

Czech University of Life Sciences, Prague

Faculty of Economics and Management

Department of Statistics



Diploma Thesis on

**Big Data Analysis for predicting Road Traffic Accident
Severity**

Prepared by: Ametu Negash Redi

Thesis supervisor: Ing. Tomáš Hlavsa, Ph.D

© 2021 CULS Prague

CZECH UNIVERSITY OF LIFE SCIENCES PRAGUE

Faculty of Economics and Management

DIPLOMA THESIS ASSIGNMENT

B.Sc. Ametu Negash Redi

Systems Engineering and Informatics
Informatics

Thesis title

Big Data Analysis for Predicting Road Traffic Accident Severity

Objectives of thesis

The main objective of this thesis is to develop detailed analysis for detecting the viable models to predict the occurrence and the level of traffic accidents severity.

Having this as a primary objective, there are also specific goals that the study aims to achieve such as:

- I. Defining the determinant factors of traffic accident.
- II. Understand the relations of the dependent and independent variables of traffic accident
- III. Identifying the possible suitable model in predicting the severity of traffic Accident

Methodology

To achieve the objectives, the development of this thesis involves a literature review of Big Data, of its characteristics and the comparisons of different statistical models used in predictive analytics. The practical part will deploy the methods i.e. Decision tree, Linear Regression, and Neural network. To identify the best predictive model, the following characteristics such as Root Mean Square error (RMSE), receiver operating characteristics (ROC) curve or confusion matrix will be checked.

The proposed extent of the thesis

60 – 80 pages

Keywords

Big Data, Predictive Modelling, exploratory data analysis, linear regression, decision tree, neural network, traffic accident.

Recommended information sources

- ABBOTT, D. Applied Predictive Analytics : Principles and Techniques for the Professional Data Analyst. Praha: John Wiley & Sons, Incorporated, 2014. ISBN 9781118727935.
- CUKIER, K. Data, data everywhere: A special report on managing information, 2010. The Economist <http://www.economist.com/node/15557443>, Accessed on June 15, 2015.
- EUROPA, U Road safety statistics: What is behind the figures, 2017. [Online]. Available: http://europa.eu/rapid/press-release_MEMO-17-675_en.htm. [Accessed: 20 – Apr- 2017].
- LARSON, R. – FARBER, E. *Elementary statistics : picturing the world*. Boston: Pearson Prentice Hall, 2015. ISBN 9780321693624.
- ROBSON, L. F. C & CHRISTOS, F. & CAETANO, T. Jr. Data Mining in Large Sets of Complex Data: Data Mining and Knowledge Discovery. Springer-Verlag London, [e-book], Springer, 2013, ISBN 978-1-4471-4890-6.
- SUMATHI, S.– SIVANANDAM, S.N. Introduction to data mining and its applications. Springer. 2006. ISBN 978-3-540-34350-9.
- TUFFERY, S. Data Mining and Statistics for Decision Making. UK, West Sussex: Wiley, 2011. ISBN 978-0-470-68829-8.

Expected date of thesis defence

2020/21 SS – FEM

The Diploma Thesis Supervisor

Ing. Tomáš Hlavsa, Ph.D.

Supervising department

Department of Statistics

Electronic approval: 23. 11. 2020

prof. Ing. Libuše Svatošová, CSc.

Head of department

Electronic approval: 24. 11. 2020

Ing. Martin Pelikán, Ph.D.

Dean

Prague on 17. 03. 2021

Declaration

I declare that I have worked on my diploma thesis titled " Big Data Analysis for predicting Road Traffic Accident Severity" by myself and I have used only the sources mentioned at the end of the thesis. As the author of the diploma thesis, I declare that the thesis does not break copyrights of any other person.

In Prague on 31 March 2021

A handwritten signature in blue ink, consisting of several overlapping loops and lines, positioned above a horizontal dotted line.

Ametu Negash Redi

Acknowledgement

I would like to express my sincere gratitude to my Supervisor Ing. Tomáš Hlavsa, Ph.D for the constructive advices and for being very collaborative throughout the process of preparing this diploma thesis. I would like to thank my parents and my brother Dr. Sakin for always supporting and believing in me. Finally, I would like to thank my friends for all the unconditional supports.

Big Data Analysis for predicting Road Traffic Accident Severity

Abstract

With the recent growth of urban transportation, Frequent road traffic accident and other issues must be addressed. Understanding the causes of traffic accident and developing an early alarm model for drivers and pedestrians would be critical in resolving traffic accident issues. To achieve the end goal and scopes, this study relies on Big Data as a primary source. It exhibits typical characteristics and issues that are arisen when dealing with Big Data. The focus of the thesis is to develop an in depth analysis of traffic accident severity and to determine which of the predictive modeling used is the best one for predicting occurrence and the level of accident severity as target variable. The three models are selected for practical part i.e., Logistic Regression, Decision tree, and Neural network and to identify the best predictive model, models are tested using different performance measurement parameters Root Mean Square error (RMSE), receiver operating characteristics (ROC). Based on the model comparison in SAS model studio, the test results show that the decision tree seemed to perform better than logistic regression and neural network models. This research study shows that the algorithm can predict accidents with 87% accuracy. The dataset was taken from the UK traffic accident from 2005 to 2015 retrieved from Kaggle source. The aim of analyzing this dataset was for academic purposes only, and the data would not be used for any other purpose. It includes 1,048,575 records of accidents with various information about traffic accident.

Keywords: Traffic accident, Big data, predictive modeling, Decision tree, Logistic Regression and Neural network.

Analýza velkých dat pro předpovídání závažnosti dopravních nehod

Abstraktní

S nedávným růstem městské dopravy je třeba řešit časté dopravní nehody a další problémy. Pochopení příčin dopravních nehod a vývoj modelu včasného varování pro řidiče a chodce by bylo rozhodující při řešení problémů dopravních nehod. K dosažení konečného cíle a rozsahů se tato studie opírá o Big Data jako primární zdroj. Vykazuje typické vlastnosti a problémy, které vyvstávají při práci s Big Data. Cílem této práce je vyvinout hloubkovou analýzu závažnosti dopravních nehod a určit, které z použitých prediktivních modelů je nejlepší pro predikci výskytu a úrovně závažnosti nehod jako cílové proměnné. Tyto tři modely jsou vybrány pro praktickou část, tj. Logistická regrese, Rozhodovací strom a Neuronová síť a pro identifikaci nejlepšího prediktivního modelu jsou modely testovány s použitím různých parametrů měření výkonu Kořenová střední chyba kvadratury (RMSE), provozní charakteristiky přijímače (ROC). Na základě porovnání modelů v SAS model studio, výsledky testů ukazují, že rozhodovací strom fungoval lépe než logistická regrese a modely neuronových sítí. Tato výzkumná studie ukazuje, že algoritmus dokáže předpovědět nehody s přesností 87%. Soubor dat byl převzat z dopravní nehody ve Velké Británii z let 2005 až 2015 a získán ze zdroje Kaggle. Cílem analýzy této datové sady bylo pouze pro akademické účely a data by nebyla použita k žádnému jinému účelu. Zahrnuje 1 048 575 záznamů o nehodách s různými informacemi o dopravní nehodě.

Klíčová slova: Dopravní nehoda, Velká data, prediktivní modelování, Rozhodovací strom, Logistická regrese, Neuronová síť.

Contents

1. INTRODUCTION.....	1
2. Objectives and Methodology.....	3
2.1 Objectives	3
2.2 Methodology	3
3. Literature Review.....	5
3.1 Big data	5
3.1.1 Characteristics of Big Data.....	5
3.1.2 Statistical Methods for Big Data Analysis.....	7
3.1.3 What are the challenges of Big Data analysis?.....	8
3.2 Data Mining Techniques and Modelling	10
3.3 What differentiates predictive analytics from other types of analytics?	11
3.4 Performance Indicator for practically applied Methodologies	21
3.5 What is Roads Traffic accident and its burden	22
3.5.1 Factors affecting road traffic accident.....	24
3.6 Related works	25
4. Practical Part.....	31
4.1 Overview of the Case Study	31
4.1.1 Problem framing.....	32
4.1.2 Preparing the data.....	32
4.2 Explanatory Data Analysis	36
4.2.1 Univariate Statistical Analysis.....	36
4.2.2 Bivariate statistical analysis.....	50
4.2.3 Correlation between the dependent variable and the independent variable.....	55
4.3 Model building and training	59
4.3.1 Custom transformation of the data.....	59
5. Results and Discussions.....	61
5.1 Model building and evaluation	61
6. Conclusion.....	72
7. Reference.....	74
Appendix 1. value coding for all selected variables.....	79

List of Figures

Figure 1: The V's of Big Data [source: Data science: Surya Gutta].....	7
Figure 2: Predictive Analytics Process [source: International Journal of Computer Applications]	8
Figure 3: Conceptual classification of Big Data challenges [source: adapted from (Akerkar 2014) & (Zicari 2014)].....	10
Figure 4: Regression line Data [source: Applied Predictive Analytics: Dean Abbott].....	12
Figure 5: biological neuron versus artificial neural network: (a) human neuron; (b) artificial neuron; (c) biological synapse; and (d) ANN synapses. [source: Combining Finite Element Method and Data-Driven: Zhenzhu Meng, 2020].....	15
Figure 6: The diagrammatic illustration of activation functions used in ANN. [source: Statistics for Machine Learning: Dangeti 244, 2017).	16
Figure 7: Example of simple decision tree (Sumathi and Sivanandam, 2006).	18
Figure 8: Distribution of dependent variable [source: own]	36
Figure 9: Distribution of Age_Band_of_Driver [source: own].....	38
Figure 10: Distribution of Day of the week [source: own]	39
Figure 11: Distribution of Junction Detail [source: own]	40
Figure 12: Distribution of Junction Control [source: own].....	41
Figure 13: Distribution of Light Conditions [source: own]	42
Figure 14: Distribution of Road Type [source: own].....	43
Figure 15: Distribution of Sex_of_Driver [source: own].....	44
Figure 16: Distribution of Speed limit category [source: own]	45
Figure 17: Distribution of Vehicle_Manoevre [source: own].....	46
Figure 18: Distribution of Vehicle_Type [source: own].....	47
Figure 19: Distribution of _1st_Point_of_Impact [source: own].....	48
Figure 20: Distribution of No of Casualty [source: own]	49
Figure 21: Distribution of Number of Vehicles Involved [source: own].....	50
Figure 22: Accident severity and Sex_of_Driver [source: own].....	51
Figure 23: Accident severity and Age_Band_of_Driver [source: own]	52
Figure 24: Accident severity and Speed limit Category [source: own]	53
Figure 25: Accident severity and Light_Conditions [source: own]	54
Figure 26: Accident severity and Road_Type [source: own].....	55
Figure 27: SAS Model studio Pipeline [source: own]	61
Figure 28: t Values by Parameter [source: own].....	63
Figure 29: Accuracy Reports for Logistics regression [source: own].....	64
Figure 30: ROC Reports for Logistics regression [source: own].....	64
Figure 31: Neural Network Diagram for Accident Severity [source: own]	65
Figure 32: ROC report for Neural Network [source: own].....	66
Figure 33: Accuracy report for Neural Network [source: own].....	66
Figure 34: Decision Tree Diagram for Accident Severity [source: own]	67
Figure 35: ROC report for Decision Tree [source: own]	68
Figure 36: Accuracy report for Decision Tree [source: own]	68
Figure 37: Most important Variables for champion model [source: own].....	71

List of Tables

Table 1: Class labels of the selected attributes with their data type and description.	33
Table 2: Train and test Dataset (source: Own).....	35
Table 3: Skewness and Kurtosis of the Distribution of the Dependent Variable [source: own].....	37
Table 4: Spearman Correlation Coefficients.....	56
Table 5: Multicollinearity test using tolerance and VIF (Source: Own).....	58
Table 6: List of variables after transformation (Own source).....	60
Table 7: Variable Classifications	62
Table 8: ROC and Accuracy comparison for the three models [source: own]	69
Table 9: Model Comparison	70

1. INTRODUCTION

Information is all around us in various forms and plays an important role for both people and businesses. Every day, massive amounts of data are generated in a wide range of forms. It is important for them to have the right knowledge at the right time in order to survive in this highly competitive world. Companies need information about their clients, goods, the world, and even themselves. Working with large volume of data has many benefits, including the potential to make more rigorous in the decisions. It is important, though, that the increase in volume be matched by an increase in data quality. In different dimensions, many attentions arise. From a technical standpoint, it is critical that systems develop and become capable of handling large amounts of data in a timely and secure manner. On the business side, it becomes necessary for companies to be able to filter the information that is really important from what is not. Finally, on the social side, it is important to ensure security and confidentiality in the handling of information.

Big Data is the term used to describe the large amount of data, both structured and unstructured, that affects our daily lives activities and need specialized tools to produce significant results that would be impossible to achieve in smaller volumes. The focus should not be exclusively in dealing with high quantity of data, but on the possibility that these data offer in the intention of creating information and knowledge that can make companies and public entities more competitive, which will let them to offer better services for consumers and citizens.

To achieve the end goal and scopes, this study relies on Big Data as a primary source. It exhibits typical characteristics and issues that are arisen when dealing with Big Data. The focus of the thesis is to develop an in depth analysis of traffic accident severity and to determine which of the predictive modeling used is the best one for predicting accident severity as target variable. Road traffic injuries are reasons for many deaths globally. 2015 World Health Organization's Global Status Report asserts that 1.2 million people die every year by Road traffic accidents. (Organization, 2015) This figure has continued to rise each year, as shown by the organization's subsequent results. For example, the 2018's report of the organization shows the number of traffic injury deaths has elevated to 1.35 million people die on the roads each year. (Świdorski et al., 2018) Based on the global road safety status reports, traffic injury is the leading cause of death in children and adults under 30 years causing grinding social, economic and health problems and

Low-and middle-income countries had higher road traffic fatality rates compared to high-income countries. Even though the developed countries are technologically much greater than developing countries, still the number of death and injuries are recording day to day and resulting in huge losses at the economic and social levels.

The ability to predict future accidents (For example, where, when, and how...) is thus extremely useful not only to public safety officials, but also to transportation managers and individual travelers. A potential application of such technique would be real-time safe route recommendation for governments stakeholders for society specially drivers and pedestrians. With the rapid development of data collection techniques and the availability of big urban datasets in recent years, predicting traffic accidents has become more realistic. Detailed rainfall data, public transportation information, and motor vehicle crash reports could provide valuable information for traffic accident analysis.

2. Objectives and Methodology

This chapter provides a discussion of the research objective and methodology that was used in this study. It discusses effective approaches equipped during the research process and structured into study area and the data analysis techniques. This is important in ensuring that the study addresses the set objectives on which it is founded.

2.1 Objectives

The need to gather and use the extensive data produced every day in all sectors is increasing. One of the sectors that generate huge amount of data every day is the transportation sector where the requirements of a big data characteristics mate. Traffic flow and traffic accident data is one of the big data generated everyday throughout the world. To get a meaningful and useful systems equipped for gathering, classifying and detailing the accident and crash related information seriously improves the capacity to create, test, and implement mitigation strategies properly. The duty of recognizing accidents causality components gets unpleasantly theoretical without timely, accurate, complete, integrated, and available information that incorporates area, cause, contributing elements, and related activities associated with injuries involving personnel.

The main objective of this thesis is to develop detailed analysis for detecting the viable models to predict the occurrence and the level of traffic accidents fatality.

Having this as a primary objective, there are also specific goals that the study aims to achieve such as:

- I. Defining the determinant factors of traffic accident
- II. Understand the relations of the dependent and independent variables of traffic accident
- III. Use Statistical predictive modeling technique to predict severity of traffic Accident

2.2 Methodology

To achieve the objectives, the development of this thesis involves a literature review of Big Data, of its characteristics and also the comparisons of different statistical models. Part of the literature

review will also describe the different predictive statistical methods and techniques in general and focuses on those that will be applied on the data, such as Logistic Regression, Neural Network and Decision Tree model. The last part of the literature consultation includes an analysis of previous papers and research's on Traffic accident, the severity of traffic accidents and research-based recommendations forwarded.

The practical part will start with the main characteristics and information of the dataset used in this research. Before proceeding with the Analysis, it will be important to apply the appropriate cleaning and wrangling procedures in order to obtain normalized data. The study will also use explanatory data analysis to prepare the data for pre-processing and to get a full understanding of the variables and how each of them affects the dependent variable. This is done through the explanatory data analysis (EDA) techniques, each variable will be analyzed independently, through the Univariate Statistical Analysis. After the analysis of each variable independently, the study will use multivariate analysis to examine several variables to see if one or more of them are predictive of the dependent variables. The predictive variables are independent variables, and the outcome is the dependent variable. The multivariate analysis of variance is very important to understand the relationship of the target variable with the independent ones. The last step of the practical part will deploy the three methods i.e., Decision tree, Logistic Regression and Neural network. To identify the best predictive model, the following characteristics such as Root Mean Square error (RMSE), receiver operating characteristics (ROC) curve or confusion matrix will be checked. The study dataset has two distinct datasets accident data and Vehicle data. The data has 1,048,576 observations and 55 variables.

3.Literature Review

In this section, the researcher undertakes a systematic literature review on the works of various scholars in the field of Big data analysis to understand the theory underpinning on the area and then find empirical evidence that has been done by scholars and practitioners in the related field.

3.1 Big data

Big Data is described by (Manyika et al., 2011) as “datasets that are too large for traditional database software tools to collect, process, handle, and analyze.” It's difficult to store and retrieve large amounts of structured and unstructured data in a reasonable period of time. The term "Big Data" was coined as a result of some of these shortcomings in handling and processing large amounts of data using conventional storage techniques. Although big data has attracted attention due to the emergence of the Internet, it cannot be compared to it. The Internet, on the other hand, makes it easier to gather and exchange raw data as well as information. Big Data is concerned with how these data can be collected, interpreted, and comprehended in such a way that they can be used to forecast future actions with high precision and a reasonable time delay. In 2011, the McKinsey Global Institute described big data as "datasets that are too massive for conventional database software tools to collect, store, handle, and analyze." “Big data innovations,” according to the International Data Corporation (IDC), are “a modern wave of technologies and architectures designed to commercially derive value from very large quantities of a range of data, by enabling high-velocity capture, discovery, and/or analysis” (Gantz & Reinsel, 2011). Though big data has been defined in various forms but there is no specific definition. Few have defined what it does while very few have focused on what it is. In addition to defining big data, there is a need to understand how to make the best use of this data to obtain valuable information for decision making.

3.1.1 Characteristics of Big Data

Initially, big data was described or characterized by the dimensions mentioned below, which were referred to as the 3V model:

a) Volume: is refers to the magnitude of the data that is being generated and collected from terabytes to petabytes, it is rising at a faster pace (1024 terabytes). What can't be captured and stored right now would be possible in the future thanks to increased storage capacity. The classification of big data on the basis of volume is relative with respect to the type of data generated and time. In addition, the type of data, which is often referred as Variety, defines “big” data. For example, different data management technology can be required for text and video of the same volume (Gandomi & Haider, 2015)

b) Velocity: Velocity refers to the data generation rate. Traditional data analytics is based on periodic updates- daily, weekly or monthly. With the increasing rate of data generation, big data should be processed and analyzed in real- or near real-time to make informed decisions. The importance of time is crucial in this case(Gandomi & Haider, 2015); (Singh et al., 2012)High-frequency data is generated in a few domains, including retail, telecommunications, and finance. The data generated through Mobile apps, for instance, demographics, geographical location, and transaction history, can be used in real-time to offer personalized services to the customers. This would help to retain the customers as well as increase the service level.

c) Variety: Variety refers to different types of data that are being generated and captured. They go beyond structured data and fall into the semi-structured and unstructured data divisions (Zikopoulos & Barbas, 2012); (Singh et al., 2012), (Gandomi & Haider, 2015) The data that can be organized using a pre-defined data model are known as structured data. The tabular data in relational databases and Excel are examples of structured data and they constitute only 5% of all existing data (Mendes et al., 2010). Unstructured data, such as video, text, and audio, cannot be arranged using these pre-defined templates. Semi-organized data is data that comes in between structured and unstructured data. This group contains the Extensible Markup Language (XML).

Following that, a few more dimensions were added, which are mentioned below:

d) Veracity: Is veracity refers to the unreliability associated with the data sources (Gandomi & Haider, 2015)For instance, sentiment analysis using social media data (Twitter, Facebook, etc.) is subject to uncertainty. There is a need to differentiate the reliable data from uncertain and imprecise data and manage the uncertainty associated with the data.

e) Variability: Variability and complexity have been added as additional dimensions by SAS. Inconsistency in big data velocity often contributes to variability in data flow rate, which is referred to as variability (Gandomi & Haider, 2015). Data are generated from various sources and there is an increasing complexity in managing data ranging from transactional data to big data. Data generated from different geographical locations have different semantics (Zikopoulos & Barbas, 2012), (Forsyth & Ponce, 2012).

f) Low-Value density: Data in its original form is unusable. Data is analyzed to discover very high value (Bentolila et al., 2012). Logs from a website, for example, cannot be used in their raw form to derive business value. It must be examined in order to forecast consumer behavior.



