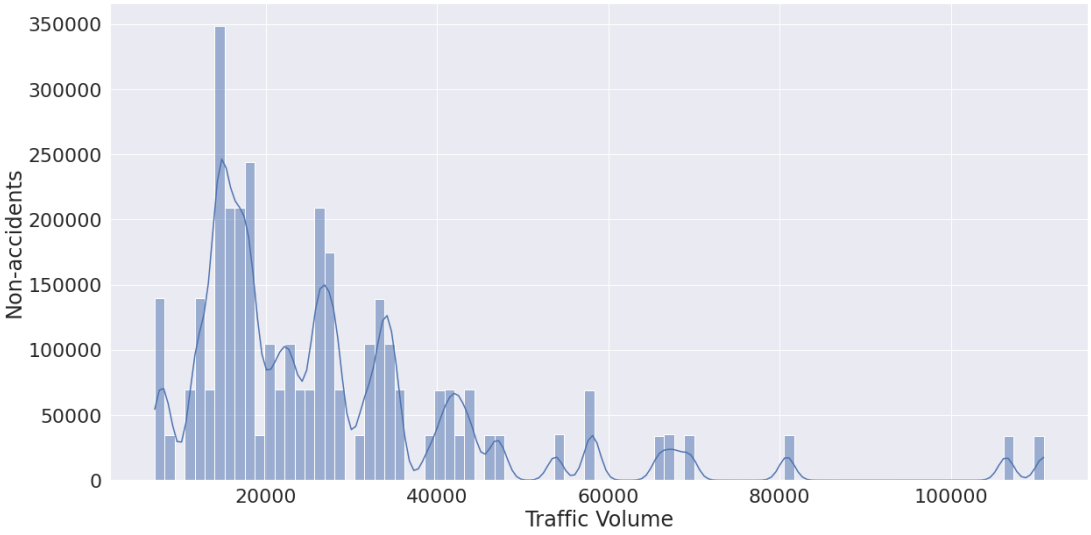
(b) Proportion of accidents in different speed limits

Figure 4.14: Accidents versus Speed limit

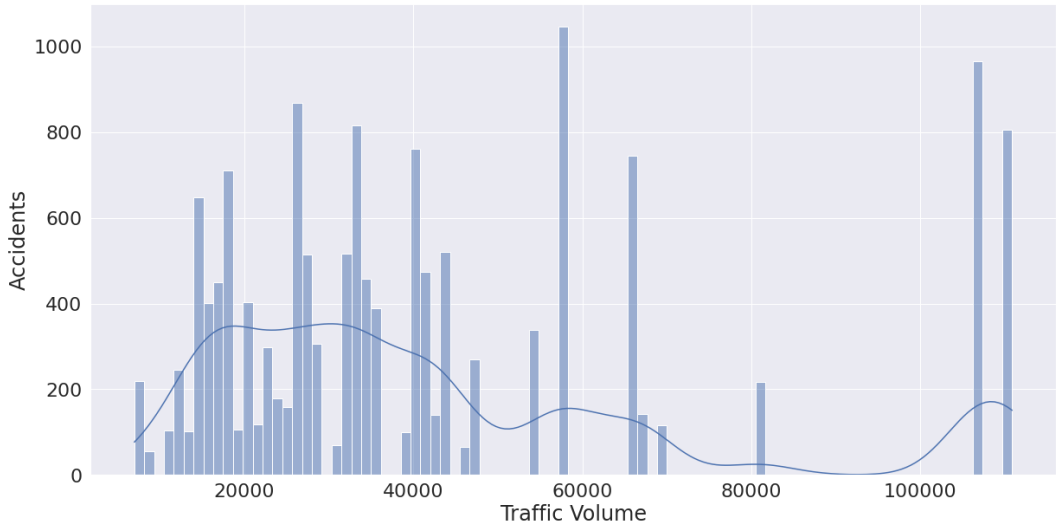
4.1.10 Traffic Volume

On one side, traffic is known to be the main concern of commuters when an inner-city trip is started, as it is considered the reason for delays and congestion, and a high traffic volume can cause accidents since the interaction between cars increases. Thus, every commuter seeks the roads with the least traffic volume. On the other side, although the broader and bigger road segments seem to carry higher traffic volumes, they may cause a lower accident rate based on the number of commuters that use those segments and the capacity of those segments. Thus, traffic volume is a controversial factor in accidents that needs to be addressed.

The blue bars in Figure 4.15a illustrate the distribution of daily traffic volume in the dataset, which is a right-skewed distribution that shows most of the roads in Calgary from 2017 to 2020 had relatively low daily traffic volume between 7,000 to 35,000 vehicles per day. In addition, the number of roads with daily traffic between 35,000 to 80,000 vehicles per day was lower than the first class, and there was only one road segment with above 100,000 average daily traffic, which was Deerfoot TR.



(a) Distribution of the Traffic Volume



(b) Distribution of accidents with respect to the Traffic Volume

Figure 4.15: Accidents versus Traffic Volume

The blue bars in Figure 4.15b show the distribution of accidents with respect to the traffic volume from 2017 to 2020 in Calgary, which is more scattered than the traffic distribution. A comparison between the accident distribution and the traffic distribution reveals that the rate of accidents in roads with higher average daily traffic is typically higher than those with lower average daily traffic. For example, by increasing the

daily traffic from 20,000 vehicles per day to 60,000, the rate of accidents increases by 4.5 times. In conclusion, the traffic volume has shown to be an effective contributing factor to road accidents and must be considered in APM development.

Thus far, different explanatory variables in the dataset are investigated, and they have been shown to affect road accidents somehow or other. Some variables have a direct relationship with accidents, such as Traffic volume and Speed limit, but others, such as relative humidity, have a reverse relationship. However, the temperature has a mixed effect, in such a way that above +15 degrees Celsius has a somehow direct linear relationship, and below that, has a reverse exponential relationship, as shown in Figure 4.6. Some categorical variables such as weather, road class, and road segment have also showed to have correlation with accidents. Accidents are correlated to the time of the day, the day of the week, and the month too, but the correlations are different and unique.

Thus, the prediction of the accidents requires a powerful model that can extract all sorts of hidden patterns and relationships between the accidents and the accident determinants. In the following section, the Logistic Regression model as well as Decision Tree-based models, in addition to the Neural Network and Deep Neural Networks, are developed to find the model with the highest prediction capabilities and the influential features are introduced.

4.2 Accident model development

In this section, five models, including Logistic Regression, Decision Tree, Random Forests, Gradient Boost, and Neural Networks, are developed utilizing the tools discussed in the previous chapter. For each model, the best hyper-parameters are introduced, the results are evaluated, and the important metrics are discussed and compared with one another. Then, the feature importance of each model is brought to show how each model uses the explanatory variables to predict the accidents based on its performance.

4.2.1 Logistic Regression

A grid search algorithm is used to find the best hyper-parameters in the Logistic Regression model, and the optimal parameters with the results of the model after training, are shown in Tables 4.1a and 4.1b. As shown in Table 4.1b, results are not very promising and show that Logistic Regression is not a good model to find the relationship between explanatory variables and the accidents. The maximum accuracy reached by this model is 66%, and the Recall is even worse, 63%. To give an insight on how good the model result is, a 50% accuracy is equal to a random guess, which means a model with no knowledge of the previous accidents that can randomly guess the occurrence of accidents, has 50% accuracy. Recall is the most important metric for

accidents prediction, as it shows how many percent of actual accidents are correctly predicted as an accident.

Table 4.1: Optimal hyper-parameters and results of Logistic regression Model

(a) Hyper-parameters

Parameter	Optimal value
penalty	'l1'
tol	1e-4
C	1.0
random_state	None
solver	'liblinear'
max_iter	100
multi_class	'auto'

(b) Model results

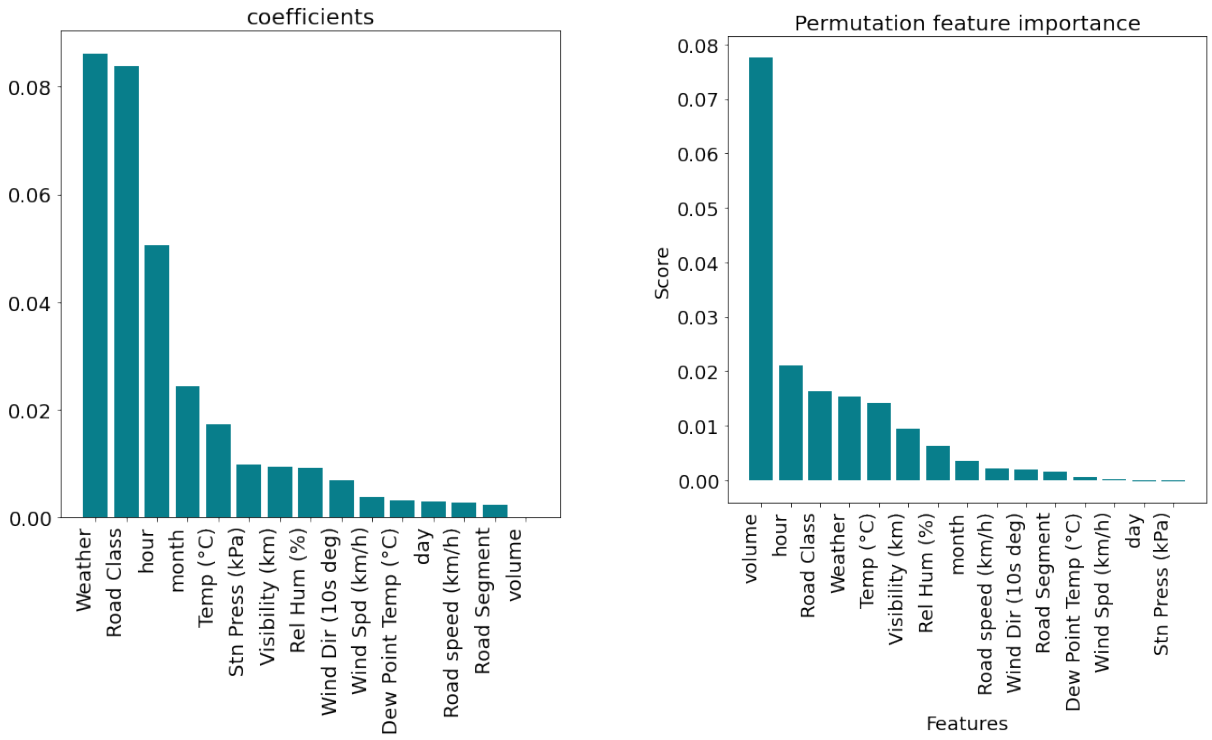
Metric	Result
Accuracy	66%
Precision	67%
Recall	63%
F1 Score	0.65

One possibility of the Logistic regression model is that by extracting the β_i or w_i, or coefficients of explanatory variables from a trained model, the importance assigned to each feature by the model can be indicated. The green bars in Figure 4.2a indicate the coefficients of the explanatory variables in the Logistic regression model. The road class has the biggest coefficient, followed by the hour and the Dow point temperature. The month, the standard pressure, and the temperature are the following features with moderate coefficients, and the rest of the features are assigned with coefficients with less than 0.01. The questionable point is that the traffic volume and the road segment are among features with low coefficients, while, from a rational point of view, these two features must be influential in accidents and receive bigger coefficients. An important outcome of this model is that the Dew point temperature as the weather attribute is among three variables with a bigger coefficient compared to others.

Since the linear combination of coefficients will be affected by the link function in a Logistic Regression model, extracted coefficients from that model cannot show how each feature ultimately affects the prediction of outputs. Therefore, a Permutation feature importance measure is performed, a model-agnostic metric to evaluate the effect of each feature on the prediction that has been introduced by Leo Breiman[89]. Permutation feature importance measures the effects of explanatory variables on the accuracy of the output, regardless of the models' structure, by randomly shuffling a single feature and measuring the drop in model accuracy. Shuffling the feature breaks the relationship between that feature and the target; consequently, the measured accuracy drop shows the target's dependency on that feature.

Green bars in Figure 4.16a illustrate each feature's importance in accidents utilizing the Permutation feature importance. The notable point is that even though the coefficient of the traffic volume variable is the smallest in Figure 4.16a, this variable has the highest impact on accident prediction based on the Permutation feature importance shown in Figure 4.16b. Following the traffic volume, The Dew point temperature is the second most important feature, then the road class and the hour. The latter three features are the features with the large coefficients in the Logistic Regression model that shows measuring the importance of the

variables based on the value of coefficients is neither accurate nor irrelevant, yet it shows some degrees of importance. The rest of the features have not been shown to contribute much to accident prediction, which is acceptable based on the model's accuracy and shows the Logistic Regression model is not powerful enough to capture all the non-linear relationships between accidents and accidents determinant.



(a) Coefficients of explanatory variables

(b) Permutation Score of all features

Figure 4.16: Feature Importance in the Logistic Regression model

Besides the coefficients of the variables in a logistic regression model, the p-value is another parameter that shows how important each variable is to determine the target variable, as shown in Table 4.2. By convention, the threshold to evaluate the p-value is 0.05, which means if the p-value of a variable is smaller than 0.05, that variable is significant in the analysis; if greater than 0.05, the variable has no significance statistically and can be excluded. As can be seen, all p-value of all the variables are less than 0.05, and they all are significant in the regression model. The coefficient sign is another important factor that shows if the relationship between the variable and the target variable is direct or reverse. For example, the temperature coefficient sign is negative, which means lower temperatures are correlated with a higher number of accidents and vice versa. Also, since the traffic volume coefficient is positive, higher traffic volumes are more correlated with a higher number of accidents.

Comparing the coefficients of some of the variables such as traffic volume with some studies in the literature such as [21] shows that the sign of the coefficient and the p-value verifies what is found in the

literature. However, the coefficient value is slightly different, which shows that traffic volume has a slightly different correlation with accidents in different places.

