

***Dissertation***

***KF7029***

***MSc Computer Science***

***and Digital Technologies Project***

Student Name: Vincent Adam Cook

Student Number: W15009923

Supervisor Name: Nanlin Jin

Second Marker Name: Alan Godfrey

Dissertation Title: Machine learning models for accident prediction 2018/2019

**Declaration**

I declare the following:

(1) That the material contained in this dissertation is the end result of my own work and that due acknowledgement has been given in the bibliography and references to **ALL** sources be they printed, electronic or personal.

(2) The Word Count of this Dissertation is 14700

(3) that unless this dissertation has been confirmed as confidential, I agree to an entire electronic copy or sections of the dissertation to being placed on the eLearning Portal (Blackboard), if deemed appropriate, to allow future students the opportunity to see examples of past dissertations.  I understand that if displayed on eLearning Portal it would be made available for no longer than five years and that students would be able to print off copies or download.

(4) I agree to my dissertation being submitted to a plagiarism detection service, where it will be stored in a database and compared against work submitted from this or any other Department or from other institutions using the service.

In the event of the service detecting a high degree of similarity between content within the service this will be reported back to my supervisor and second marker, who may decide to undertake further investigation that may ultimately lead to disciplinary actions, should instances of plagiarism be detected.

(5) I have read the Northumbria University Policy Statement on Ethics in Research and Consultancy and I confirm that ethical issues have been considered, evaluated and appropriately addressed in this research.

**SIGNED:**

**DATE: 19/09/2019**

# Abstract

This report contains an exploration into the possible uses of machine learning for creating a safer route finding model. This is achieved with the use of research into the current methods that are used for route finding as well as research into machine learning and the techniques that are used for classification and the data that could be used for the creation of such a model. After research is completed preparation of data and the methods that could be used for such are explored in order to create a dataset which will create a strong learner for predictions of accidents and a model that will be used is chosen which should produce the best results.

This is followed by the creation and testing of the model as well as some other models which will be used as a comparison in order to best assess how the model has performed when run with the training and testing data for it.

Data from these models are then captured and assessed based on the creation and the error rates that are produced in order to determine the models viability for use in prediction of if accidents and show that it has use for the purpose of aiding in finding safer routes for drivers before showing some ways in which the technology could be used for future research into the area and advance upon the field.

Contents

[Declaration 2](#_Toc19728476)

[Abstract 3](#_Toc19728477)

[1 Introduction 7](#_Toc19728478)

[1.1 Background 7](#_Toc19728479)

[1.2 Aim 7](#_Toc19728480)

[1.3 Objectives 8](#_Toc19728481)

[1.4 Research Approach 9](#_Toc19728482)

[1.5 Structure of the report 9](#_Toc19728483)

[2 Literature Review 10](#_Toc19728484)

[2.1 History 10](#_Toc19728485)

[2.1.1 History of route planning 10](#_Toc19728486)

[2.2 Machine learning use in routing 12](#_Toc19728487)

[2.3 Machine learning techniques 13](#_Toc19728488)

[2.4 Data Requirements 15](#_Toc19728489)

[3 System Design 18](#_Toc19728490)

[3.1 Programming language 18](#_Toc19728491)

[3.2 Deciding Aim for Model 18](#_Toc19728492)

[3.3 Raw data analysis 19](#_Toc19728493)

[3.3.1 Mapping Accident locations 19](#_Toc19728494)

[3.3.2 Visualisation of the causes 24](#_Toc19728495)

[3.4 Data Preparation 30](#_Toc19728496)

[3.4.1 Taking variables needed from the dataset 30](#_Toc19728497)

[3.4.2 Dealing with categorical variables 32](#_Toc19728498)

[3.4.3 Negative data 33](#_Toc19728499)

[3.4.4 Randomize data 34](#_Toc19728500)

[3.5 Choosing a model 34](#_Toc19728501)

[4 Implementation 35](#_Toc19728502)

[4.1 Model 35](#_Toc19728503)

[5 Testing 39](#_Toc19728504)

[5.1 Model Test Data 39](#_Toc19728505)

[5.2 Comparison models 41](#_Toc19728506)

[5.2.1 Creation process 41](#_Toc19728507)

[6 Results 43](#_Toc19728508)

[6.1 Model Results 43](#_Toc19728509)

[6.1.1 Relative influence for training set 43](#_Toc19728510)

[6.1.2 Error rate and optimal number of learners 44](#_Toc19728511)

[6.1.3 Partial Dependence plots 45](#_Toc19728512)

[6.1.4 Predictions with test set 46](#_Toc19728513)

[6.1.5 Predicted Vs Actual Accidents 46](#_Toc19728514)

[6.2 Other models for comparison 47](#_Toc19728515)

[6.2.1 REP Decision Tree 47](#_Toc19728516)

[6.2.2 REP Decision tree visualised error 48](#_Toc19728517)

[6.2.3 REP Decision Tree Error Summary 48](#_Toc19728518)

[6.2.4 Random Forest Summary 48](#_Toc19728519)

[6.2.5 Random Forest Details 48](#_Toc19728520)

[6.2.6 Random Forest Error Visualisation 48](#_Toc19728521)

[7 Evaluation of results 49](#_Toc19728522)

[7.1 Analysis 49](#_Toc19728523)

[7.1.1 Relative influence 49](#_Toc19728524)

[7.1.2 Model error 49](#_Toc19728525)

[7.1.3 Partial dependence 50](#_Toc19728526)

[7.1.4 Comparison to other test models 50](#_Toc19728527)

[7.2 Evaluation 51](#_Toc19728528)

[7.3 Reflection 52](#_Toc19728529)

[8 Conclusion 53](#_Toc19728530)

[8.1 Summary 53](#_Toc19728531)

[8.2 Conclusion 53](#_Toc19728532)

[8.2.1 Issues with the model 54](#_Toc19728533)

[8.2.2 Issues addressed 54](#_Toc19728534)

[8.3 Suggestions for future research 55](#_Toc19728535)

[Bibliography 56](#_Toc19728536)

[Appendix A. Research Proposal 62](#_Toc19728537)

[Appendix B. Data Licensing 79](#_Toc19728538)

[Appendix C. Category Lookup tables 80](#_Toc19728539)

[Appendix D. Table of figures 82](#_Toc19728540)

[Appendix E. Decision tree examples from gradient booster 85](#_Toc19728541)

# Introduction

## Background

Traffic collisions cost most countries 3% of their gross domestic product each year and kill around 1.35 million people in the same time period. This makes it the leading cause of death among 5-29 year olds worldwide. (World health organisation , 2018) This means road safety is a serious issue which does not have one simple solution. While most solutions explored focus on large scale efforts to teach people and to develop new ways for people to be safer on the roads there could be ways in which smaller changes to the way individuals get around the roads could help them to avoid issues.

Many people around the world use satellite navigation in order to get directions for their journeys around 108 million of these devices were sold in Europe in the years between 2007 and 2016. (Liu, 2017) However the number of these sales has been falling year on year since its peak in 2011 this is not because the technology has fallen out of favour but because of the ease of use and access to applications on phones such as google maps which has a reach of 90% of the android users due to it coming pre-installed on phones. (Liu, 2018)

With navigation programs having such high usage and being easily available to users it would be of benefit to use this technology to help reduce the risk to drivers on the roads. This report will explore the option of adding ways to aid in finding the safest routes for users to these devices allowing them the choice of sacrificing some speed for having a safer journey.

## Aim

The aim of this project is to create a machine learning model which could be added to a program to allow safer routes to be found

This will be achieved by conducting research into machine learning and into techniques used by existing programs then developing a model which can be used to predict if accidents will happen. This has changed from the proposal due to wanting to spend extra work on the model itself rather than the implementation as this part will form the actual research portion of the report as opposed to the visual implementation which would form a product but not advance on the field.

## Objectives

1. To collect to be used in the model
   1. Collect data on traffic flow on roads
   2. Collect data on collisions and other dangers on roads
   3. Collect data on whether conditions
2. To Undertake research into current forms of route planning
   1. Research into methods of modelling the traffic conditions
   2. Research into the methods of determining the safest route
   3. Research into the effect on whether conditions on safest route
3. To conduct research into machine learning models
   1. To research models used in route finding
   2. To research models which would be useful for the purpose of this report
4. Clean and organise data
   1. clean data and ensure all gaps are removed from all data sets
   2. combine data when needed as some could come from multiple sources
   3. change data to the form best for the chosen model
5. Develop a machine learning modal to determine if accident has taken place
   1. develop a model to best predict if an accident will occur
   2. develop some models to test against
   3. test models using data to see how they perform
6. Evaluate the results to determine whether the models created could act as a prediction for accidents taking place

Some objectives have changed from the original proposal as to reflect that the work undertaken will be more to do with the backend of the model which will predict the accidents and to do with the data that has been taken coming from a single source rather than many

## Research Approach

The research will be conducted mostly from online sources as with computing research a lot of information is out of date as it is published so using the internet to find the information ensures that the latest research can be found. The information will come from a number of both formal sources such as research journals and books and some informal ones such as blogs on the topic by reputable sources as these can prove valuable for the topic with the discussions that take place on them. For the model itself will be created using up to date techniques that are researched and tested using part of the data set which is used for its creation. Evaluation will be made on the results that are taken from the models such as the error rates and the decision trees that are generated.

## Structure of the report

The report will contain an abstract and introduction which will explain the topic to the user. These will be followed by a review of relevant literature which will look into the rationale for the report as well as the techniques that could be used in the creation of the model to provide the context of what is created to obtain the results. This will be followed by the design section which will contain analysis of the data used for the model to determine what will be the important factors in its use as well as how the data will be prepared for the model and how the type of model that would be used is chosen. This will be followed by a short section showing how the model was implemented. Once the model has been created some results will be taken from the model for the next section which will include the errors of the model as well as the results of the test data and some data on other models that have been created for comparison. Finally these results will be analysed and a conclusion will be provided explaining the performance of the model and concluding with future research which could be under taken.

# Literature Review

## History

Looking into the history of what this report is for will allow for extra perspective into the reasons as to why a model like the one that is to be created will be useful to society.

### History of route planning

Journey planning can trace its origins back to the mathematical problem proposed by Euler in 1736 known as the 7 bridges of Königsberg in which he proposed a problem in which a path needed to be found through 7 bridges from the islands around Königsberg without crossing the same bridge twice. (Euler, 1741) This problem began a history of using mathematics to attempt to solve problems with city planning and finding routes around maps. And laid the foundations for the mathematical field of graph theory.

Route finding for navigation however finds its nearest routes in the shortest path problem, studies have had trouble finding the earliest examples of this problem and theorise it as being something that is seen as trivial and dates back to the earliest points of civilization. (Schrijver, 2010) Because of this mathematical studies for the problem originate relatively late in history with graph problems being common in the late 1800s such as (Wiener, 1873) and heuristic approaches which are common today being studied as late as the 1950s with Dijkstra's algorithm which uses a graph based system calculating the shortest path using an open and closed set of nodes which are used to find the shortest path to each node by calculating the closest node each step of the way adding the rest to the closed set then the none calculated nodes from the new one to the open and repeating this operation until the path is found (Dijkstra, 1959) and its variants such as A\* (Hart, et al., 1968 ) still in common usage today.

Route planning in a practical sense for the use of transport can be dated back to the 19th century with the creation of timetable compilations for railways to enable people to plan their journeys across the multiple railways in the country before the use of universal timetables and planning among the different railways around the country. This is most notable with the compilation of Bradshaw’s guide beginning in 1939 which created a standard way of collecting and printing the data from the sources (Bradshaw, 1939) this shows that as the public need to travel larger distances a need is created for better ability to plan and coordinate these journeys to make travelling more efficient. As computing systems advanced these compiled timetables have been improved to have advanced systems for planning journeys accounting for multiple varying factors. Computing systems have been used widely in industries such as rail and freight transportation.

With the increase of processing power and the availability of home computers on the rise in the 1990s several systems were created for use on home computers beginning with the first home release of the system created for Dutch railways on diskette (Tulp, 1991) these releases were powerful enough to use route finding algorithms to calculate the average walking distances between platforms and stations as opposed to only including the information available on the company timetables like the older compilations.

Many companies through the 90s and early 2000s released systems such as this with more and more complex systems as time went on allowing for plans on what roads to use and allowing for combining multiple forms of public transport walking and road use.

As the internet became more common the sales of systems like these dropped as people created free systems that could be accessed through web browsers and eventually through phone apps.

The rise in smartphone use and GPS in recent years has allowed for the large scale collection of data showing the ways in which we travel and use the roads applications such as google maps (Google, 2019) make use of this data to create more and more complex algorithms for allowing people to find the routes to take on their journeys.

This modern data collection includes many complex sources which could help to determine the path taken, as well as the GPS data other data sources such as traffic figures, speed camera positions and collisions information can be used with various machine learning models to give accurate predictions on what traffic will be like in the area as well as in some cases real time updates for when you will encounter trouble on the roads. (McDeed, 2016)

## Machine learning use in routing

Some problems have been very hard to solve with traditional mathematics due to them being considered np-hard (non-deterministic polynomial-time hard) sue to their solutions taking exponentially more time as more points are added. One of the most famous of these is a routing problem known as the traveling salesman problem which was first described mathematically by (Flood, 1956) this problem describes the issue with finding the shortest possible route a salesman would need to take in order to drop packages off at multiple locations in the shortest amount of time. The reason this is such a difficult issue is that whenever a new location is added to the list the times need to be calculated from each past destination to the new one and from the new location to every other destination to find where it would fit in the route still giving the shortest time between them.

This problem has only become more important in the modern day with the popularization of the internet and online companies such as amazon needing to deliver millions of packages in a day and needing to optimize their routes to do so in the most efficient way possible.

Modern approaches to this are looking to machine learning and cloud computing to find solutions as these allow for large amounts of data to be processed and routes found in a shorter time when adding new destinations to the map as no definitive solution has been found people have attempted and found various different methods that could help in improving the finding of routes. Approaches include the use of algorithms such as the genetic algorithm (Grefenstette, et al., n.d.) And the AS (ant swarm) (Dorigo & Gambardella, 1997) algorithm these solutions don’t provide a complete way to find the best route but they do increase the efficiency and time which has been valuable for industries needing ways to increase the optimization of their routes even if they do not provide a complete mathematical solution.

The problem with most of these solutions however when applied to routing for companies has been that they require specific inputs that the company may not actually have in order to optimise the route, most specifically the travel time that it would take between each location. This has led to people needing to research into finding methods of estimating these also using machine learning methods such as the research of (Lazovskiy, 2018) in which they look into ways of applying machine learning to find the estimated time of routes based on taxi data prior to using the genetic algorithm to find the route in order to remove the need of inputting data.

## Machine learning techniques

In order to choose which machine learning algorithm to use for the problem first an understanding of the types of algorithm available are needed.

The types of machine learning algorithm can be separated into a few main categories based on the ways that they achieve their results and how they are trained with data, they are categorised as;

* Supervised algorithms – mostly classification algorithms where data is mapped from data sources with known results for the model to learn how the outcomes were found so that it can make predictions off data with unknown results later.
* Unsupervised algorithms – models where the data is given as a set of inputs with no labels and the algorithm will decide for itself how they will be distributed.
* Semi-supervised models – where the data comes with both labelled examples and none labelled ones and the algorithm has to decide using a combination of the two
* Reinforced learning – where the model learns based on observations of the world gaining knowledge from what it sees to provide feedback
* Transduction models – which are similar to supervised models but rather than creation of a function try to predict new outputs based on inputs given
* Learn to learn algorithms – where the algorithm is trained to learn based on previous experiences. (Zhang, 2010)

Based on these categories and the data that is available it can be seen that there is a lot of labelled categorized data that can be used for the model. It would seem that this would be best used with a supervised learning approach. With this a classification route will be taken.

The simplest approach to this is decision tree algorithms, there are a large number of these algorithms available however all work in a similar way. A decision tress is a method of visualising the data that is available as a number of choices based on how they each interact in a way similar to a flowchart. These are really good for showing the way that classified data leads to a particular result using each connection as a node in the chart and each of the destinations as an outcome of that particular choice. (Brid, 2018 ) Machine learning algorithms based on decision trees take the data that is given to them and pass it through the algorithm created to decide on how each of the sets of data should be separated and the result that should come from them based on the past data.

Using this method means that when the algorithm has used the data to build how the result was gained using the data that you provide it once you provide it with more data with an unknown result in the end it should be able to find that result from the tree.

As classifiers decision tress offer a number of benefits such as;

* Allowing complex decisions to be approximated using smaller local decisions at various branches of the tress
* Reduce computation time by only testing variables against a subset of the classes at a time (safavian & Landgrebe, 1990)

However these still com with potential issues that should be taken notice of;

* The number of terminals can end up much larger than the number of classes especially in large trees taking up more memory space and search time when needed to be accessed.
* As trees get larger there is more chance for errors due to the accumulation of them from level to level in the trees
* The performance depends on how the tree is designed and with an algorithm deciding how the trees are built it could lead to low performance. (safavian & Landgrebe, 1990)
* Single trees are prone to overfitting and underfitting data. (Sayad, 2019)

To fix some of the issues you have with decision trees a number of more advanced algorithms have been designed which improve on the way that they work such as random forest algorithms which generate a number of trees rather than a single one and obtain their results from a combination of the predictions made by the trees. This relies on a principle known as the wisdom of crowds – that when you ask for an estimate from a large enough sample of people their answers will converge on the original answer (HALTON, 2019) – to provide a close enough estimate for the correct answer. This is useful for weaker data sets to improve on their results. (Yiu, 2019)

Amore advanced version of this technique is gradient boosting models, these models work by iterating through a large number of decision trees in the same way as a random forest classifier however each new tree that is generated is fit into a modified version of the original data set based on the previous trees increasing the weights of the observations as it iterates through the trees which reduces the error rate of the model with each iteration and improves the predictions that are made. (Singh, 2018)

These models help to improve the predictions made by decision trees as well as remove a lot of the issues with overfitting of the data and work better with larger data sets with many different factors to the decisions that need to be made.

As the data that is collected for this report is complex with a lot of known factors the gradient boosting model seems to fit the data most if looking to find an accurate and fitting model for classification.

## Data Requirements

The requirements for the project will come from the causes of accidents that could occur on the roads that could be used to determine the data needed for calculations on the road that should be taken. For the purpose of this project it has been decided that statistics will be used from the UK due to this being the locale of the author. Because of this it will be important to remember that as the reasons will likely be the same in other locales the distribution across these reasons will change dependant on the road conditions and any final product which came from this research would need to account for these discrepancies.

The data used for the prototype will be taken from the road safety data provided by ( Department for Transport, 2018), the data is provided under open government license. (Figure 47)

Before using the data found first it must be determined what data from the set will be important to the project. For this some analysis of the aggregate data on traffic accidents provided by the (Department of transport, 2019) from the dataset “Reported road accidents by contributory factor, region and country, Great Britain, 2017” which gives a detailed breakdown of all recorded reasons for accidents. We can narrow down the reasons for accidents. (Table 1)

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Reason** | **England** | **Wales** | **Scotland** | **Great Briton** |
| **Road environment** | 9675 | 687 | 952 | 11314 |
| **Vehicle defects** | 1351 | 107 | 81 | 1539 |
| **Injudicious action** | 16396 | 769 | 1115 | 18280 |
| **Driver/Rider error or reaction** | 57338 | 2872 | 3979 | 64189 |
| **Impairment or distraction** | 11658 | 612 | 703 | 12973 |
| **Behaviour or inexperience** | 18212 | 737 | 1333 | 20282 |
| **Vision affected by external factors** | 7984 | 424 | 575 | 8983 |
| **Pedestrian only (casualty or uninjured)** | 10103 | 425 | 798 | 11326 |
| **Special Codes** | 3857 | 140 | 212 | 4209 |
| **Totals** | 83004 | 4043 | 6078 | 93125 |

Table 1: number of accidents by grouped cause and country

Using this data the factors can be found that will make the main contribution to accidents on the roads and determine the weighting of the factors that will form part of the finished program and which datasets will be needed to concentrate on.

The data shows that by far the largest contributing factor is driver errors or reactions which includes various factors where the driver has driven dangerously such as pulling out from junctions too early or failing to signal. As such these factors will be needed from the dataset and given high importance when determining the chances of an accident on the roads. It is possible that further data could be found on this from the police reports on dangerous driving (Police UK, 2019) . Which can be used to find the instances of dangerous driving that were reported and investigated in each area even when they did not lead to an accident to show how dangerous the area is.

The next largest factor is the behaviour or inexperience of the drivers involved. Similar to the other category this can largely be factored into the drivers behaving dangerously such as aggressive driving instances which can be found in a similar way to the previous category. Though a small part of this category included is also the age and experience of the driver, some data on this can be found through the driver licence data that is provided by the (Department of transport , 2019) this data includes breakdowns of the ages of drivers and types of licenses that are currently held in the UK including breakdowns by postcode area which will give a view of the general experience of the drivers that will be on the roads in a particular area.

Road environment as a factor is a major condition which can be further narrowed down to the individual road and so give a more accurate prediction this category broadly encompasses the conditions and layouts of the particular roads in which the accidents take place.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Reason** | **England** | **Wales** | **Scotland** | **Great Briton** |
| **Poor or defective road surface** | 465 | 25 | 49 | 539 |
| **Deposit on road** | 884 | 79 | 95 | 1,058 |
| **Slippery road** | 6,163 | 320 | 603 | 7,086 |
| **Inadequate or masked signs or road markings** | 326 | 10 | 37 | 373 |
| **Defective traffic signals** | 110 | 3 | 6 | 119 |
| **Traffic calming** | 86 | 6 | 5 | 97 |
| **Temporary road layout** | 191 | 4 | 20 | 215 |
| **Road layout** | 1,869 | 288 | 204 | 2,361 |
| **Animal or object in carriageway** | 690 | 44 | 68 | 802 |

Table 2: breakdown of number of accidents based on road conditions

For this data some information can be found in the department of transport statistics breaking the data up as far as the types of road and the areas in which It takes place however by the nature of the types of accident in this section a better solution would be to have a system that could take in up to date information on the roads in a similar way that traffic information is updated in apps such as google maps (Google, 2019) to give the conditions at that time of day. This will need to be considered but may be left as an idea for future improvements.

Another factor that needs to be taken into consideration with the impact on road conditions is the weather

# System Design

## Programming language

For organising data small data sets will be modified and organised using MATLAB (MathWorks , 2010) which will be used whenever possible due to the multiple uses for the suit of software provided and the ease of use. With any larger datasets which become too big for using in a program like a language such as R (R Core Team, 2013) or PySpark (Nandi, 2015) may be used due to their extra ability for organising large datasets and even the possibility for the need of cloud computing solutions.

## Deciding Aim for Model

An important step with the end goal is to decide what the model being used is to come to a decision in what it is that is needed for the output for the model. With the end goal in mind of creating a route planning software that will use the model to find the safest route to take there are a couple of approaches that could be used.

Given the data that is being used for the model there are two main sections that could be taken for this decision. The data on the accidents themselves – how severe the accidents were, the number of vehicles involves etc. this would give a view based on passed information of how dangerous a road has been in the past and the data involved showing the conditions at the time of the accident such as the weather, road conditions etc.

The first of those approaches could be used to decide based on past history whether each road is dangerous however with the changes in conditions this could be more unreliable if changes had been made to improve the roads over time.

The second could be used to create a predictive model for the road based on the current conditions by feeding the conditions at the time of each accident into a model and allowing it to learn the conditions of when an accident happens as opposed to the conditions when it has not the machine learning algorithms can be used to determine if there is a higher chance of an accident based on the road conditions at the time by taking data available and running it to predict the outcome.

Deciding this determines which data will be needed for the model when preparing it for use.

## Raw data analysis

### Mapping Accident locations

One of the first things to do when making a design for the final product is to analyse the raw data to find any patterns and links that may be different from the analysis mad by other parties that was found during the literature review.

The first step taken in this was to map all of the accidents to get a visual representation of where they take place.