Figure 1: The V's of Big Data [source: Data science: Surya Gutta]

3.1.2 Statistical Methods for Big Data Analysis

Big Data Analytics involves, to a large extent, collecting data from different sources, making it available to analysts and finally delivering data products that are useful to the organization of business, the foundation of Big Data Analytics is turning vast volumes of unstructured raw data from various sources into a data product that is beneficial to organizations.

Increased and more advanced technologies have contributed to data growth over the past few decades, leading to greater data volume, velocity, and complexity. In terms of knowledge expansion and better decision making, this has brought immense opportunities in all fields and

all industries. However, together with the opportunities, there are significant challenges. In addition to the general challenges, there are also statistical challenges for analyzing Big Data. To make accurate data analysis, different categories of statistical methods have been developed. Five of them will be analyzed in detail as they are used in the practical part, but this section will also provide a complete picture of the statistical techniques and modeling that are most used

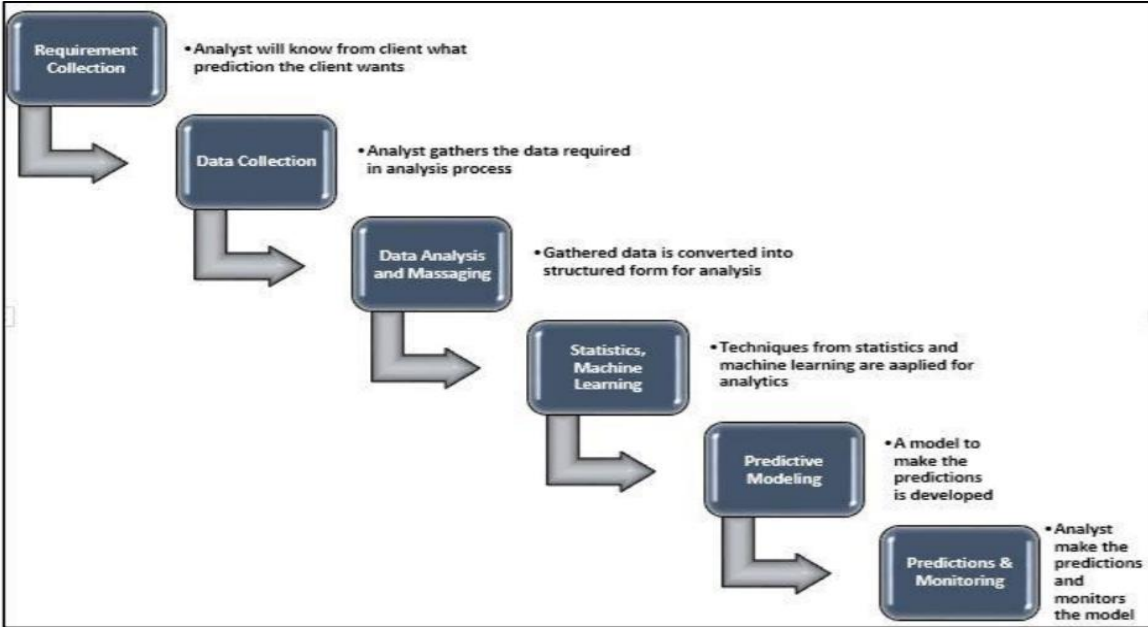


Figure 2: Predictive Analytics Process [source: International Journal of Computer Applications]

3.1.3 What are the challenges of Big Data analysis?

Although the benefits of Big Data are factual and substantial, a plethora of challenges remain to be addressed in order to fully realize the potential of Big Data. Most of these challenges are a function of the characteristics of Big Data, while others are caused by the existing measurement techniques and models, and even others are caused by the shortcomings of modern data processing systems (Jin et al., 2015). The extensive Big Data challenges studies have paid attention to the difficulties of understanding the notion of Big Data (Hargittai, 2015), deciding what data should be produced and collected (Crawford et al., 2013), Concerns about privacy (Lazer et al., 2009) and ethical considerations relevant to the mining of such data (Boyd & Crawford, 2012) One of the biggest problems with big data, for example, is the high cost of infrastructure (Wang & Wiebe,

2016). Also, with the availability of cloud computing technologies, hardware equipment remains very costly.

Furthermore, human interpretation is often used to filter through data in order to create useful knowledge.

According to a report by (Akerkar, 2014) and (Zicari, 2014), the large challenges of Big Data can be grouped into three groups based on the data life cycle: data, process, and management challenges.

The data challenges relate to the characteristics of the data themselves (example data volume, variety, velocity, veracity, volatility, quality, discovery, and dogmatism).

- Process challenges are linked to a series of techniques: how to capture data, how to integrate data, how to transform data, how to select the right model for analysis and how to deliver results.
- Management challenges include, for example, privacy, security, governance and ethical aspects.

According to them (Fan et al., 2014), In order to successfully deal with these dimensions, well-designed big data systems must strike a balance between data processing goals and the cost of data processing (i.e., computational, economical, and programming efforts) in big data systems. Big Data are distinguished by high dimensionality and a broad sample size, according to (Fan et al., 2014). These two characteristics present three distinct challenges: (i) High dimensionality combined with large sample size causes problems such as high computational expense and algorithmic instability; (ii) Massive samples in Big Data are normally aggregated from multiple sources at various time points using various technologies; (iii) Massive samples in Big Data are typically aggregated from multiple sources at various time points using various technologies. This creates issues of heterogeneity, experimental variations, and statistical biases, and requires us to develop more adaptive and robust procedures. Scale, Security, Consistency, Skilled manpower and many more are the other challenges in Big Data.

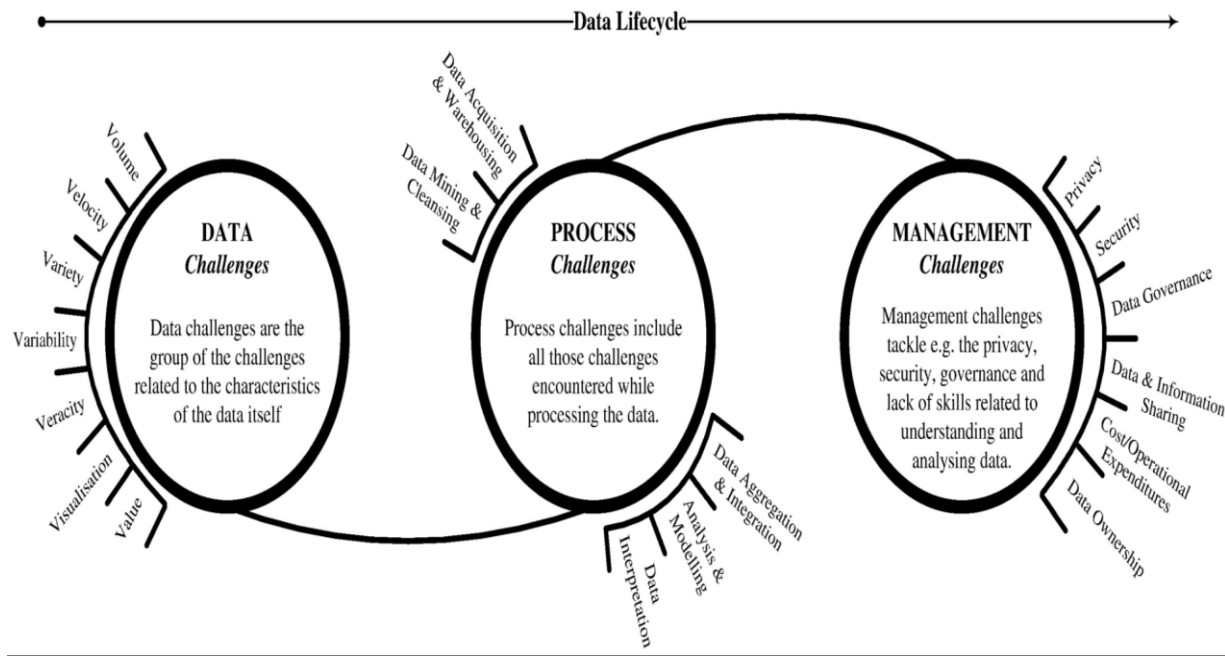


Figure 3: Conceptual classification of Big Data challenges [source: adapted from (Akerkar 2014) & (Zicari 2014)]

3.2 Data Mining Techniques and Modelling

Data mining is the set of methods and techniques for exploring and analyzing data sets (which are often large), in an automatic or semi-automatic way, in order to find among these data certain unknown or hidden rules, relationships, or tendencies; special structures output the most important aspects of usable data while reducing the amount of data. In a nutshell, data mining is the method of extracting information, or knowledge, from a collection of data. As a result, data mining can be both informative and predictive. Predictive (or explanatory) techniques are structured to extrapolate new knowledge based on the basis of current information, while descriptive (or exploratory) techniques are designed to pull out information that is present but hidden in a mass of data (as in the case of automated clustering of individuals and searches for correlations between goods or medicines). this new information being qualitative (in the form of classification or scoring¹) or quantitative (regression)(Tufféry, 2011).

Predictive modeling is one of the most used in Big Data and also differently called supervised learning. It is used in both structured and unstructured data to predict future outcomes and it is one of the most popular big data advanced analytics use cases. A predictive model is a statistical or

data-mining solution consisting of algorithms and techniques that can be used on both structured and unstructured data (together or individually) to determine future outcomes (Hurwitz et al., 2013).

Predictive analytics is a concept that is commonly applied in reference to statistical and analytical techniques. Statistics, machine learning, database techniques, and optimization techniques are all used to create this definition. Its history can be traced back to classical statistics. It makes predictions about the future by analyzing present and historical evidence. Predictive analytics models can be used to forecast future events and variable behavior. The score is mostly based on predictive analytical models. And involves several steps through which a data analyst can predict the future based on the current and historical data (Engino\uglu & Memi\cs, 2018).

3.3 What differentiates predictive analytics from other types of analytics?

First, predictive analytics is data-driven, meaning that algorithms derive key characteristic of the models from the data itself rather than from assumptions made by the analyst. Put another way, data-driven algorithms induce models from the data. The induction process can include identification of variables to be included in the model, parameters that define the model, weights or coefficients in the model, or model complexity. Second, predictive analytical algorithms automate the process of finding data patterns. Powerful induction algorithms not only discover the coefficients or weights of the models, but also the very shape of the models. Predictive modeling algorithms are supervised learning techniques, which means they search for correlations between inputs and one or more target variables. The target variable is key: Good predictive models need target variables that summarize the business objectives well (Abbott, 2014). This chapter, however, describes only some of statistical methods most used for practical part and found in predictive analytics.

¹ The statistical technique is called ‘classification’ or ‘discrimination’; the application of this technique to certain business problems such as the selection of customers according to certain criteria is called ‘scoring’.

Linear Regression

The most common algorithm predictive modelers use for regression problems is linear regression, linear regression is a statistical procedure for calculating the value of a dependent variable from an independent variable. The relationship between two variables is measured using linear regression. Linear regression analysis is the most commonly used statistical method. It is a simulation technique in which a dependent variable is estimated based on one or more independent variables.

Figure 4 shows a scatterplot with a trend line added to show the best fit of the data points using linear regression. The linear regression algorithm finds the slope of the output with respect to the input and Predictive modelers usually use linear regression in the same way other algorithms are used, primarily to predict a target variable accurately (Abbott, 2014).

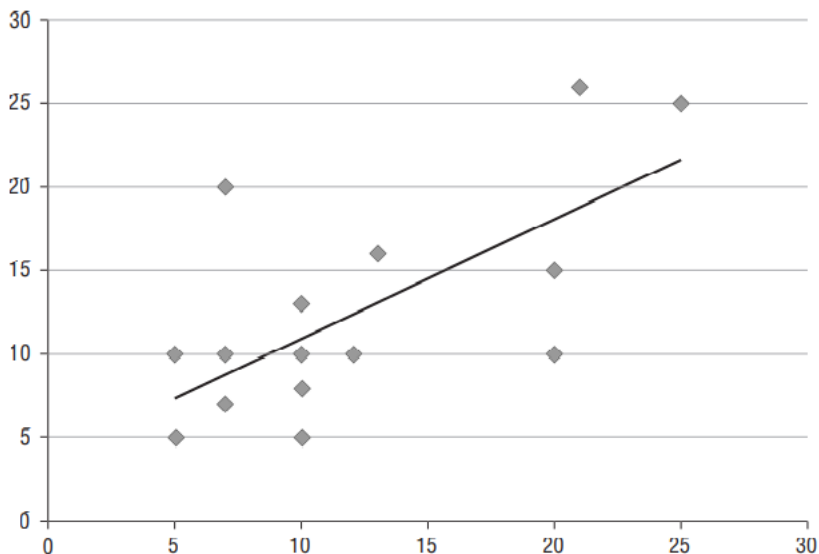


Figure 4: Regression line Data [source: Applied Predictive Analytics: Dean Abbott]

(Kutner et al., 2005). on a book of Applied Linear Statistical Models 5th edition, explain Linear Regression model as follows:

Simple Linear Regression Model with Distribution of Error Terms Unspecified

we consider a basic regression model where there is only one predictor variable and the regression function is linear. The model can be stated as follows:

$$Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i \quad \dots\dots\dots (1.1)$$

Where

Y_i is the value of the response variable in the i th trial β_0 and β_1 are parameters X_i is a known constant, namely, the value of the predictor variable in the i th trial ε_i is a random error term with mean $E\{\varepsilon_i\} = 0$ and variance $\sigma^2\{\varepsilon_i\} = \sigma^2$; ε_i and ε_j are uncorrelated so that their covariance is zero (i.e., $\sigma\{\varepsilon_i, \varepsilon_j\} = 0$ for all $i, j; i \neq j$) $i = 1, \dots, n$

Regression model (1.1) is said to be simple, linear in the parameters, and linear in the predictor variable. It is “simple” in that there is only one predictor variable, “linear in the parameters,” because no parameter appears as an exponent or is multiplied or divided by another parameter, and “linear in the predictor variable,” because this variable only occurs in the first power. A model that is linear in the parameters and in the predictor, variable is also called a first-order model

General multiple linear regression model

where

$$Y_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \dots + \beta_{p-1} X_{i,p-1} + \varepsilon_i ,$$

- $i = 1, 2, \dots, n$

- Y_i is the value of the response variable for the i th case

- $X_{i1}, X_{i2}, \dots, X_{i,p-1}$ are known constants, X_{ik} is the value of the k th explanatory variable for the i th case

- $\beta_0, \beta_1, \dots, \beta_{p-1}$ are parameters, $p - 1$ predictors, p parameters

- $\varepsilon_i \text{ iid} \sim N(0, \sigma^2)$

The authors states that this general linear model is able to fit a surprisingly large set of different data sets.

The difference between the target variable value in the data and the value predicted by the regression model, y minus \hat{y} , is called a residual and is often annotated with the letter e , as in the equation:

$$e = y_i - \hat{y}_i \dots \text{(Abbott, 2014)}$$

As part of the analysis, coefficients like p-value, standard errors, confidence interval are calculated. Simple Linear Regression and Multiple Linear Regression are divided into linear regression analysis. Although only one dependent variable and one regressor are used in the first one, the second one expands the analysis to several independent variable. It assesses not only the independent categorical variables, but also the dimensional one.

Neural network

Neural network models are being considered as universal approximators, which means by using a neural network methodology, we can solve any type of problems with the fine-tuned architecture. Hence, studying neural networks is a branch of study and special care is needed. In fact, deep learning is a branch of machine learning, where every problem is being modeled with artificial neural networks (Dangeti, 2017).

Because of the hidden layers, neural networks are stronger at predictive analytics. To make predictions, linear regression models only use input and output nodes, while neural networks use the hidden layer to improve accuracy. This is due to the fact that a neural network learns in the same way that a human does.

The structure of the neural-network algorithm consists of three layers:

(1) Layer input: This will enter past data values into the next layer. (2) The hidden layer: the key component of the neural network. It has a complex function that creates predictors. A set of nodes in the hidden layer called neurons is a math function that modifies the input data. (3) The output layer: Here, the predictions made in the hidden layer are collected to produce the final layer that is the model's prediction. So, how can it make predictions? A collection of input values is considered by each neuron. Each of them is associated with a "weight," which is a numerical value that can be extracted using either supervised or unsupervised training, such as data clustering, and a "bias," which is a value that can be derived using either supervised or unsupervised training, such as data clustering. The network chooses from the answer produced by a neuron based on its' weight and bias. The neuron computes a function by imputing each neuron having specific input with a given

amount of weight. Then an activation functions like sigmoid, tanh, rectified linear unit (ReLU) will be applied on the linear combination weighted inputs on the aggregated sum.

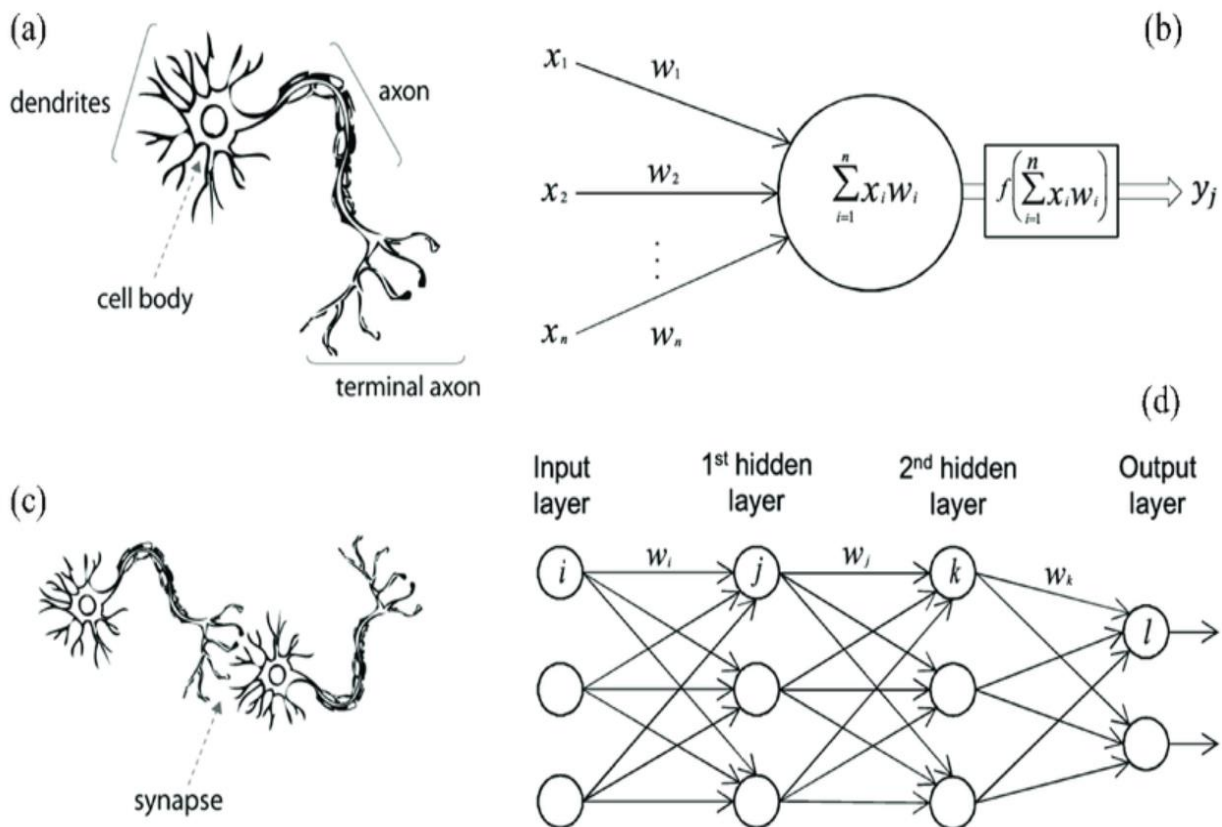


Figure 5: biological neuron versus artificial neural network: (a) human neuron; (b) artificial neuron; (c) biological synapse; and (d) ANN synapses. [source: Combining Finite Element Method and Data-Driven: Zhenzhu Meng, 2020].

The activation functions are simply the principle by which the neurons process the inputs and transform them. Sigmoid function takes any real number and change to value between 0 and 1 by using $\sigma(x) = 1 / (1+e^{-x})$, while Tanh function change the input real numbers to the range between -1 and +1 by using $\tanh(x) = (e^x - e^{-x}) / (e^x + e^{-x})$.

In the cause of rectified linear unit (ReLU) uses the function $f(x) = \max(0, x)$ and simply use the threshold at zero. Additionally, linear function $f(x) = x$ could be also used as an activation function. The prediction or classification using ANN could be implemented first by identification of the input/output data, normalizing data, establishing the network with the suitable structure, learning

the machine, testing the algorithm and applying the model generated by the machine. (Tufféry, 2011)).