Variable	Coefficient	p-value
Month	0.0264	0.000
Day	-0.0026	0.001
Temperature	-0.0184	0.000
Dew point Temperature	-0.0025	0.000
Humidity	-0.0095	0.000
Wind Direction	-0.0067	0.000
Wind Speed	0.0042	0.000
Visibility	0.0111	0.000
Pressure	-0.0441	0.000
Weather	0.0755	0.000
Hour	0.0529	0.000
Road Segment	-0.0028	0.000
Road Speed	0.0061	0.000
Traffic Volume	0.0002	0.000
Road Class	0.0841	0.000

Table 4.2: Coefficients and p-value of variables in the Logistic Regression model

4.2.2 Decision Tree

As discussed in the previous chapter, the difference threshold between training and test accuracy scores for this application is set to a maximum of 1%. Then, the optimal hyper-parameters are found using a grid search algorithm for the Decision Tree model; the best hyper-parameters and the best model results are shown in Table 4.3. Results shown in Table 4.3b are slightly better than the results of the Logistic Regression, especially, Recall, which is the most important metric, has 3% improvement from 63% in Logistic regression to 66% in the Decision tree, which shows the Decision tree performs slightly better.

Since the Decision Tree utilizes the Gini Impurity measure, this measure can show how this model has taken different explanatory variables into consideration. Green bars in Figure 4.17a illustrate the Gini score assigned to each variable. Unlike Logistic Regression, which gave greater than zero coefficients to almost all the features shown in Figure 4.16b, the Decision Tree has not captured any information in features such as the day, the weather, the wind speed, and standard pressure. Gini Impurity feature importance can

Table 4.3: Optimal hyper-parameters and results of the Decision Tree Model

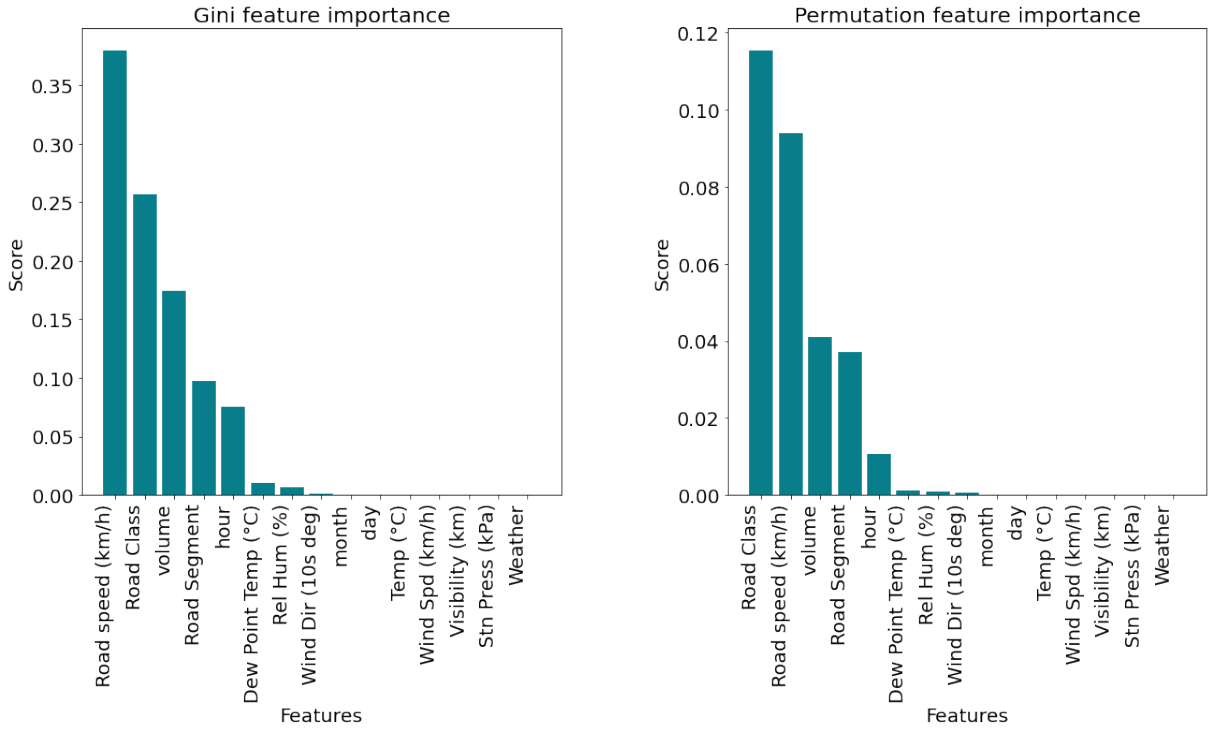
(a) Hyper-parameters

Parameter	Optimal value
criterion	'gini'
max_depth	30
cpp_alpha	0.001
random_state	0
min_impurity_decrease	0
min_sample_leaf	8
min_sample _split	20
splitter	'best'
class_weight	None
max_feature	'auto'
min_weight_fraction_leaf	0.0

(b) Model results

Metric	Result
Accuracy	68%
Precision	73%
Recall	66%
F1 Score	0.69

be misleading, especially when dealing with variables with many unique values; therefore, a Permutation feature importance is also performed to evaluate the effect of explanatory variables on the prediction using the trained Decision Tree. Based on the results in Figure 4.17b, Permutation importance verifies the Gini impurity measure; and also it shows the Decision Tree classifier failed to capture the relationship between accidents and the weather features. This classifier predicts accidents mostly based on the speed limit, the road class, the traffic volume, the hour, and the road segment.



(a) Gini importance of explanatory variables

(b) Permutation Score of all features

Figure 4.17: Feature Importance in the Decision Tree model

4.2.3 Random Forest

The optimal hyper-parameters that delivered the highest accuracy in the Random Forest model are shown in Table 4.4a. Table 4.4b shows the results of the Random forest model. The overall accuracy is improved by 4% compared to the Decision tree, and also the Recall improved from 66% to 68% in Random Forest. Random Forest has also improved other metrics such as Precision by 1% and F1 score from 0.69 to 0.71. Random Forest uses the same impurity measure as Decision Tree; thus, an impurity feature importance can be extracted from the model to show how much influence each explanatory variable has on the decision making.

Table 4.4: Optimal hyper-parameters and results of the Random Forest Model

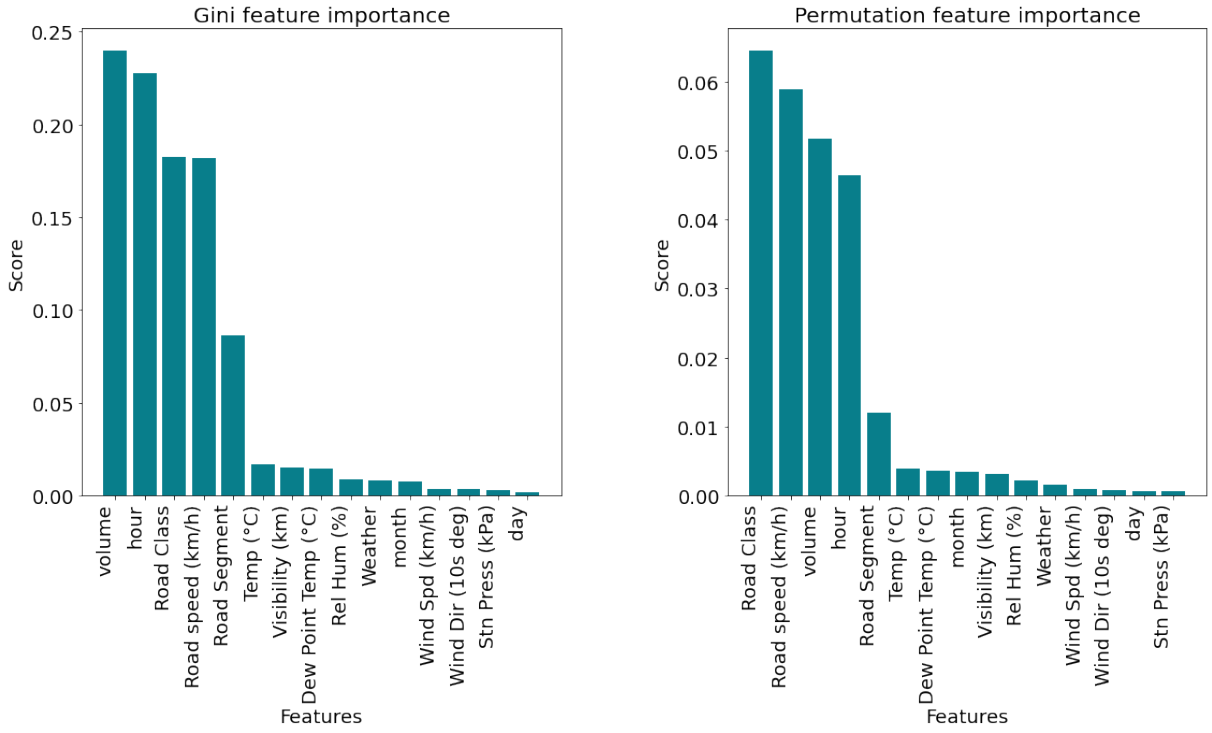
(a) Hyper-parameters

Parameter	Optimal value
criterion	'gini'
max_depth	30
cpp_alpha	0.0
random_state	0
min_impurity_decrease	0.00008
min_sample_leaf	3
min_sample _split	10
class_weight	None
max_feature	'auto'
min_weight_fraction_leaf	0.0
n_estimator	200
bootstrap	True
n_jobs	None
max_samples	None

(b) Model results

Metric	Result
Accuracy	72%
Precision	74%
Recall	68%
F1 Score	0.71

Figure 4.18a shows the extracted impurity feature importance from the model. It shows that the Random Forest makes the decisions using road characteristics and weather features. Random Forest indeed has recognized some relationships between accidents and weather attributes. Gini feature importance only shows the importance of the features in the training data; in order to show the real effect of the features in unseen data on prediction, Permutation importance is performed as shown in Figure 4.18b. Road class, traffic volume, road speed, the hour, and the road segment are the most influential factors that show Random Forest follows the Decision Tree's main principles because these first five factors are the same as the first five factors in the Decision tree model. However, after those factors, the weather attributes contribute to accident prediction; that shows that even though a single decision tree was not able to extract the dependencies between accidents and weather, a combination of trees can find them.



(a) Gini importance of explanatory variables

(b) Permutation Score of all features

Figure 4.18: Feature Importance in the Random Forest model

4.2.4 Gradient Boost

Several hyper-parameters are to be tuned when the whole ensemble of weak learners in a Gradient Boost model is constructed. The list of some of the critical hyper-parameters with the optimal values is obtained in Table 4.5a; where 'iteration' shows the maximum number of weak learners or trees that can be built, 'learning_rate' is used to specify the gradient steps, 'l2_leaf_reg' is the coefficient of L2 regularization in the cost function, 'depth' shows the depth of each tree, 'early_stopping_rounds' sets the overfitting detector type.

Table 4.5: Optimal hyper-parameters and results of the Gradient Boost Model

(a) Hyper-parameters

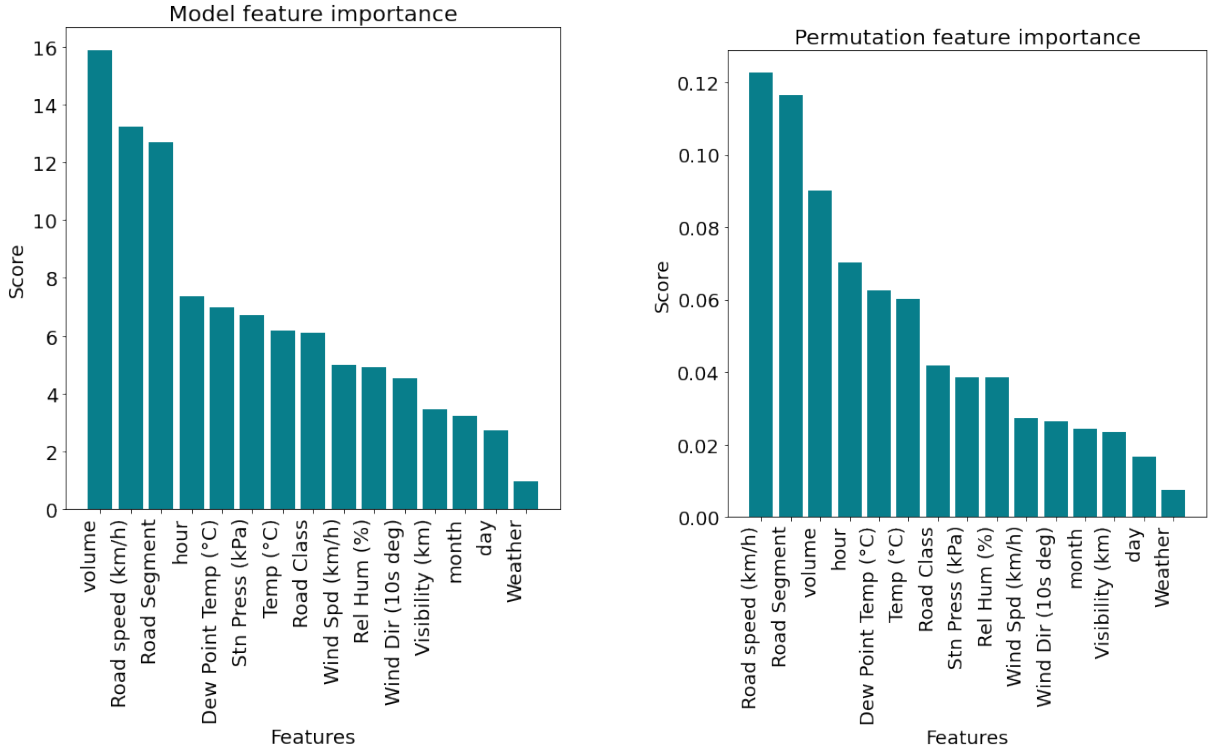
Parameter	Default value	Optimal value
iterations	1000	150
learning_rate	0.03	0.5
l2_leaf_reg	3.0	3.0
depth	6	7
early_stopping_rounds	False	False

(b) Model results

Metric	Result
Accuracy	80%
Precision	79%
Recall	80%
F1 Score	0.79

Based on the results in Table 4.5b, the Gradient boost trees are more robust than the other two previous tree-based models, Decision Tree and Random Forests. The Recall, which is the most important metric, has

been significantly improved by 12% compared to the Random Forests, and the overall accuracy has been improved by 8% from 72% in Random Forest to 80% in Gradient Boost. The robustness of the Gradient Boost has enabled the model to capture the relationships between the occurrence of accidents and weather factors besides road characteristics, which is illustrated by the internal feature importance of Gradient Boost and the Permutation feature importance in Figures 4.19a and 4.19b. The Permutation score shows that the Gradient Boost classifier utilizes Dew point Temperature and the Temperature as the fourth and fifth influential factors in accidents following the Speed limit, road segment, road traffic, and the hour of the day. Other weather features such as the standard pressure, relative humidity, wind speed and direction are also important in predictions. Besides, the least important factor is the overall weather determined by the model.