This will be done by reading the data into MATLAB (MathWorks , 2010) and creating a table containing just the longitude and latitude columns which were then used with the mapping toolbox to create points on the map of the uk.

worldmap([49 59],[-12 4]) ;

geoshow('landareas.shp');

geoshow(longlat.Latitude,longlat.Longitude,...

'DisplayType', 'Point', 'Marker', '+', 'Color', 'red') ;

Figure 1 - code for representing points on a map

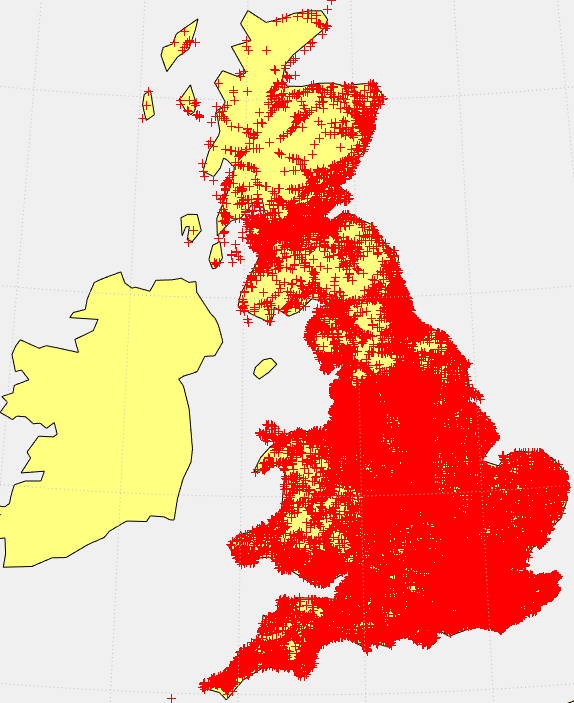


Figure 2: map of accidents 2017

From this map (Figure 2) a minimal amount of information can be taken, we can see that due to the number of accidents it can be very hard to find a pattern just from a visual representation of the data such as this as it looks as if accidents are spread across the whole country covering most areas. Looking at the blank patches of the map however shows that the cover is thicker in areas which are more populated such as the cover on the map around London and other major cities being a lot thicker than in areas with a lot of countryside such as wales, the Pennines and the Scottish highlands which have a lot more gaps and so less accidents which would suggest that the roads are safer in country areas.

This map could be made to show more data by separating the types of accident on it using the severity values from the table in order to show the locations where worse accidents happen in the country.

To do this a new table will be created to include the longitude latitude and severity and then this time using the geobubble function which allows for better displaying of complex data on a map show the points with the severity used to categorise the data so that the different severity ratings show in different colours.

As the data uses 1 for the most severe accidents the order of them needs to also be reversed in properties to give an appropriate ranking on the map.

Figure 3: code for mapping geobubble map based on severity of accidents

gb = geobubble(longlatSeverity,'Latitude','Longitude') ;

gb.SourceTable.Severity = (longlatSeverity.Severity) ;

gb.ColorVariable = 'Severity';

neworder = {'3','2','1'};

gb.SourceTable.Severity = reordercats(gb.SourceTable.Severity,neworder);

We can then narrow down the map to show just the fatal accidents to get a clearer picture of where they take place. This will be done by using the ismember function to find any with severity of 3 (Table 14) and removing them.

Figure 5: code to remove less severe accidents

longlatSeverity(ismember(longlatSeverity.Severity,3),:)=[];

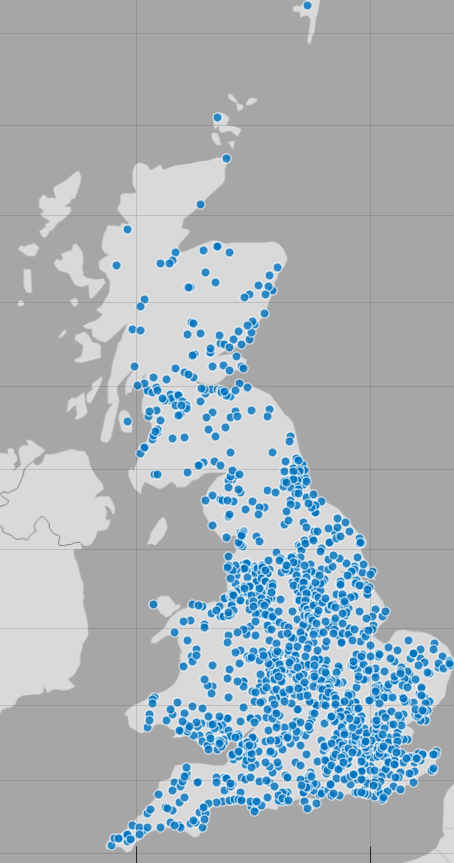


Figure 7: map showing only fatal accidents in 2017

From this map (Figure 7) you can get a clearer view of where worse accidents happen within the country. While there is still a spread across a lot of the country you can clearly see that a lot of points are more cantered in the more populated area of the country with the major cities having large amounts over the top of them, there are also clusters between these cities in lines which seem to be busy connecting routes such as motorways though this will require investigating further.

From his the next logical step would be to investigate the number of accidents that happen on a type of road, to begin with we will find the number that took place on each type which will help to show what kinds of road are most dangerous.

To do this a sum will be taken grouped by each type of road.

|  |  |  |
| --- | --- | --- |
| **DataValue** | **RoadType** | **Total number of accidents** |
| 1 | Roundabout | 8417 |
| 2 | One way street | 3386 |
| 3 | Dual carriageway | 20340 |
| 6 | Single carriageway | 93811 |
| 7 | Slip road | 1476 |
| 9 | Unknown | 2552 |

Table 3: number of accidents by road type 2017

This would suggest that most of the accidents happen on single carriageways by a large margin with dual carriageways close behind this could suggest that single carriage way roads are more dangerous however it is important to remember that it could also be that could be more commonly driven or that there are simply more of them to have accidents on.

Interestingly one way streets are missing from this list as there were no reported accidents on them which could mean that they are safest roads to travel due to there being no incidents of accidents which caused injury on them ion this time frame.

In order to analyse this further the data needs to be sorted by the actual roads the accidents take place, to do this the same process will be carried out using the road number and the type of roads as the grouping variables, both are needed due to some numbers being repeated (such as the A4 and M4) this gives a long list of results so in order to show it better it will need to be visualised on a map. For this we will need the longitude and latitude coordinates again, because these cover the entire road we will just use one point on the road to represent the results not each accidents location on the map.

After this a lot of the larger results are hidden by the many roads with only one incident in this time frame while this obscures what was wanted from this map, so the next step will be to make the more dangerous roads stand out more, to do this the data will be categorized to label the roads as more dangerous based on the number of accident’s that happen on them and then represent this data on the map using the colouring of the bubbles. This will make the larger geobubble representing the more dangerous roads stand out more than those with a low number.

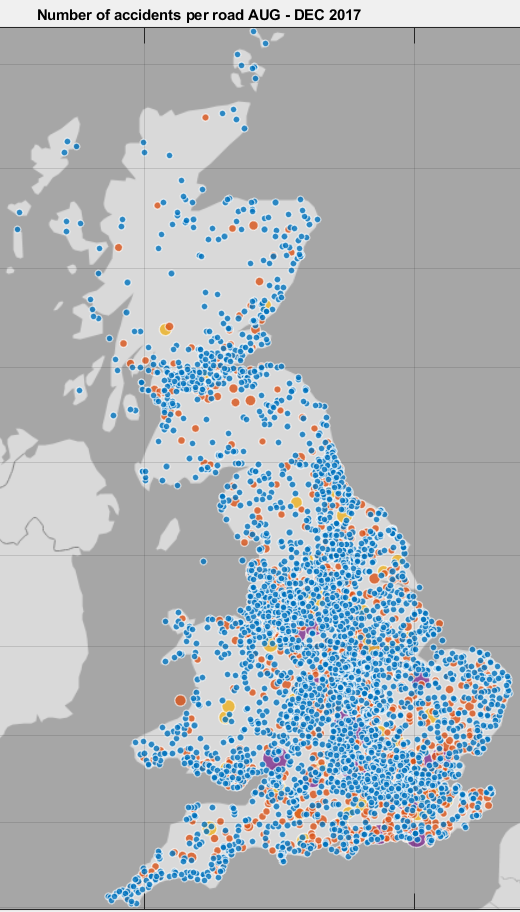
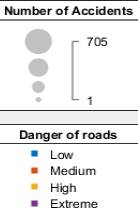


Figure 9: number of accidents per road 2017 using coloured Geobubbles

Not that there is a map that shows the data more clearly (Figure 9) we can identify that there are some roads that have serious issues with accidents which will be ones that the application designed will need to avoid. From the map it seems as if most of the problem roads are located around the coast which will need to be explored from the data itself. Busier roads are as expected the most dangerous with the A4 coming out with the largest number of accidents in the country as it acts as a major connection between London and busy areas for commuters and tourists.

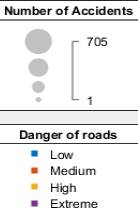
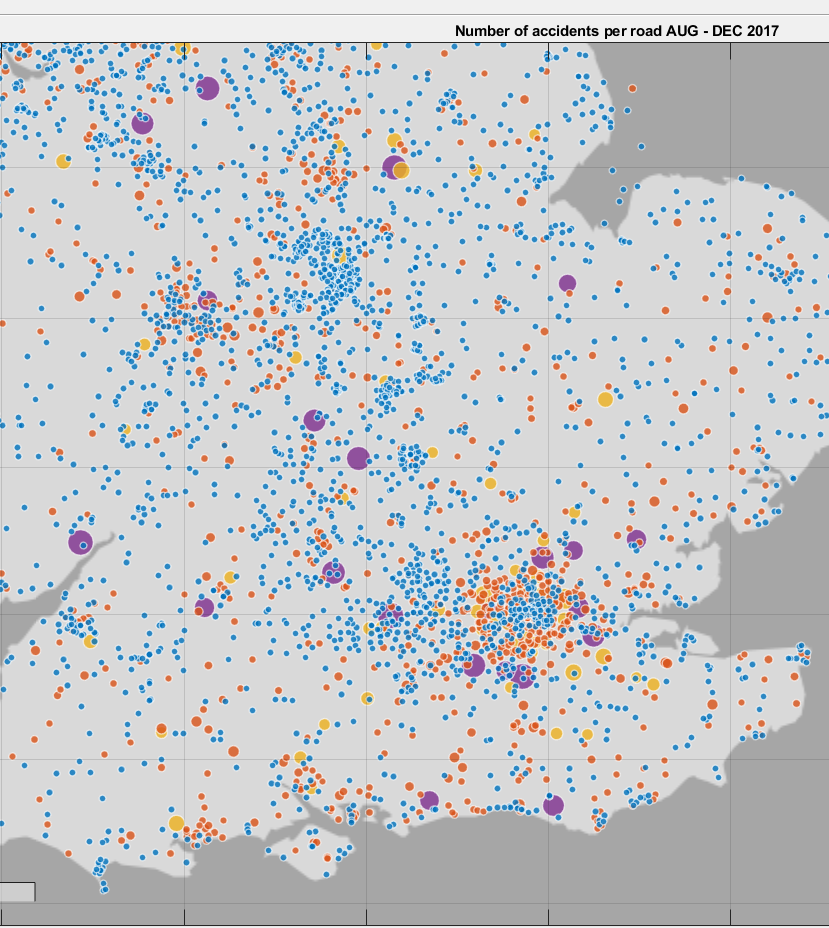


Figure 10: number of accidents south of England 2017

Looking at just the south of England (Figure 10) we can more clearly see that a lot of problem areas happen around major cities, the area of London has many more dangerous roads listed than the surrounding countryside with a lot of the most dangerous roads (those with more than 300 accidents in the time period) surrounding the capital on roads which would be used by commuters this would suggest that when testing the application we will be looking for busy commuter roads to be avoided and country roads to be labelled as safer alternatives than crossing through cities even if it would make a journey longer.

### Visualisation of the causes

Before working on the model it is good to examine the relationships between the data that will be used for the model. This will help with the design of the model and with the understanding we will have when the model has been trained and tested due to the knowledge of the relationships gained from the investigation.

The first bits of data that will be investigated is the link between the number of accidents and time. This will help gain an understanding of when the most accidents take place and to see if any patterns emerge of the most dangerous times.

The first relationship between them that will be explored is the number of accidents over the months in the year, this will be important as different times of year bring different dangers due to the changes in the weather and the differing lengths of the days which mean less light available to the drivers.

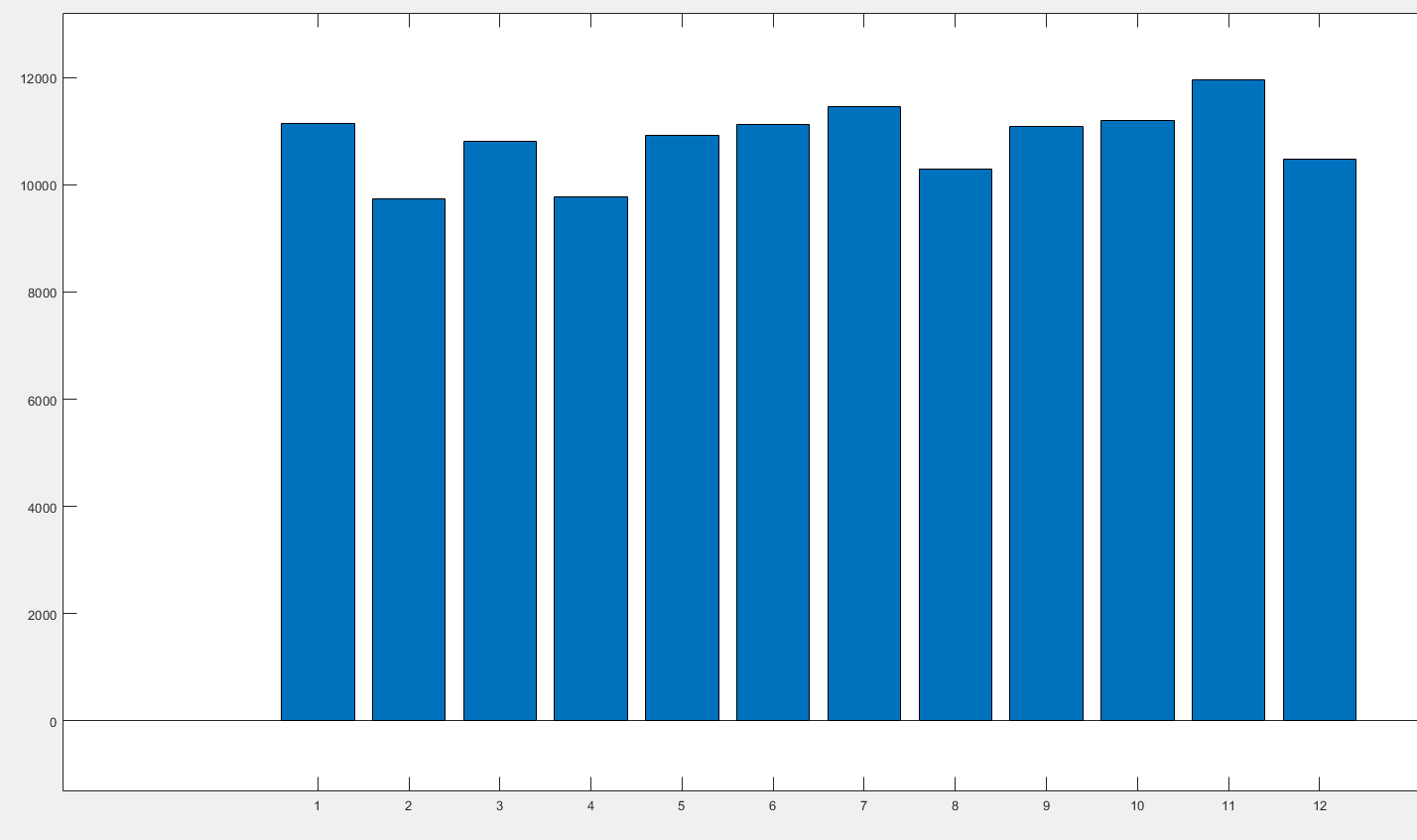


Figure 11: number of accidents by month 2017

From this we can see that there is a relatively stable amount of accidents with few large drops in number however there are some spikes in the number that we have especially in November as winter starts and more of a drop after January as we start to come out of winter. On top of this there is also some higher numbers in the middle of summer between the months of May and July this could be due to people traveling longer distances and driving on unfamiliar roads due to summer holidays or due to people going on more leisurely drives if the weather increases such as people taking motorbikes out during summer months when there is better weather.

Due to the variance of the number of accidents in each month it will be good to assume that the month of the year will make a good variable to use in the model as the time seems to have an effect on the number of accidents that will happen.

With these observations it would be good to see the severity of the accidents over time with this to see if there is a relationship with the severity and the time that the accident takes place.

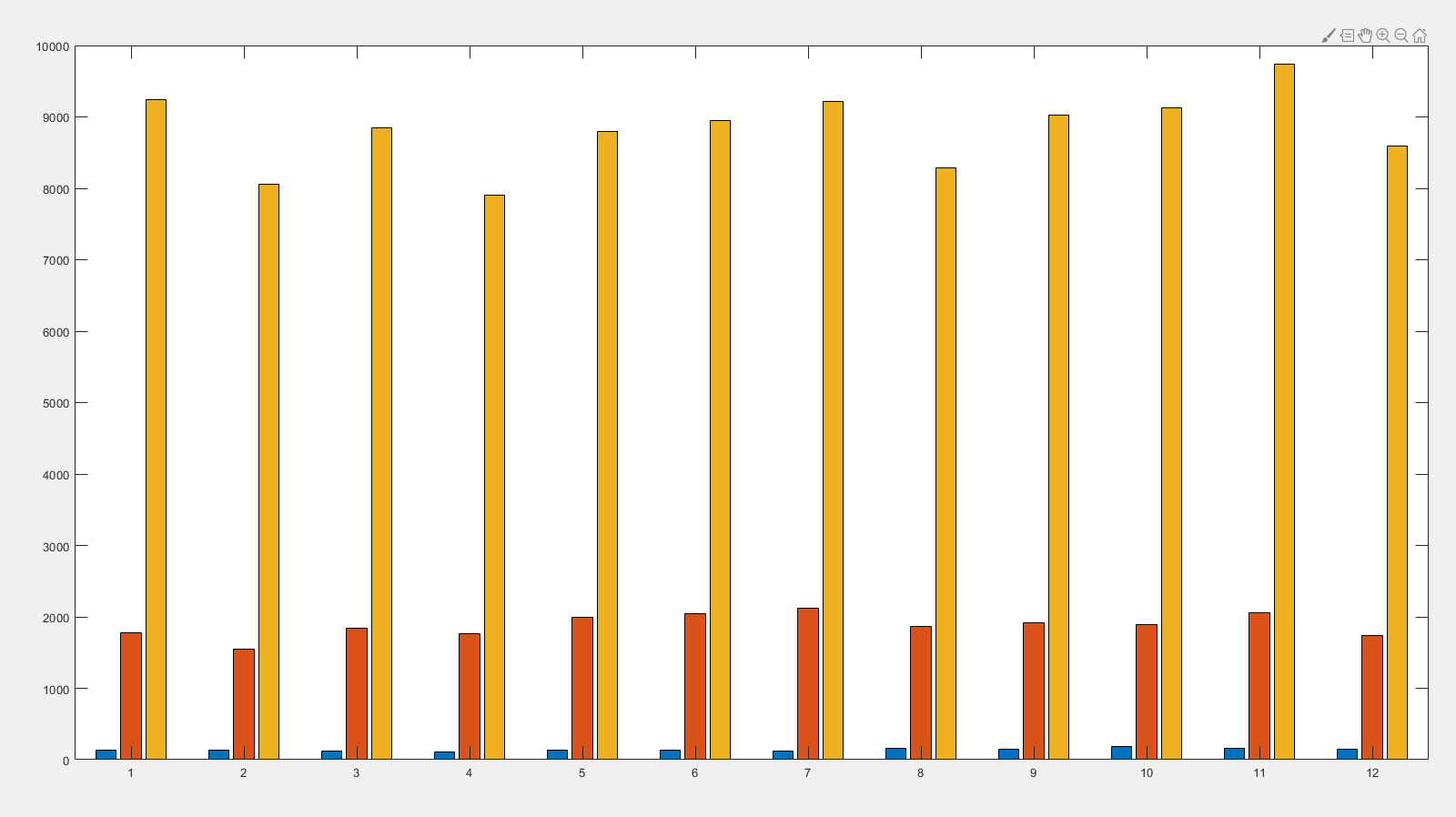


Figure 12: number of accidents of each severity by month 2017

From this we can see that as noted in (Figure 4) the number of non-serious accidents vastly outweigh the number of more serious ones in any given month as they do through the year we can also see from this however that in most months that the rise in total accidents the number of fatal accidents most months stays relatively stable in comparison while the severe but none fatal accidents mostly rise and fall with the none serious ones as well as this the peak in the number of accidents comes in November while the peak in fatal accidents comes in October.

What can take from this is that the number of accidents total in a month does not necessarily have a correlation with the number of fatal accidents and knowing that it would be reasonable to say that we can treat all accidents as the same from the model as the number of fatal accidents is relatively stable it would be better to concentrate on avoiding accidents at all which would potentially have an effect.

Finally in the investigation of the accidents over time we can look at the number of vehicles involved in the accidents overtime to give further insight into the types of accident that appear in each month.

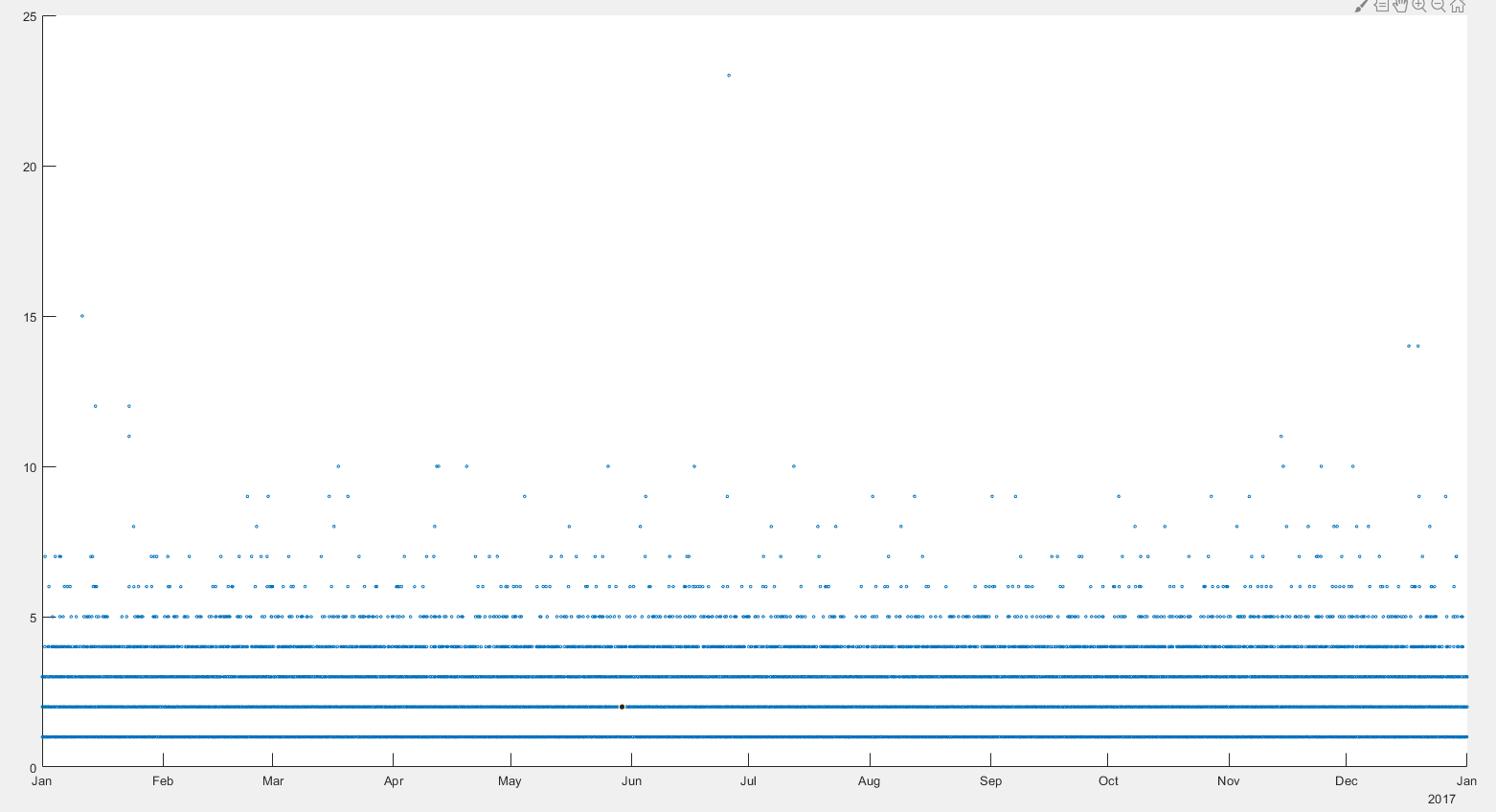


Figure 13: number of vehicles involved in accidents over time

From this information we can take that the number of vehicles involved in each accident is generally low as the number of points thin out for each month as the number of vehicles climbs as would be expected this means that there are fewer accidents involving a lot of vehicles.