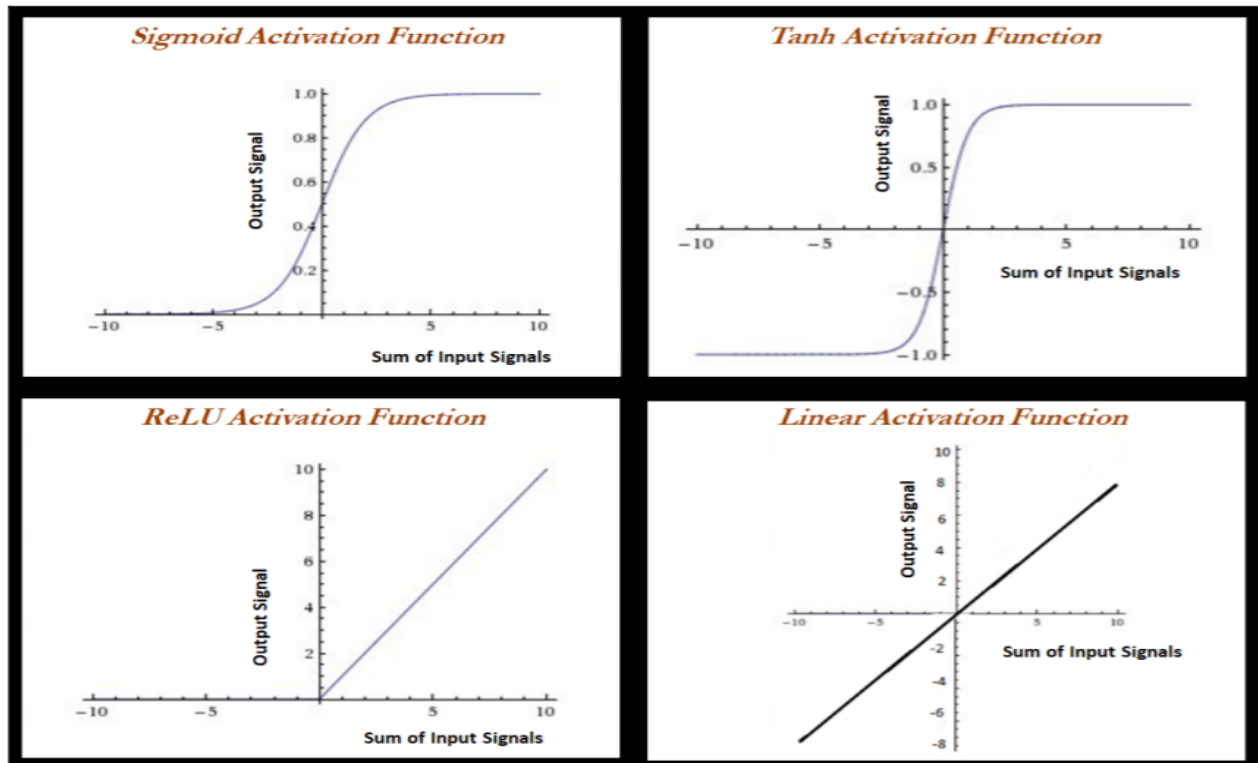


Figure 6: The diagrammatic illustration of activation functions used in ANN. [source: *Statistics for Machine Learning: Dangeti 244, 2017*).

So, why aren't neural networks used by everyone? For starters, they necessitate vast quantities of computing power, making them prohibitively expensive. Furthermore, neural networks perform better when trained on incredibly large data sets, which your company can lack. But with IT tech getting cheaper, the first hurdle may soon disappear. Soon, technology like ANNs will mean that there will be no more “unpleasant surprises”.

Decision tree

Decision tree learning is widely used in data mining. The output of this method is a classification model that predicts the value of a target attribute based on the input attributes. The decision tree constructs classification models in the form of trees. Each interior node in these trees represents one of the input variables, and it has a number of branches equal to the number of possible values of that input variable. Each leaf node holds a value of the target attribute. The leaf node represents

the decision made based on the values of the input variables from the root to the leaf. The decision tree algorithm was used to understand existing data and to predict the severity of the new accidents.

Decision trees require less data preparation than many other techniques because they are equally capable of handling continuous and categorical variables. Categorical variables, which pose problems for neural networks and statistical techniques, are split by forming groups of classes. Continuous variables are split by dividing their range of values. Because decision trees do not make use of the actual values of numeric variables and they are not sensitive to outliers and skewed distributions of numeric variables, because the tree only uses the rank of numeric variables and not their absolute values (Berry & Linoff, 2004).

The use of decision tree algorithm has the following main advantages:

- ✚ Decision tree learning algorithms are very efficient and scale well as the number of records or fields in the modeling data increase. (Abbott, 2014)
- ✚ Because of their inner workings, decision trees handle large datasets and increased volumes of data without halting their prediction speed or losing their high accuracy. This makes them extremely useful for big-data problems.
- ✚ The models developed are easy to understand and by non-experts also can be easily visualize.
- ✚ No need of preprocessing like normalization and standardization of feature is needed as each feature is processed separately (Wild et al., 2017).
- ✚ Any extreme values will be separated in small nodes and will not affect the classification problem (Tufféry, 2011).

The basic goal of a decision tree is to split a population of data into smaller segments. There are two stages to prediction. The first stage is training the model this is where the tree is built, tested, and optimized by using an existing collection of data. In the second stage, you use the model to predict an unknown outcome.

When the decision tree algorithm completes the building of a tree, terminal or leaf nodes are collectors of records that match the rule of the tree above the terminal node. Every record in the

training dataset must ultimately end up in one and only one of the terminal nodes. For each of the decision tree algorithms described, the algorithm steps are as follows (Abbott, 2014)

1. For every candidate input variable, assess the best way to split the data into two or more. Select the best split and divide the data into the subgroups defined by the split.
2. Choose one of the subgroups and repeat Step 1 (this is the recursive part of the algorithm). Repeat for each subgroup.
3. Continue splitting until all records after a split belong to the same target variable value or another stop condition is applied. A stop condition may be as sophisticated as a statistical significance test, or as simple as a minimum record count.

Decision tree building algorithms begin by trying to find the input variable that does the best job of splitting the data among the desired categories. At each succeeding level of the tree, the subsets created by the preceding split are themselves split according to whatever rule works best for them. The tree continues to grow until it is no longer possible to find better ways to split up incoming records(Berry & Linoff, 2004).

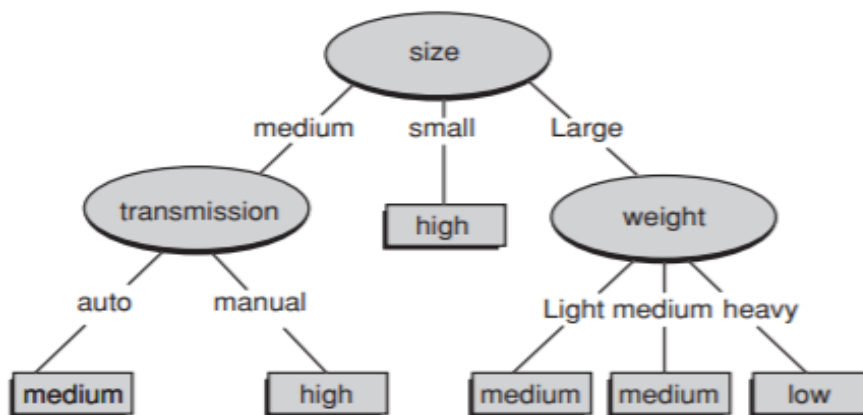


Figure 7: Example of simple decision tree (Sumathi and Sivanandam, 2006).

Random forest

Another type of supervised learning algorithm used for prediction is random forest. It is also one of the most commonly used in Machine Learning. What it does is margining many decision trees and creating ‘a forest’ to give a better and more accurate prediction. Same as with the Decision

Tree, it is also used for Classification and Regression problems which form most of Machine Learning systems. Each decision tree has access to only a random subset of training data and uses only a random subset of attributes at each node of each decision tree. Having such a wide variety of estimation from each decision tree, the random forest takes an average of all of them and makes the prediction.

Logistic Regression

Like all other regressions, Logistic regression is also part of predictive modeling and is foremost used to model a binary (0,1) variable based on one or more other variables, called predictors. The binary variable being modeled is generally referred to as the response variable, or the dependent variable. I shall use the term “response” for the variable being modeled since it has now become the preferred way of designating it. For a model to fit the data well, it is assumed that

- The predictors are uncorrelated with one another.
- That they are significantly related to the response.
- That the observations or data elements of a model are also uncorrelated.

In the case of logistic regression, the response is binary (0,1), a model predicts the probability that the response has a value of 1 given a specific set of predictor values (1= if the phenomena is happened, and 0 if not) (Hilbe, 2016)

Generally logistic regression can be assumed unique due to its ability to perform the following three different purposes (Hilbe, 2016);

- For predicting the probability of response variable equals 1
- For categorization of outcomes or predictions
- For accessing the add or risk, which is associated to the models of predictors

The ability of logistic regression to perform these three different tasks and its applicability in the binary variables make it from other types of regression. The model assumes: No correlation of predictor variables, Significant correlation of predictors to the response variable and No correlation of the observation or data elements of the model

Logistic regression is a non-parametric technique, which classify the probability of attribute distribution. In the application of logistic regression, the log of probability of odds will be matched with linear combination of selected characteristic variables (Gough et al., 2017). Logistic regression applies a logistic transformation to the input and restrict the output ranging from $-\infty$ to $+\infty$ to the probability between 0 and 1. Therefore as there are only two outputs in the model.

To interpret logistic regression models, you'll examine: (Abbott, 2014)

- Coefficient: Sometimes shown as the Greek symbol beta (β). This is the parameter found by the logistic regression algorithm; one coefficient per input variable plus one constant (bias term).
- Standard error of the coefficient (SE): A measure of certainty of the coefficient found by logistic regression. Smaller values of standard error imply a smaller level of uncertainty of the coefficient value.
- Confidence Interval (CI): The range of values the coefficient is expected to fall between. The CI is computed as: $CI = \pm\beta SE$.
- z statistic or Wald test: z is calculated by the equation:

$$z = \frac{\beta}{SE}$$

The larger the value of z, the more likely the term associated with the coefficient is significant in the model. The larger the value of z, the more likely the term associated with the coefficient is significant in the model.

“Let us look at a logistic regression and how it differs from normal or ordinary linear regression. Recall that a regression attempts to understand a response variable on the basis of one or more predictors or explanatory variables. This is usually symbolized as

$$\hat{y}_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \dots + \beta_p x_{ip}$$

where \hat{y} , or y^* , is the sum of the terms in the regression. The sum of regression terms is also referred to as the linear predictor, or $x\beta$. Each βx is a term indicating the value of a predictor, x , and its coefficient, β . In linear regression, which is based in matrix form on the Gaussian or normal

probability distribution, \hat{y} is the predicted value of the regression model as well as the linear predictor. j indicates the number of predictors in a model. There is a linear relationship between the terms on the right the predicted or fitted values of the model and -hand side of above Equation the linear predictor. $\hat{y} = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \dots + \beta_p x_p$. This is not the case for logistic regression. The linear predictor of the logistic model is

$$\hat{y} = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \dots + \beta_p x_{ip} \quad \text{'' (Wiggins \& He, 2016)}$$

3.4 Performance Indicator for practically applied Methodologies

Selecting the best model from a set of candidates for a given set of data is obviously not an easy task. This portion of the chapter is very important to determine the efficiency of the model in the practical section, based on the results of the metrics used conclusions will be described with respect to the best model to use for a particular dataset of Road traffic accident. To identify the best predictive model, the following characteristics such as Root Mean Square error (RMSE), receiver operating characteristics (ROC) curve and confusion matrix will be discussed.

Root Mean Square Error (Root Mean Square Error)

Both the root mean square error (RMSE) and the mean absolute error (MAE) are regularly employed in model evaluation studies (Willmott & Matsuura, 2005). The root mean square error (RMSE) has been used as a standard statistical metric to measure model performance in meteorology, air quality, and climate research studies. The mean absolute error (MAE) is another useful measure widely used in model evaluations. They have both been used to assess model performance (Chai & Draxler, 2014). In many of the model sensitivity studies using RMSE alone, a detailed interpretation is not critical, as variations of the same model will have similar error distributions. When evaluating different models using a single metric, differences in the error distributions become more important. The underlying assumption when presenting the RMSE is that the errors are unbiased and follow a normal distribution. For other kinds of distributions, more statistical moments of model errors, such as mean, variance, skewness, and flatness, are needed to provide a complete picture of the model error variation. (Tukey, 1977; Huber and Ronchetti, 2009).

RMSE as the name suggests, represents the standard deviation of the differences between the predicted values (theoretical values) and the observed ones and the most used metrics in the regression for continuous variables. It defines how well the model performs, and it always has a non-negative value. A value of zero would mean that the model is neither overfitting nor underfitting and that the data perfectly fits. This never happens in reality. The lower the RMSE, the better the model performs; moreover, there is no set specific range of values for RMSE since it is depending on the dependent variable. Its mathematical calculation is;

$$\text{RMSE} = \sqrt{\frac{1}{n} \sum_{j=1}^n (y_j - \hat{y}_j)^2}$$

Receiver Operating Characteristics (ROC)

ROC curves are generally used when statisticians wish to determine the predictive power of the model. The ROC curve is understood as the optimal relationship of the model sensitivity by one minus the specificity. When using ROC analysis, the analyst should look at both the ROC statistic as well as at the plot of the sensitivity versus one minus the specificity. Values from 0.65 to 0.80 have moderate predictive value and any logistic models fit into this range. Values greater than 0.8 and less than 0.9 are generally regarded as having strong predictive power. Values of 0.9 and greater indicate the highest amount of predictive power, but models rarely achieve values in this range. The model is a better classifier with greater values of the ROC statistic, or area under the curve (AUC). (Hilbe, 2016)

ROC curves are also used for classification purposes and is used in predictive modeling to distinguish the difference between correct positive and false favorable rates. So, to do this, the model must include both true positives and true negatives. The way the ROC curve does this is by plotting the sensitivity in the y-axis; the probability of predicting that a true positive is a real true positive and by plotting the specificity in the x-axis; the probability that a true negative is as such (Dan Lans 2019) .

3.5 What is Roads Traffic accident and its burden

A road traffic accident is an accident involving a vehicle that has caused harm or damage to an individual, animal, another vehicle or property on a road or in a public area.

Traffic accident can be described as an event that results in death, injury, and material loss due to the crash of one or more vehicles moving on roads. Traffic accidents are one of the most pressing international health and development concerns in the world. The main cause of traffic accidents may be human factors and individual performance(Aslan et al., 2012). Traffic accidents are unfortunately very common in all over the world and the majority of these road crashes are caused by human error. Thousands of lives are lost each year as a result of these terrible car accidents, some of which are relatively minor. Because our lives should be put in jeopardy if we drive recklessly, it is important that we drive cautiously and obey all traffic laws. However, just because you are careful does not mean that you can assure that all other drivers on the road will do the same thing. If you are involved in a car accident, it may not be your fault, and you should not be held liable for the damages caused by the ignorance or mistakes of other drivers. In such cases, you should consider protecting yourself by filing a car accident claim.

Research's and WHO's report proof that Road traffic accident is a burning issue until today and couldn't find solution even though the causes and consequences are very known. This global problem needs more attention to reduce the severity and the frequency of accident occurrence. The historical data on past incidents provides a valuable opportunity for researchers to determine the most important factors in such accidents, which will help them find appropriate ways to minimize the problem in the future.

“According to the 2018 Global Status Report on Road Safety, the situation is worsening. The number of people injured in road accidents has grown to 1.35 million every year. That equates to nearly 3700 people dying every day on the world's roads. Every year, tens of millions more people are injured or disabled, people who experience life-altering disabilities with long-term repercussions. These losses are having a major impact on family and communities. Emergency response, health insurance, and human grief all come at a high price. Rapid urbanization, low road practices, lack of compliance, persons driving intoxicated or fatigued, those under the influence of drugs or alcohol, speeding, and failing to wear seatbelts or helmets are all factors contributing to this trend. One of the most heart-breaking statistics in this report is that road traffic injuries are the leading cause of death for people aged between 5 and 29 years of age. No child should die or be

seriously injured while walking, cycling or playing” (Dr. Tedros Adhanom Ghebreyesus, Director of General world health organization, adopted from Global status report on road safety 2018)

Who’s 2018 report clarifies that, Road traffic injury is now the leading cause of death for children and young adults aged 5 to 29 years, signaling a need for a shift in the current child health agenda, which has generally ignored the importance of road safety. It has surpassed HIV/AIDS, tuberculosis, and diarrheal diseases as the eighth leading cause of death among all age groups. Road traffic accidents and fatalities disproportionately affect disadvantaged road users and those residing in low- and middle-income countries, where the rising number of deaths is fuelled by increasingly motorized transportation. There were no declines in the number of road traffic deaths in any low-income country between 2013 and 2016, but there were reductions in 48 medium- and high-income countries. During this time, the number of deaths rose in 104 countries. Those tangible report facts leads to focus on frequent traffic accidents, factors and other problems need to be improved and understanding the causes of traffic accidents and make an early alarm model for the drivers and pedestrians to be crucial to solve traffic accident-related problems in some way to prevent the problem.

3.5.1 Factors affecting road traffic accident

Road traffic accident is one of the most common reasons of deaths and injuries in the world

Traffic safety and accident studies have been extensive in the field of research, as the rise in accidents has been alarming throughout the world. From the work done by the researchers, it can be said that traffic accidents are mainly caused by three factors, i.e., Personal, or human behavioral factors, Road and Environmental factors, Vehicle factors.

Personal or human factors include, in particular, the age of the driver or victim, the gender of the victim, whether he was drunk while driving, etc. Some of environmental factors, weather condition, lighting conditions of road, day or night, etc. Road geometric factors are junction or intersection type, then curves and horizontal slope, etc.

Due to their high volumes and negligence of road traffic designs, vulnerable road users (VRUs) which could be pedestrians, cyclists and Powered Two-Wheelers (PTWs) accounts for more than half of the fatalities. Over the past few decades, the rate of traffic accidents and fatalities has gone

down considerably, but the decrease in traffic accident injuries for VRUs is low (Global status report on road safety, 2015; Road safety statistics, 2018; Road safety statistics, 2016). Pedestrians are estimated to be 284 times more likely than motorists to be killed or injured in a traffic crash (Robertson et al., 2017). Failure to wear seatbelts and helmets as well as people driving fatigued (Road safety statistics, 2018), the increasing use of mobile phones and other portable devices globally has become a new cause of traffic accidents nowadays because of many types of distractions they brought (Cukier, 2010). Drivers and VRUs use their phones for talking, texting, and listening to music while driving and walking. This results in an inattention which eventually becomes the reason for many traffic accidents (Terzano, 2013). Traffic accident injuries and fatalities have an immense impact on the physical, mental, and health issues of individuals as well as financial and social consequences for individuals and their families (Másilková, 2017). At the national level, economic developments of low- and middle-income countries is being hampered by traffic accidents. It cost governments approximately 3% of their Gross Domestic Product (GDP) (Global status report on road safety, 2015). Various accident protection mechanisms have been applied to minimize traffic accidents. Increasing visibility of roads and planning bumpers to reduce their speed are among passive measures for VRUs traffic accident reduction. Automatic braking is no more limited to luxury cars. Many contemporary vehicles are fitted with this feature for mitigation of forward-collisions. Moreover, passive measures like educating traffic safety and setting strict law enforcement are also among the most crucial VRU traffic accident prevention mechanisms. In recent days, active traffic accident protection steps that involve road user detection, collision prediction, alerting and collision avoidance have got a lot of attention. However, the problem is still apparent, and traffic accidents have continued to be reasons for deaths of many people. Those who didn't die of traffic accidents suffer permanent physical and mental consequences.

3.6 Related works

In 2008, (Cyprus, 2013) presented factors affecting Accident Severity Within and Outside Urban Areas in Greece, which included an analysis of accident occurrence, recognition of associated determinants, and quantification of their consequences, as well as a distinction between within and outside urban areas using logistic regression analysis. The data for all road injuries reported in

Greece in 2008 was disaggregated for this macroscopic analysis. Occupants from all vehicle type were examined (cars, trucks, buses, motorcycles, etc.), including both front and back seat occupants. Furthermore, the author research aims to estimate the probability of fatality/severe injury versus slight injury and he calculated the odds ratios (relative probabilities) for various road accident configurations. Appropriate logistic regression models were developed and their significance was tested by applying a hypothesis testing technique. As a result, only 2 severity outcomes were considered:

(1) killed or severely injured (KSI)

(2) slightly injured (SI).

The database consisted of 16,426 injured people (3327 KSI and 13099 SI). He found Ten variables to have a significant effect on severity outside urban areas: weather, time, light (<50 cc) and heavy motorcycles(>50cc), buses, older road users(>60years), and 4 collision types (side, sideswipe, rear-end, and head-on). Severity inside urban areas was found to be influenced by the following variables: light and heavy motorcycles, bicycles, buses, time of the accident, location, young (18–30 years) and older (>60 years) road users, and 3 collision types (sideswipe, fixed object, rear-end).

According to the author result conclusion, it was also found that more severe accidents tend to occur at night. The odds ratio was lower in urban areas (0.724) than outside urban areas (0.816). In addition to alcohol usage, another possible explanation is that people drive faster or less carefully because roads are less congested at night. In addition, insufficient illumination of roads and junctions especially outside urban areas in Greece could explain this increased accident severity.

Yannis T.H. also suggested that, Authorities should improve the road lighting conditions and provide more systematic police enforcement at night to check for drinking and driving offenders. The author examined only accident severity which indicates there is still open research issues for further studies.