(a) Gradient Boost internal importance of explanatory variables

(b) Permutation Score of all features

Figure 4.19: Feature Importance in the Gradient Boost model

4.2.5 Neural Networks

As discussed in the previous chapter, hyper-parameter tuning in Neural Networks includes two steps; finding the best architecture of the model and finding the optimal internal hyper-parameters. In the first stage, the aforementioned recommendations were taken to develop the first Neural Network, and after hyper-parameter tuning, the prediction results are shown in Table 4.6. The first model contains two hidden layers, the first

with 100 neurons, and the second is a dropout layer to help reduce overfitting with a 10% drop rate. The first model results are promising compared to the previous models; the overall accuracy improved to 88% from 80% in the Gradient Boost model, and the Recall has improved roughly 3% from 80% to 83%. Thus far, the Neural Network is superior to the previous models; however, the Recall score has more room to improve since that is a critical metric in accidents prediction.

Metric	Gradient Boost	First Neural Network
Accuracy	80%	88%
Precision	79%	94%
Recall	80%	83%
F1 Score	0.79	0.89

Table 4.6: Result comparison between GB and NN

In the second stage, two more layers are added to the Neural network model, making it a deep Neural Network. The results shown in Table 4.7, show a slight improvement in all metrics over the first model and show the capacity of a Deep Neural Network. Although the initial extremely imbalanced aggregated accidents data is undergone resampling techniques to have a balanced number of accidents and non-accident events; the preliminary experiments showed the Recall score is slightly lower than other scores in the first two models; which led to utilize a technique to assign weights to classes during the training process. Thus, the last and the best model is trained using four layers and weighted classes, improving the Recall score by 4%, but has slightly worsened other metrics as a trade-off. Table 4.7 shows the final results of the best model, in which the overall accuracy is 92%, and the Recall is 92%.

Metric	2 hidden layer	4 hidden layers	4 hidden layers with weighted classes
Accuracy	88%	93%	92%
Precision	94%	96%	90%
Recall	83%	88%	92%
F1 Score	0.89	0.93	0.91

Table 4.7: The results of different Neural Network Architectures

Figure 4.20 illustrates the structure of the best Deep Neural Network (DNN) model with six layers; An input layer that includes 199 neurons that represent 15 features, an output layer that outputs the probability of the accident, and four hidden layers. The first and the third hidden layers include 110 neurons each, and the second and fourth layers are dropout layers with a 10% drop as a regularization to control overfitting.

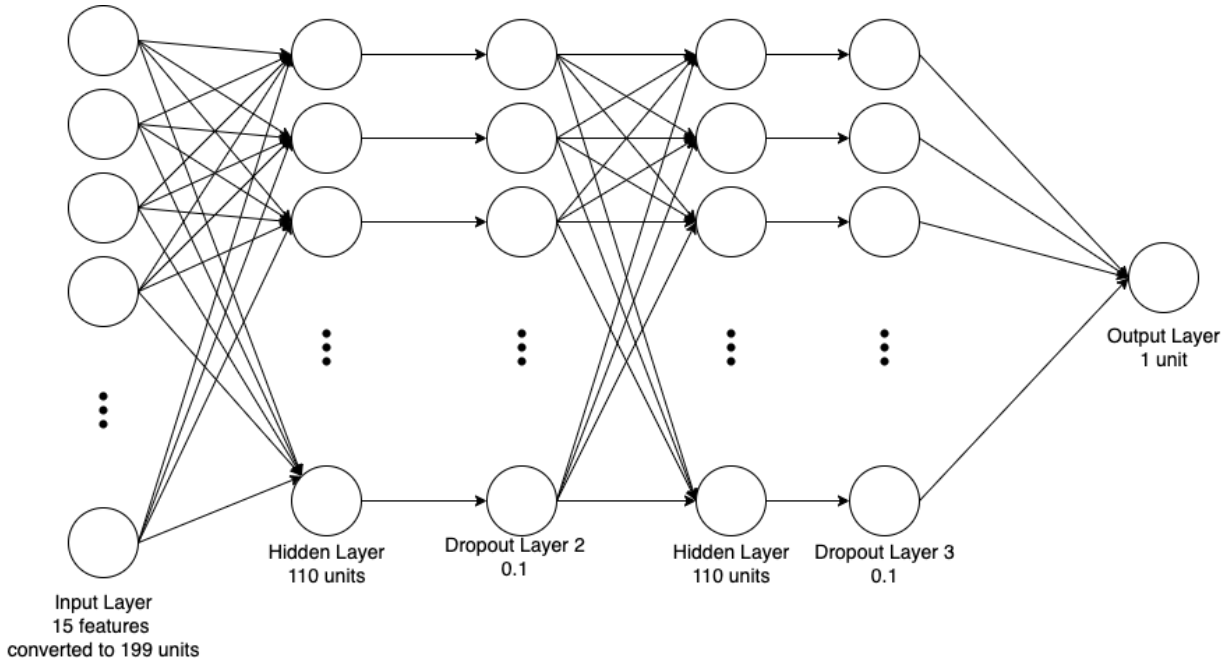


Figure 4.20: Final Neural Network structure

Neural Networks are known as black boxes that cannot be interpreted. There are too many parameters in a Deep Neural Network that are optimized to give the most accurate output, and it seems to be nearly impossible to track the effect of each explanatory variable through all those parameters and connections and activation functions; however, one model-agnostic method can be used to evaluate the effect of those input variables in prediction. Figure 4.21 illustrates how important the input variables are in predicting accidents, utilizing the Permutation feature importance.

The proposed DNN model is able to find non-linear relationships and hidden patterns in the dataset. Unlike the previous ML and statistical models, the DNN uses all the features with high degrees of importance. For instance, the features, such as the road segment, the hour, the day, the speed limit, and the month somehow have the same effect on accident prediction, after the road segment. Following those features, all the weather attributes, the road class, and the traffic volume correlate to accidents somehow the same.

In order to show how much the weather attributes are affecting the prediction, in other words, how correlated the weather attributes are with the accident occurrence, all the weather features are removed, and the same DNN is trained without the weather features. Without considering the weather feature, the results showed that the maximum accuracy the model can reach is 67%, which is 25% less than when all the weather features are included in the model training.

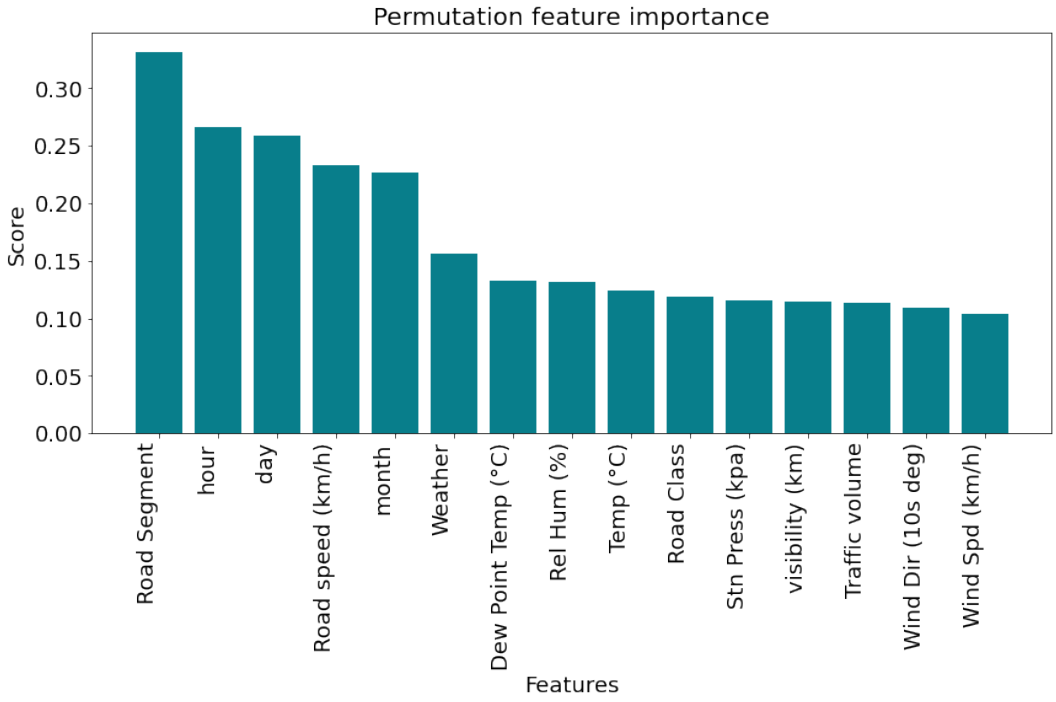


Figure 4.21: Deep Neural Network Permutation feature importance

Chapter 5

Application

This chapter provides an application of the proposed deep learning model to show the applicability of the thesis. As discussed in the previous chapters, different factors have been identified that can cause or influence accidents in different circumstances. One of those factors is the road speed limit, which was recognized as the fourth determinant of the accidents in Calgary by the proposed deep learning model. Unlike speeding, a potential risk factor in driving, the speed limit's effect remained controversial. One study found that although driving at a higher speed increases the braking distance, lowering the average speed may cause more accidents because interactions between vehicles are more at lower speeds[90]. Some researchers found a positive relationship between the frequency of accidents and speed using statistical techniques, while others found the opposite[91]. Hence, speed has remained a controversial factor, and there are still debates about it among transportation researchers and engineers.

The proposed deep learning model was trained on historical accidents data with distinctive road characteristics and weather features. The speed effect on five highway road segments is investigated by altering the speed limit and investigating the rate of accidents prediction as to the model's output. By doing so, the effect of increasing and decreasing the road speed limit will be found on the number of accidents, and new speed limits can be proposed to lower the number of accidents accordingly. This analysis is only conducted on the highways, since the exploratory analysis shows that in lower speed limits on residential roads, there is no evident correlation between the posted speed limit and the occurrence of accidents; therefore the model cannot be used to propose changes on those road segments. In addition, the posted speed limit on the residential roads are mostly determined by the safety policies and legislation.

5.1 Speed limit effect

The blue, Orange, and red curves in Figure 5.1 show the accident prediction results on Stoney Trail using various speed limits ranging from 80 to 110 km/h. This freeway is divided into Stoney TR NE, Stoney TR NW, and Stoney TR SE. Even though these three curves are slightly different due to their unique traffic volume, road characteristics, demographic and driving behavioural differences, they follow the power model shown in Figure 5.2, which Goran Nilsson introduced to show the relationship between average speed and the number of crashes [92]. This power model is one of the primary and basic foundations and principles of road safety.

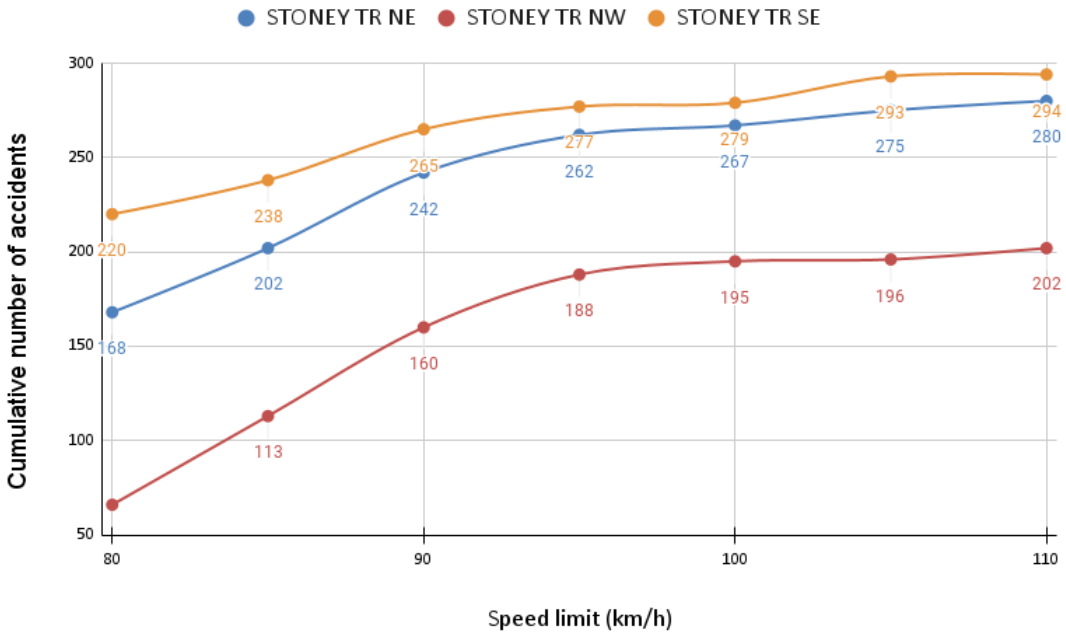


Figure 5.1: Predicted number of accidents on Stoney TR segments

Reducing the speed limit from 100 to 90 km/h caused an 18% (from 195 to 160), 9% (from 267 to 242), and 5% (from 279 to 265) reduction in the predicted number of accidents on the freeway's northwest, northeast, and southeast segments, respectively. The orange curve shows the changes in predicted accidents in the southeast part of the freeway. It has the lowest change by lowering the speed limit among other segments due to differences in demographics, driving, and road characteristics. On the other side, more accidents are predicted if the speed limit is increased; for instance, 10 km/h increment results in 15, 13, and 7 more predicted accidents on Stoney TR SE, NE, and NW, respectively.