While there is a scattering of accidents involving more than 5 vehicles across the year the highest numbers are spread between December and January meaning that accidents in winter months could involve more vehicles than those at other points in the year. This could be due to the weather conditions or the lighting of the roads during those times. There is one exception to this with a single accident involving most vehicles happening around the end of June however this is a single occurrence and so does not change the patterns seen in the other months.

Due to the data on accidents showing that there seems to be a correlation between the number of accidents and the time of the year it would be best to look into the effects of weather and the light conditions at the times of accidents as this will give a view of whether these are the main factors in the accidents occurring at different times of year or if more details will need to be looked into.

First it would be good to map the weather conditions by time as this will give a view of when the weather conditions changed in the year to get a view of this compared to the number of accidents over time.

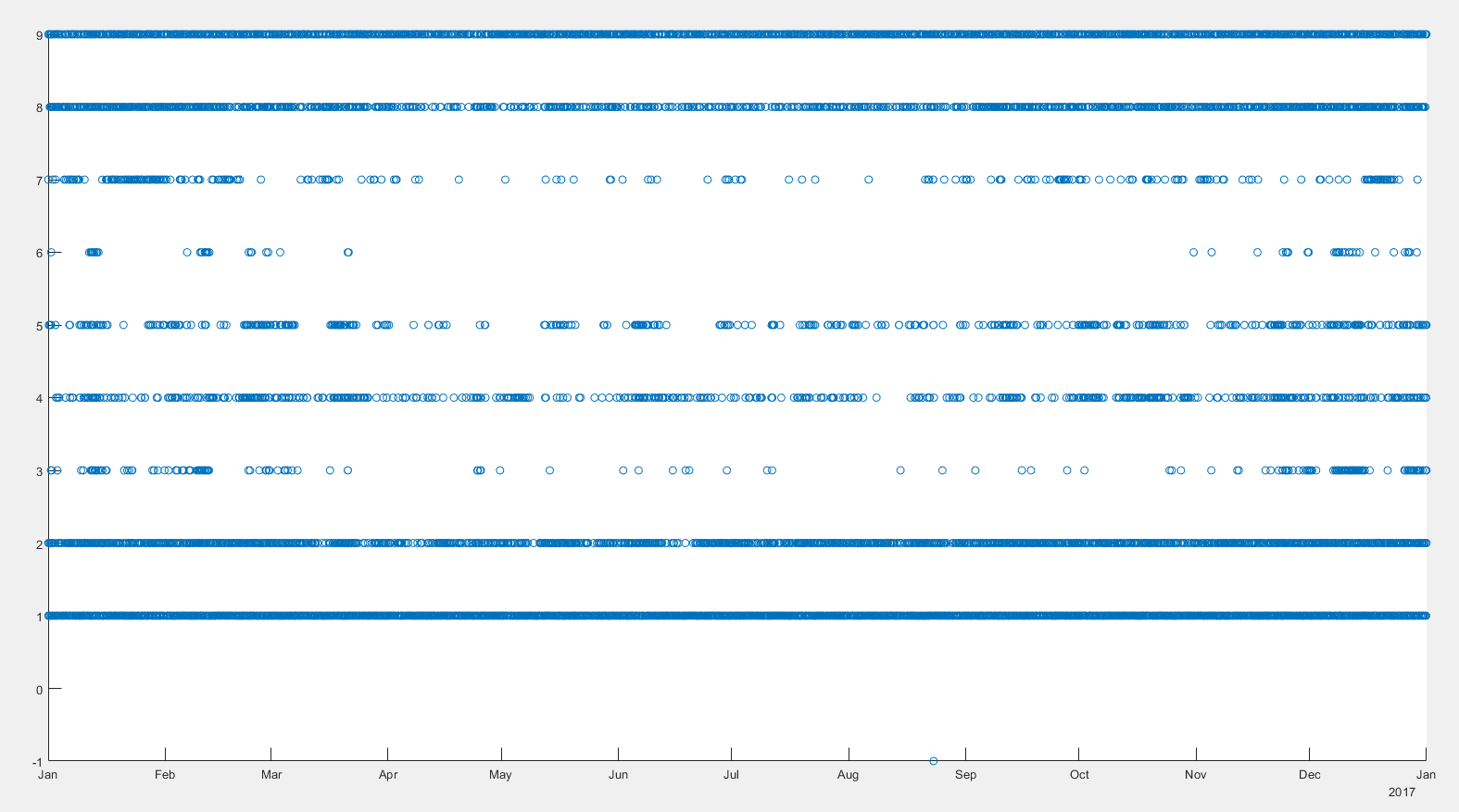


Figure 14: weather conditions over time 2017

With the representation of weather over time (Figure 14) a few patterns can be seen with the types of weather in various months. From the categories given (Table 12) the accidents taking place in times when the weather was fine and raining are evenly spread across the year as would be expected with a climate such as Britain’s. Times when there is high wind is also spread evenly across the year however with patches in summer when there is less instances so would be slightly more dangerous during the winter months and could be a small factor in why the number of accidents rises during that time. The same could be said for rain and high winds and foggy conditions combined which is spread through the year with even more gaps in summer the two types that stand out as causing more accidents in winter however are snow and snow with high winds as despite small patches with snow in the summer months these are more confined to the winter and as expected would be likely causes for more accidents in those times.

From the graph it can be seen that there were a large amount of accidents due to snow in late November, early December and mid-January and February which as can be seen in (Figure 11) are when some of the higher spikes in accidents were confirming the possible link between the different weather conditions and the accidents which means that they are worthwhile using as a cause in the model.

The next thing to check is the light conditions, for this the light conditions will be checked against the number of accidents that occur under each condition as this will show any relationships between the number of accidents and the conditions.

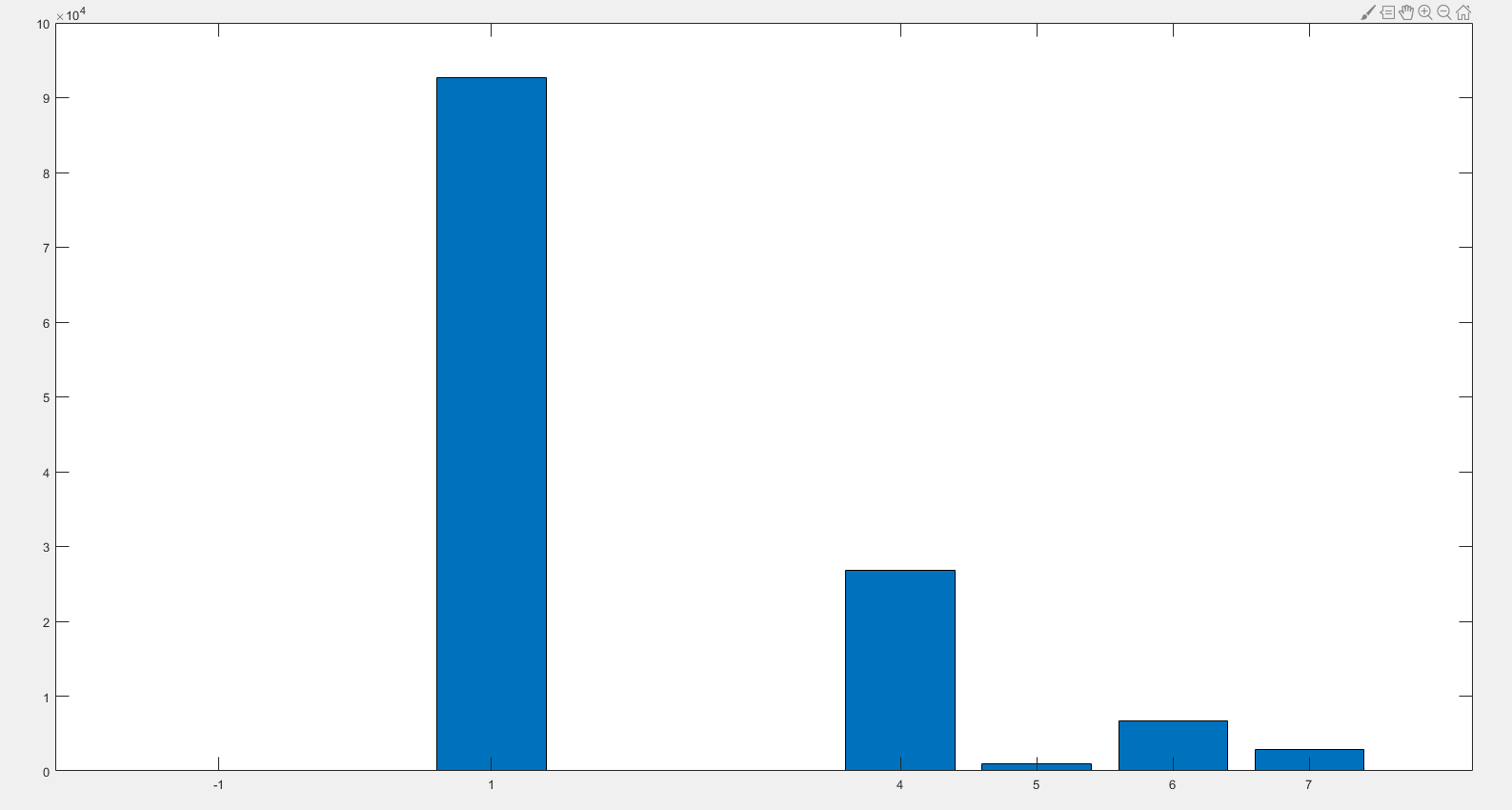


Figure 15: number of accidents for each light condition

As with the weather conditions the light conditions are categorical with numbers assigned to the type of variable (Table 13). From the graph (Figure 15) we can see that most accidents happen in daylight which isn’t surprising given the numbers of people that would be driving during the day would be much higher. For the accidents that took place in the dark most of them took place where there were lighting followed by places where no lights were present. This means that there does not seem to be an issue with accidents being caused by faulty streetlights as proportionately very few accidents were caused during times when the lights were out.

With these results I would find this to only be a minor factor in the accidents happening not a major one and would expect the model to reflect that.

To further explore the effect of the light and to look into the other factors that could come with the times of the day outside this the relationships between the number of accidents and the time of the day should be explored

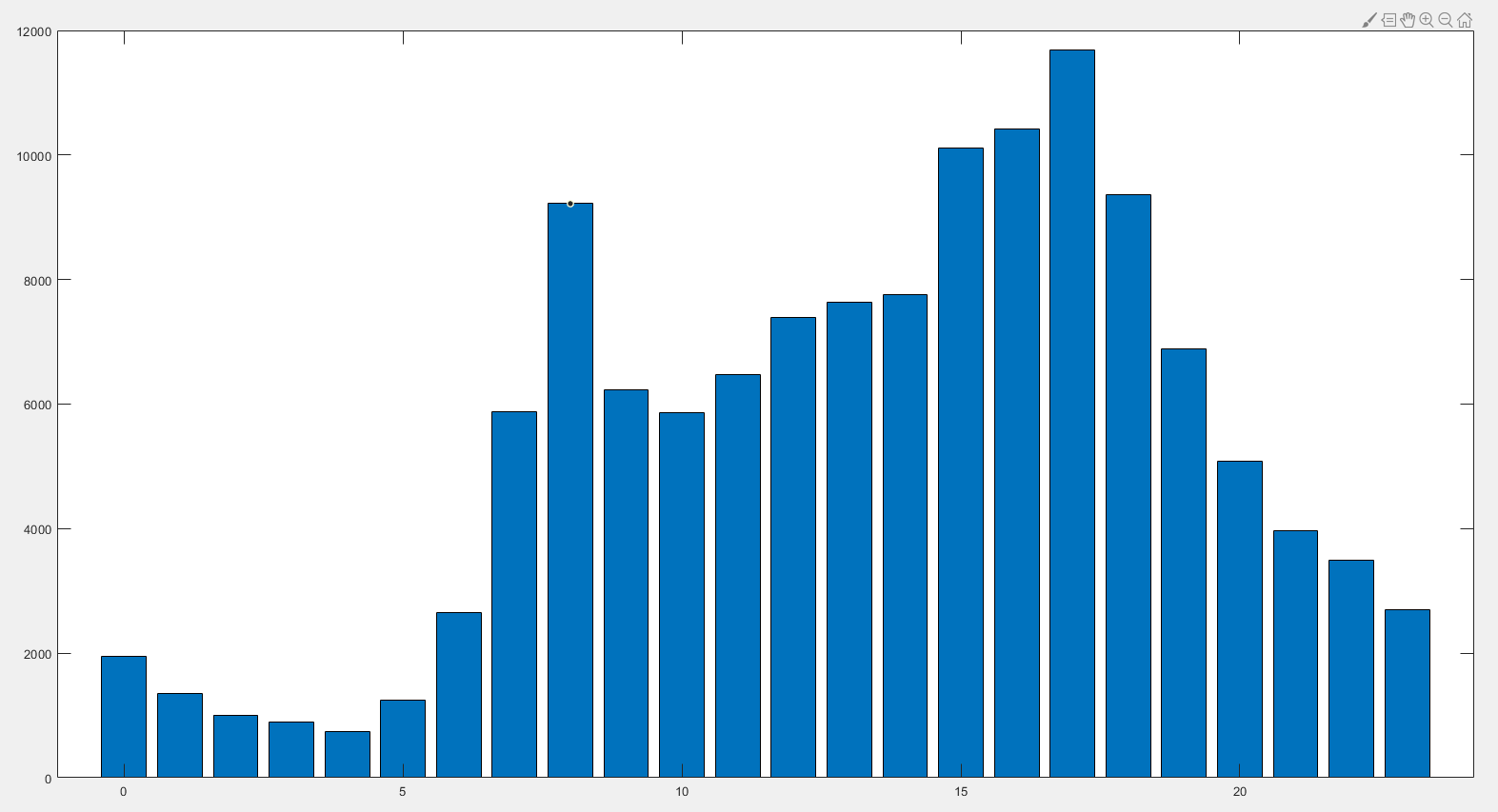


Figure 16: number of accidents by hour

With the hours of the day we can see a lot of relationships between them and the number of accidents. As seen in the light conditions we can confirm that most accidents happen during daylight hours with a gradual drop as it gets later into the night having lows at around 4am. This would be expected as there will be less traffic on the roads at that time of night which would mean less accidents.

Through the day there is a gradual bell curve to the number of accidents from morning to afternoon with the exception of large spikes in the numbers at 8 am and 5pm which would be accidents occurring at rush hours. This means that a large factor in the number of accidents on the roads is the number of cars which are there at any time as quiet times have very few accidents where rush hour has many.

The next data that would be good to compare is the number of accidents on the various types of road to see what effect the kind of road has on the number of accidents that happen.

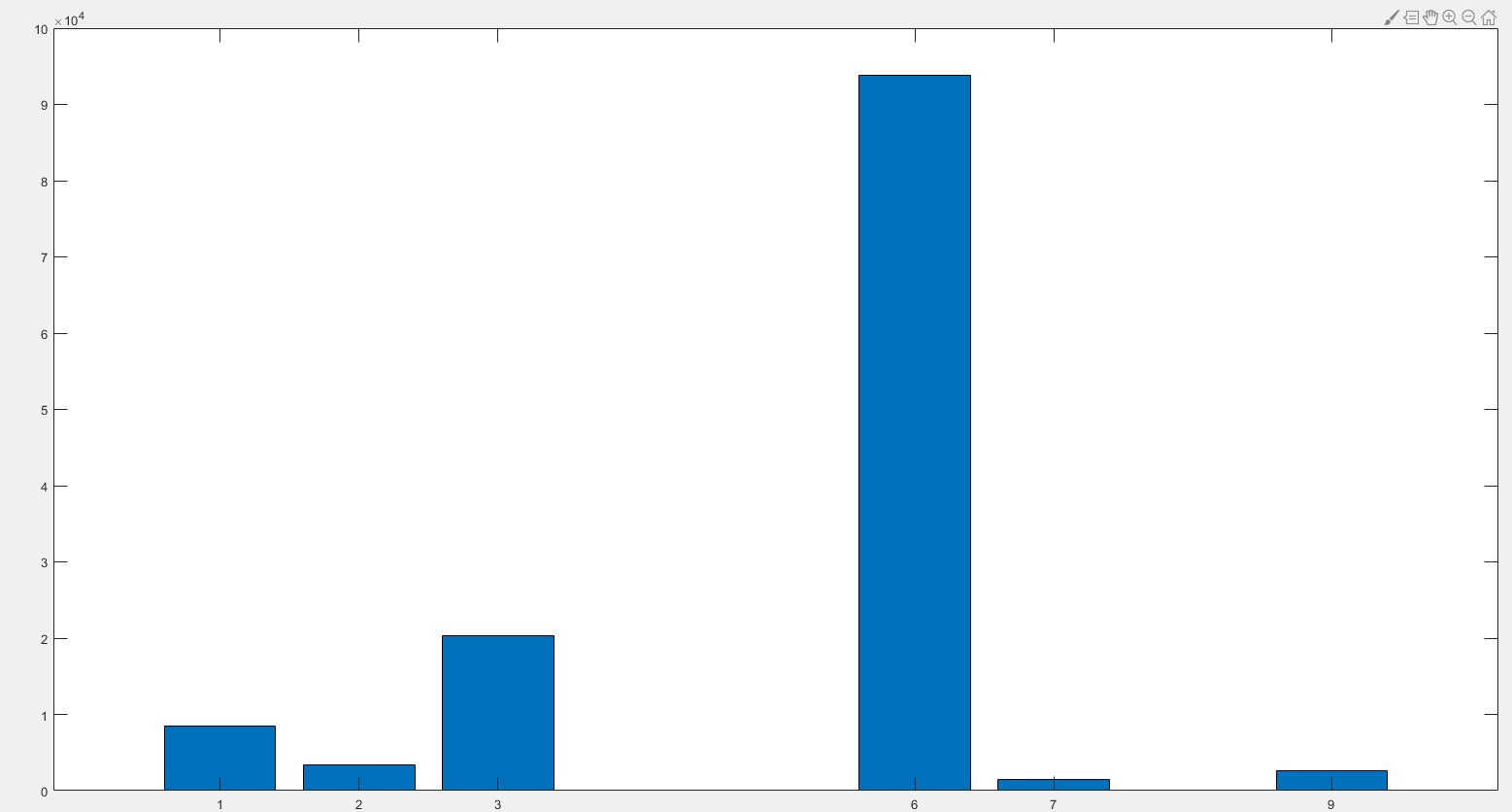


Figure 17: number of accidents by road type

The data for the road type is also a categorical variable with what the numbers represent found in (Table 14) from this data we can see that the majority of accidents happen on single carriage way roads which is unexpected as the highest number on individual roads are on major roads which are usually motorways or dual carriage ways this would suggest that while individually the most dangerous roads are larger ones overall single carriageway roads are more dangerous. This is followed though by a long way by dual carriageways with roundabouts in third. Due to the large differences in numbers it would suggest that the type of road can have a large effect on whether an accident will take place.

Also of note is that the category represented by 12 shows is for one way streets and slip roads which there is no accident data for in this time period showing this to be the safest type of road to travel on.

With roundabouts showing as one of the higher road types it would also be good to look into the accidents that take place near or on junctions. It would be expected that there would be more accidents on junctions with cars having to turn and give more potential to crash.

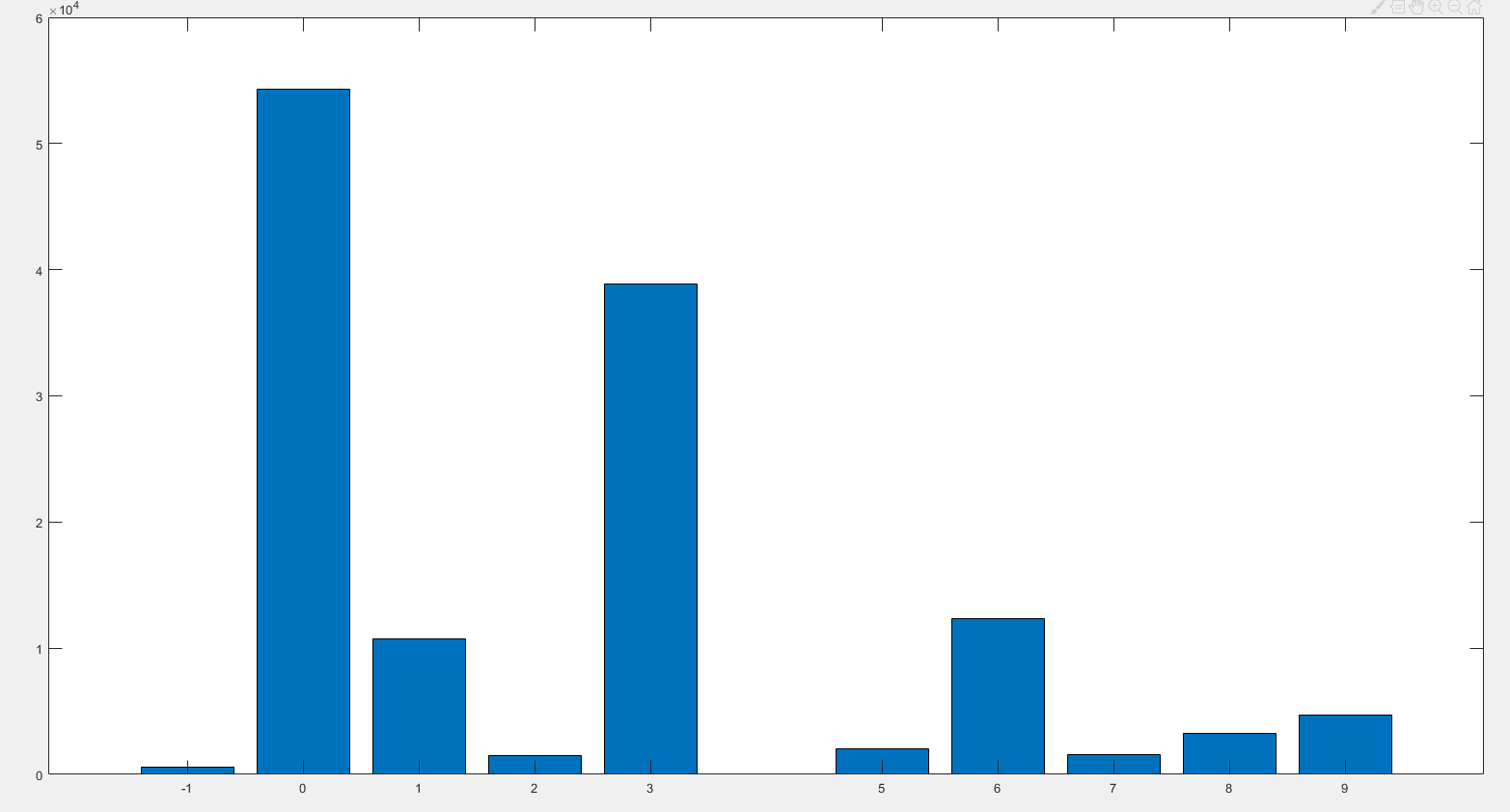


Figure 18: number of accidents by junction type

Category data can be found in (Table 15). As we can see from the graph (Figure 18) the highest number of accidents happen away from junctions however the total number of accidents across all junctions would outweigh this by a large number. Among the junction types the most accidents happen on T junctions or staggered junctions which means that the accidents take place when drivers have to cross a junction where others could be turning at the same time, this makes sense as it would increase the risk of the cars crossing each other’s path raising the odds of getting in an accident. Following from this is accidents on crossroads which would have the same issues as T junctions.