(Papantoniou et al., 2019) published a research paper to understand the factors which are responsible for a road accident. The experiment was carried out on a group of 95 participants from

different age groups. These participants were requested to drive under various distraction in rural and urban street condition, and in addition in both low and high traffic conditions. The authors likewise inform that gender, age, driving experience and education were appeared to have the most noteworthy impact on driving mistakes.

(Deme & Bari, 2018) study Research Paper was conducted to classify the major causes of road traffic accidents, with an emphasis on the Addis-Adama expressway (in Ethiopia), to project causes of traffic accidents on the expressway, and to investigate the causes of traffic accidents on the expressway. To determine the causes of road accidents, descriptive and inferential statistical analyses were used. The expressway is 76.3 kilometers long in total, and the authors used secondary data obtained from Ethiopian Toll Road Enterprise. According to the survey, 417 crashes were registered on average between September 2014 and February 2016. According to the writers, as a result of these crashes, over 672 accidents were reported on the expressway, resulting in 285 human injuries and 387 property damage. Of the total number of human injuries, 37 were fatal, 65 were moderate, and 183 were minor. Day of the Week, Collision Types, Vehicle Types, and Plate Number Region were all major factors in traffic accidents. From four black spot regions on the expressway, most of the accidents were occurred at main gate.

The authors concluded that Despite small number of vehicles operating, the level of crash accident recorded in Ethiopia made the country one of the tops in the world.

Similar record was observed on the newly constructed Addis Ababa-Adama Expressway. Traffic accident was higher on working days (Monday to Friday) of the week mostly between 6:01 to 12:00 AM local time. Unethical behavior of drivers, over speed, sight distances, phone usage during driving, drug usage etc. were the main causes of all crash accidents. Day of the Week, Collision Types, Types of Vehicles and Plate Number Region highly determine the accident.

(Liu & May, 2012), Traffic Injury Prevention studied was about Analysis of Risk Factors Affecting the Severity of Intersection Crashes by Logistic Regression and it points that Intersections are one of the most dangerous locations in a roadway network, because they are not only a convergence point for vehicles and pedestrians traveling on conflicting paths, but they also often impose significant responsibility on road users to make successful gap judgements. Furthermore, Intersections often cause traffic congestion and have a high rate of serious side-impact collisions.

The authors used Logistic regression as methodology, and as a result, seven risk factors were found to be significantly associated with the severity of intersection crashes, including driver age and gender, speed zone, traffic control type, time of day, crash type, and seat belt usage.

In this report, male drivers, as well as those aged 65 and over, were found to have a higher chance of being involved in fatal intersection accidents. Intersection accidents that happened between midnight and early morning (12:00 a.m. to 5:59 a.m.), in 100 km/h speed zones, or with no traffic control have a greater risk of a deadly result than the other categories.

Intersection collisions involving pedestrians or a driver who was not wearing a seat belt were also more likely to result in a fatality. In general, identification of risk factors and the discussion of the odds ratio between levels on the impact of the intersection crash severity would be beneficial for road safety stakeholders to develop initiatives to reduce the severity of intersection crashes.

The authors put as a limitation on study was, Due to the limited range of variables involved in there data set, it was not possible to examine the effect of a number of other potential risk factors (e.g., geographical features, the presence of safety cameras, road surface friction, etc.) that may have influenced the crash.

(Meshram & Goliya, 2013) were presented an analysis of accidents on small portion NH-3 Indore to Dhamnod. The data for this study was gathered from September 2009 to September 2011. Faulty road geometry caused more accidents in the Manpur area. Accidents in the urban portion of the country (Indore) account for more than 35% of all accidents per year. According to the authors prediction, this may due to high speeds and more vehicular traffic and they mentioned that; In the present study area, the frequency of fatal accidents ware two in a week and six for minor accidents in a week. There are more incidents between the hours of 6 p.m. and 8 p.m. since more buses run between villages and the city at that time. In the study region, one fatality and five casualties occur per km per year. The number of trucks passing through the study corridor is growing year after year. Due to the construction of by passes in Rajendra Nagar, traffic has been reduced since 2000.

(Parab et al., n.d.) were presented an analysis Severe Traffic Accidents in United Kingdom. The author worked for analyzing the accidents and finding a correlation between the factors which are involved during the calamity and to identify the factors which are involved behind a road accident

by taking into consideration age, gender, vehicle details, climatic conditions, road conditions, speed limit, license attempts and locality. On analyzing the data, it is observed that most of the citizens in UK apply for a driving license as soon as they are legally permitted to have one; that is between the age group of 16 to 20. Moreover, Male drivers are majorly involved in road accidents and also on further analyzing on which day of the week accident occur the most, the author result justifies that maximum accidents take place over the weekends, Total number of fatalities in United Kingdom in 2017, when compared to 2016, London and South East United Kingdom had an increase by 13% and 15% respectively, drivers of age group 21 to 45 are involved in maximum accidents.

Hence, the author highly recommended to introduce a system in which candidates between the age group of 21 to 45 must renew their driving license after a particular period of time. The number of accidents which are happening in urban areas needs to be controlled. Improvement in road safety is essential and hence a law must be proposed to renew driving license after a particular year's span. On detailed review over the sentiments, more negative sentiments are observed and hence a phobia is likely to develop within people.

As discussed on the above scholars, types of vehicles, weather, working area, gender, age and day of a week are the main variable factor that are briefly discussed in all scholars. Some of them also mentions driving experience and education level, plate number of regions are among factor of road traffic accidents. From the above researchers we can easily understand that Occurrence of accident on weekdays (Monday to Friday) is higher than weekends and a non-seat belt-wearing driver were more likely to lead to a fatal outcome.

most of the research investigating indicates that the contributing factors of road accidents are made by young and older drivers and some of researchers agreed that male drivers have more probability for being case of accident than female drivers. According to (Manyika et al., 2011), the major causes of all crash deaths were unethical driving, excessive speed, sight distances, phone use while driving, drug use, and so on. Day of the Week, Collision Types, Types of Vehicles and Plate Number Region highly determine the accident almost all the above author idea agree with this.

The common recommendations from the scholars are the number of accidents which are happening in urban areas needs to be controlled, Improvement in road safety is essential and hence a law must be proposed to renew driving license after a year's span, should improve the road lighting conditions and provide more systematic police enforcement at night to check for drinking and driving offenders, and state governments should implement new system and give special concentration for it.

4. Practical Part

Introduction

This chapter presents analysis of the study as set out in the research methodology. The dataset was analyzed in line with the objectives. The analysis results were presented in graphs and tables.

4.1 Overview of the Case Study

The dataset was extracted from a Kaggle source of UK traffic incidents from 2005 to 2015. Kaggle is the world's biggest data science community with powerful tools and resources. After retrieving the traffic accident data from the site, the data exploration and cleaning activities has been done. The data cleaning is started using excel and SAS studio university Edition. From the given data only 15 variables have been selected any duplication has been removed for this research and also defining their values was done before doing any of the analysis part. The aim of analyzing this dataset was for academic purposes only, and the data would not be used for any other purpose. It includes 1,048,575 records of accidents with various information about traffic accident (Accident related, vehicle related and environment related information). It is important to understand the benefit of this study before analyzing the data and its variables. The occurrence and the level of distraction of traffic accident in the world is increasing both in property damage and human life. Therefore, understanding and analyzing the factors that contribute to accident severity and generating a predictive model that can help decision makers and academicians forecast severity of accidents from a set of data has paramount importance to both groups.

It would be possible to recognize some of the Big Data challenges listed in the literature review Throughout the step-by-step study. In addition, several steps may be taken to resolve each of these problems. After discovering these challenges, different statistical methods will be applied to the data set for the primary goal, that is to come to a conclusion on which of the techniques is the best one in predicting accident severity. Throughout the process, secondary insights will be obtained. A Machine Learning model will be used to achieve the key objective. Machine Learning is the science (and art) of programming computers so they can learn from data. Why machine learning for this specific dataset and goal? Applying machine learning techniques to dig into large amounts of data can help discover patterns that were not immediately apparent. This is called data mining. (Géron, 2017). Different models will be used and will be able to learn from this data set and from

there, using performance measure metrics, a conclusion can be made on which of the models is the best one when dealing accident datasets.

There are 14 qualitative variables selected from the dataset for this specific study. These variables will be analyzed in detail in the further sections. The tool used for the analysis of the dataset is SAS Studio 9.4. SAS university edition is an open-source software where data is run on SAS studio online and results can be generated it's a powerful software to analyses big data.

4.1.1 Problem framing

It's important to understand how the traffic accident data set is analyzed before looking at each variable and how it affects the dependent one. As a result, defining the type of method, the algorithms that will be applied to the dataset, the right performance measure and the assumptions that need to be verified.

Based on how many and what kind of supervision they get during testing training, Machine Learning systems can be categorized. There are four major categories: supervised learning, unsupervised learning, semi-supervised learning, and reinforcement learning (Daassi-Gnaba et al., 2017)). This system is supervised since the training set of this dataset contains labeled solutions. Furthermore, since the dependent variable to be predicted is qualitative variable, (the level of accident severity from the given accident), this task is a regression task. The algorithms used for this study are Logistic Regression, Decision Tree and Neural network which have been described theoretically in the literature review.

4.1.2 Preparing the data

Table 1 Class labels of the selected attributes with their data type and description.

Attributes Name		Data type	Description
Accident related Attribute	Day	Numeric	The day of Accident
	Day of a Week	Numeric	Days Monday to Sunday
	Time	Numeric	The accident occurred at What time of the day
	Number of Casualties	Numeric	Number of Casualties Happened at a time
	Number of Vehicle	Numeric	Number of Vehicle on accident
	Vehicle type	Numeric	Vehicle type that case an accident
	Age band of driver	Numeric	Age of the driver and injured persons
Driver vehicle Related Attributes	Sex of driver	Numeric	Gender of driver
	Vehicle type	Numeric	Type of Vehicle case the accident
	Causality type	Numeric	Cause for Accident. (Cyclist, Auto, pedestrian, ...)
	Causality class	Numeric	victim of Causality (E.g. driver, passenger, Pedestrian)
	Journey purpose	Numeric	To/from work, vacation or other
Road related Attribute	Road type	Numeric	Is the road Roundabout, one-way Slip road, other...
	Junction detail	Numeric	On what junction area of accident happen
	Junction Control	Numeric	Who was controlling Auto traffic sign, Authorized person, Give way or uncontrolled, Other
	Light condition	Numeric	Lighting condition of the road at the time of accident
	Weather	Numeric	The weather conditions
	Road Surface Condition	Numeric	Whether the surface of the road was dry, wet, sandy, or oily
	Speed limit	Numeric	The road speed limit
	Road class	Numeric	Types of road designed for. E.g. Motorway, class A,B,C or Other

Note: see appendix for variable code value.

One of the key challenges before preparing a dataset for machine learning algorithms is to split the dataset between train and test set. The best approach is to randomly pick 20 percent of the dataset and put it aside. The disadvantage of this is that whenever the program runs, it takes different test sets, until it uses the whole dataset. However, for simplicity and study purpose, the test set here is defined by randomly picking 20 percent.

This is achieved by generating a split train/test function and adding it to the dataset. Since this has the drawback that generates new test and train set every time it runs, the data frame train and test will be saved as a comma-separated values (CSV) files and used in the next steps of the analysis. From this step, all the analysis will be done on the train _set data frame. The command used to generate the train/test dataset in SAS Studio 9.4 is as follows.

```
proc surveyselect data=AMETU.IMPORT rat=0.8  
  
out= AMETU.IMPORT_select outall  
  
method=srs;  
  
run;  
  
data AMETU.IMPORT_train AMETU.IMPORT_test;  
  
set AMETU.IMPORT_select;  
  
if selected =1 then output AMETU.IMPORT_train;  
  
else output AMETU.IMPORT_test;  
  
run;
```

The train set contains 838,860 Observations and the test set contains 209,715 Observations. The data has 14 variables with Accident Severity as the dependent variable, and all the other 13 variables as independent.

Variable	Train Dataset		Test Dataset	
	N	N Miss	N	N Miss
Day_of_Week	838860	0	209715	0
Vehicle_Manoeuvre	838860	0	209715	0
Vehicle_Type	838860	0	209715	0
Sex_of_Driver	838860	0	209715	0
Age_Band_of_Driver	838860	0	209715	0
Accident_Severity	838860	0	209715	0
NumberofVehiclesInvolved	838860	0	209715	0
NoofCasualty	827241	11619	206834	2881
Road_Type	838860	0	209715	0
Speedlimitcategory	838860	0	209715	0
Junction_Detail	838860	0	209715	0
Junction_Control	838860	0	209715	0
Light_Conditions	838860	0	209715	0
_1st_Point_of_Impact	838860	0	209715	0

Table 2: Train and test Dataset (source: Own)

Before starting the EDA, it is important to check the duplicates. In case there are, they need to be removed by creating a number of Accident index. From the figure above, there are no duplicates in the train set and even in the whole dataset as the selection was made on purpose to pull only distinct Accident indexes.

It is also important to define the type of variables that are present so to understand them and the influence on the dependent variable and also to determine the statistical methods that will be used. The categorical variables of the dataset are: *Day_of_Week*, *Vehicle_Manoeuvre*, *Vehicle_Type*, *Sex_of_Driver*, *Age_Band_of_Driver*, *Accident_Severity*, *NumberofVehiclesInvolved*, *NoofCasualty*, *Road_Type*, *Speedlimitcategory*, *Junction_Detail*, *Junction_Control*, *Light_Conditions* and *_1st_Point_of_Impact*.

4.2 Explanatory Data Analysis

4.2.1 Univariate Statistical Analysis

It is important to provide a thorough understanding of the variables and how each of them affects the dependent variable before preparing the data for pre-processing. This is accomplished using EDA (explanatory data analysis) techniques. The Univariate Statistical Analysis will be used to evaluate each variable separately at first.

Dependent variable: Accident_Severity

```
proc sgplot data=AMETU.IMPORT_TRAIN;  
  
    vbar Accident_Severity / fillattrs= (color=CX089b26  
    transparency=0.75)  
  
    datalabel;  
  
    yaxis grid;  
  
run;
```

The accident severity has been portrayed in table below and it shows that 86.9%,12.1% and 1% of injuries occur in slight, serious and fatal injuries respectively. From the graph we learned that most of the accidents end up by creating slight accident injuries.

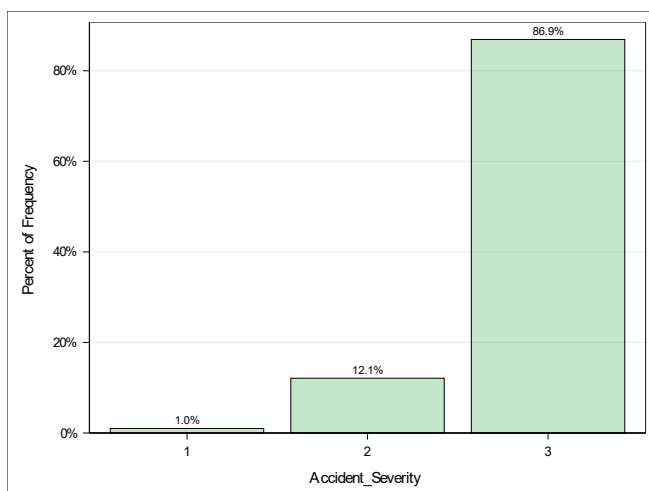


Figure 8: Distribution of dependent variable [source: own]

Using a PROC GENMOD command the dependent variable Accident Severity has a normal distribution. Skewness and kurtosis are used to get a better picture. Skewness is a measure of the asymmetry of the probability distribution of the mean random variable. If it is between 0.5 and 1, this means that the distribution is moderately skewed. The accident severity variable has three values i.e., slight, serious and fatal as it can be seen the most Sevier accident is slight (miner) while only few incidents are registered as fatal accidents.

```
proc means data=AMETU.IMPORT_train chartype mean std n vardef=df
skewness kurtosis;
var Accident_Severity;
run;
```

Analysis Variable: Accident_Severity				
Mean	Std Dev	N	Skewness	Kurtosis
2.8404954	0.4020222	838860	-2.4745503	5.5950011

Table 3: Skewness and Kurtosis of the Distribution of the Dependent Variable [source: own]

(Gould et al., 2010) and (Hair et al. 2010) asserts that, the data is considered normal if the Skewness is between 2 and +2 and the Kurtosis is between 7 and +7. The skewness and kurtosis index were used to determine the data's normality. The result suggested that the data divergence from normality was not significant, as the skewness and kurtosis index values were less than 3 and 10, respectively (Jaff et al., 2011) As a consequence, it is possible to assume that the distribution is moderately skewed. From the described method above, the mean of the Accident Severity is 2.8404954, which measures the center of the distribution of the data.

Variable: Age_Band_of_Driver

This variable is expressed in numerical values rather than actual ones, with each one representing a different age group. From the graph it can be explained that out of the 11 age groups the highest accident-prone age groups are 6, 7 and 8 while quit a significant number of observations show missing data on age group of drivers. When we look at the details of each age group, they represent 26-35 ,36-45 and 46-55 years old respectively. In general, we can conclude that the highest accident-prone age band is between 26-55 years of age.

```

proc sgplot data=AMETU.IMPORT_TRAIN;

    title height=10pt "Distribution of age band of Driver";

    vbar Age_Band_of_Driver / fillattrs=(color=CX089b26 transparency=0.75)

        datalabel;

    yaxis grid;

run;

```

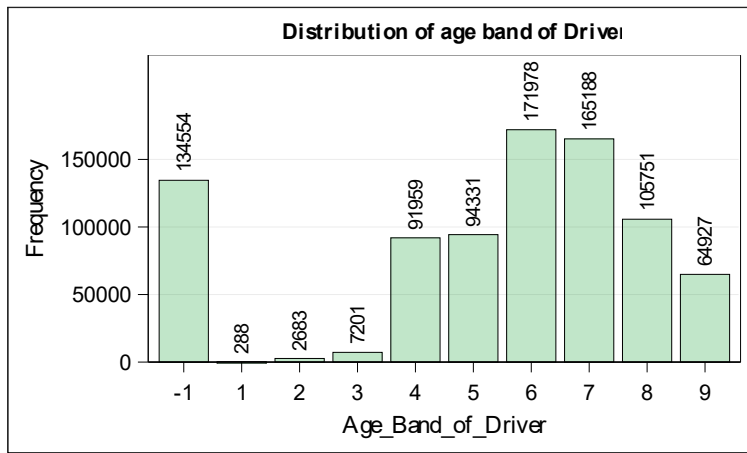


Figure 9: Distribution of Age_Band_of_Driver [source: own]

Variable: Day of the week

The data distribution for the day of the week where accidents occur shows that Friday registered the highest accidents while Sunday records the lowest accident. This is indicative that in most European nations Friday is a good day for relaxation and enjoyment and previous research suggest that in this day's people commit driving while drunk and accident and drunk driving has a direct prerequisite. From the graph we can observe that the distribution of the data on day of accident is evenly distributed.

```

proc sgplot data=AMETU.IMPORT_TRAIN;

    title height=10pt "Distribution of Day of the Week";

    vbar Day_of_Week / fillattrs=(color=CX089b26 transparency=0.75) datalabel;

    yaxis grid;

run;

```

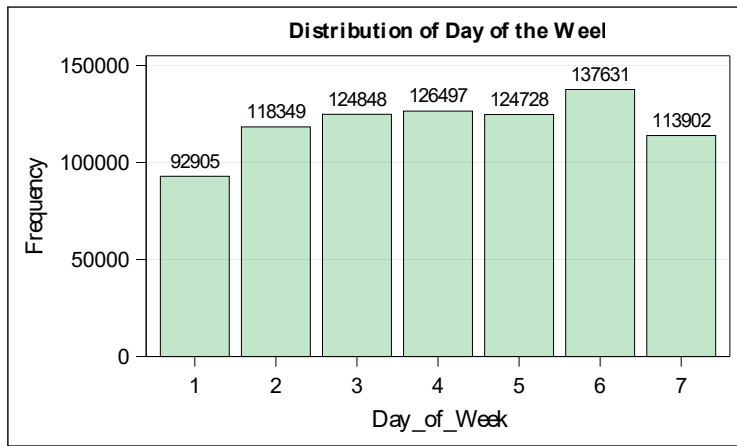



Figure 10: Distribution of Day of the week [source: own]

Variable: Junction Detail

This variable gives the junction details where the accidents happen and have ten categorical values. From the chart below, the highest rate of accident occurs away from the junction or within 20 meters of the junction. On the other hand, the second highest accident occurs on T or staggered junctions while the rest of the accidents happen at the different junction details. This tells us that most of the accidents occur around or close to the junction point or where staggered junction detail exists.

```
proc sgplot data=AMETU.IMPORT_TRAIN;
    title height=10pt "Distribution of Junction_Detail";
    vbar Junction_Detail / fillattrs=(color=CX089b26 transparency=0.75) datalabel;
    yaxis grid;
run;
```