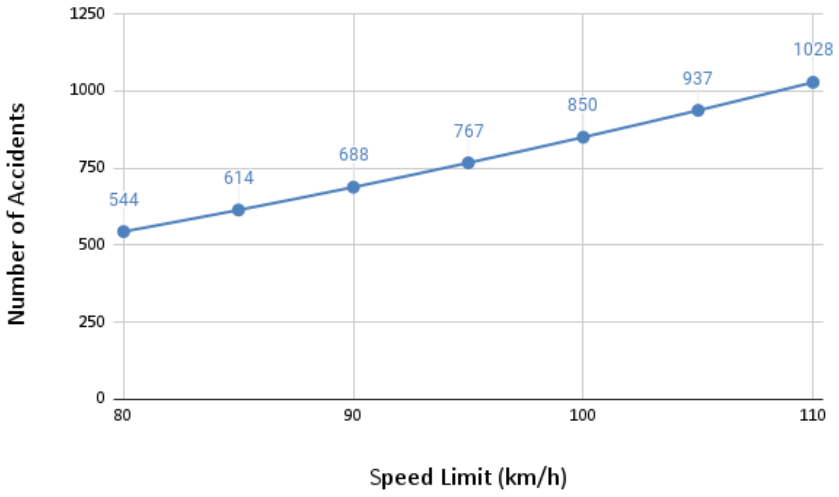


Figure 5.2: Power model introduced by Nilsson to show the relationship between accidents and average speed

The blue and the red dots in Figure 5.3 show the predicted number of accidents on the Deerfoot trail using different speed limits from 80 km/h to 110 km/h. By reducing the speed limit from 100 km/h to 90km/h on this freeway, the estimated accident reduction is 6% (from 804 to 751) on the north part and approximately 5% (from 964 to 911) in the south part. Furthermore, if the speed limit is increased from 100 km/h to 110 km/h, the predicted increase in accidents will be roughly 3% (from 804 to 829) and 9% (from 964 to 1,049) on north and south segments, respectively.

It can be observed that the effect of speed limits on road accidents in different road segments are not exactly the same, and they can behave slightly differently based on the other characteristics of the road segment, although they all follow the same pattern.

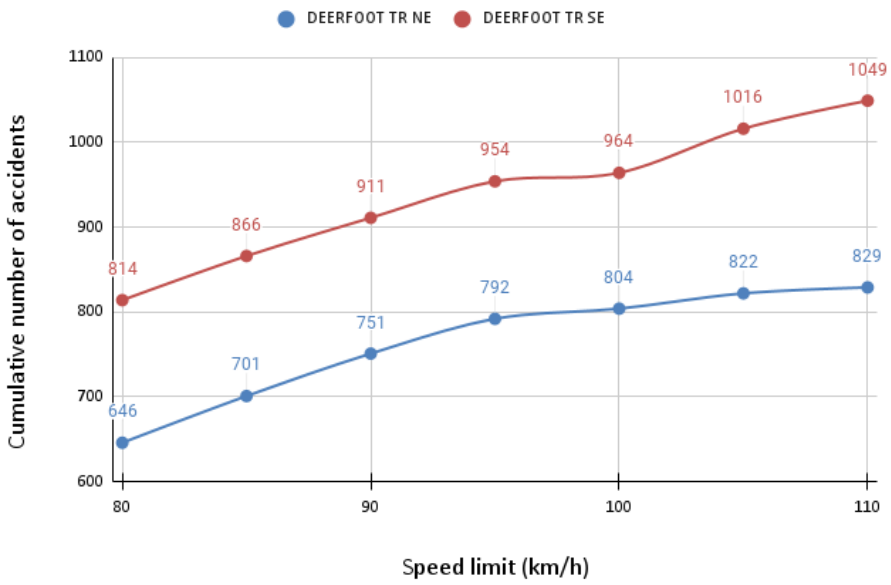


Figure 5.3: Predicted number of accidents on Deerfoot TR segments

5.2 Level of Service

What can be derived from Figures 5.1 and 5.3 is that lowering the speed limit will result in fewer predicted accidents. Thus, the number of accidents can be decreased by proposing lower speed limits, but how far the speed limit can be lowered. The primary purpose of freeways is high mobility. As a part of Calgary's Ring Road, fast and reliable travelling and good movements are two crucial factors for building and expanding freeways within and around Calgary [93]. Therefore, considering the level of service (LOS) on a freeway is essential in any proposed changes, which is defined as "a quality measure describing operational conditions within a traffic stream, generally, in terms of such service measures as speed and travel time, freedom to maneuver, traffic interruptions and comfort and convenience"[94]. Based on the Highway Capacity Manual by Transportation Research Board, six distinctive categories can be defined and measured on each highway or freeway segment as service level: A, B, C, D, E, and F [94].

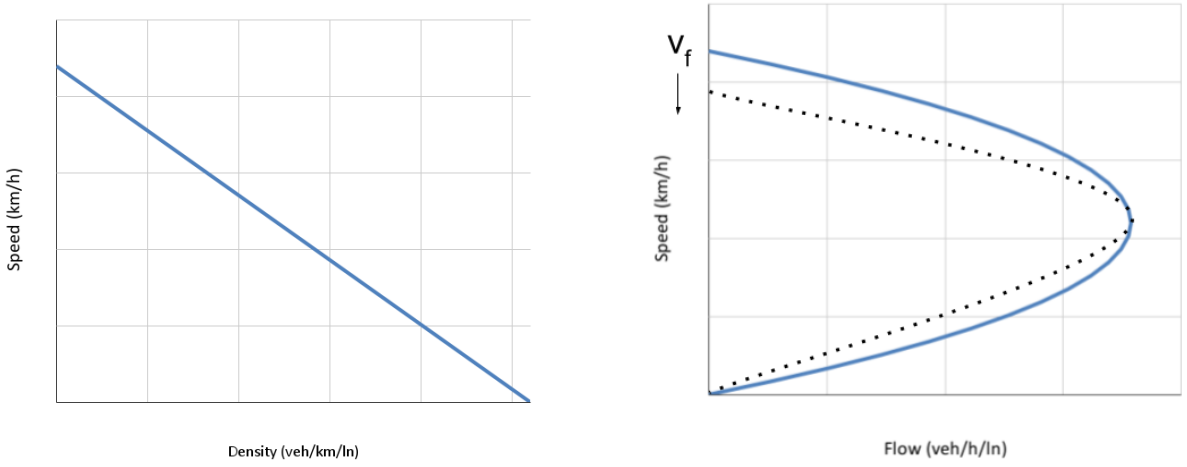
Different service levels are defined as follows:

- LOC A: Traffic flows at or above the speed limit, the ability to maneuver and change lane is high, driving is very comfort.
- LOC B: Traffic flows fairly well at the posted speed limit, change lane is slightly restricted, driving is comfort.

- LOC C: Traffic flow is stable, lane change is noticeably restricted, the flow is close to capacity.
- LOC D: Traffic flow is unstable, the average speed is slightly below the speed limit, lane change is much more limited, driving with some delays and higher total travel time.
- LOC E: Traffic flow is unstable, speed varies rapidly, small disruption can create shock waves affecting traffic upstream.
- LOC F: Broken down traffic flow, demand is more than capacity, cars move in lockstep with the car in front.

As a rule of thumb, any changes in the average speed can affect traffic attributes, and as a result, service level might change too; since the LOC is measured by speed, density, and service flow rate/demand volume. The relationships between speed, density, and flow rate are defined by the Fundamental Diagrams(FD) proposed by Greenshield. A linear relationship between speed and density was assumed by Greenshield, as shown in Figure 5.4a, which follows Equation (5.1), where v is the average speed, k is density, v_f is the free-flow speed, and k_j is the jam density. Furthermore, based on the Greenshield model, the relationship between speed and flow can be represented as the blue parabolic shape in Figure 5.4b, and the flow can be found as a product of speed and density.

The speed-flow curve shows the overall quality of the traffic operation of a road segment, which is divided into two regimes. The upper side of the parabola is called the unsaturated regime and represents the LOS A, B, C, and D. The lower side of the parabola is called the saturated regime and represents the LOS E and F, which is a broken-down traffic flow.



(a) Speed-density relationship

(b) Speed-flow relationship

Figure 5.4: FD by Greenshield

$$v = v_f - [\frac{v_f}{k_j}]k \tag{5.1}$$

Reducing the speed limit implies a reduction of the free-flow speed(v_f) ; consequently, the relationship between speed and flow shown in Figure 5.4b changes which is shown by the black dotted line. Since the speed is reduced and the flow rate/demand volume is constant, the density will be increased based on Equation (5.1), which is used to determine the service level. Figure 5.5 shows the constructed speed-flow curves for different free-flow speeds with the related density and additional service level criteria for a freeway road section.

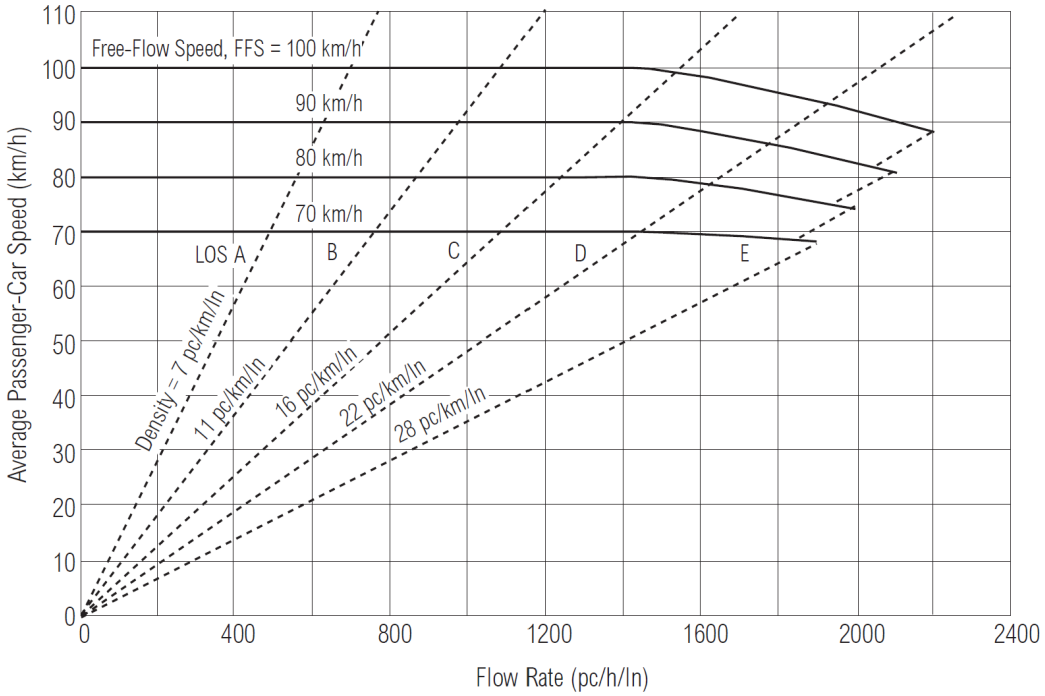


Figure 5.5: Speed-Demand flow curves with LOS Criteria

Service flow rate can be calculated using Equation (5.2); where v_p is the demand flow rate (pc/h/ln), V is the demand volume (Average Annual Hourly Traffic), PHF is the peak-hour factor, N is the number of lanes (in one direction), f_{HV} is the factor of the presence of heavy vehicles, and f_p is the adjustment factor for unfamiliar driver population.

$$v_p = \frac{V}{PHF * N * f_{HV} * f_p} \tag{5.2}$$

5.3 Proposed Speed Limit

Deerfoot TR NE and Deerfoot TR SE are examined using the proposed predictive model and speed-flow curves mentioned in the previous section to see whether reducing the speed limit will change the traffic patterns and drop the service level or not.

Based on the HCM, the capacity of a freeway segment is the number of lanes multiplied by $2,300$ vehicles per hour per lane for six or more lanes(in two directions). The maximum annual average weekday traffic in Deerfoot TR NE is $165,000$ vehicles per day(in two directions). The current speed limit is 100 km/h, therefor if the service flow rate is calculated with Equation (5.2), with the use of the speed-flow curves, the current level of service can be determined.

The below parameters are assumed:

- The free-flow speed, the mean speed of passenger cars during a moderate flow without disruptions, assumed the speed limit.
- The f_p or the adjustment factor for the unfamiliar driver population is assumed to be one since Calgary is not an attractive tourist place and the road users are the residents of Calgary.
- The PHF or the peak hour factor is assumed to be one.
- The f_{HV} or heavy vehicle factor is calculated to be 0.8, since based on the Canadian Vehicle Survey the P_T or the proportion of heavy vehicles in Alberta is roughly 0.07[95], and based on the HCM the E_T or the passenger car equivalency for trucks and busses is 1.3; Using Equation (5.3), the f_{HV} can be calculated.

$$f_{HV} = \frac{1}{1 + P_T(E_T - 1)} \tag{5.3}$$

The demand volume is roughly $82,500$ vehicles per day in one direction in 3 lanes. Using the above assumptions, the service flow rate is calculated to be $1,450$ passenger cars per hour per lane. In this case, the density is equal to 15 passenger car per kilometer per lane. using the speed-flow curve the level of service is determined to be in class C as shown by point 1 in Figure 5.6.