The other large number that comes out is from roundabouts with the others being more minor amounts of accidents. Given the difference between the various amounts it would seem like the type of junction that the driver is on has a significant effect on whether an accident will happen or not.

## Data Preparation

To do this data preparation MATLAB (MathWorks , 2010) will be used.

### Taking variables needed from the dataset

With the data being used already having been chosen the next step that is needed is to prepare it for use in the model. The data to be processed contains a lot of variables of various types which being based on the police reports taken from accidents there is a lot of information that is not needed.

With all of this data on police forces etc. the first step needed is to remove the variables that will be of no use to the model. The first items that will be removed are anything to do with the police forces and agencies involved with the collection of the data as this is of no use to the model. The data to do with the location will be removed also with the exception of the road class and road number which will be used to identify the locations of the roads for the purposes of this as the exact locations of the accidents aren’t of as much importance to the model being designed as the road in which they take place as a whole. As the model will be based on the causes of accidents the main causes will be kept with the data on the accidents themselves also being removed as with the predictions that will be made there is no possible way to know how severe the accident is or how many vehicles will be involved when reading in the data.

|  |  |  |
| --- | --- | --- |
| **Variable Name** | **Type** | **Notes** |
| Datetime | DateTime | Date and time |
| Day\_of\_Week | double | 1-7 starting at Sunday |
| 1st\_Road\_Class | double | categorical |
| 1st\_Road\_Number | double | Number on UK road systems |
| Road\_Type | Double | categorical |
| Speed\_limit | Double |  |
| Junction\_Detail | Double | categorical |
| Light\_Conditions | Double | categorical |
| Weather\_Conditions | Double | categorical |
| Road\_Surface\_Conditions | Double | categorical |
| Urban\_or\_Rural\_Area | double | Categorical includes ones not yet classified as either |

Table 5: simplified data dictionary of data kept for model

As well as removing variables at this stage the date and time variables which were stored separately were also combined prior to importing into MATLAB for the processing, this was done using Microsoft excels functions due to the way that MATLAB tables and datetime variables are processed as this will make them easier to work with at a later stage (Figure 19: function to combine date and time in excel The index variable was also removed from the dataset in this operation as the indexes given to each accident were found to not be unique and so not of the same use for processing the data instead our own index will be given when needed as an integer increasing by one for each value.

Figure 19: function to combine date and time in excel

=TEXT(K4,"dd/mm/yyyy ") &TEXT(M4,"hh:mm:ss")

This gave a smaller dataset to work with than before with just the variables that would be needed for the model. (Table 6)

### Dealing with categorical variables

A lot of the variables in this data are categorical in some ways using integer variables to represent the data that is given in various ways. This data will work with machine learning models but as it would be impossible for the model to know what the data meant the model would be unreliable and give out more errors as the model would assume that higher values were better rather than knowing the correct meanings for them.

This means that it will be better to represent these variables in a better way for the model to understand to do this the method that will be to take the categorical variables and turn them into multiple binary variables using a technique known as one-hot encoding as this gives the models two branches that they could make from each of the variables given for yes and no rather than a number of branches that it does not know the meaning of. This should increase the performance of the machine learning model once the data is run through it giving more accurate results. (Vasudev, 2017)

|  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| fine\_Weather\_Conditions | Rain\_Weather\_Conditions | | snow\_Weather\_Conditions | windy\_Weather\_Conditions | rainAndWind\_Weather\_Conditions | snowAndWind\_Weather\_Conditions | fog\_Weather\_Conditions | other\_Weather\_Conditions | unknown\_Weather\_Conditions | fine\_Weather\_Conditions |
| 0 | | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 0 | | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 0 | | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 0 | | 0 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 0 | | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 0 | | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 0 | | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 0 | | 0 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |

Table 6: example of one-hot encoded variable

As well as the categorical variables in the dataset the date and time will also be separated in this way to give a better view in the model of when the accidents take place as this will be useful in predicting the chance of an accident happening to do this the hour of the day and month of the year will be split off as categories and encoded in the same way.

### Negative data

The final major issue with the data for the model is that every example that there currently is to feed into the model is a positive result as the only data in the set currently is data on the accidents that took place. Feeding this data into the model will train it however it will only have positives to read from this will mean that the model cannot accurately predict when an accident will occur due to it not having any information given to it on when they do not meaning the resulting model will be inaccurate.

To solve this we need some negative data adding to the dataset. For data like this gaining this data is relatively easy because as this is a complete set of all the accidents in the time period which resulted in an injury any time not in the dataset for a particular road is a time when there was not an accident so complete dataset would be the accidents recorded and every time not recorded as negative data. However this comes with its own problems as a. the data set would be impossibly large including every minute of the year but also it would create a large class imbalance which would make the model not able to make a prediction as the number of none accidents would vastly outweigh the number of accidents. To solve this only a number of times on each road when this does not occur will be taken.

To do this we will randomly pick times and a road then check against the accidents to make sure it is not in there before adding the other data variables to it as appropriate.

### Randomize data

Finally once the data has been prepared the data will need to be randomized this is so that the model gets more than a linier set of results that are positive and negative and ensures that the results from the models prediction are as accurate as possible.

## Choosing a model

The final step that needs to be taken before creation of the machine learning model is to choose the right type of model for the project. What is known is that the model needs to predict what the outcome will be given a set of parameters and that the data is mostly categorical. Given this the best type of model would seem to be a decision model.

The simplest of these is a decision tress which is a model that makes choices between the data in steps deciding between each variable which has influence over the end result.

Decision trees by themselves are simple models which are a good way to represent data in a format that is easy to understand for the viewer however on their own they have a tendency to over fit or under fit the data because of this many more advanced models iterate through various trees and the result comes from a comparison of all of the trees at the end to reduce the errors in the outcomes. The gradient boost model works in such a way this provides much more accurate predictions as an end result and so is the kind of model that will be used.

# Implementation

## Model

Once all the data is prepared and the model chosen the next step is to create the machine learning model.

For the creation of the model the language R (R Core Team, 2013) will be used as it allows for a greater toolset for statistical work including a lot of open source packages that can be used for working with machine learning.

First using R a number of packages will be needed.

Figure 20: code for loading packages in R

library(rsample)

library(caret)

library(ggthemes)

library(scales)

library(wesanderson)

library(tidyverse)

library(gbm)

library(Metrics)

library(here)

The majority of these are just to allow access to various plots and data manipulation to make working with the model easier rather than needing to write functions to deal with the organisation of the data when needed. The important library here however is the gbm package which provides the gradient boosting model that will be used.

Once the packages are loaded the first thing that will be needed is to separate the data into two pieces one to use as a training set and a second to use as a testing set. This will give us data with which the model will be created and can learn from and another set to which we know they outcomes that can be used to run tests on how the model predicts the accidents after. The package rSample has a function which makes this easy.

Figure 21: code to split data into training and testing sets

set.seed(123)

RoadSplit <- initial\_split(roads, prop = .7)

RoadTrain <- training(RoadSplit)

RoadTest <- testing(RoadSplit)

This code takes the initial set and splits it with 70% becoming the training data and the other 30 being stored for using as the testing data. The seed setting beforehand is to seed the random number generator that is used for selecting the data to be separated.

Once this data has been split the gbm package is used to generate the model using the training data and a set of input parameters.

Figure 22: code for generation of model

road\_fit\_1 <- gbm::gbm(IsAccident ~.,

data = RoadTrain,

verbose = TRUE,

shrinkage = 0.01,

interaction.depth = 10,

n.minobsinnode = 5,

n.trees = 1000,

cv.folds = 10

)

This code generates the model, the parameters used for this are first what it is we wish to find which is whether an accident has taken place, next is the data that will be used for the models creation which is the training data that has been separated off.

Next we give a true or false for verbose which increases the amount of error data that will be given by the model for any troubleshooting which is needed.

The shrinkage value is the step-size reduction that is applied to each tree in the model meaning the amount of information that the model processes using the trees is reduced, lower numbers require more trees to get a good result however will maintain the accuracy of the model.

The interaction depth value gibes the number of maximum depth that is allowed in each tree which has been set to 10 due to some of the variables having a large number of Boolean values for them which will need to interact.

Minobsinnode gives the minimum number of observations which can be in the terminal nodes of each of the trees.

The number of trees gives the number of times that the model will iterate through decision trees, more of these will give a higher accuracy but at the expense of computing time. Once the model has run however plots can be made which will show us the ideal number to use.

Finally the value for cv.folds is the cross validation variable which gives the amount of validation which will be applied to the model to allow the calculation of a generalization error for seeing how the model is working once it is complete.

Once the model is run some data can be taken from it to check how it has gone. The first of these is the error rates from the model, as the model runs iterations of 1000 trees to find the best fit we need the error rate across all of the runs for it.

Figure 23: plots a graph of the error from the models using the cross fold validation

perf\_gbm1 = gbm.perf(road\_fit\_1, method = "cv")

From this we get a graph which shows the error using the cross fold validation. (Figure 36) this graph also shows us the optimal number of trees to use for the model which can be used when using the testing data and other outputs to ensure that more iterations aren’t run than are needed.

Once the error rates have been collected the next piece of information needed from the initial model is the dependencies of each of the variables in the model. This can be obtained by printing a summary from the gbm package as a table.

Figure 24: code for getting a summary showing the relative dependencies of the variables

roadEffects <- tibble::as\_tibble(gbm::summary.gbm(road\_fit\_1,

plotit = FALSE))

roadEffects %>% utils::head()

The data from this comes out as a long list showing how much the model relies on each of the variables that were fed into it which allows us to see which of them are most important and so would be the biggest factors in who has an accident from what was fed in.

As well as the list the code plots a graph of them showing each of the variables importance. Due to the number of variables a smaller graph plotting just the top 10 was also created as this will allow for an easier interpretation of the data.

Now that we have data on all of the variables as a collection we can see which ones are the most dependant and pull out some data which will show better how the variables interact, this is done with the use of a partial dependence plot.

Figure 25: code to plot a number of partial dependencies

gbm::plot.gbm(road\_fit\_1, i= 'November\_DateTime')

gbm::plot.gbm(road\_fit\_1, i= 'December\_DateTime')

gbm::plot.gbm(road\_fit\_1, i= 'Daylight\_Light\_Conditions')

gbm::plot.gbm(road\_fit\_1, i= 'CrossRoad\_Junction\_Detail')

These graphs show us the correlation between the variables in this case how each of the variables taken correlate to whether an accident was caused according to the model.

A number of these plots are created to show some important variables to the model and how they interact, the code from (Figure 25) shows the creation of plots for the top 3 variables as the model is most dependant on them as well as a variable that the model is less dependent on to give a view of why.

Once data has been taken from the model showing what we can from the initial training the next thing needed is to feed in the test data and make some predictions based off of it to give a better idea of how the model is performing.

# Testing

## Model Test Data

Feeding test data into the model will make some predictions on that data towards the outcome and allow the checking of how the model is performing. This is done with the data that was separated off before the creation of the model (Figure 21) as this is data which we know the outcomes from and so can check the results of the model against the actual data and is also a large amount of data in which to check the integrity of the model.

Figure 26: code for predictions using test data

n.trees = seq(from=100 ,to=1000, by=100)

predmatrix<-predict(road\_fit\_1,RoadTrain,n.trees = n.trees)

dim(predmatrix)

test.error<-with(RoadTrain,apply( (predmatrix-IsAccident)^2,2,mean))

head(test.error)

plot(n.trees , test.error , pch=19,col="blue",xlab="Number of Trees",ylab="Test Error", main = "Perfomance of Boosting on Test Set")

abline(h = min(test.error),col="red")

legend("topright",c("Minimum Test error Line for Random Forests"),col="red",lty=1,lwd=1)

This code (Figure 26) creates a prediction matrix from the data based on a sequence from the trees and the test data. Then plots a graph base on the testing which shows its performance by plotting the test error given by the model against the number of trees run through the model to give how the error changes over time while the model is running through its iterations. This gives us a view of how much of an error the model will give over time while being used for predictions. This graph also includes a line which gives the minimum test error to allow for comparisons.

In order to see how the model predicted data the predictions will be taken and then compared with the actual results to allow us to inspect how the model performed against the known results of the test data.

Figure 27: code for plotting of actual results against the predicted results

RoadTest$predicted <- base::as.integer(predict(road\_fit\_1,

newdata = RoadTest,

n.trees = perf\_gbm1))

ggplot(RoadTest) +

geom\_point(aes(y = predicted,

x = IsAccident,

color = predicted - IsAccident), alpha = 0.7) +

theme\_fivethirtyeight() +

theme(axis.title = element\_text()) +

scale\_x\_continuous(labels = comma) +

scale\_y\_continuous(labels = comma) +

scale\_color\_continuous(name = "Predicted - Actual",

labels = comma) +

ylab('Predicted Accidents') +

xlab('Actual Accidents') +

ggtitle('Predicted vs Actual Accidents')

Plotting the results in this way gives us a graph with the points placed for the actual and test results which will allow us to see the correlation between the two and see how the model has performed.

Finally in order to compare what’s happening in the model it may be handy to take out the individual trees from the model. As this data is usually just for debugging purposes there is no way to visually represent the trees but they can be accessed as a data frame which shows how they are taken out, the data will be taken for the first and last tree in the model as well as one from the middle to give an over view of how the model progressed between them.

Figure 28: code for taking out the data on the formation of individual trees

tree1 <- pretty.gbm.tree(road\_fit\_1,i.tree = 1)

tree1000 <- pretty.gbm.tree(road\_fit\_1,i.tree = 1000)

tree500 <- pretty.gbm.tree(road\_fit\_1,i.tree = 500)

## Comparison models

To be able to test the model and better see the way that it is working some other models will be made in which to compare the model to and see how it performs in comparison to them.

As they were discussed as part of the decision to choose the main model (2.3) the two models that will be used as a comparison for this one are a Decision tree classifier and a Random forest classifier. The decision tree to show how the gradient boosting model compares to the base form of this type of model and a random forest to show its performance in comparison to another model built on a similar principle.

As the models are for comparison purposes they will be created using Weka (Eibe , et al., 2016) while coding with R gives a more flexible approach to building the models WEKA allows for them to be created more efficiently using prebuilt model explorers available for the program.

To ensure that the models are equal in their creation the same data set will be used with a set up as similar as can be achieved with the models as this will create the best comparison for the original model.

### Creation process

To create the model using weka the first step is to open the program and access the file through the interface.

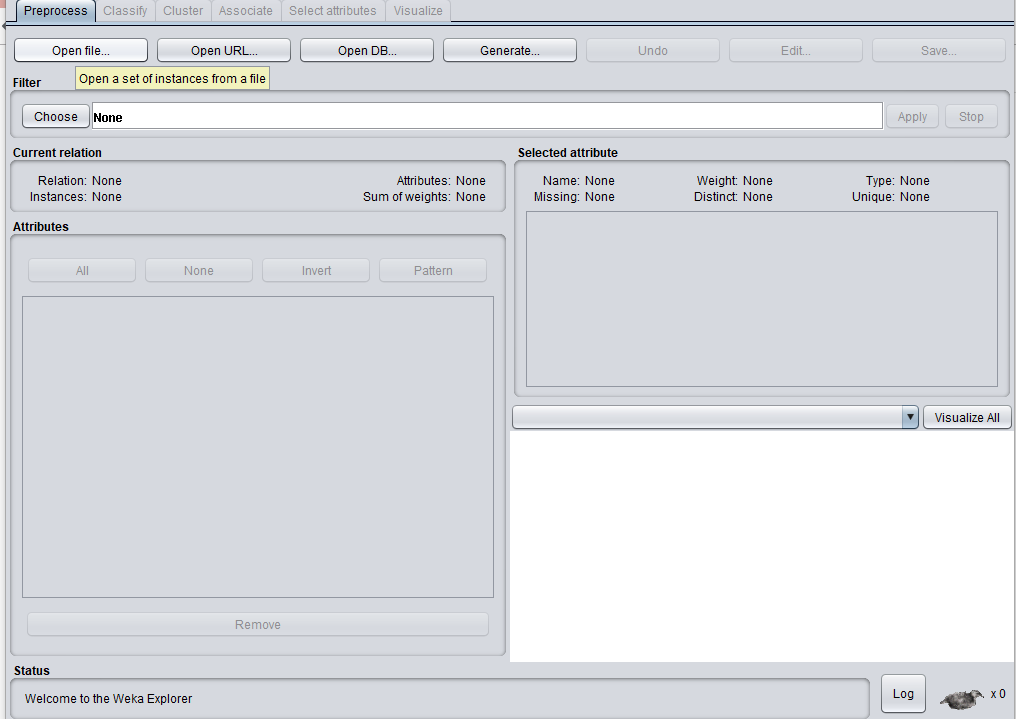


Figure 29: open file in weka interface

Weka allows the access too many forms that the data can take in this case we open as CSV and access the data that was processed earlier.

Once the file has been opened the interface allows for a number of filters to be added and the data to be manipulated for the model creation, the majority of this data is for creation of models from the raw data using the program without the previous data manipulation.

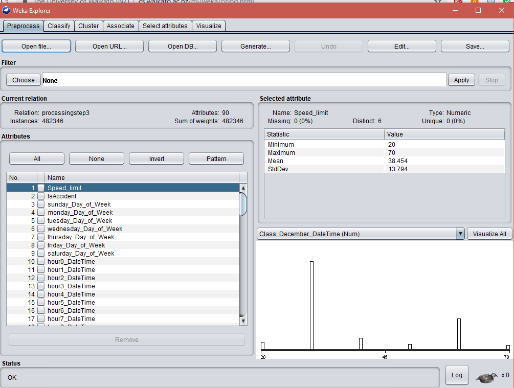


Figure 30: weka interface data loaded

As the model being used has no extra manipulation from this data before creation of the model and the data has been pre-processed this step will be skipped and instead access the classify tab in weka to create a model.

Once in this new tab a model needs to be chosen from the choose section at the top of the interface in order to select the tree that will be used for its creation. Given the data we have is all categorical and sorted as numbers the tree that will be used is the REP decision tree as others such as J48 are created to sort none numerical data.

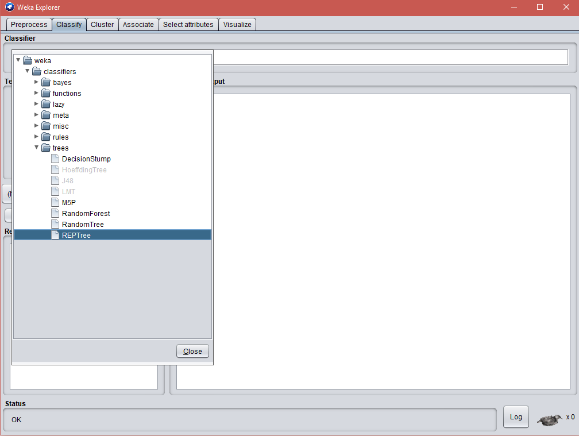


Figure 31: choosing a tree in weka

After this we build the model using the training data. Once the model has run the interface will give out a large amount of data. Given the number of variables in the dataset being used the majority of this data is the way that the tree is formed as text but more importantly the bottom of this data contains a summary which includes the Error rates for the model. This data will be important for comparisons with the model being used for this report.

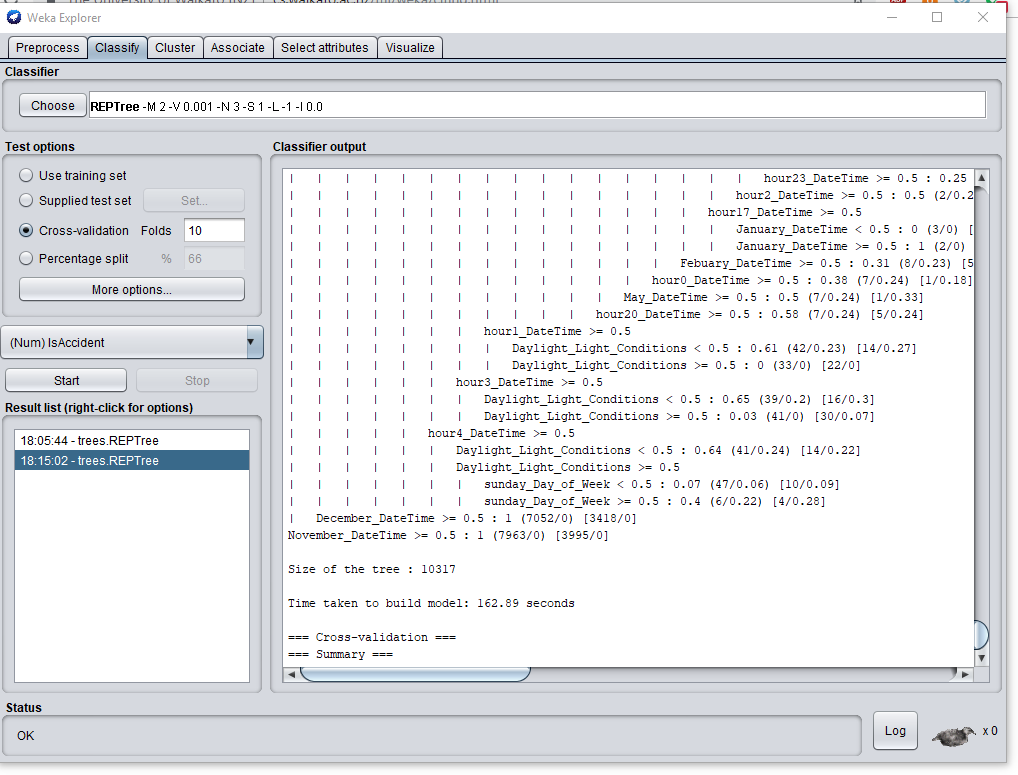


Figure 32: weka interface after model creation

From here it is also possible to visualise the tree that the model has created as well as create graphs for the error output for the tree which will be useful for comparing to the visual output for the gradient boosting model. Especially for the Random forest as this creates many trees in the same way that the gradient boosting model does and so will not be able to give the trees as outputs.