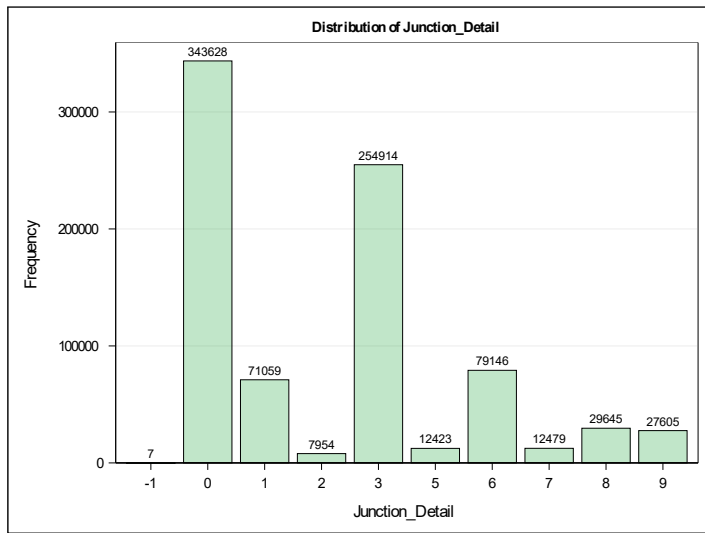


Figure 11: Distribution of Junction Detail [source: own]

Variable: Junction Control

This variable gives the information whether the road has a junction control or not and it has six categorical values where accident happen. The graph depicts that the highest rate of accident occurred with category 6 which is a junction control of Give way or uncontrolled junction control. Right-of-way at Unregulated Intersections occurs at the uncontrolled junction. When two cars arrive at the intersection at roughly the same time, the driver on the left must give way to the driver on the right. When making a left turn, yield to all oncoming traffic even if you were the first one to enter the intersection this is why in our data the highest number of accidents occur at this give way junction control. Large number of observations from the graph also indicates that junction control data has been missed from the data set.

```
proc sgplot data=AMETU.IMPORT_TRAIN;
    title height=10pt "Distribution of Junction_Control";
    vbar Junction_Control / fillattrs=(color=CX089b26 transparency=0.75) datalabel;
    yaxis grid;
run;
```

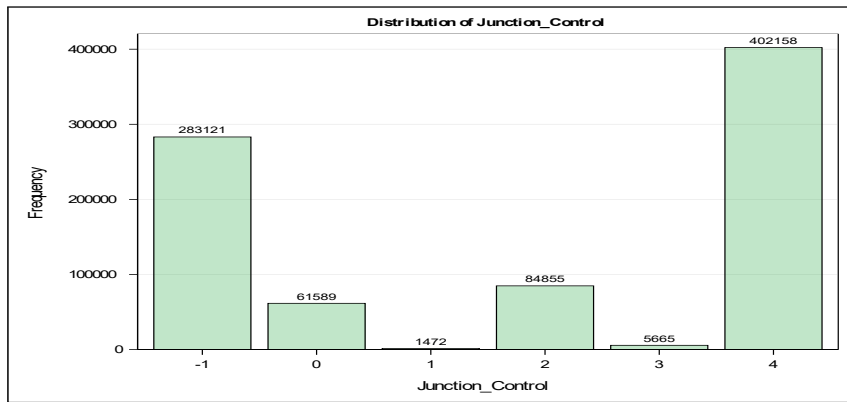


Figure 12: Distribution of Junction Control [source: own]

Variable: Light Conditions

The lighting state on the road during were accidents occurred is expressed with the variable light condition with five distinct light condition values. From the graph it can be explained that accidents by large happen with daylight condition and the lower accident happen on Darkness - lights unlit. From the graph we cannot create conclusive remark with the accidents and the light condition since in a broad day light the number of accidents occurred are quite large and the accidents that happen during darkness is small. This needs further investigation to understand the occurrence of accident and the light condition, yet this is not the intention of this study.

```
proc sgplot data=AMETU.IMPORT_TRAIN;

    title height=10pt "Distribution of Light_Conditions";

    vbar Light_Conditions / fillattrs=(color=CX089b26 transparency=0.75) datalabel;

    yaxis grid;

run;
```

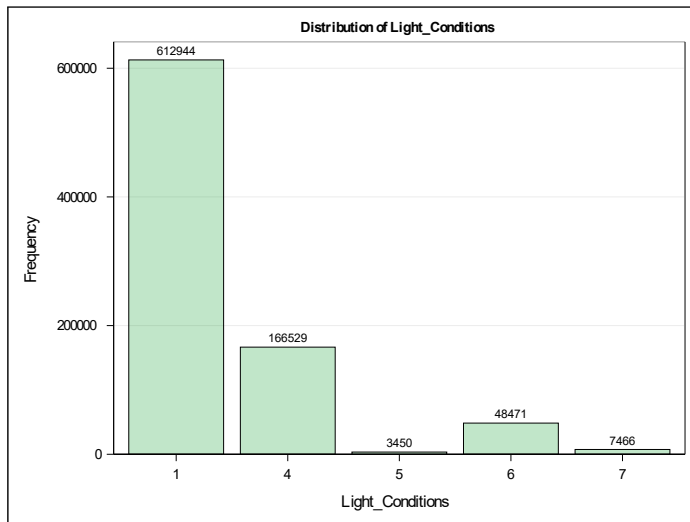


Figure 13: Distribution of Light Conditions [source: own]

Variable: Road Type

This variable shows the road type information with six categorical values. It gives information in what kind of road the accident happens. Based on the graph accidents are very high in Single carriageway road types which is a road with one, two or more lanes arranged within a single carriageway with no central reservation to separate opposing flows of traffic. while the lowest accident registered in an Unknown road type.

```
proc sgplot data=AMETU.IMPORT_TRAIN;

    title height=10pt "Distribution of Road_Type";

    vbar Road_Type / fillattrs=(color=CX089b26 transparency=0.75) datalabel;

    yaxis grid;

run;
```

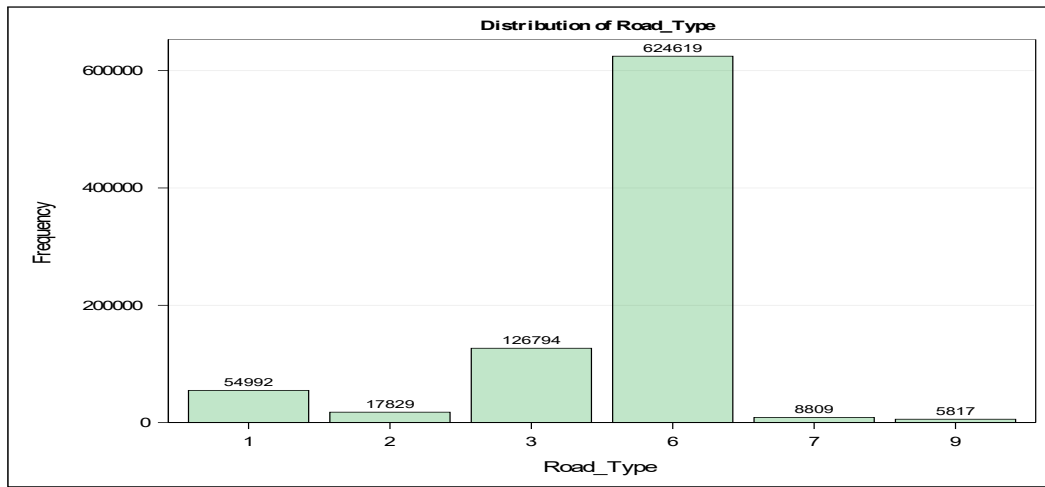


Figure 14: Distribution of Road Type [source: own]

Variable: Sex_of_Driver

Gender information of drivers is explained with four categorical values. By eliminating non expressive categories the variable is left with two generic gender values i.e., male and female. From the chart below, male drivers commit the major accidents. This shows the distribution of gender independently from the accident severity in the next section, this variable will be part of the bivariate analysis as well, where more insights will be gained on who contributes for the different levels of accident severity more.

```
proc sgplot data=AMETU.IMPORT_TRAIN;
    title height=10pt "Distribution of Sex_of_Driver";
    vbar Sex_of_Driver / fillattrs=(color=CX089b26 transparency=0.75) datalabel;
    yaxis grid;
run;
```

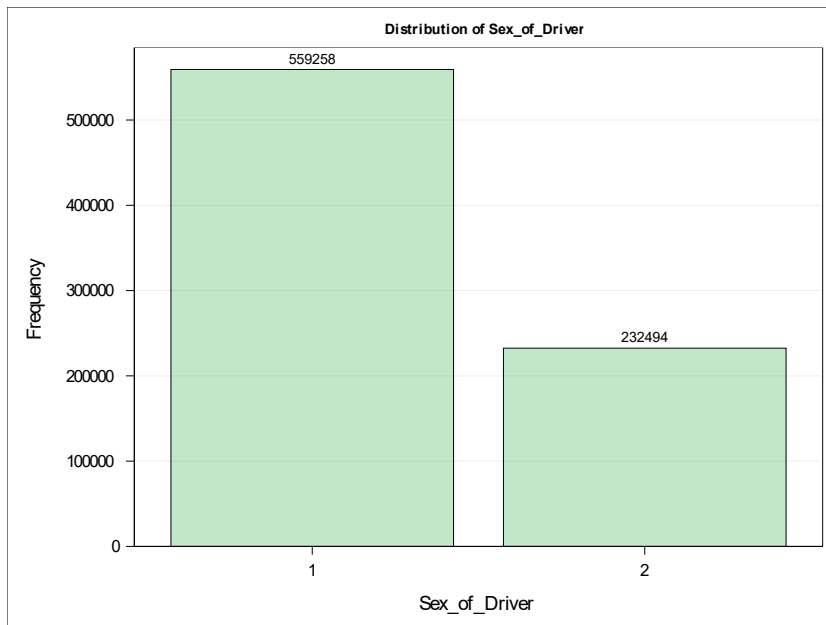


Figure 15: Distribution of Sex_of_Driver [source: own]

Variable: Speed limit category

This variable shows the speed limit in the specific road where the accident happened. The speed limit has eight categories of speeds. From the graph it can be stated that the major accident happened with the speed limit category of 4 where the speed is 30 mph which shows violations of the minimum speed limit to exhibit accidents. A speed limit of 30 miles per hour (mph) or 48 kilometers per hour (km/h) is usually applied in UK where the data is generated. The second frequent accident prevailing speed is 70mph which is the nations (UK) highest speed limit.

```
proc sgplot data=AMETU.IMPORT_TRAIN;
    title height=10pt "Distribution of Speed limit category";
    vbar Speedlimitcategory / fillattrs=(color=CX089b26 transparency=0.75) datalabel;
    yaxis grid;
run;
```

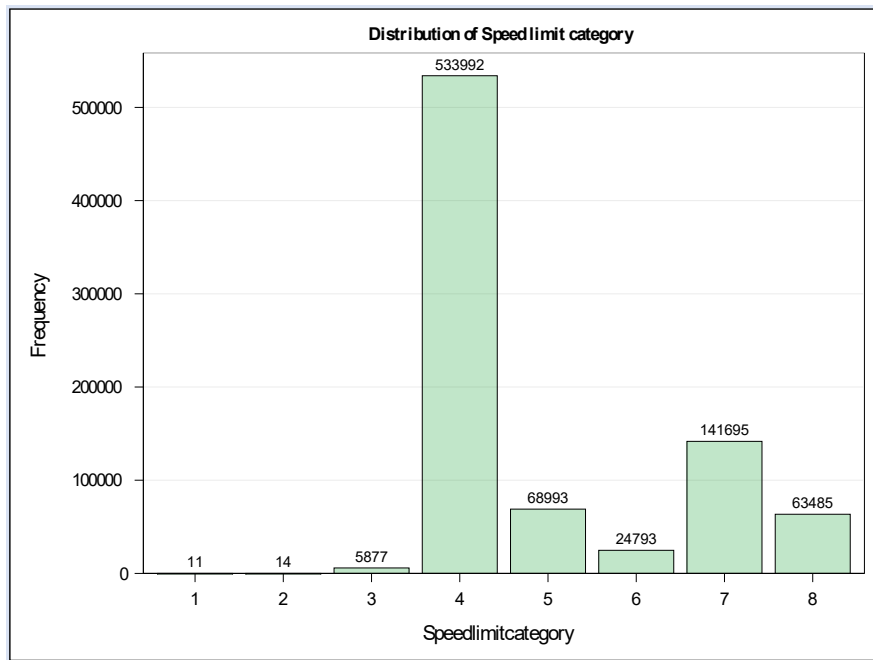


Figure 16: Distribution of Speed limit category [source: own]

Variable: *Vehicle_Manoevre*

A variety of different crash types can occur when vehicles are maneuvering. The Vehicle maneuver consists of 19 categories with various maneuvers to the vehicle. From the graph the majority of accidents occurred at the category of 19 which is Going ahead which is the driver is driving straight ahead and less maneuvering in the vehicle has happened.

```
proc sgplot data=AMETU.IMPORT_TRAIN;
    title height=10pt "Distribution of Vehicle_Manoevre";
    vbar Vehicle_Manoevre / fillattrs=(color=CX089b26 transparency=0.75) datalabel;
    yaxis grid;
run;
```

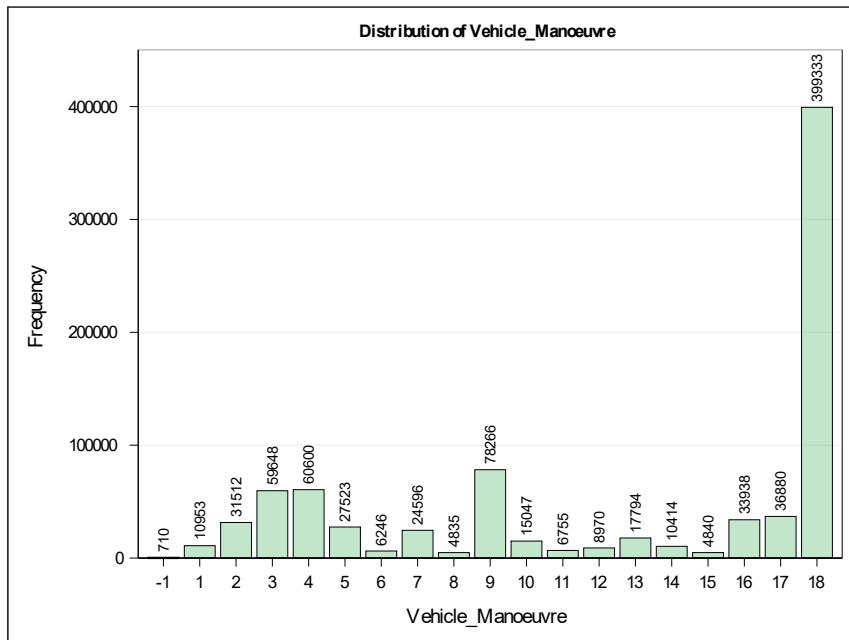


Figure 17: Distribution of Vehicle_Manoevre [source: own]

Variable: Vehicle_Type

The variable information describes the type of vehicle involved in the accident with 21 categorical values. From the graph most of the accidents happen with category value 9 which reflects vehicles types of Car. Cars are the major accident creating vehicle types.

```
proc sgplot data=AMETU.IMPORT_TRAIN;
    title height=10pt "Distribution of Vehicle_Type";
    vbar Vehicle_Type/ fillattrs=(color=CX089b26 transparency=0.75) datalabel;
    yaxis grid;
run;
```

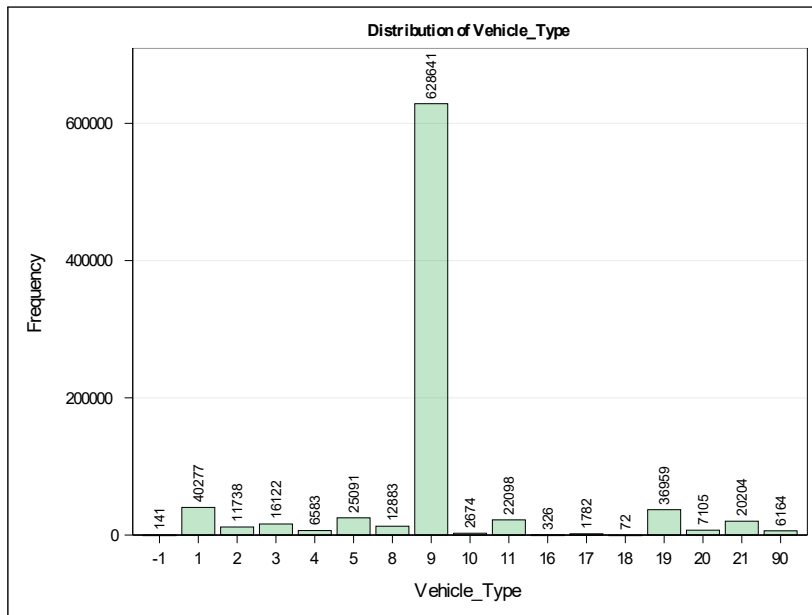



Figure 18: Distribution of Vehicle_Type [source: own]

Variable: *_1st_Point_of_Impact*

The first point of impact is the key factor for responsibility cognizance in related traffic accidents. Typically, the first crash point location between the car and the pedestrian is determined by examination by the police or other accident investigators. The variable has six categorical values and the highest point of impact in the accidents observed at the front. The remaining observations fall under Back, Offside and Nearside respectively with varying proportion.

```
proc sgplot data=AMETU.IMPORT_TRAIN;
    title height=10pt "Distribution of _1st_Point_of_Impact";
    vbar _1st_Point_of_Impact/ fillattrs=(color=CX089b26 transparency=0.75) datalabel;
    yaxis grid;
run;
```

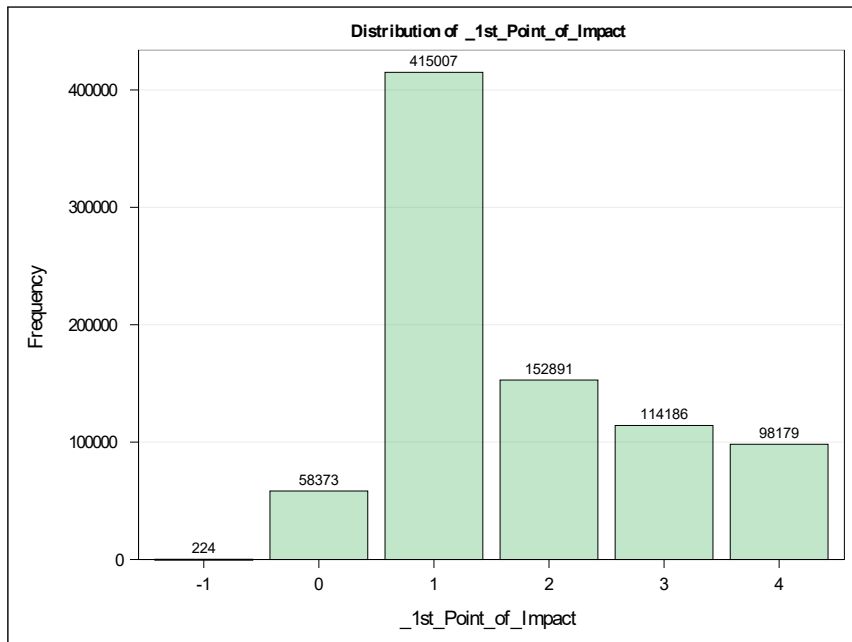


Figure 19: Distribution of *_1st_Point_of_Impact* [source: own]

Variable: No of Casualty

This variable gives the information about the no of causality during the accident. The no of causality variable has 5 categorical values. From the graph the accidents with the major category of no causality are one which shows causality that involves up to 5 individuals. The other categories are very small in observation and hence we can conclude that most of the accidents are resulting in no causality that involves up to five individuals.

```
proc sgplot data=ametu.import_train;

    title height=10pt "Distribution of NoofCasualty";

    vbar NoofCasualty / fillattrs=(color=CX089b26 transparency=0.75) datalabel;

    yaxis grid;

run;
```

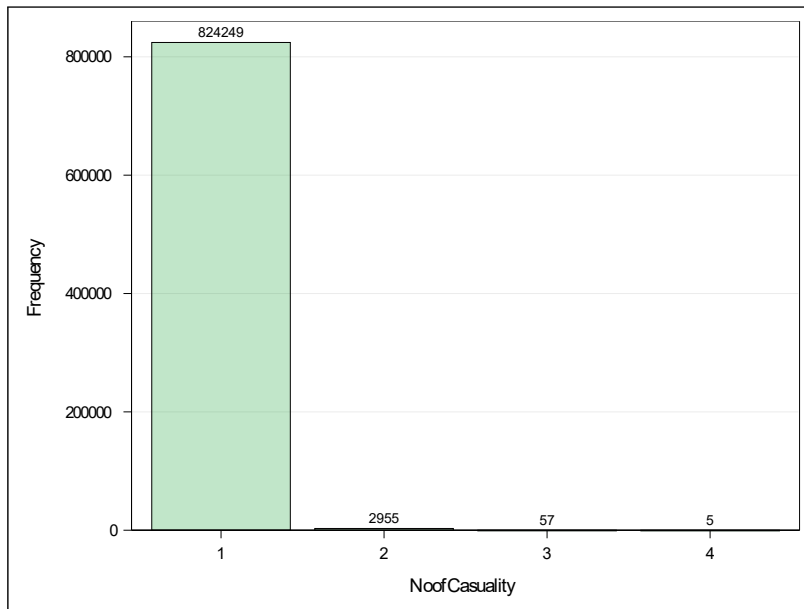


Figure 20: Distribution of No of Casualty [source: own]