If the speed limit is reduced to 90 km/h with the same service flow rate, which is $1,450$ passenger cars per hour per lane, the density becomes 17 passenger cars per kilometre per lane, and consequently, the traffic status moves from point 1 to point 2 in Figure 5.6. As can be seen, point 2 is in the LOC D, in which the interactions between cars increase and lane changes and maneuvers are more restricted, which can cause more accidents. In conclusion, even though reducing the speed limit from 100 to 90 km/h in Deerfoot TR NE might theoretically reduce the number of accidents by 6% based on the results of the prediction shown

in Figure 5.3, it will drop the service level from C to D; Therefore no speed limit reduction is recommended for this road segment.

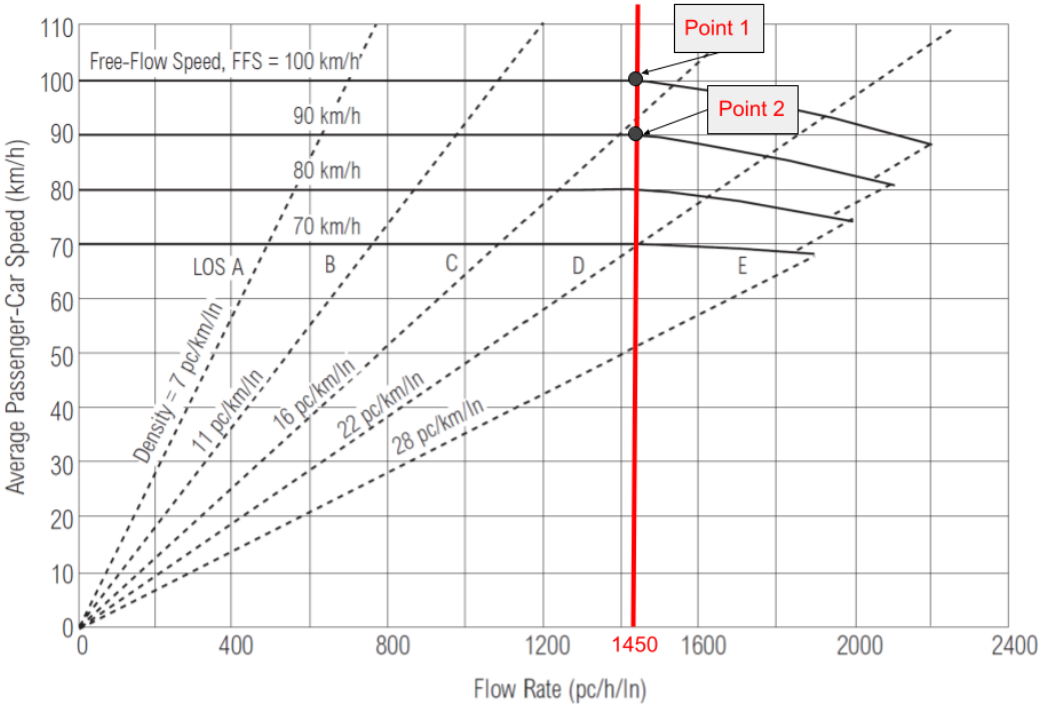


Figure 5.6: Deerfoot TR NE LOS

The same analysis is conducted on the Deerfoot TR SE, which has a maximum annual average weekday traffic of 158,000 vehicles per day(in two directions).

The current speed limit in this road segment is 100 km/h, and considering the same assumptions as of the northern part of the highway, the service flow rate is calculated to be 1370 passenger cars per hour per lane. Therefore, the density equals 14 passenger cars per kilometres per lane, and the traffic status is shown by point 1 in Figure 5.7, which is in service level C. If the speed limit is reduced to 90 km/h with the same service flow rate, which is $1,370$ passenger cars per hour per lane, the density becomes 16 passenger cars per kilometre per lane, and consequently, the traffic status moves from point 1 to point 2 in Figure 5.7. As shown in Figure 5.7, both point 1 and point 2 are located in LOS C, which shows that lowering the speed limit from 100 to 90 km/h on this road segment will maintain the service level. In conclusion, a reduction of 10 km/h in the speed limit from 100 to 90 km/h in Deerfoot TR SE is recommended, resulting in a 5%(from 964 to 911) reduction in predicted accidents with maintaining the level of service in that road segment.

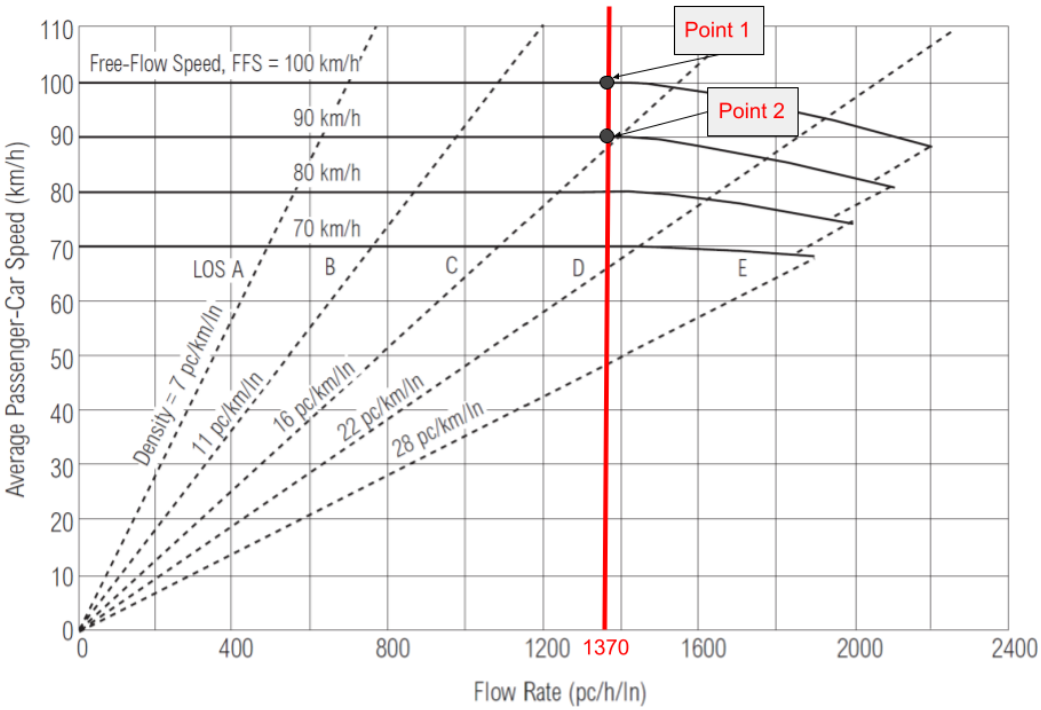


Figure 5.7: Deerfoot TR SE LOS

The effect of increasing and decreasing the speed limit on Deerfoot TR was examined utilizing the proposed Deep Neural Network model. The results showed that decreasing the speed limit from 100 to 90 km/h in Deerfoot TR SE results in a 5% reduction in the predicted accidents, from 964 to 911, with maintaining the level of service on that road segment.

The recommended reduction in the speed limit on Deerfoot TR SE not only reduces the number of accidents by 5% but also reduces the severity and fatality of accidents. According to the power models introduced by Nilsson[92], shown in Equations (5.4) and (5.5), reducing the speed limit on Deerfoot TR SE decreases the injuries and fatalities by 19% and 35%, respectively; where f_1 and g_1 are the number of injuries and fatal accidents with the current speed limit, f_2 and g_2 are the number of injuries and fatal accidents with the new speed limit, V_1 is the current speed limit, and V_2 is the new speed limit.

$$f_2 = (\frac{V_2}{V_1})^2 f_1 \tag{5.4}$$

$$g_2 = (\frac{V_2}{V_1})^4 g_1 \tag{5.5}$$

Even though a 10 km reduction in the speed limit on Deerfoot TR SE with a length of 35 km increases the travel time by 2 minutes, the 5% reduction in accidents will compensate for that, as accidents are the

cause of more than 50% of the delays on freeways based on Traffic Incident Management Handbook[96].

Another aspect of road accidents is the costs imposed on not only the people involved in accidents but also the society as a whole. The costs are categorized into three types: direct, human capital, and willingness to pay (WTP). Table 5.1 shows the costs in different categories based on fatal, injury, and property damage collisions[97].

Cost	fatality	Injury	Property Damage
Direct Costs	$225,558	$48,341	$14,065
Human Capital	$2,224,580	$89,408	$0
Willingness to Pay	$6,707,228	$158,654	$0

Table 5.1: Accident Costs

According to Alberta Traffic Collision Statistics, on average, 90.8% of the collisions are property damage collisions, 9% injury collisions, and 0.2% fatal collisions[98]. In conclusion, the recommended 10 km/h speed limit reduction on Deerfoot TR SE could reduce the number of accidents by 5%, consequently saving $169,216 in property damage, $353,460 in injury, and $242,670 in fatal collisions, and save a total of roughly $765,350, yearly.

Lastly, the environmental effect of speed limit reduction is promising. The red curve in Figure 5.8 illustrates the CO2 emission with regards to various speeds, which shows a ten km/h reduction from 100 to 90 km/h will result in emitting 4.8% less CO2. Considering the daily traffic volume and the length of the Deerfoot TR SE, roughly 43 tonnes less CO2 will be emitted daily, which equates to the amount of CO2 emitted by 10 Canadians in their households per year.

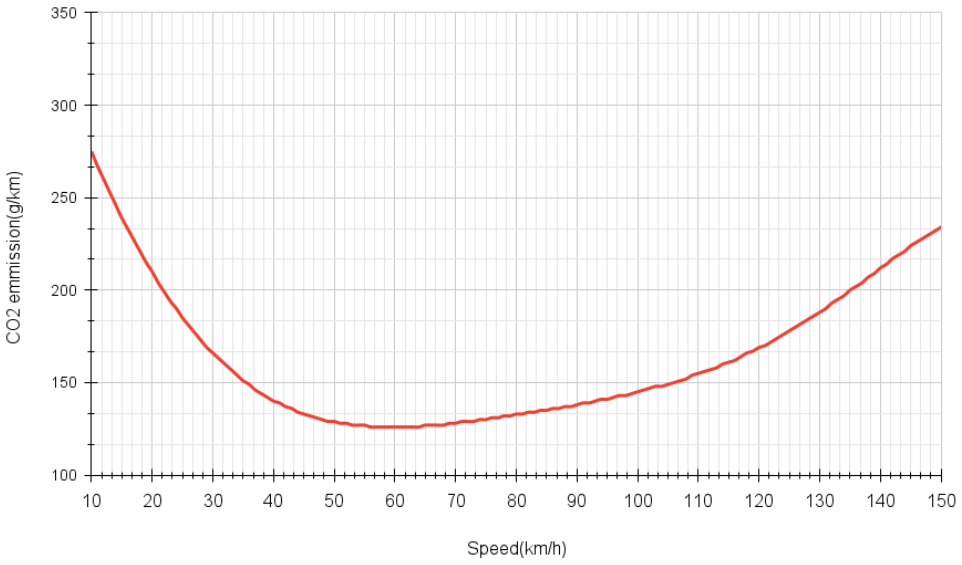


Figure 5.8: CO2 emission rate with regards to the Speed

Chapter 6

Results

This chapter discusses the results of data preparation, exploratory data analysis, and the results of Accident Prediction Model (APM) developments.

6.1 Data Preparation

In Chapter 3, the constructed dataset of accidents which was a combination of actual accident samples and a set of generated negative samples, was found to be highly imbalanced with a ratio of 361 non-accident samples to 1 accident sample. Also, the dataset contained roughly 3.5 million samples. Different resampling techniques were examined since the imbalanced datasets are harmful to the model's performance. Among those, a hybrid approach including Repeated Edited Nearest Neighbours Undersampler and Random Oversampler was taken to create a balanced dataset with 2 million samples of accident and non-accident samples distributed equally.

6.2 Exploratory Data Analysis

Feature analysis incorporating the Mutual Information Score in Chapter 3 showed that among all the variables in the feature space, road class, road speed, road segment, traffic volume, visibility, and overall weather condition had the highest correlation with accidents and the remaining features have not been shown to correlate much with accidents. Of course, the Mutual Information Score is calculated based on the Information Gain concept and was not powerful enough to capture all the dependencies. Thus, an unsupervised learning method called PCA was utilized to evaluate the effect of feature exclusion on the amount of variance in the dataset. PCA showed that in order to keep all the variance in the dataset, 14 out of 15 variables are needed,

which was a complementary result to the Mutual Information Score.

The analysis of individual explanatory variables gave insights into how accidents are correlated to those variables. Based on the overall weather analysis results shown in Figure 4.3, the weather involved ice crystals and snow accounted for the highest rate of accidents in Calgary. Driving on a surface with ice crystals is correlated to two times more likelihood of being involved in an accident, and driving on a snowy day is correlated with the increase in the probability of accidents by 55%, compared to clear or cloudy days. Another insight that can be derived from a comparison between snowy days and clear days shown in Figure 4.4 is that the number of monthly accidents is highly correlated with the number of snowy days in each month; on the contrary, the number of clear days fluctuated from month to month and had no evident correlation with accidents.

As another accident determinant, the temperature is significantly correlated with the accidents; the accidents rate below -30 degrees Celsius was three times more than between +5 to +15 degrees Celsius, as depicted in Figure 4.6. The temperature had a twofold correlation with the accidents; above +15 degrees Celsius had a direct and approximately linear correlation, and below 0 degrees Celsius had a reverse and approximately exponential correlation. Driving when the temperature is between +5 to +15 degrees Celsius had the lowest likelihood of being in an accident, which was considered the safest temperature range for driving. The colder the temperature, the higher the likelihood of accidents; for instance, the rate of accidents between -25 to -30 degrees Celsius was twice the rate of accidents between -5 to -10 degrees Celsius.

Relative humidity has been shown to have correlation with accidents, but its changes were resulted in milder changes to accidents compared to the temperature. Above 50% humidity did not have much correlation with the accidents; however, below 50% humidity did; for example, the rate of accidents when the humidity is 5 to 10% was twice the time when humidity is 30 to 40%.