# Results

## Model Results

### Relative influence for training set

#### Graph

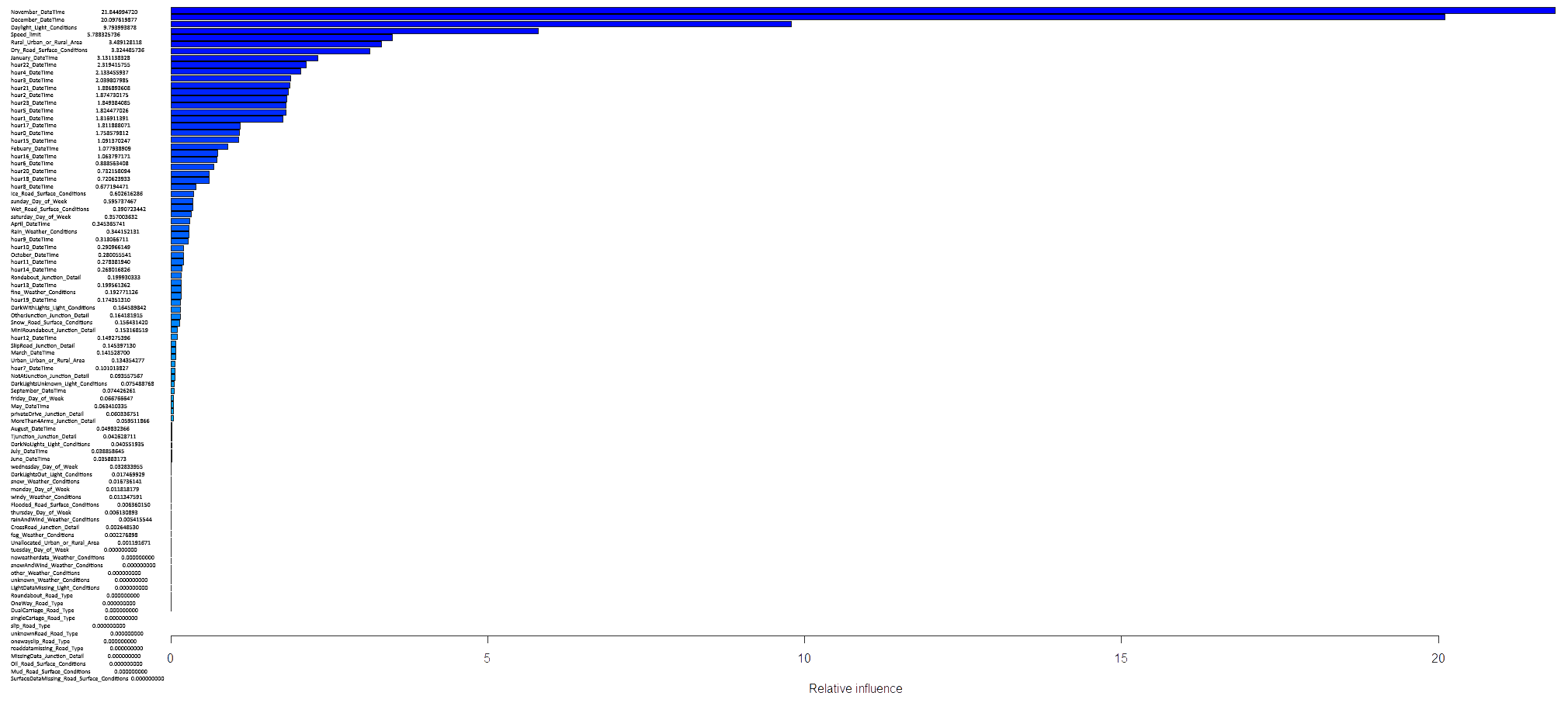


Figure 33: relative influence of model with training set

#### Top 10

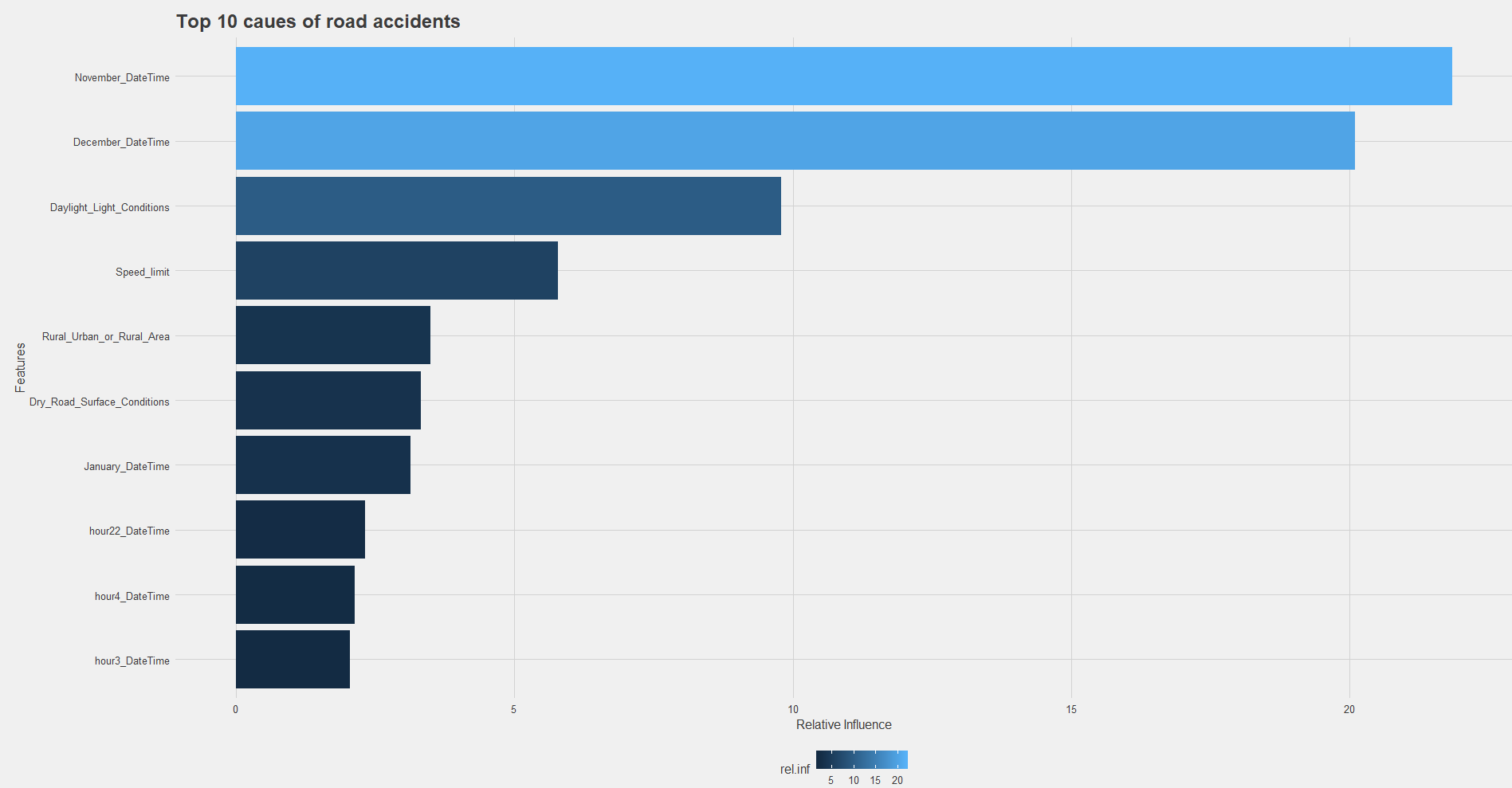


Figure 34: graph showing the top 10 from the relative influences

### Error rate and optimal number of learners

#### Graph

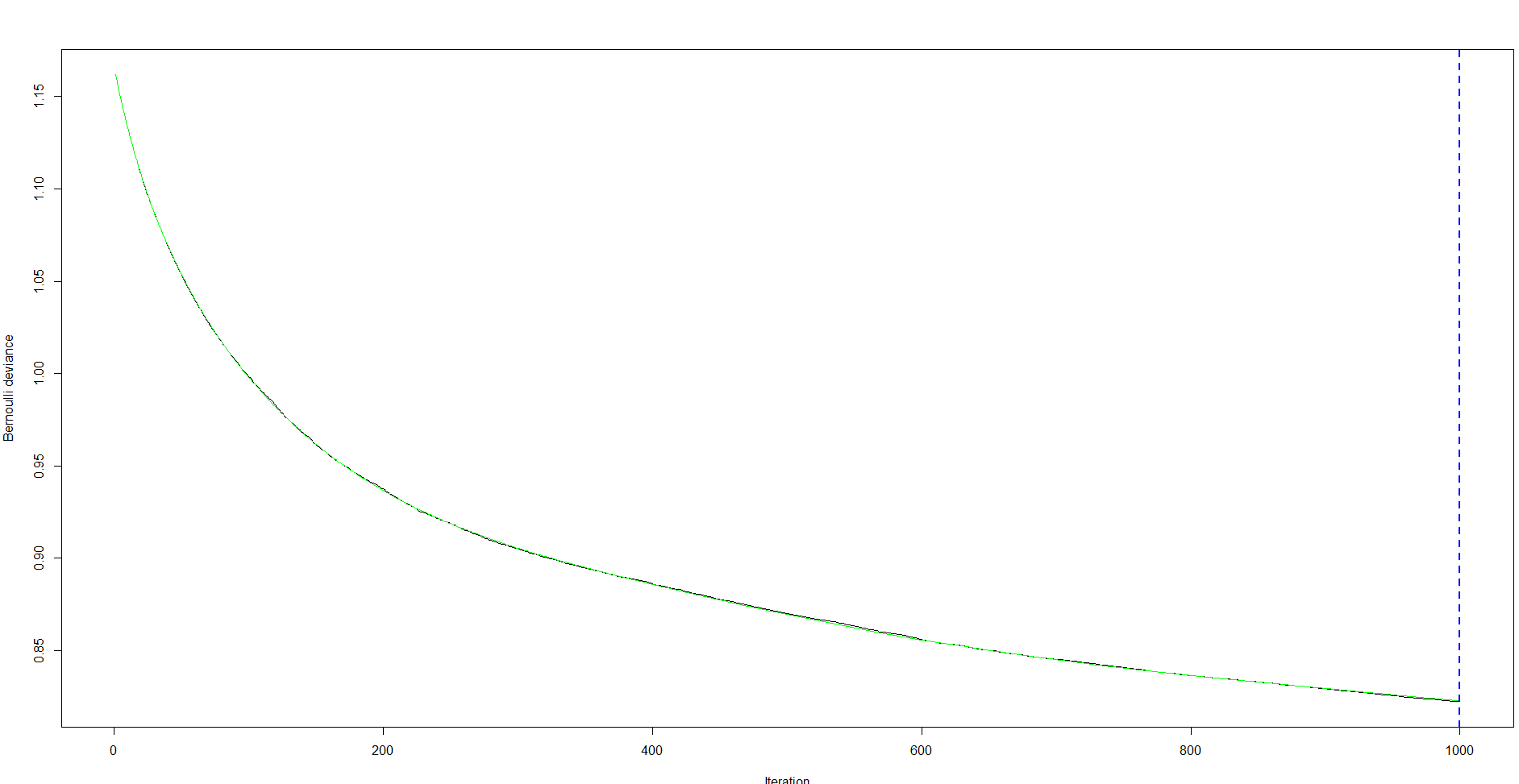


Figure 36: graph showing the error rate at each learner with the optimal number of learners