Variable: Number of Vehicles Involved

The variable represents the vehicles involved in the accidents with three categories of single, dual and multiple vehicles involved during the crash. From the graph the highest frequency of number of vehicles involved are two vehicles and the next highest is single cars that creates accidents, and the remaining accidents involve multiple vehicles during the accident.

```
proc sgplot data=ametu.import_train;
    title height=10pt "Distribution of Number of Vehicles Involved";
    vbar NumberofVehiclesInvolved / fillattrs=(color=CX089b26 transparency=0.75)
    datalabel;
    yaxis grid;
run;
```

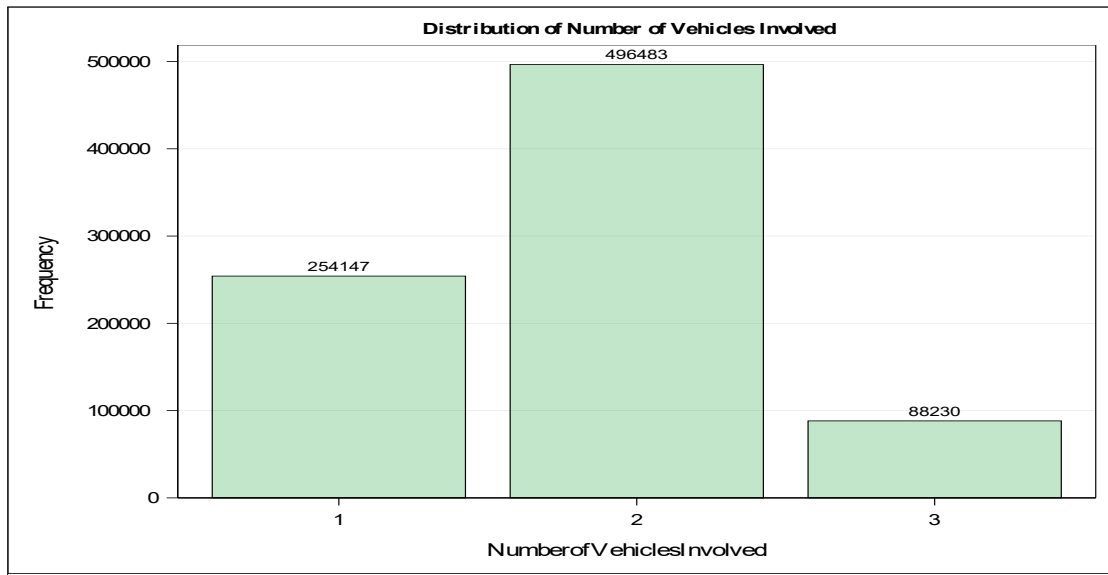


Figure 21: Distribution of Number of Vehicles Involved [source: own]

4.2.2 Bivariate statistical analysis

After analyzing each variable independently, it is very important to understand the relationship of the target variable with the independent ones. This is done through the bivariate analysis. The Accident Severity variable will be evaluated with the *sex of the driver*, *age of the driver*, *speed limit category*, *light condition* and *road type*.

Accident severity and Sex_of_Driver

From the bar chart 59.9 % accident severity resulting in slight injury are committed by male drivers while 27.1 % are committed by female drivers on the other hand the accident severity that resulted in serious injury outcomes are occurred by male drivers with 8.3 % and 3.8% with female drivers. Fatal accidents are observed in a lower proportion in the dataset and the observation resulted in only 0.7% and 0.3% with male and female respectively. The observation from the graph suggested that the accidents that resulted in all the three accident severity categories are occurred by male drivers and this is supported by multiple study done in traffic accident-related study.

```

proc sgplot data=AMETU.IMPORT;

    vbar Accident_Severity / group=Sex_of_Driver groupdisplay=cluster datalabel

        stat=percent;

    yaxis grid;

    keylegend / location=inside;

run;

```

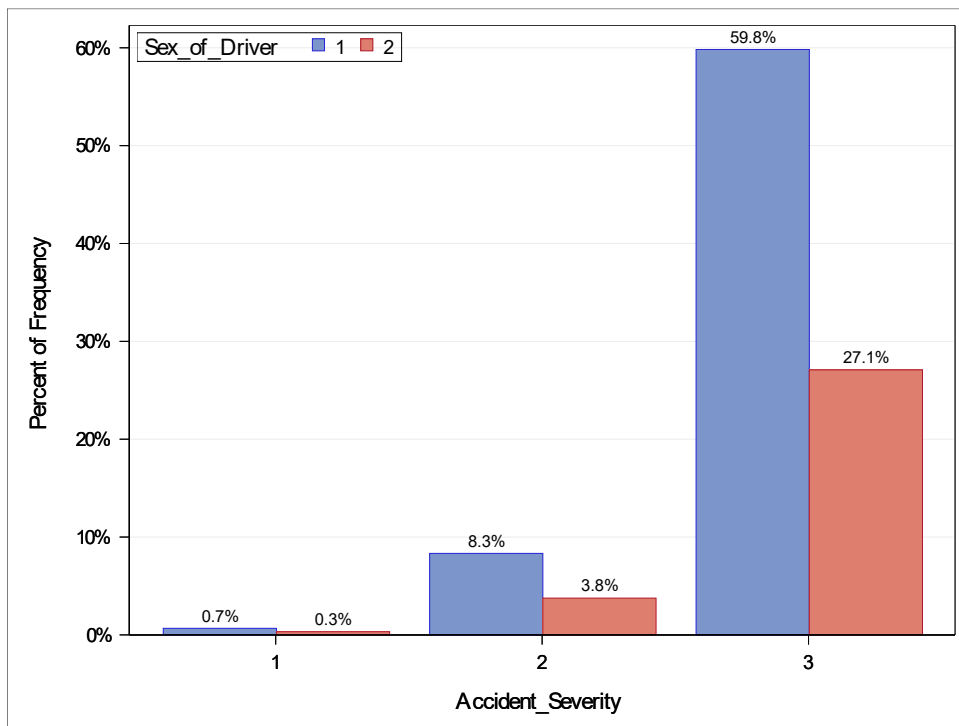


Figure 22: Accident severity and Sex_of_Driver [source: own]

Accident severity and Age_Band_of_Driver

From the bar chart below, it is obvious, that the age category with the highest number of Accidents occurred by drivers on age group between 16-55 this range will cover more than 10 % of accidents resulted in slight accidents. Also, the total number of accident severity in which that has recorded the largest number is age group between 26-45 with close to 19% of happening. The remaining two categories of accident severity that has serious and fatal injuries are relatively small in percentage. But still the age group between 26-45 is the one that causes most of these injuries.

```

proc sgplot data=AMETU.IMPORT;

    vbar Accident_Severity / group=Age_Band_of_Driver groupdisplay=cluster

        fillattrs=(transparency=0.25) datalabel stat=percent;    xaxis display=(nolabel);

    yaxis grid;

run;

```

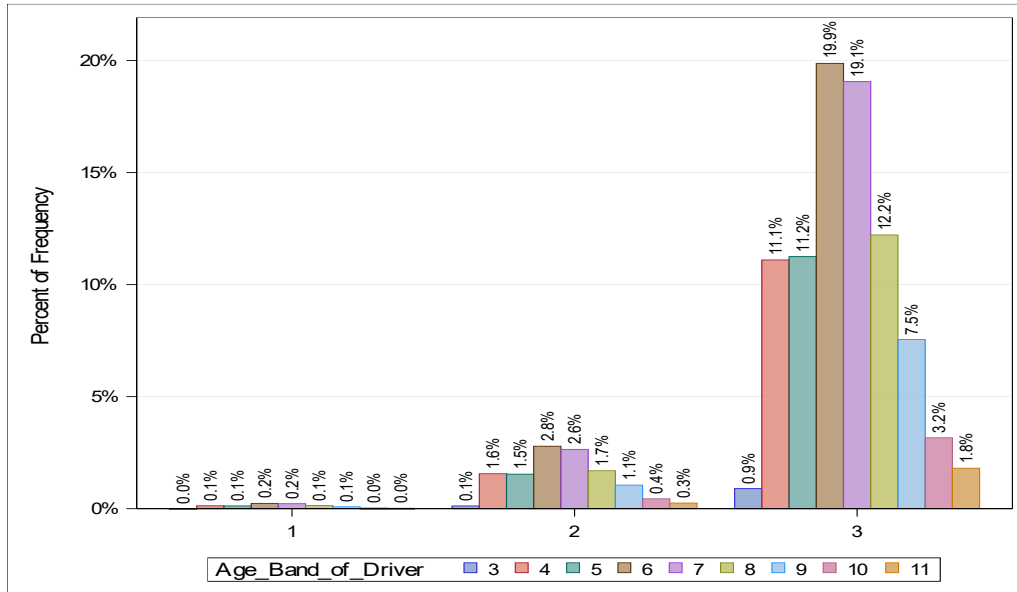


Figure 23: Accident severity and Age_Band_of_Driver [source: own]

Accident severity and Speed limit Category

When accident severity and speed limit analyzed it shows that the 30-mph speed limit category is the highest speed category that is associated with accident among the severity category that resulted in all the three severity categories i.e., slight, serious and fatal with 62.1%, 8.1% and 0.4% respectively.

```

proc sgplot data=AMETU.IMPORT;

    vbar Accident_Severity / group=Speedlimitcategory groupdisplay=cluster

        datalabel;

    yaxis grid;

run;

```

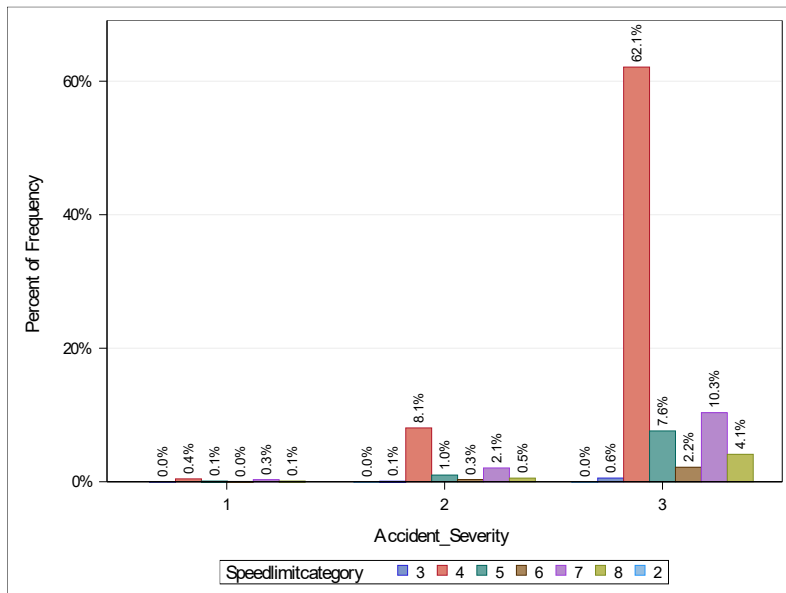


Figure 24: Accident severity and Speed limit Category [source: own]

Accident severity and Light_Conditions

Light condition during accident evaluated with the level of accident severity and resulted that out of the three categories slight accident occur 66% during day light and 18.4 % during darkness while lights lit. The two other categories of serious and fatal injuries are happened during day light with 8.4% and 0.6 % whereas during darkness with lights lit 3% and 0.3% respectively.

```
proc sgplot data=AMETU.IMPORT;
    vbar Accident_Severity / group=Light_Conditions groupdisplay=cluster
    datalabel;
    yaxis grid;
run;
```

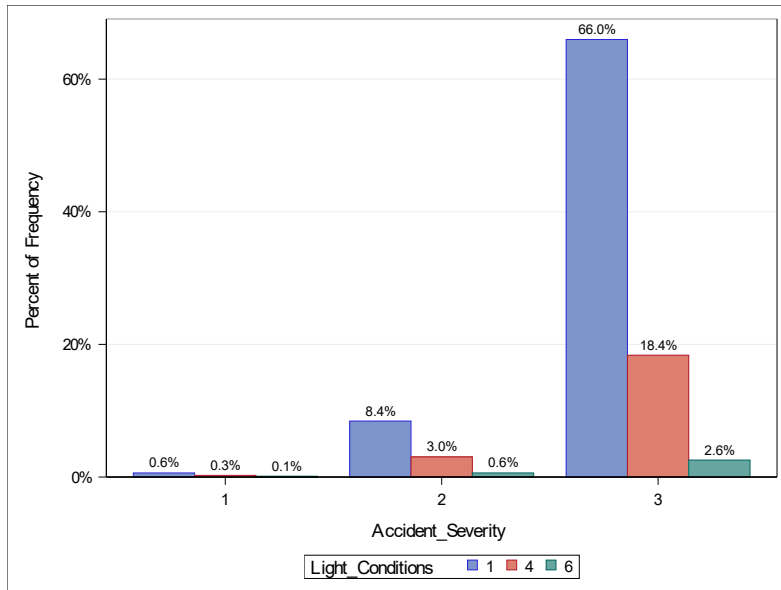


Figure 25: Accident severity and Light_Conditions [source: own]

Accident severity and Road_Type

Accident severity level varies with the road type from the graph the majority of slight accident happen with a road type of Single carriageway and also dual carriage way road type hold the second highest slight accidents. Accident severity with slight injuries occur with 64 % in Single carriageway road type and 10.7 % in dual carriageway road type. On the other hand, serious and fatal injuries happen with 9.5% and 0.7% single carriageway while and 1.4 and 0.2 % in dual carriage way road type respectively.

```
proc sgplot data=AMETU.IMPORT;

    vbar Accident_Severity / group=Road_Type groupdisplay=cluster datalabel;

    yaxis grid;

run;
```

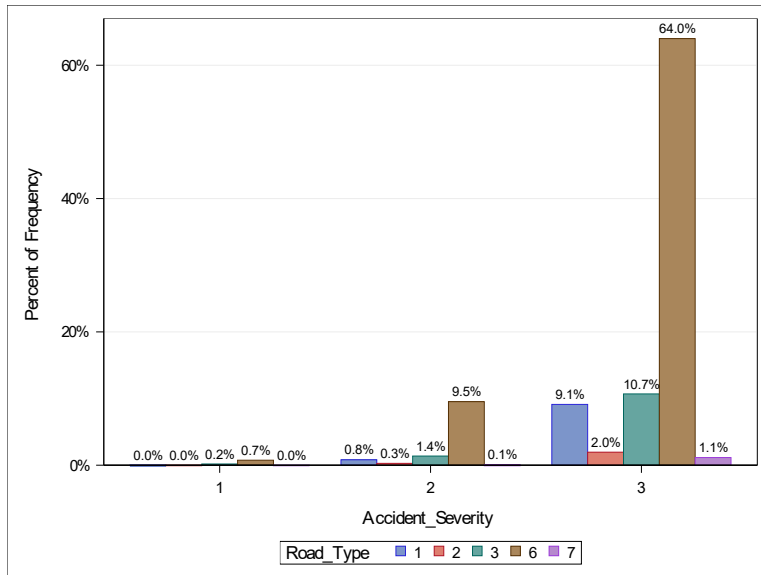



Figure 26: Accident severity and Road_Type [source: own]

4.2.3 Correlation between the dependent variable and the independent variable

Correlation analysis is a mathematical methodology for assessing the strength of relationship between two quantitative variables. A high correlation implies a close relationship between two or more variables, while a low correlation indicates that the variables are barely related. In other words, it is the method of measuring the strength of the relationship using available statistical data.

The following code is used from the correlation method to generate the table below.:

```
proc corr data=ametu.import;
    var Day_of_Week Vehicle_Manoevre Vehicle_Type Sex_of_Driver
    Age_Band_of_Driver NumberofVehiclesInvolved NoofCasualty Road_Type
    Speedlimitcategory Junction_Detail Junction_Control Light_Conditions
    _1st_Point_of_Impact;
with Accident_Severity;
run;
```

Spearman Correlation Coefficients	
Prob > r under H0: Rho=0	
Number of Observations	
	Accident_Severity
Vehicle_Manoeuvre	0.00029 0.8345
Vehicle_Type	0.00069 0.6169
Sex_of_Driver	-0.00034 0.8065
Age_Band_of_Driver	0.00058 0.6725
NumberofVehiclesInvolved	0.11224 <.0001
NoofCasualty	-0.02985 <.0001
Day_of_Week	0.00052 0.7045
Road_Type	-0.03434 <.0001
Speedlimitcategory	-0.05259 <.0001
Junction_Detail	0.00343 0.0129
Junction_Control	0.01845 <.0001
Light_Conditions	-0.05630 <.0001
_1st_Point_of_Impact	-0.00119 0.3888

Table 4: Spearman Correlation Coefficients

Surprisingly, there is no strong correlation between Accident Severity and any of the other variables. When the coefficient is equal to 1.0, the correlation is higher. There is a statistically significant negative relationship between Accident Severity and Vehicle_Type, Sex_of_Driver, NoofCasualty Road_Type, Speedlimitcategory, Light_Conditions_1st_Point_of_Impact. This is one of the many challenges in analyzing Big Data, what to do when there is no strong correlation between the target variable and any of the independent variables. One choice is to substitute empty values with the mean or median, or to delete variables with more than half of the values empty. Multicollinearity, on the other hand, is the probability of a relationship between predictors. This should not exist because it causes problems to fit the model and also to 'trust' statistically important variables.

While reviewing these results it is mandatory to see if any of the variables are statistically significant to include them in the models to be generated. As can be seen, upon review of this spearman correlation table, there are only seven variables that are statistically significant. The variables that are significant are indeed have given meaning to the fact that accident severity has to do with *NumberofVehiclesInvolved*, *NoofCasualty*, *Road_Type*, *Speedlimitcategory*, *Junction_Detail*, *Junction_Control* and *Light_Conditions*. The analysis for multicollinearity can be examined through the Variance Inflation Factor and Tolerance. This can be done by specifying the “vif”, “tol”, and “collin” options after the model statement:

```
proc reg data=ametu.import;
model Accident_Severity = Day_of_Week Vehicle_Manoevre Vehicle_Type
Sex_of_Driver Age_Band_of_Driver NumberofVehiclesInvolved NoofCasualty Road_Type
Speedlimitcategory Junction_Detail Junction_Control Light_Conditions
_1st_Point_of_Impact / vif tol collin;
title 'Accident Dataset - Multicollinearity Investigation
of VIF and Tol';
run;
```

Variable	Pr > t	Tolerance	Variance Inflation
Intercept	<.0001	.	0
Day_of_Week	0.1593	0.99963	1.00037
Vehicle_Manoeuvre	0.6396	0.98394	1.01632
Vehicle_Type	0.3241	0.99725	1.00276
Sex_of_Driver	0.5989	0.88626	1.12834
Age_Band_of_Driver	0.3581	0.88760	1.12664
NumberofVehiclesInvolved	<.0001	0.95907	1.04267
NoofCasualty	<.0001	0.99599	1.00402
Road_Type	<.0001	0.89150	1.12170
Speedlimitcategory	<.0001	0.86368	1.15783
Junction_Detail	0.0982	0.50783	1.96915
Junction_Control	<.0001	0.49366	2.02570
Light_Conditions	<.0001	0.97570	1.02491
_1st_Point_of_Impact	0.0811	0.98728	1.01289

Table 5: Multicollinearity test using tolerance and VIF (Source: Own)

As can be seen from the table the value of tolerance for any given variable which is less than 0.1 are not present in the table which reflects that there is no treat for multicollinearity.

4.3 Model building and training

4.3.1 Custom transformation of the data

Some data transformations have already been completed as part of the EDA. This one involved the drop of the variables *Day_of_Week*, *Vehicle_Manoevre*, *Vehicle_Type*, *Sex_of_Driver* *Age Band* and *1st_Point_of_Impact* as agreed that they are not statistically significant for accident severity. Other suggestions for the other variables are as follows:

- There are 6 categories with junction control which are many. By checking the frequency, they can be reduced to 3.
- There are 6 light condition categories which can be compact. By checking the frequency, they can be reduced to 4.
- There are 4 No of causality categories which can be reduced. By checking the frequency, they can be limited to 3.

Before starting the data modeling, it is a best practice to join the test set with the train set. The only reason to do this, is for not having to repeat the steps again for both datasets.

for not having to repeat the steps again for both datasets.

```
data AMETU.import_train;  
set AMETU.import_train;  
if Variables = (insignificant and missing values) then delete;  
run
```

The result after passing multiple transformation is shown below and it contains the variables that pass through multiple transformation with the SAS code that shows above.