The analysis on the time of the accidents showed 33% more accidents on the weekdays than the weekends, and Fridays had the highest number of accidents among weekdays. Furthermore, on the weekdays, accidents were highly influenced by the rush hour traffic; one in the evening from 4 pm to 6 pm, and another in the morning from 7 am to 9 am; however, the number of accidents in the evening rush hour is 50% more than the morning rush hour. On the weekends, though, the accidents were more scattered during the day; however, the evening hours from 12 pm to 7 pm had seen more accidents.

The road class also showed to be important when analyzing the accidents; for instance, Skeletal roads, roughly 24% of the roads in Calgary, accounted for 50% of the accidents from 2017 to 2020; while Arterial streets, with 23.2% of the accidents, includes 41.1% of roads in Calgary.

Besides traffic volume, speed limit, and the road class, the road segment could represent the unique characteristics of the road. The comparison between the number of accidents on the individual road segments

within the accident dataset showed that road segments followed neither traffic volume, speed limit, nor the road class entirely. For example, two road segments with the same traffic volume had a very different number of accidents; and two other road segments with the same speed limits also had a very different number of accidents. Moreover, two road segments that belonged to the same road class had also seen a significant difference in the number of accidents. Those differences show that the road segment has intrinsic characteristics related to the road's geometry, length, and width that caused those differences and can influence accidents and, therefore, have to be considered in model development.

The variability in the speed limit had a positive correlation with accidents between 50 to 100 km/h; in such a way that the rate of accidents in the speed limit of 100 km/h is 6.7 times, 2.95 times, 2.5 times, and 1.5 times more than the speed limit of 50, 60, 70, and 80km/h, respectively. In addition, high traffic volume was also shown to have a considerable influence on accidents, especially traffic volume of more than 40,000 vehicles per hour.

6.3 Accident Prediction Model results

Chapter 4 presented the development of five models, including Logistic Regression, Decision Tree, Random Forest, Gradient Boost, and Deep Neural Networks. After many preprocessing steps, the accident data was split into two main datasets, 80% for training and 20% for the test. The training set was used for model training, and the test dataset was considered a held-out set or unseen data to evaluate the model's performance. The training dataset was again split into two parts for some models: a smaller training dataset and a validation set. The k-folds Cross-Validation was used during the training stage in order to have a biased-free model since the data in the training and validation might have different distributions. Finally, the models had been undergone hyper-parameter tuning. Performance comparison of the mentioned trained models is brought in Table 6.1.

Metric	Logistic Regression	Decision tree	Random Forest	Gradient Boost	Deep Neural Networks
Accuracy	66%	68%	72%	80%	92%
Precision	67%	73%	74%	79%	90%
Recall	63%	66%	68%	80%	92%
F1 Score	0.65	0.69	0.71	0.79	0.91

Table 6.1: The results of developed predictive models

The Logistic Regression model, a special case of the Generalized Linear Model, did not perform well on the accidents data; its highest accuracy was 66%. The Recall Score or the True-Positive Score, which shows the percentage of correctly predicted accidents, was even worse, 63%, which means 37% of the actual

accidents were incorrectly predicted as non-accidents. The Decision Tree model improved the accuracy by 2% and the Recall by 3%, which is still low regarding the importance of accidents; 34% of the actual accidents was incorrectly predicted. Decision Tree showed to be slightly better than the Logistic Regression in all metrics.

The Random Forest model, which makes the decisions out of the majority votes of several Decision Trees, slightly improved the Decision Tree results. The improvement of the Recall score was only 2%, and the overall accuracy improved by 4% from 68 to 72%. However, the gradient Boost model improved all the metrics of the Decision Tree; it also outperformed the Random Forest model by far. The accuracy and Recall improved to 80% and 80%, respectively. The greatest advantage of the Gradient Boost model, thus far, was the higher capability of predicting the actual accidents; the Recall score of the Gradient Boost model was higher than other metrics, compared to the results of the Logistic Regression, Decision Tree, and the Random Forest. Regardless of the Gradient Boost model's better performance, 20% of the actual accidents were incorrectly predicted as non-accidents. Although 20% is a good performance compared to 34% from the Decision Tree and 37% from the logistic Regression, it was still too high.

The Deep Neural Network (DNN) showed superiority over previous models in different metrics. The highest accuracy achieved by that model was 92%, 12% improvement from the Gradient Boost model, and 26% from the Logistic Regression model. Moreover, the Recall score also reached 92%, 12% higher than the Gradient Boost, and 29% higher than the Logistic Regression model. Using the DNN model, only 8% of the accidents were misclassified as non-accidents, which significantly improved compared to the other investigated models. The DNN had the capacity to distinguish the accident and non-accident events, by learning the evident and hidden relationships between the explanatory variables and the accidents, better than other models. The heat-map in Figure 6.1 illustrates a comparison between the Permutation Feature Importance scores of the investigated models. The Permutation scores indicate how important each feature is when the model uses them for prediction; in other words, how each feature influences the accident prediction. The colours in the heat-map range between dark blue to white; dark blue shows higher scores, and white shows lower scores.

The Logistic Regression model, with 66% accuracy, has only found a strong correlation between traffic volume and accidents. Furthermore, although it was influenced by the hour of the day, the road class, the weather, the temperature, and the visibility, it did not find a significant correlation between accidents and the remaining features. The Decision Tree and the Random Forest models, with 68 and 72% accuracy, have captured more substantial dependencies between accidents and road class, road speed, traffic volume, the hour of the day, and road segment; at the same time, the rest of the features did not seem to contribute much to accidents in those models. However, the Gradient Boost model showed a higher capacity to find more

hidden and complex relationships. This model, with 80% accuracy, has discovered dependencies between the accidents with all features except the day and the weather condition; but has given the highest attention to the road speed, road segment, and traffic volume. Among all the investigated models, the Deep Neural Network, with 92% accuracy, was able to capture the relationship between accidents and every single explanatory variable in the dataset, as shown in Figure 6.1.

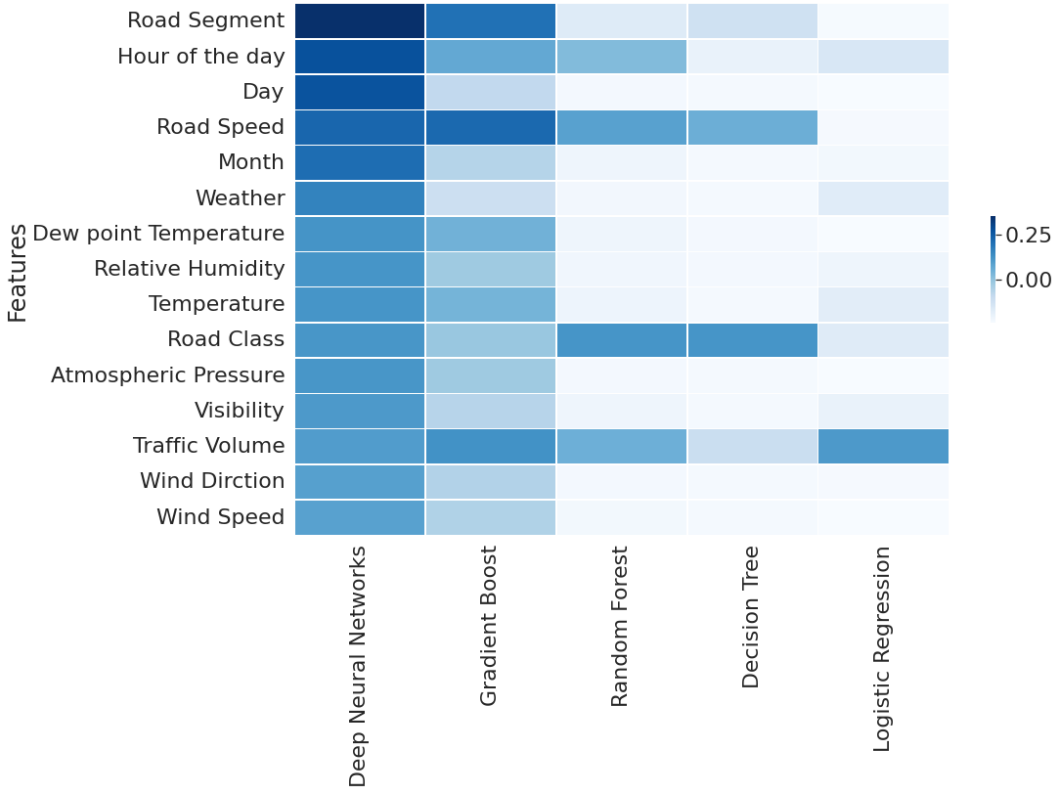


Figure 6.1: Result comparison for the Permutation feature importance of the investigated models

The red bars in Figure 6.2 illustrate the importance percentage of the explanatory variables given by the DNN model. This model has given the highest importance to the road segment, 14%. The rate of accidents on different road segments was discussed in Chapter 3, which showed that the number of accidents on different road segments does not follow the traffic volume, road class, and road speed' patterns entirely, and inherent characteristics of the road segments is also correlated with the number of accidents. Those unique inherent characteristics can be the length, width, the number of lanes, demographic conditions, sloop, curvature, the sun angle, or other external factors, which is a broad area of research.

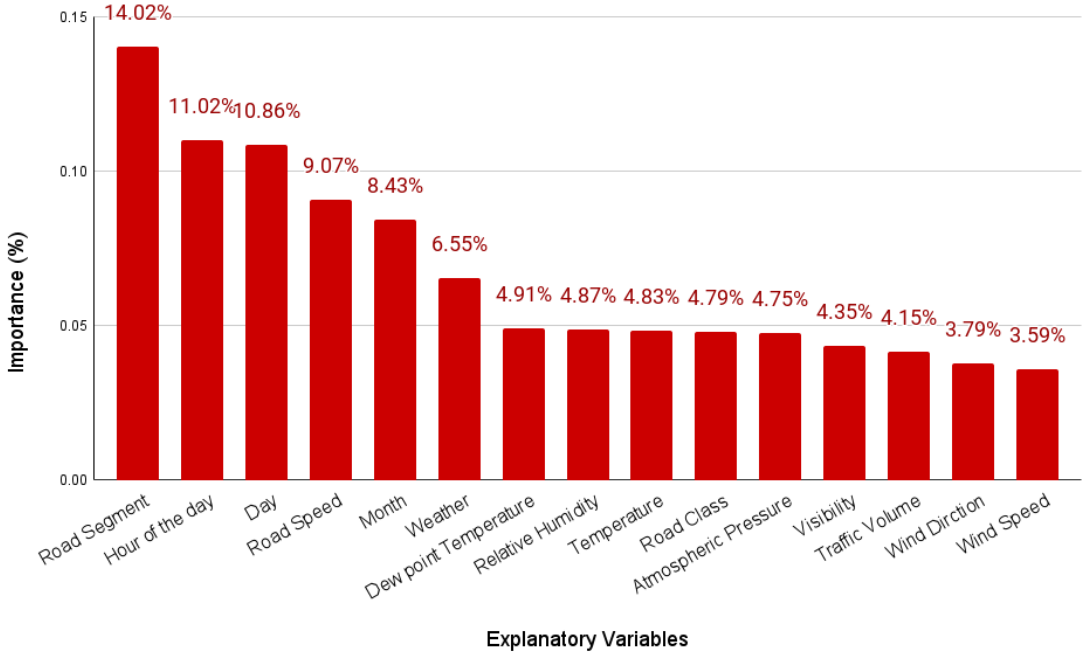


Figure 6.2: Features importance percentage of the Deep neural Networks

The second and the third important accident determinants in Calgary are the hour and the day because the traffic is different in different days and hours, with 11% and 10.86% importance, respectively. As examined in Chapter 4, on the weekdays, most of the accidents happened around rush hours; and on the weekends in the evenings; furthermore, Fridays had the highest number of accidents, and accidents were more probable on weekdays than weekends. The fourth factor, with a 9.07% effect on prediction, is the road speed, which got high scores in the Gradient Boost, Random Forest, and Decision Tree too. The month and the overall weather condition are the fifth and sixth important factors, with 8.43% and 6.55% importance; the remaining variables somehow have the same effect on accidents, between 3.59 to 4.91%.

The aggregated results of the importance of the features show the importance of the main factors that contribute to accidents in Calgary. As shown with the red bars in Figure 6.3, accidents in Calgary from 2017 to 2020 were caused, 37.66% by the weather features, 30.31% by the time factors, 18.81% by the road characteristics, 9.07% by the road speed limit, and 4.15% by the traffic volume. The weather features have shown to be the number one accident determinant in Calgary, two times more important than the road characteristics and nine times more important than the traffic volume. The results show the importance of localization in developing the Accident Prediction Models(APMs) since the area-specific features can drastically change other factors' contributions in accidents, as shown in this thesis.

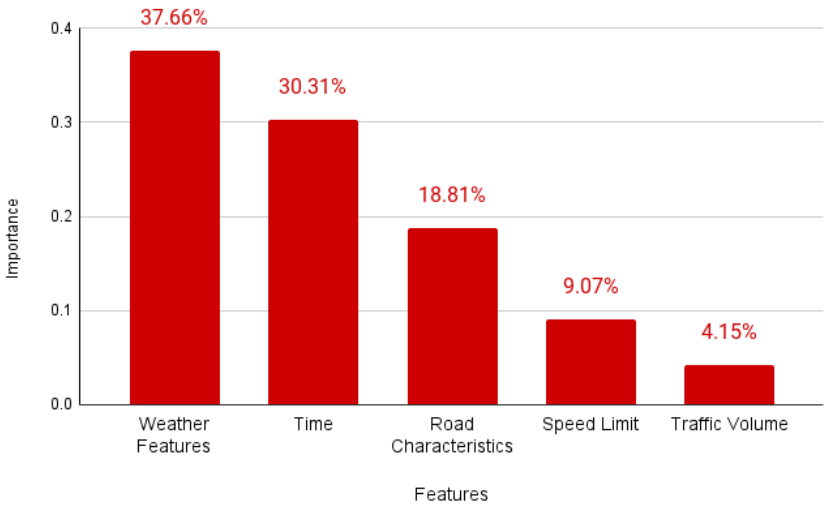


Figure 6.3: Cumulative features importance percentage of the Deep neural Networks

Chapter 7

Conclusion

7.1 Conclusion

Toward the sustainability of our future cities, provincial and local governments envision zero fatalities and serious injuries in road collisions. The first step to reaching that goal is to identify the factors that cause or affect those accidents and develop models to predict and decrease the number of accidents or alleviate their effects. Since the relationship between accidents and accident determinants can be non-linear and very complicated, especially in Calgary with harsh Winter weather, standard techniques such as Mutual Information and correlation analysis are not powerful enough to capture all those dependencies to study those relationships. However, Machine Learning and particularly Deep learning techniques have been shown to have the capabilities to capture those complex dependencies.