#### Results

Iter TrainDeviance ValidDeviance StepSize Improve

1 1.1618 nan 0.0100 0.0018

2 1.1582 nan 0.0100 0.0018

3 1.1547 nan 0.0100 0.0017

4 1.1514 nan 0.0100 0.0017

5 1.1481 nan 0.0100 0.0016

6 1.1449 nan 0.0100 0.0016

7 1.1418 nan 0.0100 0.0016

8 1.1388 nan 0.0100 0.0015

9 1.1358 nan 0.0100 0.0015

10 1.1330 nan 0.0100 0.0014

20 1.1073 nan 0.0100 0.0011

40 1.0690 nan 0.0100 0.0008

60 1.0399 nan 0.0100 0.0006

80 1.0171 nan 0.0100 0.0005

100 0.9986 nan 0.0100 0.0004

120 0.9827 nan 0.0100 0.0004

140 0.9681 nan 0.0100 0.0003

160 0.9557 nan 0.0100 0.0002

180 0.9459 nan 0.0100 0.0002

200 0.9376 nan 0.0100 0.0001

220 0.9289 nan 0.0100 0.0001

240 0.9218 nan 0.0100 0.0002

260 0.9153 nan 0.0100 0.0002

280 0.9098 nan 0.0100 0.0001

300 0.9052 nan 0.0100 0.0001

320 0.9006 nan 0.0100 0.0001

340 0.8967 nan 0.0100 0.0001

360 0.8930 nan 0.0100 0.0001

380 0.8895 nan 0.0100 0.0001

400 0.8862 nan 0.0100 0.0001

420 0.8831 nan 0.0100 0.0001

440 0.8797 nan 0.0100 0.0001

460 0.8765 nan 0.0100 0.0001

480 0.8732 nan 0.0100 0.0001

500 0.8701 nan 0.0100 0.0001

520 0.8674 nan 0.0100 0.0001

540 0.8649 nan 0.0100 0.0001

560 0.8617 nan 0.0100 0.0001

580 0.8591 nan 0.0100 0.0001

600 0.8560 nan 0.0100 0.0001

620 0.8536 nan 0.0100 0.0000

640 0.8512 nan 0.0100 0.0000

660 0.8492 nan 0.0100 0.0000

680 0.8470 nan 0.0100 0.0001

700 0.8454 nan 0.0100 0.0000

720 0.8437 nan 0.0100 0.0000

740 0.8419 nan 0.0100 0.0001

760 0.8401 nan 0.0100 0.0000

780 0.8381 nan 0.0100 0.0000

800 0.8366 nan 0.0100 0.0001

820 0.8351 nan 0.0100 0.0000

840 0.8337 nan 0.0100 0.0000

860 0.8324 nan 0.0100 0.0000

880 0.8309 nan 0.0100 0.0001

900 0.8295 nan 0.0100 0.0000

920 0.8279 nan 0.0100 0.0001

940 0.8264 nan 0.0100 0.0001

960 0.8249 nan 0.0100 0.0000

980 0.8237 nan 0.0100 0.0000

1000 0.8224 nan 0.0100 0.0000

Figure 37: error rate raw results

### Partial Dependence plots

#### November Date Time

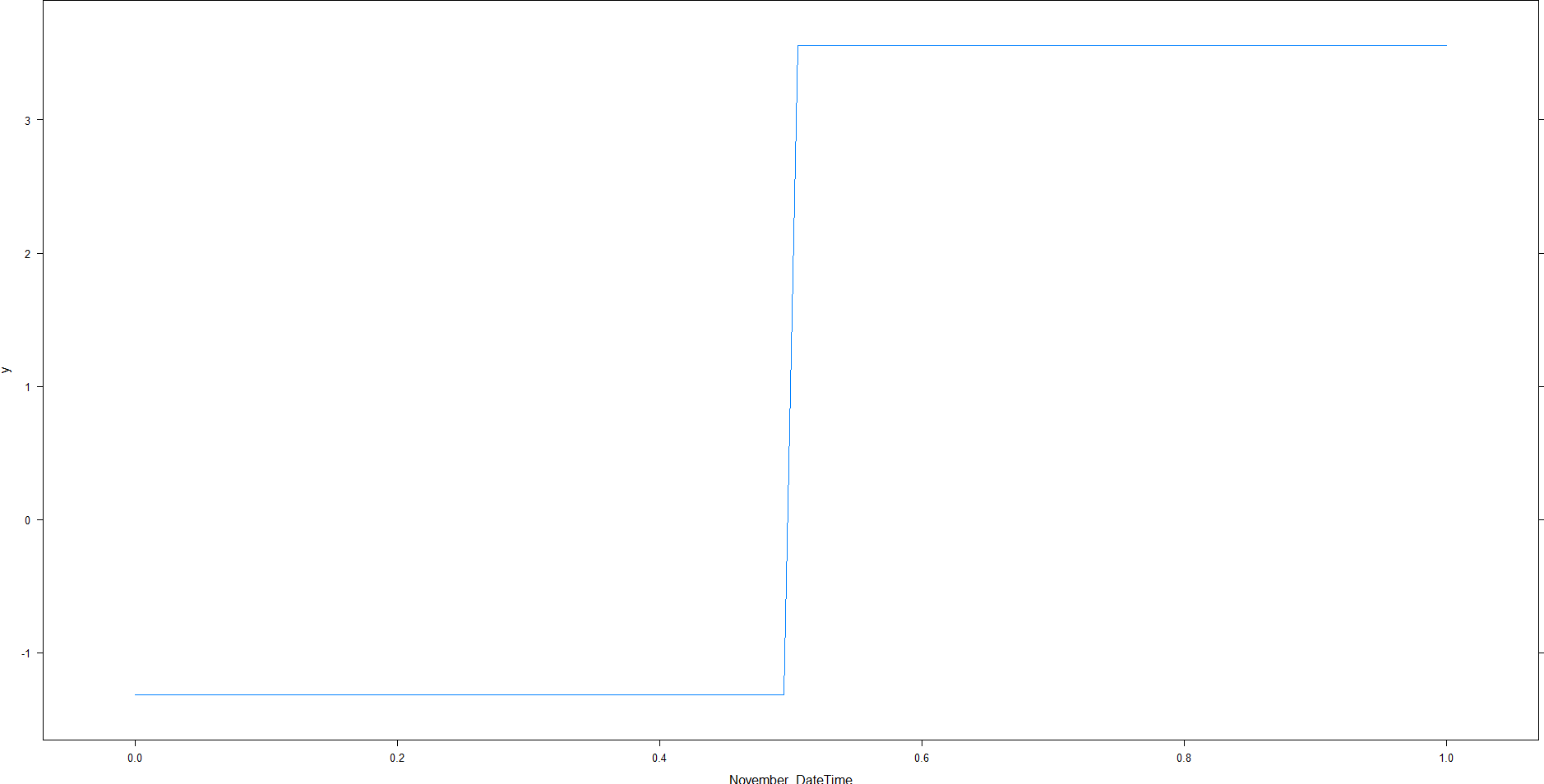


Figure 38: partial dependence plot for November date

#### December Date Time

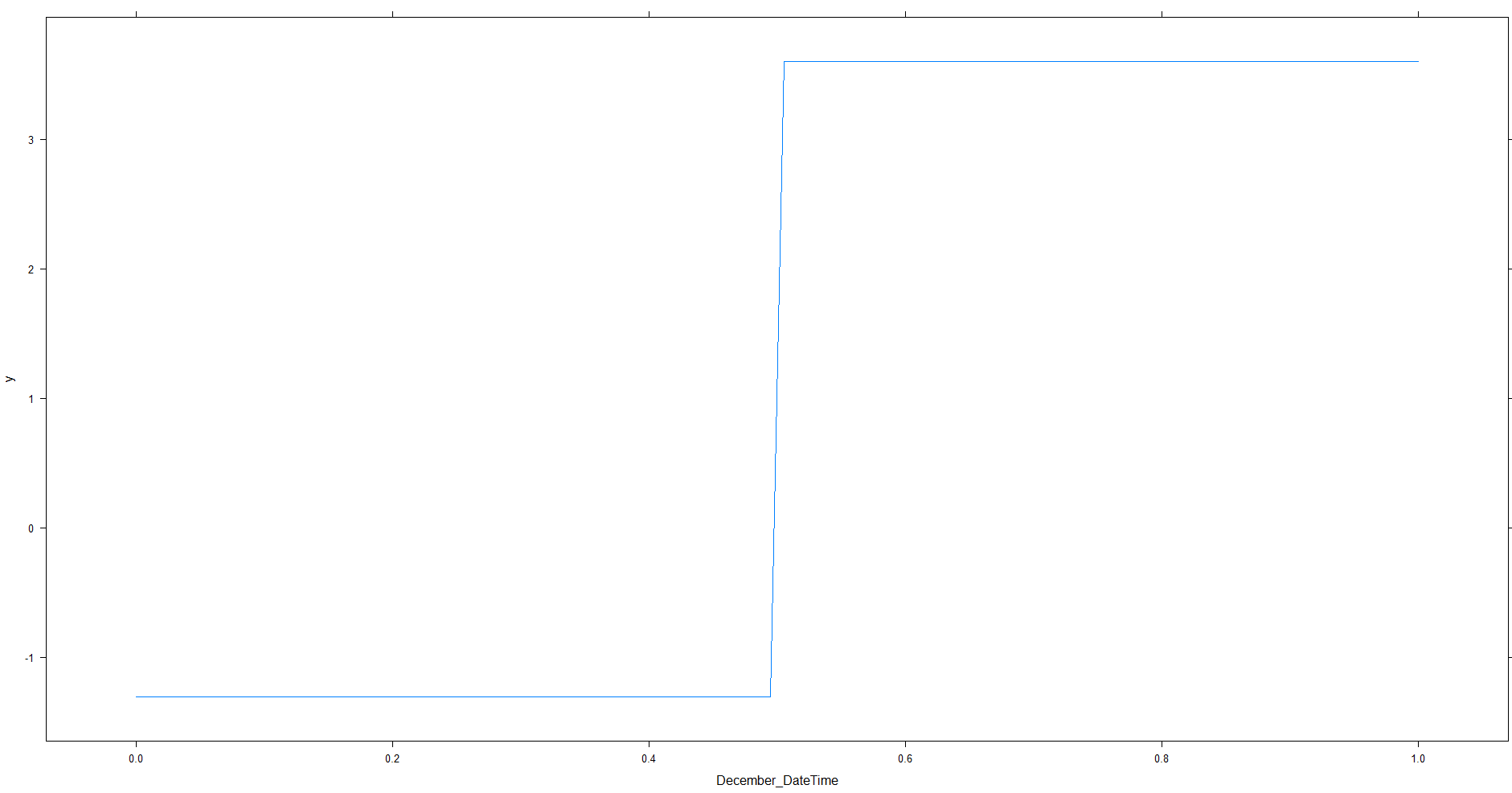


Figure 39: Partial dependence plot for December date

#### Daylight

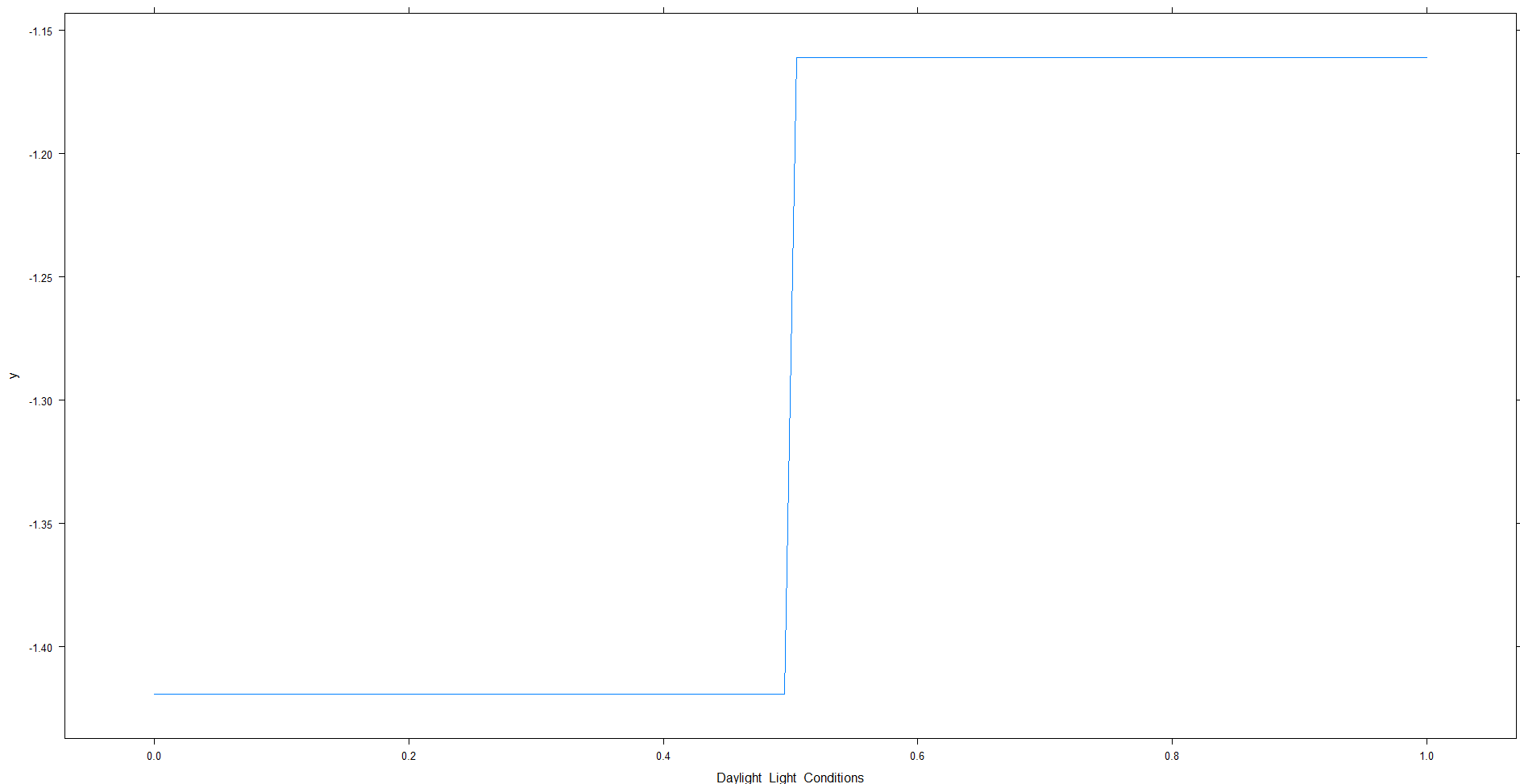


Figure 40: partial dependence for daylight conditions

#### Crossroads

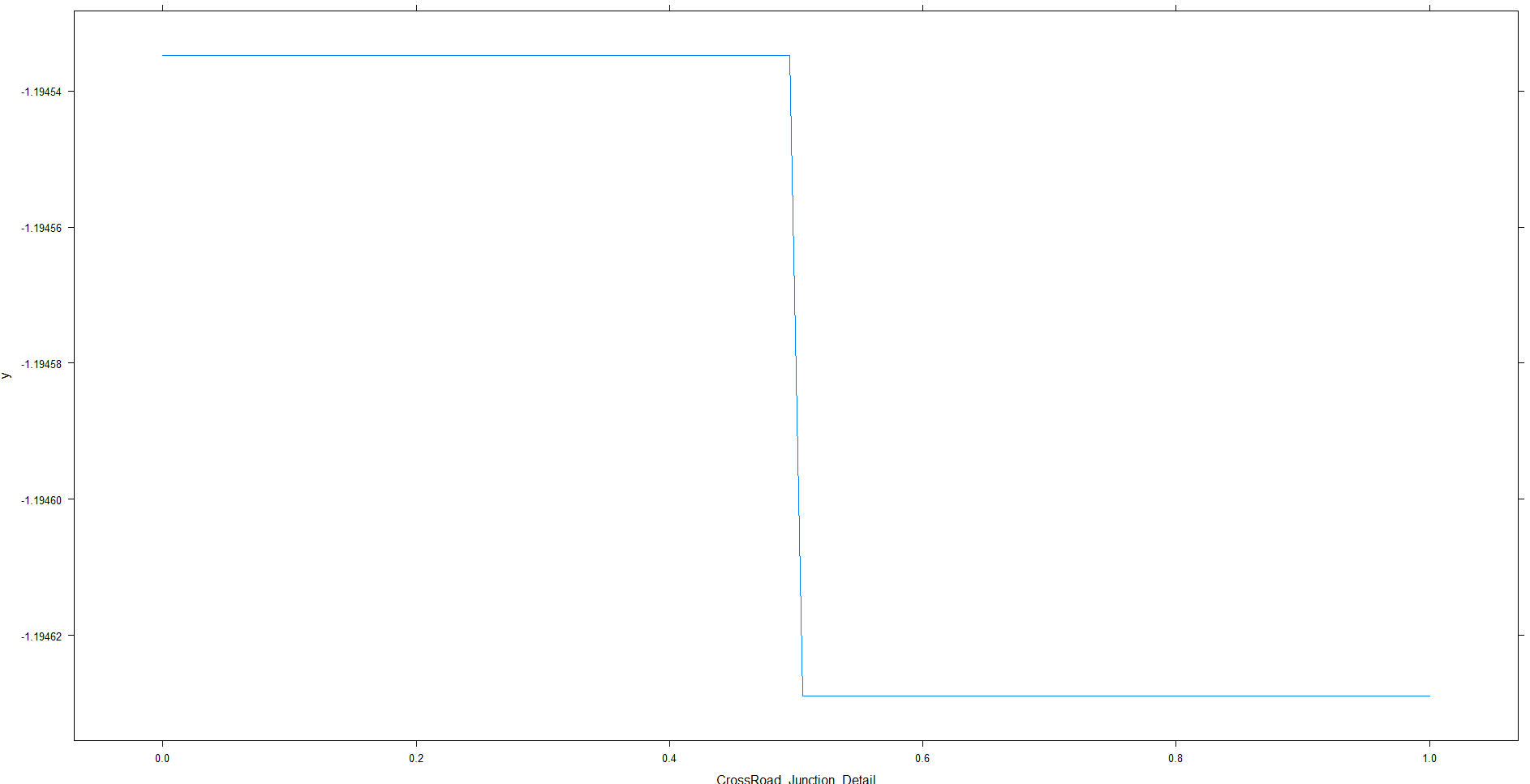


Figure 41: partial dependence plot for crossroads

### Predictions with test set

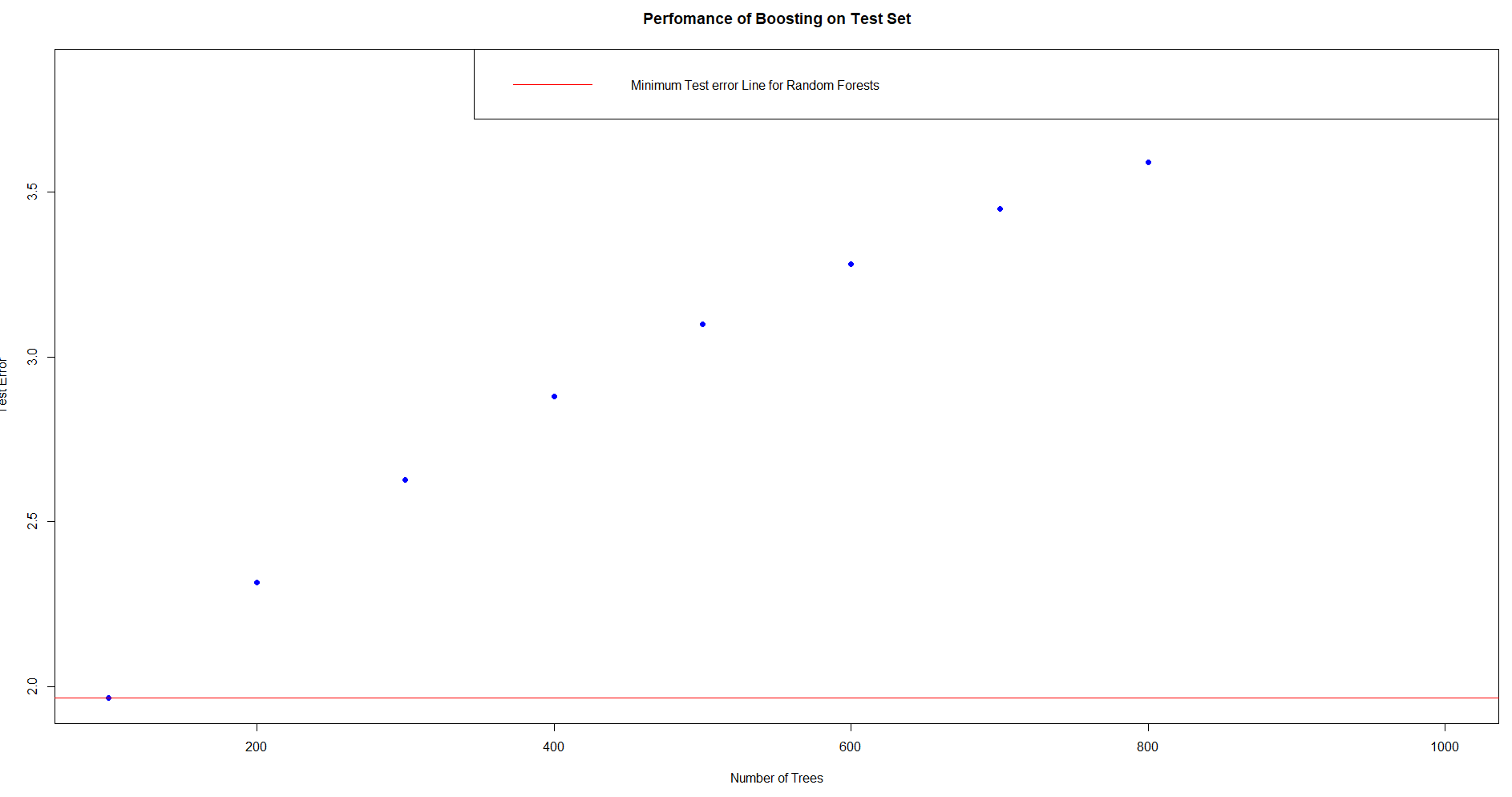


Figure 42: graph showing the performance of boosting using the test data with a line for the minimum test error

### Predicted Vs Actual Accidents

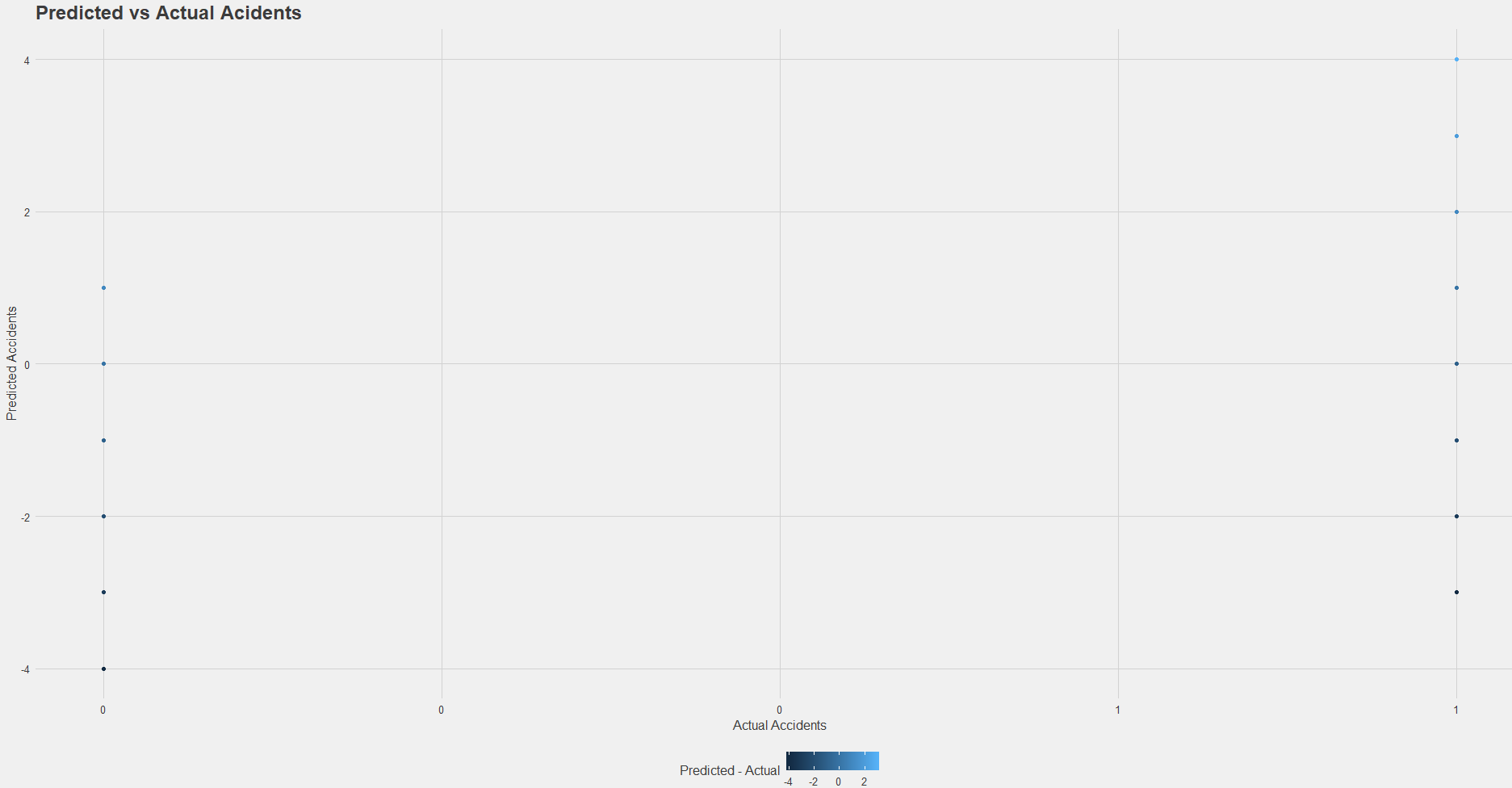


Figure 43: graph showing the predicted results against the actual results

## Other models for comparison

### REP Decision Tree

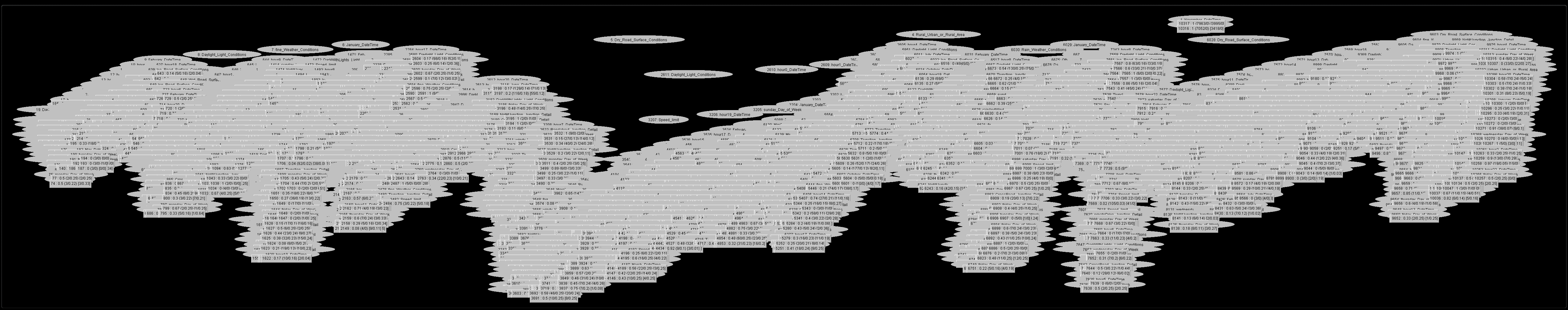


Figure 44: REP decision tree diagram

### REP Decision tree visualised error

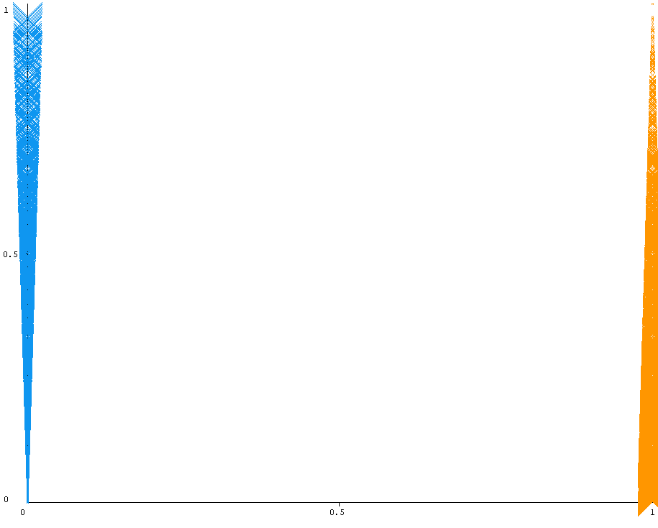


Figure 45: REP decision tree visualised error graph

### 

### REP Decision Tree Error Summary

Correlation coefficient 0.6359

Mean absolute error 0.2345

Root mean squared error 0.3424

Relative absolute error 59.5595 %

Root relative squared error 77.1748 %

Total Number of Instances 482346

### Random Forest Summary

Correlation coefficient 0.9961

Mean absolute error 0.0031

Root mean squared error 0.0153

Relative absolute error 7.184 %

Root relative squared error 10.5263 %

Total Number of Instances 482346

### Random Forest Details

RandomForest

Bagging with 100 iterations and base learner

weka.classifiers.trees.RandomTree -K 0 -M 1.0 -V 0.001 -S 1 -do-not-check-capabilities

Time taken to build model: 484.31 seconds

### Random Forest Error Visualisation

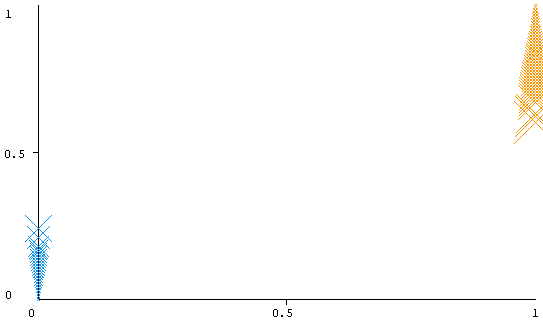


Figure 46: random forest error visualisation graph

# Evaluation of results

## Analysis

### Relative influence

From the results taken from the model showing the relative influence of the (Figure 35Figure 37) it can be seen that there are a number of variables that did not have any influence on this version of the model, these variables form part of the one shot encoded variables that were taken from the raw data showing some of the categories that were included in the data which the model takes to have no influence on passed years. It can also be seen that the variables that have the most influence in the models are the “November\_DateTime” and “December\_DateTime” variables which as was expected to have an impact based on the visualisation performed (Figure 11). These being highest on the list is not unexpected as the other factors in an accident were found to have many connections to the time of the year as the month effects the light level as well as the weather conditions that the drivers have to deal with which will increase the chances.

The next highest factor that was seen was the Daylight\_Light\_Conditions again from the visualisation (Figure 15) this was expected as the number of accidents in these conditions were much higher than the other categories.

The top 10 graph (Figure 34) created gives the best view of what had the most influence on the model as with all the results viewing some of the smaller influences can be difficult due to the high influence of the top conditions and the number of variables that are plotted on the graph, this shows the conditions that are most needed in the model and in any future iterations which would need to be looked at closer.

### Model error

From the graph produced which shows the error rates of the model (Figure 36) we can see that the models error rate starts off low meaning that the data used and model type must be relatively accurate for what is being calculated with each iteration the error rate drops with a smooth curve down being generated showing that the iteration model is the correct way to go with this calculation. The optimal number of iterations which was produced on this graph also shows that the best number to have is just below the 1000 mark which was used for the initial production and so the model will not need to be recreated due to a mistake in the number of trees that was used to keep the model in an optimal state.

### Partial dependence

From the partial dependence plots we can see that the two taken from the months that give the highest influence on the model (Figure 38 & Figure 39) both have a positive correlation when plotted against the final result of having an accident. This shows that when the model has run the majority of the trees these variables lead to an accident in the majority of cases and are a factor in the cause. The same can be found with the plot for the daylight conditions (Figure 40). When looking at the crossroads condition further down the list of influences (Figure 41) however we see that this condition has a negative correlation showing that from our results the model has generated that crossing a crossroad in fact has a negative influence compared to the rest and leads more to not having an accident. This could be due to the generation of the random negative data or because of the lower instances of accidents on crossroads compared to other junctions.

### Comparison to other test models

The test models that were created for decision tree and random forest models give some results that can be compared to the model and some others that could be analysed in their own right. As we can see from (Figure 44) a decision tree with this number of variables causes a tree that when visualised is hard to see what is going on, the number of branches in the tree become huge with so many different variables that are dependent on each other. With the gradient boosting mode trees as can be seen with the data in (Table 9, Table 10 & Table 11) the iterative nature of the gradient boosting model results in much smaller trees with less branches which iterate through thousands of times instead of fitting all the data to one tree which means that they are less likely to cause issues with over fitting of the data from the large number of branches.

From the error rate of the model a big difference can be seen from the graph showing the error for the gradient model (Figure 36) we can see that the error rate is low in the beginning and drops steadily with each iteration of the model because of the way the model uses multiple trees in order to fit the data better to the decision tree while the error rates for the decision tree model (6.2.3) we can see that the error is significantly higher due to the many branches and deviations that are needed with a model like that. Relative errors of over 50% come from the model which would be worse than the error for just flipping a coin to determine the results. For the random forest we get some better results in terms of error rates from the model as would be expected from a model using multiple trees which work on a principle such as the wisdom of crowds. For this model we get rates of around 7% which is a lot closer to the gradient boosting model. This is because a model like the random forest generates a lot of trees to do the same calculation and works on the averages from the results that it finds which means that a lot of trees from a weak learner together reach a closer result to what is desired.

## Evaluation

Using the gradient boosting model to classify the results has resulted in creation of a model with a low error rate especially when compared to other classification models which use a similar type of algorithm to classify data such as the random forest or decision tree models.

The main issue with the results gained from this model however is that they are hard to practically analyse as a lot of the inner workings of the model are hidden behind the working algorithm in a way that the simpler models such as a decision tree are not and so can be hard to access and read once the model is created.

From the results it is believed that it is clear that the best type of model that could be used for a system such as this would be an iterative model such as the gradient boost or even the random forest as the error rate for that model remained at a similar level to the other.

For testing the model real negative data may have been a better in order to achieve some better results however for a prediction model like this it would give a much higher rate of no results as there are near infinite amounts more results for none accidents than there are for accidents and a model needs a good mix of data.

## Reflection

While it is believed that the model that has been created is good for the examining of the data that was used for its creation there are a number of ways in which it could have been improved given more work and time for the creation. While the data that has been used is good to find some of the root causes for accidents they do not provide enough data to the algorithm that would identify the roads that could be causes of more accidents in a way that would be good for the purpose of this model and instead are just more causes to do with the conditions on any road at a given time such as weather. With the complex nature of the problem more factors could have been looked into such as the geography of the roads and distractions around them that could lead to accidents which would require a lot more data from different sources.

For the testing and running of the model the techniques used in splitting the data into training and testing sets in order to check the results of the model and assess error rates are a standard in the way models are usually tested. Another way that could have been used would be to test the data against those of another year’s data or other data source however this could have been unduly influenced by another source that could affect the rates of the accidents that happen based on the year which isn’t accounted for in the data.

Testing and evaluating against other models that use a similar process is another way in which I believe that the testing is a good idea as well as it give a chance to compare how the technique used operates against another to give some clarification to the meaning of the results that have been gained from the models analysis.

# Conclusion

## Summary

During the course of this research a model as created in which predictions can be made as to whether an accident will take place on a road given a number of possible factors that could lead to the accident. In order to create this research was done into the possible uses for the model as well as the common causes of accidents on the roads and the techniques that are used in machine learning to classify results which can be used to make predictions.

Using the research conducted the best technique to used was found and a model created which can be used to make the predictions when given the needed data based on a dataset of the accidents that occurred in the UK over a given time period.

This model was then tested against others which were created using the same dataset in order to evaluate its usefulness.

## Conclusion

Within this project the model created did show some promising results for the prediction of accidents. The original aim of the report was to find a model which could predict accidents for use within a route finding machine in future which would allow people to find safer routes when traveling and avoiding accidents. It is believed that the model created achieves this goal to a degree however it is only a step towards fully creating a system for aiding in avoiding accidents. The researcher believes that the current model could be applied in the way that was intended but also could be improved and added to in a number of ways as part of an on-going process due to the complexity of the issue that is being investigated.

The research carried out and the analysis of the data that was found supports this as with the literature the investigation showed some of the causes of accidents and how much of an issue they are at the current time the data that was used looks into a number of these causes and allows for predictions to be made with them but due to the data that was used only certain causes were investigated which leaves room for improvement in the model using other data sets. The data exploration carried out on the dataset also shows that the models findings have merit as they show that the factors that had the most effect on the model are indeed the ones that were present in the highest proportion of accidents which point towards the models results being useful and accurate.