Simple Statistics						
Variable	N	Mean	Std Dev	Sum	Minimum	Maximum
Accident_Severity	525898	2.85916	0.37539	1503625	1.00000	3.00000
NumberofVehiclesInvolved	525898	1.85316	0.55912	974575	1.00000	3.00000
NoofCasualty	520662	1.00323	0.05673	522343	1.00000	2.00000
Road_Type	525898	5.05667	1.75417	2659292	1.00000	7.00000
Speedlimitcategory	525898	4.70378	1.26759	2473708	2.00000	8.00000
Junction_Detail	525898	3.50548	2.45440	1843527	0	9.00000
Junction_Control	525898	3.24943	1.34857	1708868	0	4.00000
Light_Conditions	525898	1.81497	1.45305	954488	1.00000	6.00000

Table 6: List of variables after transformation (Own source)

As can be seen from the code, there are a few variables which have missing values or out of range values. They are removed using the above code so at last all missing values are removed from the data set. All these transformations can be done also automatically through the transformation pipelines found in SAS studio trial package, which helps with these automatic and sequential transformations. After all the necessary changes has been applied, the dataset will be split again into test set and train set and validation set using for training the models.

5. Results and Discussions

Introduction

This chapter shows the results obtained from SAS Model studio for all the three models. Results of all models from SAS Model studio is recorded and an analysis is carried out to compare the prediction Capabilities of the three competing models in accordance with three important measures Accuracy, Root Mean Squared Error and Area under ROC.

5.1 Model building and evaluation

The models are tested using different performance measurement parameters notably Accuracy, Root mean square error and ROC. Accuracy is the proportion of observations that are correctly classified as either an event or non-event, calculated at various cutoff values. Root Mean Square Error (RMSE) is the standard deviation of the residuals (prediction errors). The RMSE is a measure of how distributed the residuals are. In other words, it shows how closely the data is along the line of best fit. A receiver operating characteristic curve (ROC curve) is a graph that shows how well a classification model performs over all classification thresholds.

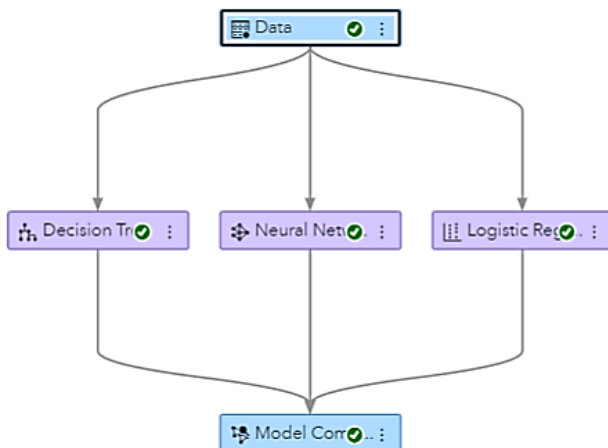


Figure 27: SAS Model studio Pipeline [source: own]

Descriptions of the variables are provided in Table 7. The data used in this model were derived from accident dataset used in the previous chapter. The data passes through data cleaning and data pruning. Using spearman correlation statistically significant variables were selected. Among the explanatory variables (independent variables), road related, Vehicle related, and accident-related

classifications were made. The 8 variables that have significance and fall under the three classification has been explained in the table below.

The data set has been split as follows a common split is 50% for training, 25% for validation, and 25% for testing.

Classification	Description	Category coding value
Vehicle Related	NumberofVehiclesInvolved	1="Single", 2="Double" and 3="Multiple"
	Speedlimitcategory	2="15 mph", 3="20 mph", 4="30 mph" 5="40 mph", 6="50 mph", 7="60 mph" 8="70 mph"
Accident Related	NoofCasualty	1="up to 5" and 2="6 to 15"
Road Related	Junction_Control	0="Not at junction or within 20 meters" 2="Auto traffic signal" 4="Give way or uncontrolled"
	Junction_Detail	0="Not at junction or within 20 meters" 1="Roundabout", 2="Mini-roundabout" 3="T or staggered junction", 5="Slip road" 6="Crossroads", 7="More than 4 arms" 8="Private drive or entrance", 9="Other junction"
	Road_Type	1="Roundabout", 2="One way street" 3="Dual carriageway", 6="Single carriageway" 7="Slip road"
	Light_Conditions	1="Daylight", 4="Darkness - lights lit" 6="Darkness - no lighting"

Table 7: Variable Classifications

Logistics regression

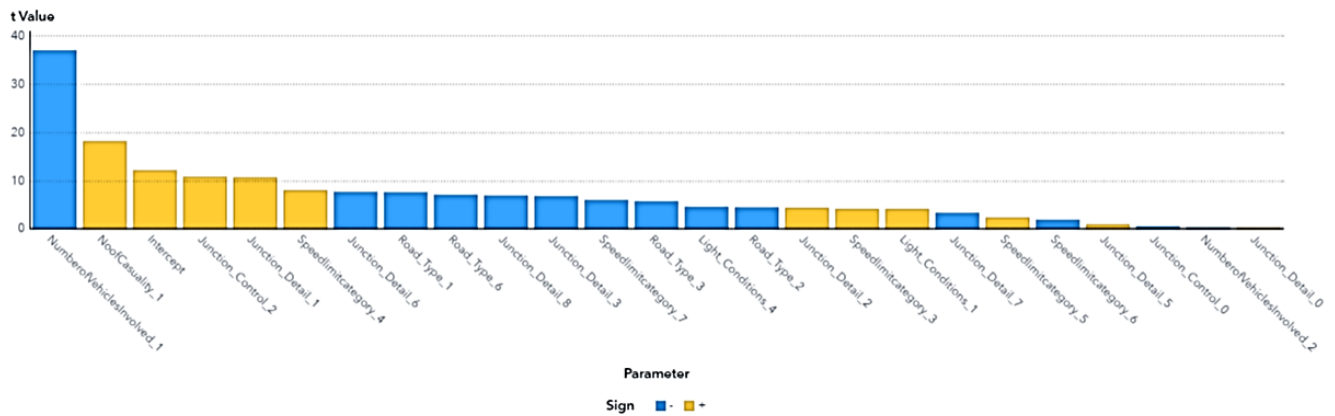


Figure 28: t Values by Parameter [source: own]

This plot displays the absolute value of the t value for each parameter estimate in the logistic regression model. Larger values indicate more significant parameters. The bar that represents the parameter is colored by the sign of the estimate. Bars that are colored as positive (+) correspond to a positive parameter estimate, which indicates an increase in the predicted probability of the event as the parameter value increases. Bars that are colored as negative (-) correspond to a negative parameter estimate, which indicates a decrease in the predicted probability of event as the parameter value increases. The most significant parameter is NumberofVehiclesInvolved_1 which involves only single vehicle in the accident with a t value of -37.0318 which means level accident severity will decrease in the predicted probability of event when the no of vehicle involved is only one. The trailing parameters were NoofCasualty_1 (Up to 5 peoples), Junction_Control_2 (Auto traffic signal) and Junction_Detail_1(Roundabout) with t values of +18.158, +0.810 and +10.6474.

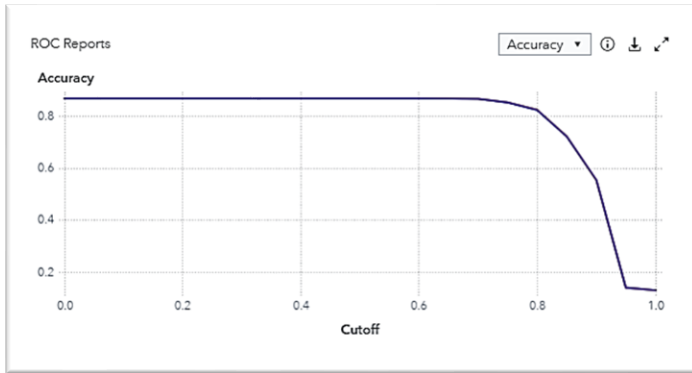


Figure 29: Accuracy Reports for Logistic regression [source: own]

At each cutoff value, the predicted target classification is determined by whether P_Accident_Severity3 (non-critical incident), which is the predicted probability of the event "3" for the target Accident_Severity, is greater than or equal to the cutoff value. For this model, the accuracy in the TEST partition at the cutoff of 0.5 is 0.8691 for the TRAIN partition at the cutoff of 0.5 is 0.8691 and for the VALIDATE partition at the cutoff of 0.5 is 0.8692. Overall, the accuracy of the model to predict the probability of accident severity is on the acceptable range

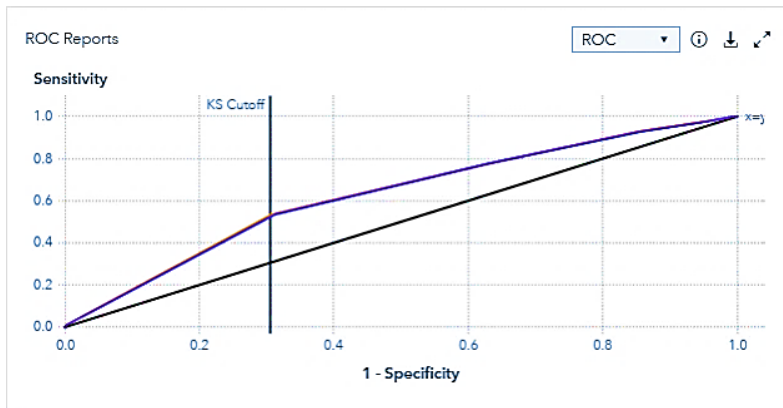


Figure 30: ROC Reports for Logistic regression [source: own]

The shape of the curve contains a lot of information, including what we might care about most for a model, the expected false positive rate, and the false negative rate. Smaller values on the x-axis (KS cutoff) of the plot indicate lower false positives and higher true negatives.

Higher true positives and lower false negatives are shown by larger value on y-axis (plot accuracy) values. when we predict a binary result, it is either correct prediction (true positive) or incorrect

(false positive) There is a conflict between these options, just as there is between true negative and false negative.

For the VALIDATE partition the KS Cutoff line is drawn at the cutoff value 0.9, where the 1-specificity value is 0.306 and the sensitivity value is 0.534.

The ROC curve that rapidly approaches the upper-left corner of the graph, where the difference between sensitivity and 1-specificity is the greatest, indicates a more accurate model. A diagonal line where sensitivity = 1-specificity indicates a random model.

From the results, it looks the prediction is quite accurate. However, this has to be compared with the other models, to be able to come to a conclusion on which model is the best one for predicting such variable.

Neural Network

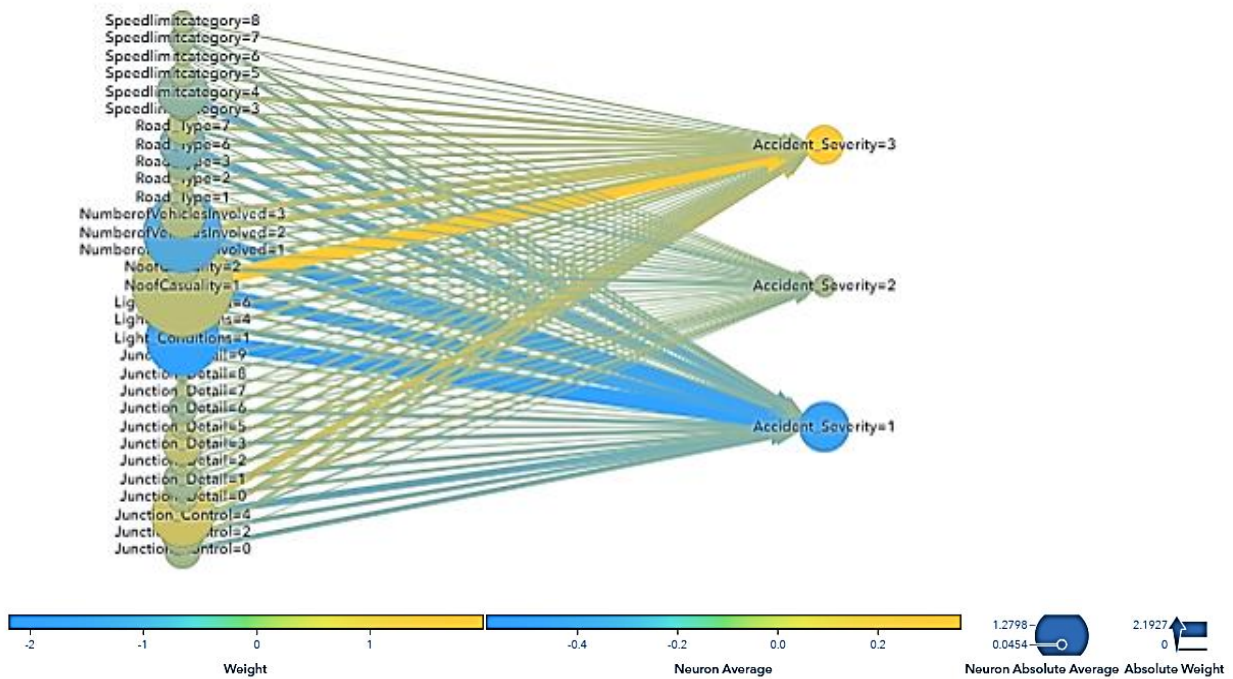


Figure 31: Neural Network Diagram for Accident Severity [source: own]

The Neural Network node creates a statistical model that is designed to mimic the biological structures of the human brain. Neural networks consist of predictors (*Speed limit category, Road type, number of vehicles involved, no of causality, light condition, junction detail and junction*

control), hidden layers, an output layer (Accident severity 1,2and 3), and the connections between each of those.

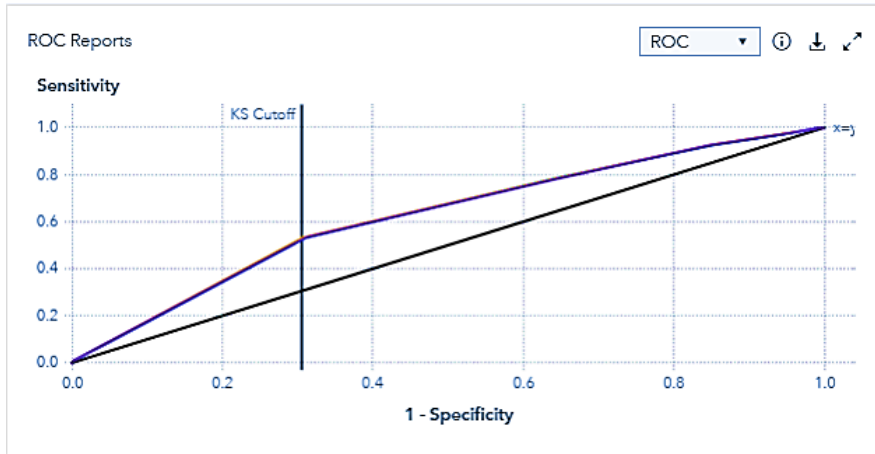


Figure 32: ROC report for Neural Network [source: own]

For the VALIDATE partition. The KS Cutoff line is drawn at the cutoff value 0.9, where the 1-specificity value is 0.306 and the sensitivity value is 0.532. From the results, it looks the prediction is lower than the logistic regression. However, this has to be compared with the remaining model, and also comparing using other performance measuring parameters to be able to come to a conclusion on which model is the best one for predicting such variable.

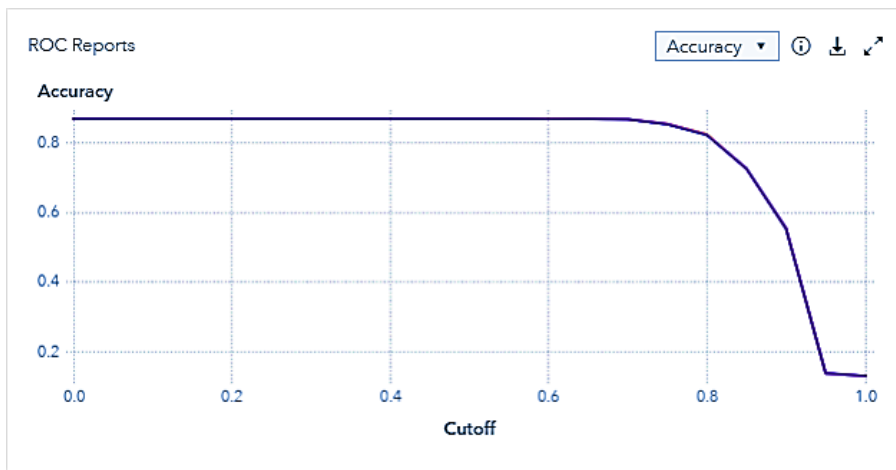


Figure 33: Accuracy report for Neural Network [source: own]

For this model, the accuracy in the TEST partition at the cutoff of 0.5 is 0.8692, for the TRAIN partition at the cutoff of 0.5 is 0.8691 and for the VALIDATE partition at the cutoff of 0.5 is 0.8692. The accuracy level is better than logistic regression but yet conclusive decision is not attain at this level and further analysis will be carried out to get the champion model.

Decision tree

Decision tree classifiers are one of the most popular and used classification techniques because the tree is constructed from the given data based on simple equations and uses the attribute selection measures such as a gain ratio measure, which ranks the attributes and determines the most useful attribute, and accordingly the researcher can recognize the most efficient attributes on the predicted purpose (Han *et al.*, 2011).

This figure shows how the misclassification rate changes for subtrees, which are created by pruning the full decision tree to various numbers of leaves. The training error decreases as the number of leaves increases, so the VALIDATE partition can be used to prune the tree to prevent overfitting. For this decision tree model, the selected subtree based on the pruning options has 8 leaves with a misclassification rate of 0.1209 for the VALIDATE partition.

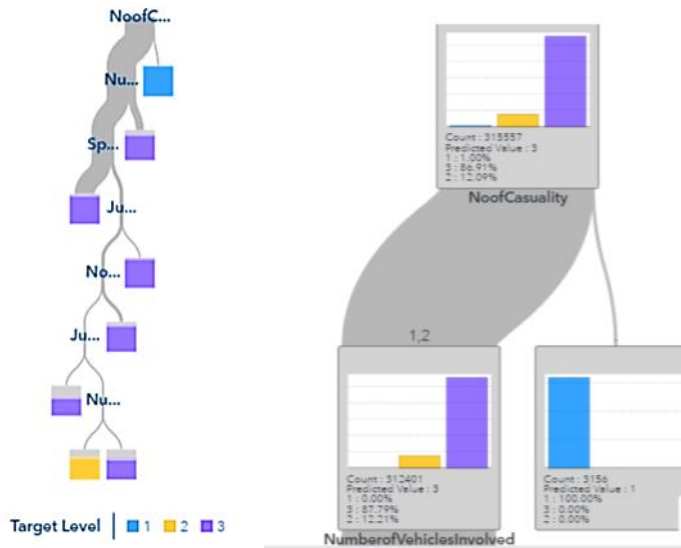


Figure 34: Decision Tree Diagram for Accident Severity [source: own]

For the VALIDATE partition the KS Cutoff line is drawn at the cutoff value 0.85, where the 1-specificity value is 0.461 and the sensitivity value is 0.704. From the results, it looks the prediction

is superior compared with the previous two models. However, this has to be compared with the models using other performance measuring parameters to be able to come to a conclusion on which model is the best one for predicting such variable.

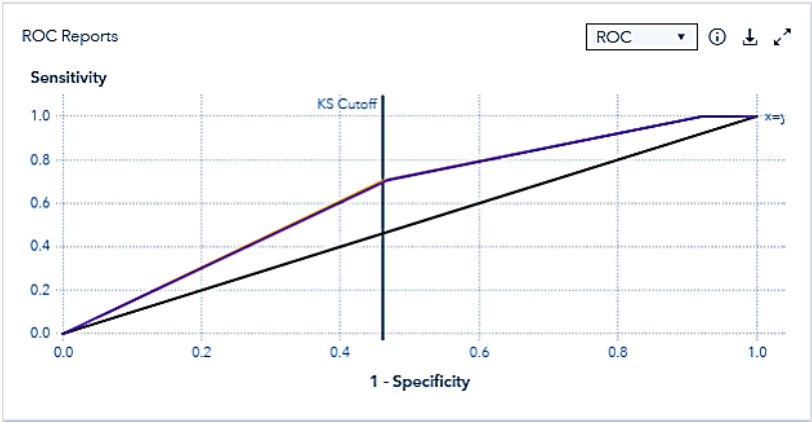


Figure 35: ROC report for Decision Tree [source: own]

For this model, the accuracy in the TEST partition at the cutoff of 0.5 is 0.8789, for the TRAIN partition at the cutoff of 0.5 is 0.8791 and for the VALIDATE partition at the cutoff of 0.5 is 0.8791. Also with this test the model shows a superior accuracy than both logistic regression and neural network models.

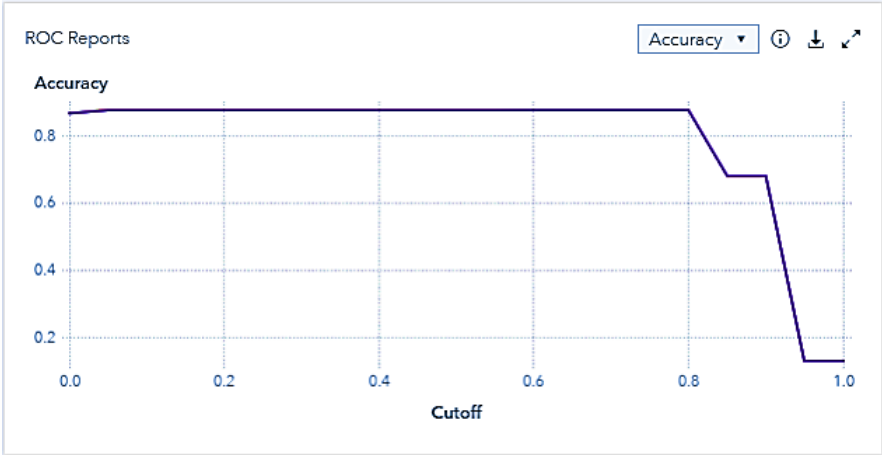


Figure 36: Accuracy report for Decision Tree [source: own]

From the figure above, both accuracy and ROC show higher model for the decision tree so far as it has the highest one. However, there is still needed to run a Model Comparison using various model performance measures which will be checked in the same way.