Thus, this thesis investigated road accidents from 2017 to 2020 in Calgary, Alberta, Canada utilizing Machine Learning and Data mining approaches in order to find the most influential accident determinants and proposed APMs to predict the number of accidents on different road segments. The analysis was conducted in three main parts:

(1) **Chapter 2** presented the existing literature on the accident prediction models and techniques.

(2) **Chapter 3** presented the preprocessing stage including accident data preparation, data augmentation, and the methodology.

(3) **Chapter 4** explored the correlation between road accident and explanatory variables in the feature space to give insights into how external factors influenced past accidents and show how complex those correlations could be. A logistic Regression model and four ML models including Decision trees, Random Forests, Gradient Boost Decision Trees, and Deep Neural Networks are developed in that chapter.

(4) **Chapter 5** presented an application for the proposed models in order to show the applicability of the thesis, an analysis of the road speed limits was carried out in that chapter, and a new speed limit was recommended for Deerfoot TR SE.

(5) **Chapter 6** discussed the results of the exploratory data analysis, as well as a comparison of the results of all the developed models .

Among all the investigated available features that correlate to accidents, weather attributes have been shown to have the highest correlation with accidents in Calgary. Accidents are correlated 37.66% by the weather features, including overall weather, temperature, relative humidity, wind speed, wind direction, Dow point temperature, pressure, and visibility. For example, driving on a surface with ice crystals has two times more likelihood of being involved in an accident and driving on a snowy day increases the probability of accidents by 55%, compared to clear or cloudy days.

The time and day of the week are also correlated to the likelihood of being involved in an accident, as the analysis shows accidents are correlated 30.31% by the time the commuters head the road.

After the model development stage, a Deep Neural Network with four hidden layers and weighted classes outperformed other investigated models such as Logistic Regression, Decision Trees, Random Forest, and Gradient Boost, with 92% accuracy.

After finding the effect of the accident determinants in Calgary, an analysis was conducted in Chapter 5 in order to find a way to reduce the number of accidents in the short term. Except for the road speed limit, the other factors seemed complicated to change to reduce accidents; for example, it is nearly impossible to force commuters to stay home in rush hours or force them not to travel through the city when snowy. Improving the infrastructure and factors that affect the traffic volume was also seemed to be out of the scope of the study. Thus the analysis was conducted on the effect of changing the road speed limit.

As the result of the above speed limit analysis, a ten km/h reduction is recommended on Deerfoot TR SE, resulting in a 5% reduction in the number of predicted accidents on that road segment with maintaining the service level. The recommended speed limit reduction will save $765,350 yearly in direct costs, human capital, and willingness to pay. In addition, the environmental effect of the recommended speed limit reduction is promising; the daily CO2 reduction as the result of the speed limit reduction equates to the amount of CO2 emitted by 10 Canadians in their households per year.

The proposed DNN model showed how Machine Learning techniques could be used to develop APMs in the city of Calgary; the same model can be used in other areas and cities in Canada with slightly different weather conditions; however, the model needs to be updated and trained with new data from those cities to be able to predict accident in the target areas.

In conclusion, the main contributions of this thesis are:

(1) Firstly, multi-source data, including road characteristics, collision data, and historical climate data are brought together to enhance the set of datasets available for traffic accident prediction modelling.

(2) Secondly, exploratory data Analysis is conducted focusing on Canadian weather features to give insight into how different external factors such as weather attributes are correlated to road accidents based on the available data in Calgary, Alberta, Canada.

(3) Lastly, Accident Prediction Models are developed utilizing state-of-the-art Machine Learning techniques to identify the collision hot spots and the influential factors to inform policy on road traffic safety in Calgary, Alberta, Canada.

7.2 Recommendation for Future Works

This thesis utilized the capabilities of the Deep Neural Networks to show how significant the effect of weather attributes could be on the occurrence and prediction of road traffic collisions. Thus, the same methodology could be used in future works to examine and find the influential factors on individual road segments in order to recommend specific solutions to individual road segments for accident reduction.

Below is a shortlist of future work recommendations:

- Using the LSTM-RNN technique to evaluate and predict the accidents on individual road segments by assuming the accidents as a sequence of events such as time series.
- Dividing the road segments into smaller sections and developing a predictive model to predict the secondary accidents based on the smaller sections.
- Predicting the accidents on intersections and mid-block road segments separately.
- Using the same techniques and models to predict the severity of the accidents in Calgary concerning the weather attributes.

Bibliography

[1] World Health Organization. Road traffic injuries. https://www.who.int/news-room/fact-sheets/detail/road-traffic-injuries, 2021. Accessed: 2022-03-9.

[2] Shaw-Pin Miaou. The relationship between truck accidents and geometric design of road sections: Poisson versus negative binomial regressions. *Accident Analysis & Prevention*, 26(4):471–482, 1994.

[3] Stijn Daniels, Tom Brijs, Erik Nuyts, and Geert Wets. Explaining variation in safety performance of roundabouts. *Accident Analysis & Prevention*, 42(2):393–402, March 2010.

[4] Jonathan Aguero-Valverde and Paul P Jovanis. Analysis of road crash frequency with spatial models. *Transportation Research Record*, 2061(1):55–63, 2008.

[5] Miao Chong, Ajith Abraham, and Marcin Paprzycki. Traffic Accident Analysis Using Machine Learning Paradigms. page 10.

[6] Hassan T Abdelwahab and Mohamed A Abdel-Aty. Artificial neural networks and logit models for traffic safety analysis of toll plazas. *Transportation Research Record*, 1784(1):115–125, 2002. Publisher: SAGE Publications Sage CA: Los Angeles, CA.

[7] Li-Yen Chang. Analysis of freeway accident frequencies: Negative binomial regression versus artificial neural network. *Safety Science*, 43(8):541–557, October 2005.

[8] The top 10 coldest countries in the world. https://myfunkytravel.com/coldest-countries-in-the-world.html. Accessed: 2022-03-9.

[9] Camilo Gutierrez-Osorio and César Pedraza. Modern data sources and techniques for analysis and forecast of road accidents: A review. *Journal of traffic and transportation engineering (English edition)*, 7(4):432–446, 2020.

[10] Jonathan J. Rolison, Shirley Regev, Salissou Moutari, and Aidan Feeney. What are the factors that contribute to road accidents? An assessment of law enforcement views, ordinary drivers' opinions, and road accident records. *Accident Analysis & Prevention*, 115:11–24, June 2018.

[11] Mobility and transport European Commission. Factors related to the vehicle. https://ec.europa.eu/transport/road safety/specialist/knowledge/powered two wheelers/contributory factors/factors related to the vehicle en, October 2016. Accessed: 2021-07-24.

[12] Hafiz Mohkum Hammad, Muhammad Ashraf, Farhat Abbas, Hafiz Faiq Bakhat, Saeed A. Qaisrani, Muhammad Mubeen, Shah Fahad, and Muhammad Awais. Environmental factors affecting the frequency of road traffic accidents: a case study of sub-urban area of Pakistan. *Environmental Science and Pollution Research*, 26(12):11674–11685, April 2019.

[13] Lasmini Ambarwati and Amelia Indriastuti. Optimization of safe pedestrian facilities and traffic management, a case study of Malang, Indonesia. *Journal of Economics and Engineering*, 4:8–14, January 2010.

[14] City of Calgary. City of calgary's open data portal. https://data.calgary.ca/stories/s/About/mgd4-4snb, 2021. Accessed: 2022-03-9.

[15] Environment and Climate Change Canada. Historical Climate Data - Climate - Environment and Climate Change Canada. https://climate.weather.gc.ca/index-e.html, October 2011. Last Modified: 2021-03-25.

[16] George Yannis, Anastasios Dragomanovits, Alexandra Laiou, Thomas Richter, Stephan Ruhl, Francesca La Torre, Lorenzo Domenichini, Daniel Graham, Niovi Karathodorou, and Haojie Li. Use of accident prediction models in road safety management–an international inquiry. *Transportation research procedia*, 14:4257–4266, 2016.

[17] Frank Gross, Bhagwant Naraine Persaud, and Craig Lyon. A guide to developing quality crash modification factors. Technical report, United States. Federal Highway Administration. Office of Safety, 2010.

[18] F Gross and A Hamidi. Investigation of existing and alternative methods for combining multiple cmfs. *Highway Safety Improvement Program Technical Support. Task A*, 9, 2011.

[19] Ali Farhan, Lina Kattan, and Richard Tay. Collisions on local roads: model development and policy level scenario analysis. *Canadian Journal of Civil Engineering*, 47(1):77–87, 2020.

[20] Poul Greibe. Accident prediction models for urban roads. *Accident Analysis & Prevention*, 35(2):273–285, 2003.

[21] Alfonso Montella, Lucio Colantuoni, and Renato Lamberti. Crash prediction models for rural motorways. *Transportation Research Record*, 2083(1):180–189, 2008.

[22] Mohamed Abdel-Aty, Nizam Uddin, Anurag Pande, M Fathy Abdalla, and Liang Hsia. Predicting freeway crashes from loop detector data by matched case-control logistic regression. *Transportation Research Record*, 1897(1):88–95, 2004.

[23] Tao Lu, Zhu Dunyao, Yan Lixin, and Zhang Pan. The traffic accident hotspot prediction: Based on the logistic regression method. In *2015 International Conference on Transportation Information and Safety (ICTIS)*, pages 107–110. IEEE, 2015.

[24] Darçin Akin and Bülent Akba. A neural network (NN) model to predict intersection crashes based upon driver, vehicle and roadway surface characteristics. *Sci. Res. Essays*, page 11, 2010.

[25] Baher Abdulhai, Stephen G Ritchie, and Mahadevan Iyer. Implementation of advanced techniques for automated freeway incident detection. 1999.

[26] Yuejing Lv, Haixia Zhang, Xing-lin Zhou, Ming Liu, and Jie Li. Research on accident prediction of intersection and identification method of prominent accident form based on back propagation neural network. In *2010 International Conference on Computer Application and System Modeling (ICCASM 2010)*, volume 1, pages V1–434–V1–438, October 2010. ISSN: 2161-9077.

[27] F Rezaie Moghaddam, Sh Afandizadeh, and M Ziyadi. Prediction of accident severity using artificial neural networks. *International Journal of Civil Engineering*, 9(1):9, 2011.

[28] Yisheng Lv, Shuming Tang, Hongxia Zhao, and Shuang Li. Real-time highway accident prediction based on support vector machines. In *2009 Chinese Control and Decision Conference*, pages 4403–4407, Guilin, China, June 2009. IEEE.

[29] Rongjie Yu and Mohamed Abdel-Aty. Utilizing support vector machine in real-time crash risk evaluation. *Accident Analysis & Prevention*, 51:252–259, March 2013.

[30] Tatiana Tambouratzis, Dora Souliou, Miltiadis Chalikias, and Andreas Gregoriades. Combining probabilistic neural networks and decision trees for maximally accurate and efficient accident prediction. In *The 2010 International Joint Conference on Neural Networks (IJCNN)*, pages 1–8, July 2010. ISSN: 2161-4407.

[31] Qi Shi and Mohamed Abdel-Aty. Big data applications in real-time traffic operation and safety monitoring and improvement on urban expressways. *Transportation Research Part C: Emerging Technologies*, 58:380–394, 2015.

[32] Hany M Hassan and Mohamed A Abdel-Aty. Predicting reduced visibility related crashes on freeways using real-time traffic flow data. *Journal of safety research*, 45:29–36, 2013.

[33] Salahadin Seid Yassin and Pooja. Road accident prediction and model interpretation using a hybrid K-means and random forest algorithm approach. *SN Applied Sciences*, 2(9):1576, August 2020.

[34] Lu Wenqi, Luo Dongyu, and Yan Menghua. A model of traffic accident prediction based on convolutional neural network. In *2017 2nd IEEE International Conference on Intelligent Transportation Engineering (ICITE)*, pages 198–202. IEEE, 2017.

[35] Amani Thaduri, Vijayakumar Polepally, and Swathy Vodithala. Traffic Accident Prediction based on CNN Model. In *2021 5th International Conference on Intelligent Computing and Control Systems (ICICCS)*, pages 1590–1594, May 2021.

[36] Le Yu, Bowen Du, Xiao Hu, Leilei Sun, Liangzhe Han, and Weifeng Lv. Deep spatio-temporal graph convolutional network for traffic accident prediction. *Neurocomputing*, 423:135–147, January 2021.

[37] Da-Jie Lin, Mu-Yen Chen, Hsiu-Sen Chiang, and Pradip Kumar Sharma. Intelligent Traffic Accident Prediction Model for Internet of Vehicles With Deep Learning Approach. *IEEE Transactions on Intelligent Transportation Systems*, pages 1–10, 2021. Conference Name: IEEE Transactions on Intelligent Transportation Systems.

[38] Fahim Ahmed Shakil, Sayed Muddashir Hossain, Rifat Hossain, and Sifat Momen. Prediction of road accidents using data mining techniques. In *Proceedings of International Conference on Computational Intelligence and Emerging Power System*, pages 25–35. Springer, 2022.