### Issues with the model

One issue that will be found from the creation of the model is that the use of one dataset relating to the accidents may not be enough to calculate all of the causes for an accident in order to give an actual prediction for the purposes of the model and future iterations could do with more data being explored and added to give better results. This is due to the amount of research that has been conducted into the many causes with people finding many different reasons that would have to be accounted for in a final version of this kind of model. Some issues may be impossible to look into in any practical way such as the research stating that accidents are commonly caused by use of items behind the wheel such as mobile phones etc. but others such as roadside distractions could be accounted for given the proper dataset such as distractions such as billboards by adding categories establishing the distance of accidents from such things.

Another issue with the model is the time that it would take to build, given the number of factors involved any changes made to it take multiple hours to build, while this could be reduced by making the dataset smaller an including less variables or by using a simpler model doing either of these things would reduce the usefulness of the model.

### Issues addressed

With developing any piece of research based on datasets there are a number of issues that have to be addressed. The main issues with a model such as this could have are related to the data that would have to be kept for its operation.

Data for both ethical and legal reasons is pretty well regulated to ensure that the personal information on people is kept out of the wrong hands. To address this in the creation of this research only information that was freely available on the government under open licencing was used in the generation of the models and data analysis that was used. All of this data is kept anonymous before it was even accessed meaning that there was never any trouble with whether there would be any troubles caused by its handling.

When operating with machine learning models there can be issues from a social standpoint with models targeting specific demographics which are built on pre-existing prejudices with the data, these also had no need to be addressed because of the data containing no identifying information on the people that were involved in the accidents.

From an ethical standpoint it is believed that this research is on strong grounds, while it pertains to something which ethically is sensitive such as accidents and casualties which could possibly be caused by unlawful activities on the parts of the drivers in some cases especially with the data being generated from police reports the aim of the processing of the data is to reduce accidents that may happen and to help reduce the possibility of casualties in the future meaning the use of this data is for the common good. As well as this with the data being what is available publically there is no identifying features for anyone involved and no possibility of using the data for any means which could compromise anything to do with the persons involved or any investigation.

## Suggestions for future research

The purpose of this research was to create a model for use in a route finding machine. This would suggest that one of the main points that would be available for future research is the adaptation or creation of such a route finder which would aid in people staying safe during their journeys.

The model created here could also be advanced on in a number of ways as part of a future project as mentioned in the conclusion there are a number of factors that could affect accidents which were not accounted for and extending the model to use them would be valuable to ensuring the most accurate results possible. For example an investigation into the geographical features of the roads which could have aided in the cause of accidents could be investigated using geospatial research in combination with data from this dataset such as the longitude and latitude.

Research into other types of model that could be used for this kind of operation could also be useful as with the data used here the idea of classification models makes sense due to the categorical nature of the data that was used however with other data sources models could be built which use techniques such as linier regression which could provide valuable other ways in which could be used for predictions to be made.

# Bibliography

Department for Transport, 2018. *Road Safety Data.* [Online]   
Available at: https://data.gov.uk/dataset/cb7ae6f0-4be6-4935-9277-47e5ce24a11f/road-safety-data  
[Accessed 06 August 2019].

ISO/IEC , 2017. *ISO International Standard ISO/IEC 14882:2017(E) – Programming Language C++..* 5 ed. s.l.:ISO.

Bradshaw, G., 1939. *Bradshaw's Railway Time Tables, and Assistant to Railway Travelling with Illustrative Maps and Plans.* First ed. London: Sheperd and Sutton, and Wyld.

Brid, R. S., 2018 . *decision-trees-a-simple-way-to-visualize-a-decision.* [Online]   
Available at: https://medium.com/greyatom/decision-trees-a-simple-way-to-visualize-a-decision-dc506a403aeb  
[Accessed 11 September 2019].

Department of transport , 2019. *GB Driving Licence Data.* [Online]   
Available at: https://data.gov.uk/dataset/d0be1ed2-9907-4ec4-b552-c048f6aec16a/gb-driving-licence-data  
[Accessed 18 July 2019].

Department Of Transport, 2018. *Reported road casualties in Great Britain quarterly provisional estimates year ending June 2018,* London: HM Government Department of transport.

Department of transport, 2019. *Contributory factors for reported road accidents (RAS50).* [Online]   
Available at: https://www.gov.uk/government/statistical-data-sets/ras50-contributory-factors  
[Accessed 18 July 2019].

Dijkstra, E. W., 1959. A Note on Two Problems in Connexion with Graphs. *Numerische Mathematlk*, pp. 269 - 271.

Dorigo, M. & Gambardella, M. L., 1997. Ant Colony System: A Cooperative Learning. *IEEE TRANSACTIONS ON EVOLUTIONARY COMPUTATION,* 1(1), pp. 53-65.

Eibe , F., Hall, M. A. & Witten, I. H., 2016. *The WEKA Workbench. Online Appendix for "Data Mining: Practical Machine Learning Tools and Techniques".* Fourth Edition ed. s.l.:Morgan Kaufmann.

Euler, L., 1741. *Solutio problematis ad geometriam situs.* [Online]   
Available at: https://scholarlycommons.pacific.edu/euler-works/53

Flood, M. M., 1956. The Traveling-Salesman Problem. *Operating research,* 4(1), pp. 1-137.

Google, 2019. *Documentation - API picker.* [Online]   
Available at: https://developers.google.com/maps/documentation/api-picker  
[Accessed 08 August 2019].

Google, 2019. *Google maps.* [Online]   
Available at: https://www.google.com/maps/  
[Accessed 18 July 2019].

Grefenstette, j., Gopal, R., Rosmaita, B. & Van Gucht, D., n.d. *Genetic Algorithms for the traveling salesman problem.* [Online]   
Available at: https://www.researchgate.net/profile/Dirk\_Van\_Gucht/publication/201976371\_Genetic\_Algorithms\_for\_the\_Traveling\_Salesman\_Problem/links/0deec53aeb247f1ec0000000.pdf  
[Accessed 08 August 2019].

HALTON, C., 2019. *Wisdom of Crowds.* [Online]   
Available at: https://www.investopedia.com/terms/w/wisdom-crowds.asp  
[Accessed 11 September 2019].

Hart, E. P., Nilsson, J. N. & Raphael, B., 1968 . A Formal Basis for the Heuristic Determination of Minimum Cost Paths. *IEEE Transactions on Systems Science and Cybernetics,* 4(2), pp. 100 - 107.

Lazovskiy, V., 2018. *Travel Time Optimization With Machine Learning And Genetic Algorithm.* [Online]   
Available at: https://towardsdatascience.com/travel-time-optimization-with-machine-learning-and-genetic-algorithm-71b40a3a4c2  
[Accessed 08 August 2019].

Liu, S., 2017. *Portable navigation devices (PND) sold in Europe\* from 2007 to 2016 (in millions).* [Online]   
Available at: https://www.statista.com/statistics/217933/market-size-of-pnds-in-europe-since-2007/  
[Accessed 19 July 2019].

Liu, S., 2018. *Navigation devices & usage - Statistics & Facts.* [Online]   
Available at: https://www.statista.com/topics/2221/navigation-devices-and-usage/  
[Accessed 19 July 2019].

MathWorks , 2010. *MATLAB. version 7.10.0..* Natick, Massachusetts:: The MathWorks Inc.

McDeed, P., 2016. *The Big Data Driving Google Maps.* [Online]   
Available at: http://ltd.edc.org/big-data-driving-google-maps  
[Accessed 28 March 2019].

Mishchenko, N., 2013. *The importance of route planning.* [Online]   
Available at: http://nelmitravel.com/importance-of-route-planning/  
[Accessed 28 March 2018].

Nandi, A., 2015. *Spark for python developers.* s.l.:Packt Publishing.

Norah, J., 2018 . *Travel Research: How Much Travel Stress Do People Experience on Vacation?.* [Online]   
Available at: https://independenttravelcats.com/travel-research-how-much-travel-stress-do-people-experience-on-vacation/  
[Accessed 28 March 2019].

open street map , 2019. *open street map.* [Online]   
Available at: https://www.openstreetmap.org/#map=6/54.910/-3.432  
[Accessed 08 August 2019].

OSRM, 2019. *Open Source Routing Machine.* [Online]   
Available at: http://project-osrm.org/  
[Accessed 08 August 2019].

Police UK, 2019. *Data downloads.* [Online]   
Available at: https://data.police.uk/data/  
[Accessed 18 July 2019].

Polson, N. G. & Sokolov, V., 2017. Deep learning for short-term traffic flow prediction. *Transportation Research Part C Emerging Technologies ,* Volume 79, pp. 1-17.

R Core Team, 2013. *R: A language and environment for statistical.* Vienna, Austria: http://www.R-project.org/.

Raghavendra, R., 2017. *Improving Traffic Prediction Using Weather Data.* San Francisco, Spark Summit.

Rong, Y. et al., 2015. Comparative analysis for traffic flow forecasting models with real-life data in Beijing. *Advances in Mobility Theories, Methodologies, and Applications,* 7(12).

ROSPA, 2017. *Safter Journey Planner.* [Online]   
Available at: https://www.rospa.com/rospaweb/docs/advice-services/road-safety/drivers/safer-journey-planner.pdf  
[Accessed 28 March 2019].

safavian, S. R. & Landgrebe, D., 1990. *A survey of Decision tree classifier Methodology,* West Lafayetten Indiana : School of electrical engineering Purdue university.

Sayad, S., 2019. *Decision Tree - Overfitting.* [Online]   
Available at: https://www.saedsayad.com/decision\_tree\_overfitting.htm  
[Accessed 11 September 2019].

Schott, P., 2017. *E-Commerce Drives Global Package Shipping Volume to Rise 48% in Two Years.* [Online]   
Available at: https://www.ttnews.com/articles/e-commerce-drives-global-package-shipping-volume-rise-48-two-years  
[Accessed 28 March 2019].

Schrijver, A., 2010. On the History of the Shortest Path Problem. *Documenta Math. ,* pp. 155-167.

Shamir, S., 2018. *Why Route Planning is Important For Your Business.* [Online]   
Available at: https://www.scmr.com/article/why\_route\_planning\_is\_important\_for\_your\_business  
[Accessed 28 March 2019].

Singh, H., 2018. *Understanding Gradient Boosting Machines.* [Online]   
Available at: https://towardsdatascience.com/understanding-gradient-boosting-machines-9be756fe76ab  
[Accessed 11 September 2019].

Smith, B. L. & Demetsky, M. J., 1997. Traffic Flow Forecasting: Comparison of Modeling Approaches. *Journal of Transportation Engineering,* 123(4).

Trueblood, ]. D., 1952. The effect of travel time and distance on freeway usage. *Public Roads,* p. 241–250..

Tulp, E., 1991. *Searching Time-table Networks,* Amsterdam: VRIJE UNIVERSITEIT.

Vasudev, R., 2017. *What is One Hot Encoding? Why And When do you have to use it?.* [Online]   
Available at: https://hackernoon.com/what-is-one-hot-encoding-why-and-when-do-you-have-to-use-it-e3c6186d008f  
[Accessed 01 September 2019].

Wiener, C., 1873. Ueber eine Aufgabe aus der Geometria situs.. *Mathematische Annalen,* Volume 6, p. 29–30.

World health organisation , 2018. *Road traffic injuries.* [Online]   
Available at: https://www.who.int/news-room/fact-sheets/detail/road-traffic-injuries  
[Accessed 19 July 2019].

Yiu, T., 2019. *Understanding Random Forest.* [Online]   
Available at: https://towardsdatascience.com/understanding-random-forest-58381e0602d2  
[Accessed 11 September 2019].

Zhang, Y., 2010. *New Advances in Machine Learning.* s.l.:InTech.

1. Research Proposal

**Aim**

To produce a model that will expand on the current machine learning models used to plan journeys to find the safest route for the driver as well as the fastest using collision data sets.

**Background, MOTIVATION and relevance**

**Background**

The idea of route planning in a practical way can be traced back as far as the 19th century and the compilation of timetables for the railways to enable people to properly plan their journeys across multiple railways most notably in the publishing of Bradshaw’s guides beginning in 1939 which created a standardised way to collect and print the data from multiple sources. (Bradshaw, 1939) This shows that since people gained the ability to travel large distances in a reasonable time there has been a need to plan the best ways to make those journeys quickly and efficiently.

In mathematics the theoretical problems associated with this can be traced back even further to the Seven Bridges of Königsberg problem devised by Leonhard Euler in 1736 which laid the foundation for graph theory (Euler, 1741) which lead to theoretical algorithms for computing such as Dijkstra's algorithm which is used to find the shortest distances between nodes on a graph. (Dijkstra, 1959)

With advances in computer technology the ability to create systems which can be used to plan more complicated journeys based on more and more complicated factors. These systems have been widely used in businesses such as the rail industry from as early as the 1970s to allow them to find trips available between two locations which could operate over multiple rail lines when selling tickets to customers.

In the early 1990s with the increase in processing power with computers systems were created which could be released for public use beginning with the system created for Dutch railways to be released on diskette. The systems at this time were more robust being able to work with the computations that would allow them to calculate things such as the average walking distances that would be needed etc. as opposed to the systems based entirely on the companies timetables previously. (Tulp, 1991)

Following this advancements have been made in the technology constantly over time with better algorithms and more features being included along with the advancements in technology that allow them. With the rise of the internet companies have begun to publish free versions of their software on the internet which later advanced into mobile apps following the release of the smart phone.

The use of smart phones for this technology has allowed for the use and collection of large amounts of data that can be used to improve the algorithms used as well as create new ways of determining the routes that should be taken entirely. Modern technology has allowed companies access to the GPS data for their users to see what routes they use and accurate times that they travel along them as well as data that has been gathered from many publically available sources such as traffic figures from traffic cameras and collisions data which can be used along with various machine learning models to give accurate predictions on what the traffic will be like in the area as well as in some cases create real time updates. (McDeed, 2016)

**Motivation**

Route planning is important for many reasons. Traveling can be stressful and research has determined that the most stressful part of any trip is the planning stage. (Norah, 2018 ) The use of a route planner for the travel can help to alleviate this stress by allowing software to work out the best way to get from point A to point B. Use of this can also help to reduce the stress of the travel itself as with the use of software which includes things such as estimated road conditions it can help to predict problems that couldn’t be seen through planning with methods such as timetables. It can also be helpful in helping you to add stops to a journey and plan using smaller locations such as petrol stations that wouldn’t be included on a time table system. (Mishchenko, 2013)

Route planning is also important for businesses, with the rise of e-commerce meaning that more packages are being delivered growing year on year. (Schott, 2017) The postal industry relies on systems to make the routes made by their drivers more efficient in terms of both time spent on the road and fuel costs. (Shamir, 2018) This means that software that is better at predicting the road conditions is a valuable asset to the companies.

With 1770 deaths on the road in the last year and a similar level every year since 2012, (Department Of Transport, 2018) there is always a concern with safety when it comes to driving and one of the main tips that is provided for being safe on the roads is to plan the journey you are going to take. The reasons for this are that when making the plans you can avoid at risk times and plan in extra breaks and stops into your route beforehand breaking up long journeys. (ROSPA, 2017)

With this having a system included in a journey planner such as the one included in googles maps programs which would allow you to plan the journey around the safest route as well as the quickest could be beneficial to businesses and people, with businesses with taking safety into account they could reduce their risks and offer better working conditions to their drivers and for people it could reduce the chances of an accident and the stress related to travel.

**Relevant research**

Traffic prediction has been researched deeply over recent years due to the relevance in many sectors which require efficiency in travel times. Comparative research has been made in the area, finding the better method between statistical and machine learning models which could be used due to the limited effect of previous models (Smith & Demetsky, 1997). Papers such as this have continued to be published over the years with people looking into better models that could be used. (Rong, et al., 2015) Other research conducted has looked into single models and attempts to create a more accurate solution using a combination of techniques. (Polson & Sokolov, 2017) Others have looked into the inclusion of other aspects into the algorithms such as whether to improve the predictions. (Raghavendra, 2017) these along with others show that there is a deep amount of research being done within this area, algorithms are improving all the time and being widened to include many different factors making them more complex.

**Scope, objectives and risk**

**Scope**

For this project data will be taken from various sources, the data used will be narrowed down to a specific area so that it all corresponds to a specific set of roads which will be used to create the models.

From here the data will be cleaned and the relevant parts taken then a variety of machine learning methods will be applied to give predictions of the routes that you could take which would be safe and quick in balance.

As this project is for the creation of new models for determining the best route taken a full route planner will not be made but rather new models that could be applied to existing route planners to give another option for people planning journeys.

**Objectives**

1. To collect Data from various sources
   1. Collect data on traffic flow on roads
   2. Collect data on collisions and other dangers on roads
   3. Collect data on whether conditions
2. To Undertake research into current forms of route planning
   1. Research into methods of modelling the traffic conditions
   2. Research into the methods of determining the safest route
   3. Research into the effect on whether conditions on safest route
3. Clean and organise data
   1. clean data and ensure all gaps are removed from all data sets
   2. combine data when needed as some could come from multiple sources
4. Develop a machine learning modal for determining the route
   1. develop a model using methods found through research for finding fastest route
   2. develop model to find a value for the safety of the routes
   3. use decision methods to choose the route best taken
5. Evaluate the results to determine whether the methods created form a worthwhile improvement on current methods.

**Risk Log**

Key

**Risk types**

F (Financial)

T (Technology)

P (People)

E (Environmental)

S (Security)

**Risk Values**

≥ 75 ***Risk very high*** - urgent action required

≥ 50 < 75 ***Risk high*** - action as soon as possible

≥ 25 < 50 ***Risk may be acceptable*** - more analysis required

< 25 ***Low risk*** - no gains expected from extra work

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| Risk Type | Risk Event | Likelihood  (1-10) | Impact  (1-10) | Risk Value  (1-100) | Risk Management Strategy | Risk owner | Commentary |
| F | Requirement for cloud computing increases cost | 2 | 8 | <25 | Keep the data values low and if in need of servers for cloud computing manage costs using free available spaces | Developer | Risk is low due to the data being used being gathered from multiple places in smaller packets means that it can be organised separately and then compiled together keeping sizes as low as possible |
| T | Computer malfunction | 3 | 5 | <25 | Keep multiple computers available that could be used for project and keep files stored online with backups to seamlessly move from one machine to another | Developer | Personal computer will be used for the majority of the project however developers have access to laptops for emergency use as well as computers on campus |
| T | Software malfunction results in loss of data | 4 | 8 | <25 | Backups will be kept at every stage both on physical media and online with GitHub used for all programming projects | Developer |  |
| S | Loss of Data being used | 1 | 1 | <25 | All data used will be public domain | Developer |  |
| P | Developer missing deadlines | 4 | 8 | ≥ 50 < 75 | Use of planning methods such as Gantt charts and developer diaries to keep on track for time | Developer |  |
| P | Developer bad at organising meetings | 4 | 8 | ≥ 50 < 75 | Developer will set up a regular meeting time to ensure that they make them | Developer | The project developer is bad at organising meetings themselves so has determined that a more regular time will help them to keep on track |
| T | Program features will keep expanding with unneeded things | 2 | 4 | <25 | A design will be made and stuck to and when revisiting the design is needed changes will only be made when it benefits the core design. | Developer | Issues with “feature creep” are always a problem in development, with a clear plan and design this can be avoided. |

**Ethics, Legal, Social, Security and Professional Issues**

**Ethics Issues**

This project will use only data which has been made public by government agencies and companies, as this data is available to all on the internet there is no ethical issue with using it in the design of new programs. As the program being created is to aid in the safety of drivers I believe the project to be ethical in its creation and purpose.

**Legal Issues**

As above all data used will be public domain and so legal to use, the software to be used will be either available on license to the university or for all to use and so available with no legal issues with the use of software’s. Al work will be that of the developer aside from the initial collection of the data involved.

When creating paths for the route the laws of the roads will have to be taken into consideration with speed limits so that the program will ensure that laws are followed by the end user and that the data provided is legal.

**Social Issues**

I foresee no social issues coming from this project, the project is for research into upgrades that could be made to the options in route planners and does not involve the public in anyway.

**Security Issues**

Working with a lot of data could at some points lead to security issues with the storage and safe keeping of that data because of that the data that will be used for this will be taken from publically available sources and so is already in the public domain. This removes any risk in data that could be considered private from being released.

Some security issues persist with the storage of work on cloud storage platforms which will be used for backups however these risks are small and the benefits of keeping backups in this way and easily accessible far outweigh the risks however to ensure that the program is accessible even at times when these services could be down a physical copy will be kept

**Professional Issues**

This project is for research purposes and not for any professional body so I foresee no professional issues from that standpoint.

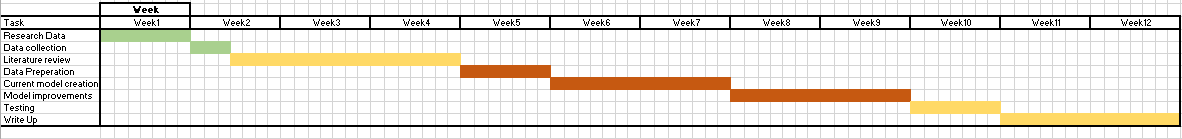
**Relevant Journals**

Journals that would publish work in this area fit into two different camps, journals which would publish research into areas surrounding travel and transport and those surrounding the machine learning principles which are used to research it. An example of a journal covering the transport side of the discussion is the *Trends in Transport Engineering and Applications*, this covers a lot of areas including the one of interest here which is intelligent transport systems and some of the research from here includes comparisons of various techniques that could be used in route planning algorithms. If the research was to come up with a new algorithm or something which could more advance the field of machine learning it could fit into the *Journal of Machine Learning Research* this would be looking more into items which could give new insight into the algorithms that are used rather than their usage. SAGE publishing’s *journal Advances in mechanical engineering* also contains some comparative studies into the comparisons in the methods used, this journal covers a large area of different fields. This could be a detriment to publishing a paper such as this which may be better in a more targeted journal.

For this paper I would be targeting a Journal such as *Trends in Transport Engineering and Applications* as it is more targeted towards the area of interest in this paper the referencing style used in this publication is to have numbered references in line with the mention and a bibliography with the details, the journal does not give a named style of numbered referencing.