Model Comparison using accuracy curve and ROC curve

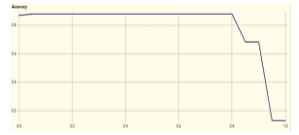
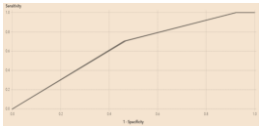

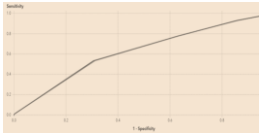
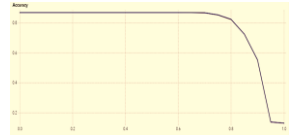
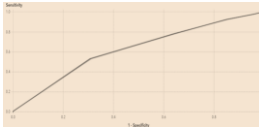
Models	Accuracy Curve	ROC Curve
Decision Tree	 The accuracy curve for the Decision Tree model shows a high accuracy of approximately 0.88 for the first 14 features. After the 14th feature, the accuracy drops sharply to about 0.25 by the 15th feature.	 The ROC curve for the Decision Tree model shows a curve that is significantly above the diagonal line, indicating good predictive performance. The area under the curve is approximately 0.75.
Logistic Regression	 The accuracy curve for the Logistic Regression model shows a high accuracy of approximately 0.88 for the first 14 features. After the 14th feature, the accuracy drops sharply to about 0.25 by the 15th feature.	 The ROC curve for the Logistic Regression model shows a curve that is significantly above the diagonal line, indicating good predictive performance. The area under the curve is approximately 0.75.
Neural Network	 The accuracy curve for the Neural Network model shows a high accuracy of approximately 0.88 for the first 14 features. After the 14th feature, the accuracy drops sharply to about 0.25 by the 15th feature.	 The ROC curve for the Neural Network model shows a curve that is significantly above the diagonal line, indicating good predictive performance. The area under the curve is approximately 0.75.

Table 8: ROC and Accuracy comparison for the three models [source: own]

The champion model for this project is Decision Tree. The model was chosen based on the KS (Youden) for the Test partition (0.24). 87.89% of the Test partition was correctly classified using the Decision Tree model. If we are looking closely the three performance measures i.e., the RMSE, Misclassification rate, KS and the Area under ROC curve we learned that the four most important factors are NoofCasualty, NumberofVehiclesInvolved, Speedlimitcategory, and Junction_Detail.

Model Comparison

Performance Measures	Model Name		
	Decision Tree	Logistic Regression	Neural Network
Misclassification Rate	0.121075551	0.130868846	0.13084983
Average Squared Error	0.06938452	0.074901175	0.074870785
Misclassification Rate	0.121075551	0.130868846	0.13084983
KS (Youden)	0.237753987	0.22343634	0.222359482
F1 Score	0.93487844	0.92998128	0.929990741
Area Under ROC	0.630098447	0.624155287	0.622885697
Root Average Squared Error	0.263409415	0.27368079	0.273625263
Accuracy	0.878924449	0.869131154	0.86915017

Table 9: Model Comparison

From the table above the performance measures score different values the root average squared error and misclassification rate lower values reflect that the model is better in this aspect decision tree has the lowest score on both measures on the other hand the area under the ROC and KS highest value reflects that the model is best and decision tree gets the highest on both accounts. From the analysis and a step wise comparison using the different performance measures the best model happen to be the decision tree.

Variable Importance

The variable importance plot is a crucial output of the Decision tree model. This plot shows the role of each variable in classifying the data for each variable in the matrix and on the y-axis is each variable, and on the x-axis is its important. The most critical variables are mentioned first, followed by the least important. The mean decrease in accuracy is how much the model fit decreases when a variable is dropped. The greater the drop, the more significant the variable.

The top two feature variables in the Decision tree importance list (figure 37) are the No of Causality, which is an Accident-related factor, and No of vehicles involved, which is a vehicle related factor.

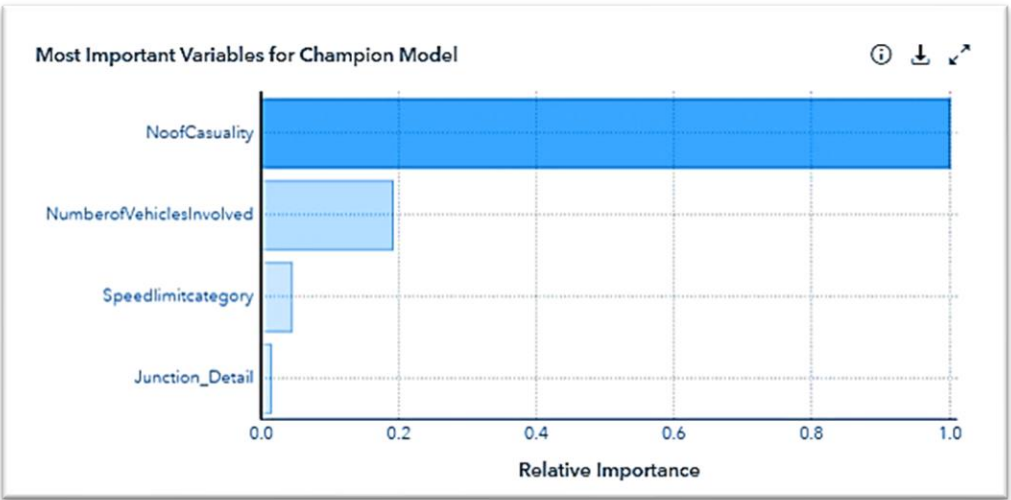


Figure 37: Most important Variables for champion model [source: own]

These results presume the probability of accident severity to be higher when multiple vehicles are involved. The variable that is ranked third is speed limit, which is also vehicle related factor. The fourth ranked variable is junction detail which is a road related factor. These four top-ranked variables for accident severity are among the three factors (vehicle Related, road related, and accident related). Figure 37 shows the importance values for the feature variables in decision tree model.

The most important input for this model is NoofCasualty. The input NumberofVehiclesInvolved has a relative importance of 0.19, for example, which means it is 0.19 times as important as NoofCasualty.

6. Conclusion

The personal experience in my native country, where traffic accidents are one of the most pressing issues and their effects in human life, including physical injury, property damage, and overall economic effect, prompted me to choose this topic for this diploma thesis. The machine learning option gives chance to create a predictive model that can has a capacity to predict the occurrence of accident severity. In most developing nations having access to large volume of data to generate predictive models is impossible. The data used in this research is found from Kaggle data science community. The data that is used in this study has accident related, road related, vehicle related and driver related data. While having access and visibility to an immense amount of data, there was a strong personal interest to develop deeper analysis to make data more useful and powerful.

As mentioned at the beginning of this thesis, the main goal of this study is to develop an in-depth analysis of traffic accident severity and to determine which of the predictive modeling used is the best one for predicting accident severity as target variable.

The research included both theoretical and practical part. The study of literature begins with a general overview of Big Data the common challenges of it. This is covered because the dataset used for the study is Big Data, which has all of its challenges and potentials. Characteristics such as Volume, Velocity and Variety are covered and pointed as three of the five challenges when dealing with Big Data.

This study investigated the performance of the three machine learning algorithms to build classifiers that are precise and reliable. This includes the Logistic regression (LR), Neural Network (NN), and decision tree (DT) algorithms. Based on the model comparison node in SAS model studio, the test results show that the decision tree seemed to perform better than the other models.

This research study shows that the algorithms can predict accidents with 87% accuracy.

The study can help provide useful information for highway engineers and transportation designers to design safer roads. Further studies should be done to collect related information and investigate the impacts of these factors. It is recommended that Decision tree (the best model for predicting accident severity) to be applied in monitoring Fatal, serious and slight injuries. The recommended predictive model can be used to rapidly and efficiently identify the Key factor causing accident severity. One limitation of the current study is that some of the factors (i.e., characteristics of the

driver, passenger, and pedestrian, along with traffic conditions) may have possible effects on accident severity and duration, which are not considered because of the lack of suitable data.

7. Reference

- Abbott, D. (2014). *Applied predictive analytics: Principles and techniques for the professional data analyst*. John Wiley & Sons.
- Akerkar, R. (2014). *Introduction to artificial intelligence*. PHI Learning Pvt. Ltd.
- Aslan, A., Karakoyun, O., Güler, E., Aydin, S., Gök, M. V., & Akkurt, S. (2012). Evaluation of bone mineral density, osteoporosis prevalence and regional risk factors in Turkish women living in Kastamonu: KASTÜRKOS study. *Eklem Hastalıkları ve Cerrahisi= Joint Diseases & Related Surgery*, 23(2), 62–67.
- Bentolila, S., Heller, W. P., Sun, T., Babina, A. M., Friso, G., van Wijk, K. J., & Hanson, M. R. (2012). RIP1, a member of an Arabidopsis protein family, interacts with the protein RARE1 and broadly affects RNA editing. *Proceedings of the National Academy of Sciences*, 109(22), E1453--E1461.
- Berry, M. J. A., & Linoff, G. S. (2004). *Data mining techniques: for marketing, sales, and customer relationship management*. John Wiley & Sons.
- Boyd, D., & Crawford, K. (2012). Critical questions for big data: Provocations for a cultural, technological, and scholarly phenomenon. *Information, Communication & Society*, 15(5), 662–679.
- Chai, T., & Draxler, R. R. (2014). Root mean square error (RMSE) or mean absolute error (MAE)?--Arguments against avoiding RMSE in the literature. *Geoscientific Model Development*, 7(3), 1247–1250.
- Crawford, V. P., Costa-Gomes, M. A., & Iriberry, N. (2013). Structural models of nonequilibrium strategic thinking: Theory, evidence, and applications. *Journal of Economic Literature*, 51(1), 5–62.
- Cukier, K. (2010). Data, data everywhere. *Economist*, 394(8671), 3–5.
- Cyprus, U. (2013). Yannis G. Yatracos Cyprus U. of Technology February 20, 2014. *ArXiv Preprint ArXiv:1304.4929*.

- Daassi-Gnaba, H., Oussar, Y., Merlan, M., Ditchi, T., Géron, E., & Holé, S. (2017). Wood moisture content prediction using feature selection techniques and a kernel method. *Neurocomputing*, 237, 79–91.
- Dangeti, P. (2017). *Statistics for machine learning*. Packt Publishing Ltd.
- Deme, D., & Bari, M. (2018). Journal of Equity in Science and Sustainable Development Vol. 2 (2): 13-23 Article Number: JESSD17. 11.2017 ISSN 2523-1901 (Online)
Copyright{\copyright} 2018. *Journal of Equity in Science and Sustainable Development*, 2(2), 13–23.
- Engino\u{g}lu, S., & Memi\cs, S. (2018). A Review on An Application of Fuzzy Soft Set in Multicriteria Decision Making Problem [PK Das, R. Borgohain, International Journal of Computer Applications 38 (12)(2012) 33--37]. *Proceeding of The International Conference on Mathematical Studies and Applications*, 173–178.
- Fan, J., Han, F., & Liu, H. (2014). Challenges of big data analysis. *National Science Review*, 1(2), 293–314.
- Forsyth, D. A., & Ponce, J. (2012). *Computer vision: a modern approach*. Pearson,.
- Gandomi, A., & Haider, M. (2015). Beyond the hype: Big data concepts, methods, and analytics. *International Journal of Information Management*, 35(2), 137–144.
- Gantz, J., & Reinsel, D. (2011). Extracting value from chaos. *IDC Iview*, 1142(2011), 1–12.
- Gough, D., Oliver, S., & Thomas, J. (2017). *An introduction to systematic reviews*. Sage.
- Gould, C. M., Diella, F., Via, A., Puntervoll, P., Gemünd, C., Chabanis-Davidson, S., Michael, S., Sayadi, A., Bryne, J. C., Chica, C., & others. (2010). ELM: the status of the 2010 eukaryotic linear motif resource. *Nucleic Acids Research*, 38(suppl_1), D167--D180.
- Hargittai, E. (2015). Is bigger always better? Potential biases of big data derived from social network sites. *The ANNALS of the American Academy of Political and Social Science*, 659(1), 63–76.
- Hilbe, J. M. (2016). *Practical guide to logistic regression*. crc Press.

- Hurwitz, J. S., Nugent, A., Halper, F., & Kaufman, M. (2013). *Big data for dummies*. John Wiley & Sons.
- Jaff, M. R., McMurtry, M. S., Archer, S. L., Cushman, M., Goldenberg, N., Goldhaber, S. Z., Jenkins, J. S., Kline, J. A., Michaels, A. D., Thistlethwaite, P., & others. (2011). Management of massive and submassive pulmonary embolism, iliofemoral deep vein thrombosis, and chronic thromboembolic pulmonary hypertension: a scientific statement from the American Heart Association. *Circulation*, *123*(16), 1788–1830.
- Jin, X., Wah, B. W., Cheng, X., & Wang, Y. (2015). Significance and challenges of big data research. *Big Data Research*, *2*(2), 59–64.
- Kutner, M. H., Nachtsheim, C. J., Neter, J., Li, W., & others. (2005). *Applied linear statistical models* (Vol. 5). McGraw-Hill Irwin Boston.
- Lazer, D., Pentland, A., Adamic, L., Aral, S., Barabasi, A.-L., Brewer, D., Christakis, N., Contractor, N., Fowler, J., Gutmann, M., & others. (2009). Social science. Computational social science. *Science (New York, NY)*, *323*(5915), 721–723.
- Liu, H., & May, K. (2012). Disulfide bond structures of IgG molecules: structural variations, chemical modifications and possible impacts to stability and biological function. *MAbs*, *4*(1), 17–23.
- Manyika, J., Chui, M., Brown, B., Bughin, J., Dobbs, R., Roxburgh, C., Hung Byers, A., & others. (2011). *Big data: The next frontier for innovation, competition, and productivity*. McKinsey Global Institute.
- Másilková, M. (2017). *Health and social consequences of road traffic accidents*. Retrieved February 12, 2018.
- Mendes, F. A. R., Gonçalves, R. C., Nunes, M. P. T., Saraiva-Romanholo, B. M., Cukier, A., Stelmach, R., Jacob-Filho, W., Martins, M. A., & Carvalho, C. R. F. (2010). Effects of aerobic training on psychosocial morbidity and symptoms in patients with asthma: a randomized clinical trial. *Chest*, *138*(2), 331–337.
- Meshram, K., & Goliya, H. S. (2013). Accident analysis on national highway-3 between Indore

- to Dhamnod. *International Journal of Application or Innovation in Engineering \& Management (IJAEM) Volume, 2.*
- Organization, W. H. (2015). *Global status report on road safety 2015.* World Health Organization.
- Papantoniou, P., Yannis, G., & Christofa, E. (2019). Which factors lead to driving errors? A structural equation model analysis through a driving simulator experiment. *IATSS Research*, 43(1), 44–50.
- Parab, N., Bhawkar, A., & Patel, B. (n.d.). *Strategic ICT and E-Business Implementation on Tourism of Unexplored Places in Republic of Ireland.*
- Robertson, R. D., Meister, S. R., Vanlaar, W. G. M., & Hing, M. M. (2017). Automated vehicles and behavioural adaptation in Canada. *Transportation Research Part A: Policy and Practice*, 104, 50–57.
- Singh, T. N., Kainthola, A., & Venkatesh, A. (2012). Correlation between point load index and uniaxial compressive strength for different rock types. *Rock Mechanics and Rock Engineering*, 45(2), 259–264.
- Świdorski, A., Borucka, A., & Skoczyński, P. (2018). Characteristics and assessment of the road safety level in Poland with multiple regression model. *Transport Means', Proceedings of the 22nd International Scientific Conference, Part I, Lithuania*, 92–97.
- Terzano, K. (2013). Bicycling safety and distracted behavior in The Hague, the Netherlands. *Accident Analysis \& Prevention*, 57, 87–90.
- Tufféry, S. (2011). *Data mining and statistics for decision making.* John Wiley \& Sons.
- Wang, Y., & Wiebe, V. J. (2016). Big Data Analytics on the characteristic equilibrium of collective opinions in social networks. In *Big Data: Concepts, Methodologies, Tools, and Applications* (pp. 1403–1420). IGI Global.
- Wiggins, A., & He, Y. (2016). Community-based data validation practices in citizen science. *Proceedings of the 19th ACM Conference on Computer-Supported Cooperative Work \&*

Social Computing, 1548–1559.

Wild, M., Ohmura, A., Schär, C., Müller, G., Folini, D., Schwarz, M., Hakuba, M. Z., & Sanchez-Lorenzo, A. (2017). The Global Energy Balance Archive (GEBA) version 2017: A database for worldwide measured surface energy fluxes. *Earth System Science Data*, 9(2), 601–613.

Willmott, C. J., & Matsuura, K. (2005). Advantages of the mean absolute error (MAE) over the root mean square error (RMSE) in assessing average model performance. *Climate Research*, 30(1), 79–82.

Zicari, R. V. (2014). Big data: Challenges and opportunities. *Big Data Computing*, 564, 103.

Zikopoulos, B., & Barbas, H. (2012). Pathways for emotions and attention converge on the thalamic reticular nucleus in primates. *Journal of Neuroscience*, 32(15), 5338–5350.

Global status report on road safety 2018 ISBN 978-92-4-156568-4

Appendix 1. value coding for all selected variables

<p>value Age_Band_of_Driver -1="Unknown" 1="0-5 years old" 2="6-10 years old" 3="11-15 years" 4="16-20 years old" 5="21-25 years old" 6="26-35 years old" 7="36-45 years old" 8="46-55 years old" 9="56-65 years old" 10="66-75 years old" 11="over 75"</p>	<p>value Junction_Detail 0="Not at junction or within 20 metres" 1="Roundabout" 2="Mini-roundabout" 3="T or staggered junction" 5="Slip road" 6="Crossroads" 7="More than 4 arms (not roundabout)" 8="Private drive or entrance" 9="Other junction" -1="Data missing or out of range"</p>	<p>value Speed_limit_category 1="10 mph" 2="15 mph" 3="20 mph" 4="30 mph" 5="40 mph" 6="50 mph" 7="60 mph" 8="70 mph"</p>
<p>value Day_of_Week 1="Sunday" 2="Monday" 3="Tuesday" 4="Wednesday" 5="Thursday" 6="Friday" 7="Saturday"</p>	<p>value Light_Conditions 1="Daylight" 4="Darkness - lights lit" 5="Darkness - lights unlit" 6="Darkness - no lighting" 7="Darkness-lighting unknown" -1="Data missing or out of range"</p>	<p>value No_of_Casualty 1="up to 5" 2="6 to 15" 3="16 to 30" 4="31 to 50" 5="More than 50"</p>
<p>Value 1st_Point_of_Impact 0="Did not impact" 1="Front" 2="Back" 3="Offside" 4="Nearside" -1="Data missing or out of rang"</p>	<p>value Vehicle_Manoevvre 1="Reversing" 2="Parked" 3="Waiting to go - held up" 4="Slowing or stopping" 5="Moving off" 6="U-turn" 7="Turning left" 8="Waiting to turn left" 9="Turning right" 10="Waiting to turn right" 11="Changing lane to left" 12="Changing lane to right"</p>	<p>value Vehicle_Type 1="Pedal cycle" 2="Motorcycle 50cc and under" 3="Motorcycle 125cc and under" 4="Motorcycle over 125cc and up to 500cc" 5="Motorcycle over 500cc" 8="Taxi/Private hire car" 9="Car" 10="Minibus(8-16passenger seat)" 11="Bus or coach (17 or more pass seat" 16="Ridden horse" 17="Agricultural vehicle" 18="Tram"</p>

	13="Overtaking moving vehicle - offside" 14="Overtaking static vehicle - offside" 15="Overtaking - nearside" 16="Going ahead left-hand bend" 17="Going ahead right-hand bend" 18="Going ahead other" -1="Data missing or out of range"	19="Van/Good 3.5 tonnes mgw or under" 20="Goods over 3.5T and under 7.5T" 21="Goods 7.5T mgw and over" 22="Mobility Scooter" 23="Electric motorcycle" 90="Other vehicle" 97="Motorcycle-unknown cc" 98="Goods vehicle-unknown weight" -1="Data missing or out of range"
value Road_Type 1="Roundabout" 2="One way street" 3="Dual carriageway" 6="Single carriageway" 7="Slip road" 9="Unknown" 12="One way street/Slip road" -1="Data missing or out of range"	value Sex_of_Driver 1="Male" 2="Female" 3="Not Known" -1="Data missing or out of range"	valueNumber_of_Vehicles_Involved 1="Single" 2="Double" 3="Multiple"
value Accident_Severity 1="Fatal" 2="Serious" 3="Slight"		