[39] Pei Li, Mohamed Abdel-Aty, and Jinghui Yuan. Real-time crash risk prediction on arterials based on LSTM-CNN. *Accident Analysis & Prevention*, 135:105371, February 2020.

[40] Yisheng Lv, Shuming Tang, and Hongxia Zhao. Real-time highway traffic accident prediction based on the k-nearest neighbor method. In *2009 international conference on measuring technology and mechatronics automation*, volume 3, pages 547–550. IEEE, 2009.

[41] Lei Lin, Qian Wang, and Adel W. Sadek. A novel variable selection method based on frequent pattern tree for real-time traffic accident risk prediction. *Transportation Research Part C: Emerging Technologies*, 55:444–459, June 2015.

[42] Zekun Yang, Wenping Zhang, and Juan Feng. Predicting multiple types of traffic accident severity with explanations: A multi-task deep learning framework. *Safety science*, 146:105522, 2022.

[43] Jae Hun Kim, Juyeon Kim, Gunwoo Lee, and Juneyoung Park. Machine Learning-Based Models for Accident Prediction at a Korean Container Port. *Sustainability*, 13(16):9137, January 2021. Number: 16 Publisher: Multidisciplinary Digital Publishing Institute.

[44] Chao Chen, Xiaoliang Fan, Chuanpan Zheng, Lujing Xiao, Ming Cheng, and Cheng Wang. Sdcae: Stack denoising convolutional autoencoder model for accident risk prediction via traffic big data. In *2018 Sixth International Conference on Advanced Cloud and Big Data (CBD)*, pages 328–333. IEEE, 2018.

[45] Murat Ozbayoglu, Gokhan Kucukayan, and Erdogan Dogdu. A real-time autonomous highway accident detection model based on big data processing and computational intelligence. In *2016 IEEE International Conference on Big Data (Big Data)*, pages 1807–1813, December 2016.

[46] Sahraei* MA and Tortum A. Çodur MK, Çodur MY. *Forecasting the Accident Frequency and Risk Factors: A Case Study of Erzurum*. Turkey, 2022.

[47] Yiming Gu, Zhen Sean Qian, and Feng Chen. From twitter to detector: Real-time traffic incident detection using social media data. *Transportation research part C: emerging technologies*, 67:321–342, 2016.

[48] Eleonora D'Andrea, Pietro Ducange, Beatrice Lazzerini, and Francesco Marcelloni. Real-time detection of traffic from twitter stream analysis. *IEEE transactions on intelligent transportation systems*, 16(4):2269–2283, 2015.

[49] Chengcheng Xu, Wei Wang, and Pan Liu. A Genetic Programming Model for Real-Time Crash Prediction on Freeways. *IEEE Transactions on Intelligent Transportation Systems*, 14(2):574–586, June 2013. Conference Name: IEEE Transactions on Intelligent Transportation Systems.

[50] Pantelis Kopelias, Fanis Papadimitriou, Konstantinos Papandreou, and Panos Prevedouros. Urban freeway crash analysis: geometric, operational, and weather effects on crash number and severity. *Transportation Research Record*, 2015(1):123–131, 2007.

[51] Miao M. Chong, Ajith Abraham, and Marcin Paprzycki. Traffic Accident Analysis Using Decision Trees and Neural Networks. *arXiv:cs/0405050*, May 2004. arXiv: cs/0405050.

[52] Jinming You, Junhua Wang, and Jingqiu Guo. Real-time crash prediction on freeways using data mining and emerging techniques. *Journal of Modern Transportation*, 25(2):116–123, June 2017.

[53] Francesca La Torre, Monica Meocci, Lorenzo Domenichini, Valentina Branzi, Andrea Paliotto, et al. Development of an accident prediction model for italian freeways. *Accident Analysis & Prevention*, 124:1–11, 2019.

[54] Madhar Taamneh, Sharaf Alkheder, and Salah Taamneh. Data-mining techniques for traffic accident modeling and prediction in the united arab emirates. *Journal of Transportation Safety & Security*, 9(2):146–166, 2017.

[55] Sachin Kumar and Durga Toshniwal. A data mining approach to characterize road accident locations. *Journal of Modern Transportation*, 24(1):62–72, 2016.

[56] Quanjun Chen, Xuan Song, Harutoshi Yamada, and Ryosuke Shibasaki. Learning Deep Representation from Big and Heterogeneous Data for Traffic Accident Inference. page 7.

[57] Chao Chen, Xiaoliang Fan, Chuanpan Zheng, Lujing Xiao, Ming Cheng, and Cheng Wang. SDCAE: Stack Denoising Convolutional Autoencoder Model for Accident Risk Prediction Via Traffic Big Data. In *2018 Sixth International Conference on Advanced Cloud and Big Data (CBD)*, pages 328–333, August 2018.

[58] Lu Wenqi, Luo Dongyu, and Yan Menghua. A model of traffic accident prediction based on convolutional neural network. In *2017 2nd IEEE International Conference on Intelligent Transportation Engineering (ICITE)*, pages 198–202, September 2017.

[59] Neda Kamboozia, Mahmoud Ameri, and Seyed Mohsen Hosseinian. Statistical analysis and accident prediction models leading to pedestrian injuries and deaths on rural roads in Iran. *International Journal of Injury Control and Safety Promotion*, 27(4):493–509, October 2020. Publisher: Taylor & Francis _eprint: https://doi.org/10.1080/17457300.2020.1812670.

[60] Fang Zong, Hongguo Xu, and Huiyong Zhang. Prediction for traffic accident severity: comparing the bayesian network and regression models. *Mathematical Problems in Engineering*, 2013, 2013.

[61] Chukwutoo C Ihueze and Uchendu O Onwurah. Road traffic accidents prediction modelling: An analysis of anambra state, nigeria. *Accident Analysis & Prevention*, 112:21–29, 2018.

[62] Ming Zheng, Tong Li, Rui Zhu, Jing Chen, Zifei Ma, Mingjing Tang, Zhongqiang Cui, and Zhan Wang. Traffic Accident's Severity Prediction: A Deep-Learning Approach-Based CNN Network. *IEEE Access*, 7:39897–39910, 2019. Conference Name: IEEE Access.

[63] Canada's historical climate data. https://climate.weather.gc.ca/. Accessed: 2022-01-24.

[64] Nitesh V Chawla, Kevin W Bowyer, Lawrence O Hall, and W Philip Kegelmeyer. Smote: synthetic minority over-sampling technique. *Journal of artificial intelligence research*, 16:321–357, 2002.

[65] Shivani Tyagi and Sangeeta Mittal. Sampling Approaches for Imbalanced Data Classification Problem in Machine Learning. In Pradeep Kumar Singh, Arpan Kumar Kar, Yashwant Singh, Maheshkumar H. Kolekar, and Sudeep Tanwar, editors, *Proceedings of ICRIC 2019*, Lecture Notes in Electrical Engineering, pages 209–221, Cham, 2020. Springer International Publishing.

[66] RandomUnderSampler — Version 0.8.1. https://imbalanced-learn.org. Accessed: 2022-01-9.

[67] Inderjeet Mani and I Zhang. kNN approach to unbalanced data distributions: a case study involving information extraction. In *Proceedings of workshop on learning from imbalanced datasets*, volume 126. ICML United States, 2003.

[68] Dennis L. Wilson. Asymptotic Properties of Nearest Neighbor Rules Using Edited Data. *IEEE Transactions on Systems, Man, and Cybernetics*, SMC-2(3):408–421, 1972.

[69] Ivan Tomek and others. AN EXPERIMENT WITH THE EDITED NEAREST-NIEGHBOR RULE. 1976.

[70] An Experiment with the Edited Nearest-Neighbor Rule. *IEEE Transactions on Systems, Man, and Cybernetics*, SMC-6(6):448–452, 1976.

[71] Giovanna Menardi and Nicola Torelli. Training and assessing classification rules with imbalanced data. *Data Mining and Knowledge Discovery*, 28(1):92–122, January 2014.

[72] N. V. Chawla, K. W. Bowyer, L. O. Hall, and W. P. Kegelmeyer. SMOTE: Synthetic Minority Oversampling Technique. *Journal of Artificial Intelligence Research*, 16:321–357, June 2002.

[73] Haibo He, Yang Bai, Edwardo A. Garcia, and Shutao Li. ADASYN: Adaptive synthetic sampling approach for imbalanced learning. In *2008 IEEE International Joint Conference on Neural Networks (IEEE World Congress on Computational Intelligence)*, pages 1322–1328, June 2008. ISSN: 2161-4407.

[74] Felix Last, Georgios Douzas, and Fernando Bacao. Oversampling for Imbalanced Learning Based on K-Means and SMOTE. *Information Sciences*, 465:1–20, October 2018. arXiv: 1711.00837.

[75] Hien M. Nguyen, Eric W. Cooper, and Katsuari Kamei. Borderline over-sampling for imbalanced data classification. *International Journal of Knowledge Engineering and Soft Data Paradigms*, 3(1):4–21, January 2011. Publisher: Inderscience Publishers.

[76] Erik G Learned-Miller. Entropy and Mutual Information. *Department of Computer Science, University of Massachusetts, Amherst*, page 4, 2013.

[77] Michael E. Tipping and Christopher M. Bishop. Probabilistic Principal Component Analysis. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 61(3):611–622, 1999. _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1111/1467-9868.00196.

[78] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830, 2011.

[79] Anna Veronika Dorogush, Vasily Ershov, and Andrey Gulin. Catboost: gradient boosting with categorical features support. *arXiv preprint arXiv:1810.11363*, 2018.

[80] Martín Abadi, Ashish Agarwal, Paul Barham, Eugene Brevdo, Zhifeng Chen, Craig Citro, Greg S. Corrado, Andy Davis, Jeffrey Dean, Matthieu Devin, Sanjay Ghemawat, Ian Goodfellow, Andrew Harp, Geoffrey Irving, Michael Isard, Yangqing Jia, Rafal Jozefowicz, Lukasz Kaiser, Manjunath Kudlur, Josh Levenberg, Dandelion Mané, Rajat Monga, Sherry Moore, Derek Murray, Chris Olah, Mike Schuster, Jonathon Shlens, Benoit Steiner, Ilya Sutskever, Kunal Talwar, Paul Tucker, Vincent Vanhoucke, Vijay Vasudevan, Fernanda Viégas, Oriol Vinyals, Pete Warden, Martin Wattenberg, Martin Wicke, Yuan Yu, and Xiaoqiang Zheng. TensorFlow: Large-scale machine learning on heterogeneous systems, 2015. Software available from tensorflow.org.

[81] M Madhiarasan and SN Deepa. A novel criterion to select hidden neuron numbers in improved back propagation networks for wind speed forecasting. *Applied intelligence*, 44(4):878–893, 2016.

[82] M Madhiarasan and SN Deepa. Comparative analysis on hidden neurons estimation in multi layer perceptron neural networks for wind speed forecasting. *Artificial Intelligence Review*, 48(4):449–471, 2017.

[83] Basheer Qolomany, Majdi Maabreh, Ala Al-Fuqaha, Ajay Gupta, and Driss Benhaddou. Parameters optimization of deep learning models using particle swarm optimization. In *2017 13th International Wireless Communications and Mobile Computing Conference (IWCMC)*, pages 1285–1290. IEEE, 2017.

[84] Namig J Guliyev and Vugar E Ismailov. On the approximation by single hidden layer feedforward neural networks with fixed weights. *Neural Networks*, 98:296–304, 2018.

[85] Jeff Heaton. *Introduction to neural networks with Java*. Heaton Research, Inc., 2008.

[86] WGM Vanlaar, H Woods-Fry, H Barrett, C Lyon, S Brown, C Wicklund, and RD Robertson. The impact of covid-19 on road safety in canada and the united states. *Accident Analysis & Prevention*, 160:106324, 2021.

[87] Tirf reports on the impact of the covid-19 pandemic on travel behaviour road safety. https://tirf.ca/news/tirf-reports-on-the-impact-of-the-covid-19-pandemic-on-travel-behaviour-road-safety/. Accessed: 2022-02-07.

[88] The top 10 coldest countries in the world. https://www.calgary.ca/transportation/tp/planning/transportation-planning-studies/road-classification.html. Accessed: 2022-03-9.

[89] Leo Breiman. Random forests. *Machine learning*, 45(1):5–32, 2001.

[90] David Navon. The paradox of driving speed: two adverse effects on highway accident rate. *Accident Analysis & Prevention*, 35(3):361–367, May 2003.

[91] Alyssa Ditcharoen, Bunna Chhour, Tunyarat Traikunwaranon, Nalin Aphivongpanya, Kunanon Maneerat, and Veeris Ammarapala. Road traffic accidents severity factors: A review paper. In *2018 5th International Conference on Business and Industrial Research (ICBIR)*, pages 339–343, May 2018.

[92] Göran Nilsson. Traffic Safety Dimensions and the Power Model to Describe the Effect of Speed on Safety. page 121, 2004.

[93] Calgary Ring Road. https://www.alberta.ca/calgary-ring-road.aspx. Accessed: 2021-11-11.

[94] Wachs Martin and M. Samuels John. Highway Capacity Manual 2000, October 2000.

[95] Elżbieta Macioszek. The passenger car equivalent factors for heavy vehicles on turbo roundabouts. *Frontiers in built environment*, 5:68, 2019.

[96] PB Farradyne. Traffic Incident Management Handbook. page 108, November 2000.

[97] Collision costs study. https://drivetolive.ca/research/collision-cost-study/. Accessed: 2022-02-14.

[98] Alberta traffic collision statistics. https://open.alberta.ca/dataset/25020446-adfb-4b57-9aaa-751d13dab72d/resource/982e6d4f-64d5-4167-81ca-b8c10d76fa59/download/trans-alberta-traffic-collision-statistics-2018.pdf. Accessed: 2022-02-14.