**schedule of Activities**

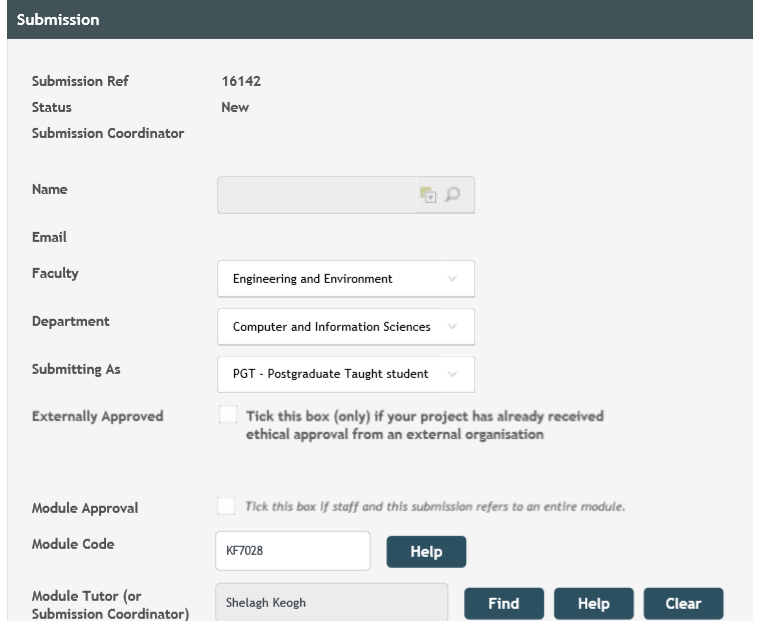
**Gantt chart**

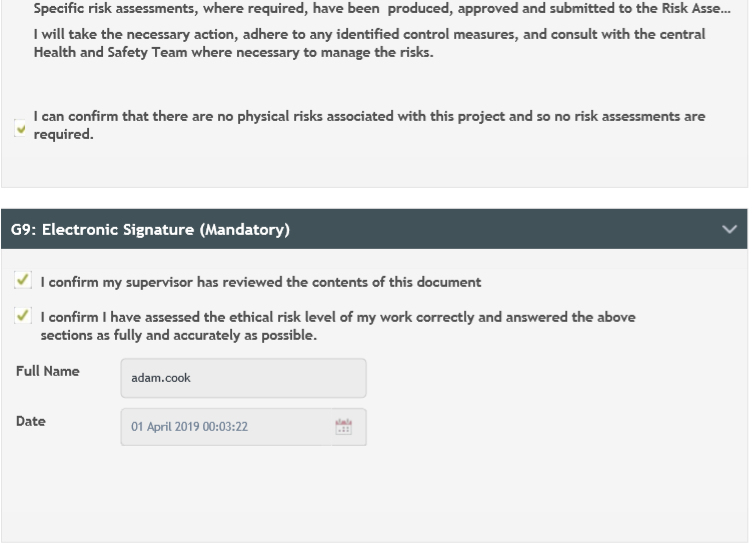
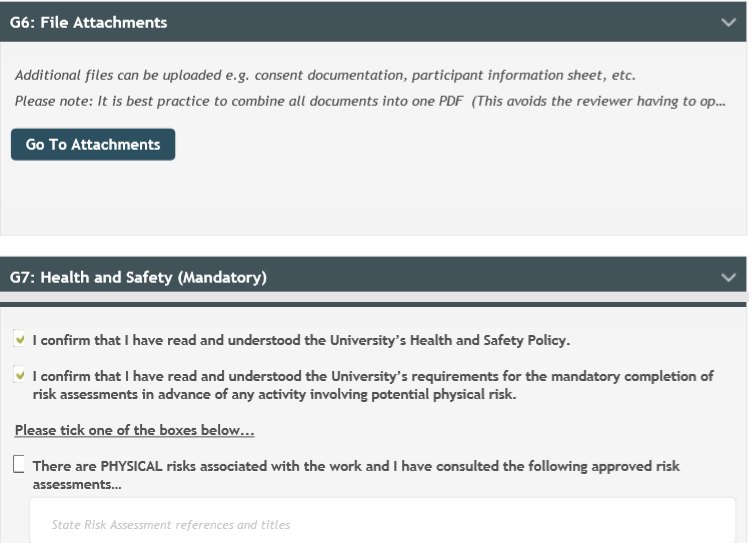
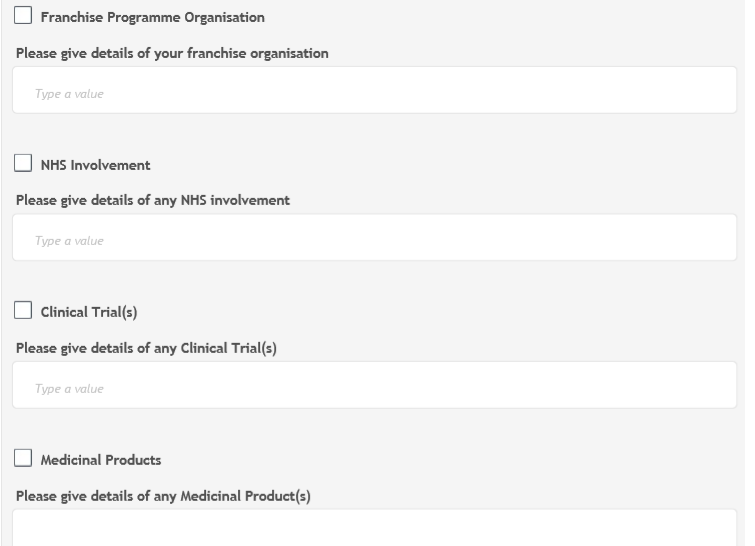
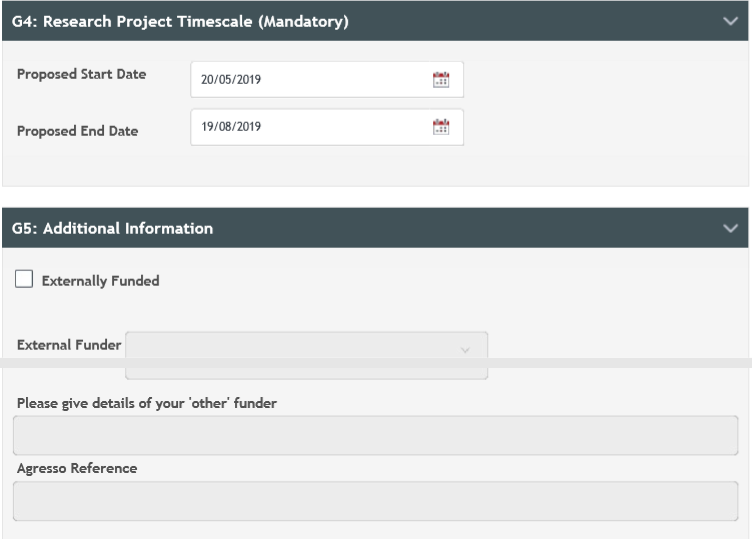
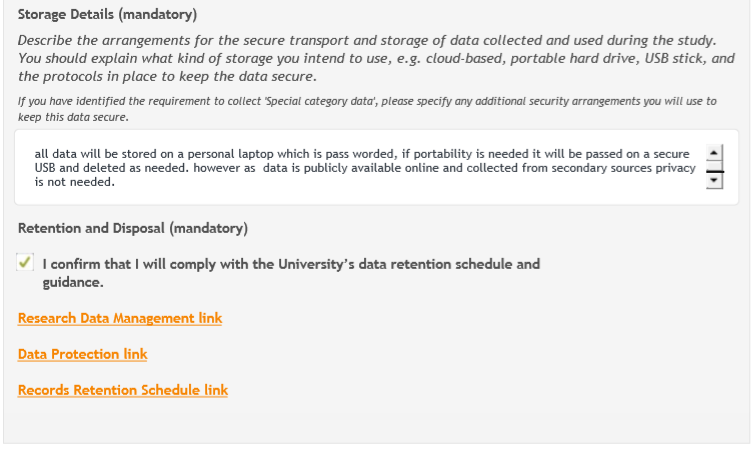
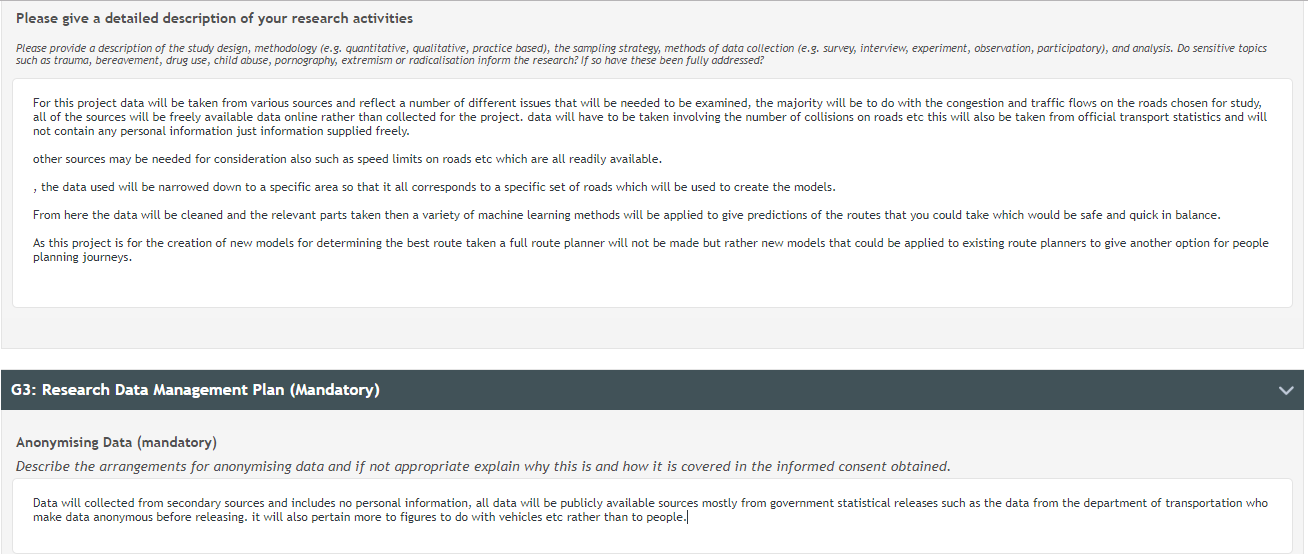
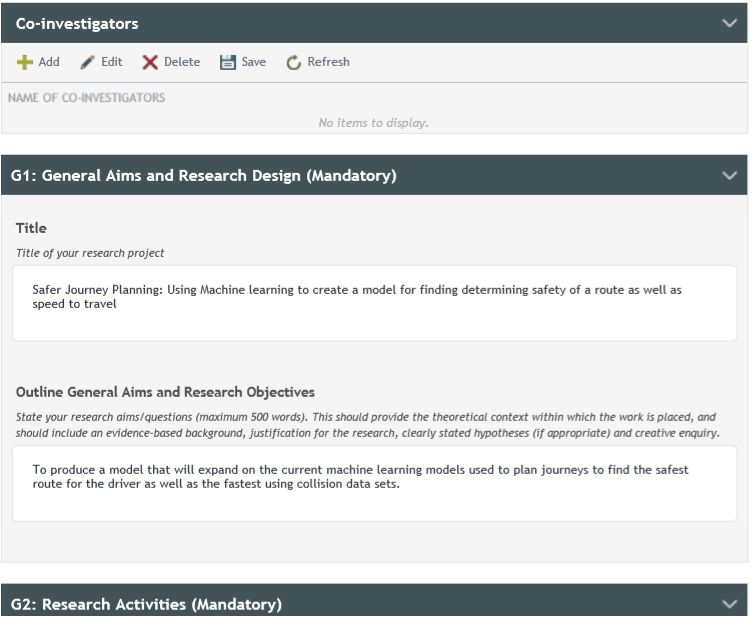
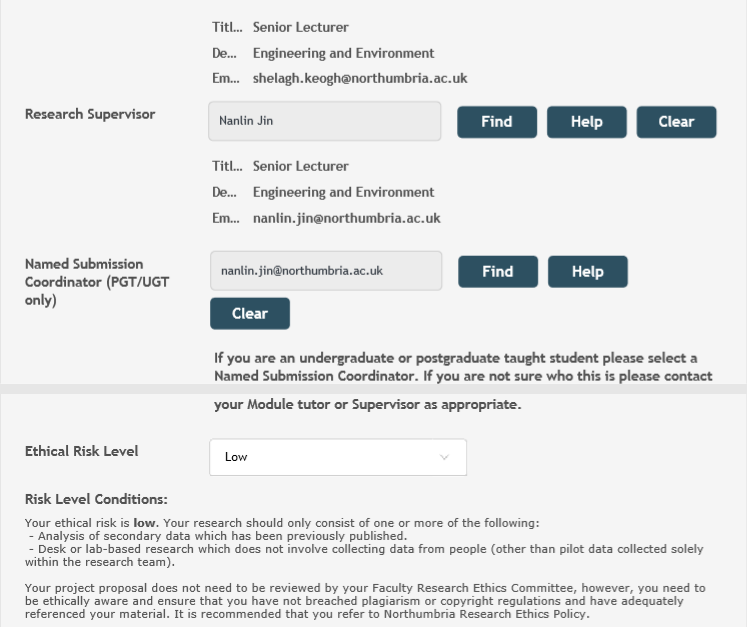


Monitoring and control

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| **Task/Objective/Milestones Deliverable** | **Planned**  **start** | **Planned**  **end** | **Actual start** | **Actual end** | **Deliverable** | **Reflections** |
| ***Research the data sources needed*** | ***Week 1*** | ***Week 1*** |  |  | ***Information on the data that will be needed to create a model*** |  |
| **Collect data for modelling** | **Week 2** | **Week 2** |  |  | **A collection of data for creation of the project** |  |
| **Research into models currently used for route planning** | **Week 2** | **Week 4** |  |  | **A literature review** |  |
| **Prepare data** | **Week 4** | **Week 5** |  |  | **Collection of tables that can be used for modelling** |  |
| **Create model using current ideas from research** | **Week 5** | **Week 7** |  |  | **Working model using the researched methods** |  |
| **Attempt to improve on the model using methods to find safest route** | **Week 7** | **Week 9** |  |  | **Extended model** |  |
| **Test model using data collected** | **Week 9** | **Week 10** |  |  | **Test data** |  |
| **Write up results and conclusion** | **Week 10** | **Week 12** |  |  | **Finished report** |  |

**Appendix**

**Ethics Form**



1. Data Licensing

<https://data.gov.uk/dataset/cb7ae6f0-4be6-4935-9277-47e5ce24a11f/road-safety-data>

Figure 47 : link to open government license

1. Category Lookup tables

Table 11: codes for weather conditions

|  |  |
| --- | --- |
| **code** | **label** |
| 1 | Fine no high winds |
| 2 | Raining no high winds |
| 3 | Snowing no high winds |
| 4 | Fine + high winds |
| 5 | Raining + high winds |
| 6 | Snowing + high winds |
| 7 | Fog or mist |
| 8 | Other |
| 9 | Unknown |
| -1 | Data missing or out of range |

Table 12: categories for light conditions

|  |  |
| --- | --- |
| **code** | **label** |
| 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 |

Table 13: road type categories

|  |  |
| --- | --- |
| **code** | **label** |
| 1 | Roundabout |
| 2 | One way street |
| 3 | Dual carriageway |
| 6 | Single carriageway |
| 7 | Slip road |
| 9 | Unknown |
| 12 | One way street/Slip road |

Table 14: severity codes

|  |  |
| --- | --- |
| **Code** | **Label** |
| 1 | Fatal |
| 2 | Serious |
| 3 | Slight |

Table 15: junction type categories

|  |  |
| --- | --- |
| code | label |
| 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 |

1. Table of figures

[Figure 1 - code for representing points on a map 23](file:///C:\Users\Adam\Documents\Masters%20Project\VincentAdamCookW15009928.docx#_Toc19647712)

[Figure 2: map of accidents 2017 24](#_Toc19647713)

[Figure 3: code for mapping geobubble map based on severity of accidents 25](file:///C:\Users\Adam\Documents\Masters%20Project\VincentAdamCookW15009928.docx#_Toc19647714)

[Figure 4: map of accidents based on severity 2017 26](#_Toc19647715)

[Figure 5: code to remove less severe accidents 27](file:///C:\Users\Adam\Documents\Masters%20Project\VincentAdamCookW15009928.docx#_Toc19647716)

[Figure 6: map showing severe and fatal accidents 2017 28](#_Toc19647717)

[Figure 7: map showing only fatal accidents in 2017 29](#_Toc19647718)

[Figure 8: map representing the number of accidents on roads 2017 31](#_Toc19647719)

[Figure 9: number of accidents per road 2017 using coloured Geobubbles 33](#_Toc19647720)

[Figure 10: number of accidents south of England 2017 34](#_Toc19647721)

[Figure 11: number of accidents by month 2017 35](#_Toc19647722)

[Figure 12: number of accidents of each severity by month 2017 36](#_Toc19647723)

[Figure 13: number of vehicles involved in accidents over time 37](#_Toc19647724)

[Figure 14: weather conditions over time 2017 38](#_Toc19647725)

[Figure 15: number of accidents for each light condition 39](#_Toc19647726)

[Figure 16: number of accidents by hour 40](#_Toc19647727)

[Figure 17: number of accidents by road type 41](#_Toc19647728)

[Figure 18: number of accidents by junction type 42](#_Toc19647729)

[Figure 19: function to combine date and time in excel 45](file:///C:\Users\Adam\Documents\Masters%20Project\VincentAdamCookW15009928.docx#_Toc19647730)

[Figure 20: code for loading packages in R 49](file:///C:\Users\Adam\Documents\Masters%20Project\VincentAdamCookW15009928.docx#_Toc19647731)

[Figure 21: code to split data into training and testing sets 50](file:///C:\Users\Adam\Documents\Masters%20Project\VincentAdamCookW15009928.docx#_Toc19647732)

[Figure 22: code for generation of model 50](file:///C:\Users\Adam\Documents\Masters%20Project\VincentAdamCookW15009928.docx#_Toc19647733)

[Figure 23: plots a graph of the error from the models using the cross fold validation 51](file:///C:\Users\Adam\Documents\Masters%20Project\VincentAdamCookW15009928.docx#_Toc19647734)

[Figure 24: code for getting a summary showing the relative dependencies of the variables 52](file:///C:\Users\Adam\Documents\Masters%20Project\VincentAdamCookW15009928.docx#_Toc19647735)

[Figure 25: code to plot a number of partial dependencies 52](file:///C:\Users\Adam\Documents\Masters%20Project\VincentAdamCookW15009928.docx#_Toc19647736)

[Figure 26: code for predictions using test data 53](file:///C:\Users\Adam\Documents\Masters%20Project\VincentAdamCookW15009928.docx#_Toc19647737)

[Figure 27: code for plotting of actual results against the predicted results 54](file:///C:\Users\Adam\Documents\Masters%20Project\VincentAdamCookW15009928.docx#_Toc19647738)

[Figure 28: code for taking out the data on the formation of individual trees 54](file:///C:\Users\Adam\Documents\Masters%20Project\VincentAdamCookW15009928.docx#_Toc19647739)

[Figure 29: open file in weka interface 56](#_Toc19647740)

[Figure 30: weka interface data loaded 57](#_Toc19647741)

[Figure 31: choosing a tree in weka 58](#_Toc19647742)

[Figure 32: weka interface after model creation 59](#_Toc19647743)

[Figure 33: relative influence of model with training set 60](#_Toc19647744)

[Figure 34: graph showing the top 10 from the relative influences 61](#_Toc19647745)

[Figure 35: relative influence raw results 62](#_Toc19647746)

[Figure 36: graph showing the error rate at each learner with the optimal number of learners 63](#_Toc19647747)

[Figure 37: error rate raw results 64](#_Toc19647748)

[Figure 38: partial dependence plot for November date 65](#_Toc19647749)

[Figure 39: Partial dependence plot for December date 65](#_Toc19647750)

[Figure 40: partial dependence for daylight conditions 66](#_Toc19647751)

[Figure 41: partial dependence plot for crossroads 66](#_Toc19647752)

[Figure 42: graph showing the performance of boosting using the test data with a line for the minimum test error 67](#_Toc19647753)

[Figure 43: graph showing the predicted results against the actual results 67](#_Toc19647754)

[Figure 44: REP decision tree diagram 79](#_Toc19647755)

[Figure 45: REP decision tree visualised error graph 80](#_Toc19647756)

[Figure 46: random forest error visualisation graph 81](#_Toc19647757)

[Figure 47 : link to open government license 113](#_Toc19647758)

[Table 1: number of accidents by grouped cause and country 18](#_Toc19647759)

[Table 2: breakdown of number of accidents based on road conditions 19](#_Toc19647760)

[Table 3: table showing the meanings of the severity codes 25](#_Toc19647761)

[Table 4: number of accidents by road type 2017 30](#_Toc19647762)

[Table 5: simplified data dictionary of raw data 44](#_Toc19647763)

[Table 6: simplified data dictionary of data kept for model 45](#_Toc19647764)

[Table 7: example of one-hot encoded variable 46](#_Toc19647765)

[Table 8: part of the one-hot encoded datetime data 47](#_Toc19647766)

[Table 9: data representation of first tree in model 68](#_Toc19647767)

[Table 10: data representation of the 500th tree in the model 71](#_Toc19647768)

[Table 11: data representation of the 1000th decision tree in the model 75](#_Toc19647769)

[Table 12: codes for weather conditions 114](#_Toc19647770)

[Table 13: categories for light conditions 114](#_Toc19647771)

[Table 14: road type categories 114](#_Toc19647772)

[Table 15: junction type categories 115](#_Toc19647773)

1. Decision tree examples from gradient booster

Table 8: data representation of first tree in model

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  | SplitVar | SplitcodePred | LeftNode | Rightnode | MissingNode | ErrorReduction | Weight | Prediction |
| **0** | 87 | 5.000000e-01 | 1 | 29 | 30 | 2276.7334 | 168821 | -4.924260e-06 |
| **1** | 88 | 5.000000e-01 | 2 | 27 | 28 | 2103.9014 | 164661 | -9.425696e-04 |
| **2** | 0 | 6.500000e+01 | 3 | 25 | 26 | 423.4967 | 160995 | -1.809036e-03 |
| **3** | 75 | 5.000000e-01 | 4 | 23 | 24 | 484.5816 | 156674 | -2.241706e-03 |
| **4** | 66 | 5.000000e-01 | 5 | 6 | 22 | 227.6339 | 91681 | 1.369027e-04 |
| **5** | -1 | -3.606260e-03 | -1 | -1 | -1 | 0.0000 | 28768 | -3.606260e-03 |
| **6** | 12 | 5.000000e-01 | 7 | 20 | 21 | 135.0429 | 62913 | 1.848525e-03 |
| **7** | 11 | 5.000000e-01 | 8 | 18 | 19 | 125.5472 | 60998 | 2.265526e-03 |
| **8** | 10 | 5.000000e-01 | 9 | 16 | 17 | 132.5045 | 59072 | 2.681654e-03 |
| **9** | 13 | 5.000000e-01 | 10 | 14 | 15 | 140.3788 | 57201 | 3.116769e-03 |
| **10** | 9 | 5.000000e-01 | 11 | 12 | 13 | 137.6722 | 55231 | 3.592033e-03 |
| **11** | -1 | 4.089340e-03 | -1 | -1 | -1 | 0.0000 | 53186 | 4.089340e-03 |
| **12** | -1 | -9.341826e-03 | -1 | -1 | -1 | 0.0000 | 2045 | -9.341826e-03 |
| **13** | -1 | 3.592033e-03 | -1 | -1 | -1 | 0.0000 | 55231 | 3.592033e-03 |
| **14** | -1 | -1.020776e-02 | -1 | -1 | -1 | 0.0000 | 1970 | -1.020776e-02 |
| **15** | -1 | 3.116769e-03 | -1 | -1 | -1 | 0.0000 | 57201 | 3.116769e-03 |
| **16** | -1 | -1.062087e-02 | -1 | -1 | -1 | 0.0000 | 1871 | -1.062087e-02 |
| **17** | -1 | 2.681654e-03 | -1 | -1 | -1 | 0.0000 | 59072 | 2.681654e-03 |
| **18** | -1 | -1.049748e-02 | -1 | -1 | -1 | 0.0000 | 1926 | -1.049748e-02 |
| **19** | -1 | 2.265526e-03 | -1 | -1 | -1 | 0.0000 | 60998 | 2.265526e-03 |
| **20** | -1 | -1.143409e-02 | -1 | -1 | -1 | 0.0000 | 1915 | -1.143409e-02 |
| **21** | -1 | 1.848525e-03 | -1 | -1 | -1 | 0.0000 | 62913 | 1.848525e-03 |
| **22** | -1 | 1.369027e-04 | -1 | -1 | -1 | 0.0000 | 91681 | 1.369027e-04 |
| **23** | -1 | -5.597041e-03 | -1 | -1 | -1 | 0.0000 | 64993 | -5.597041e-03 |
| **24** | -1 | -2.241706e-03 | -1 | -1 | -1 | 0.0000 | 156674 | -2.241706e-03 |
| **25** | -1 | 1.387904e-02 | -1 | -1 | -1 | 0.0000 | 4321 | 1.387904e-02 |
| **26** | -1 | -1.809036e-03 | -1 | -1 | -1 | 0.0000 | 160995 | -1.809036e-03 |
| **27** | -1 | 3.710893e-02 | -1 | -1 | -1 | 0.0000 | 3666 | 3.710893e-02 |
| **28** | -1 | -9.425696e-04 | -1 | -1 | -1 | 0.0000 | 164661 | -9.425696e-04 |
| **29** | -1 | 3.710893e-02 | -1 | -1 | -1 | 0.0000 | 4160 | 3.710893e-02 |
| **30** | -1 | -4.924260e-06 | -1 | -1 | -1 | 0.0000 | 168821 | -4.924260e-06 |

Table 9: data representation of the 500th tree in the model

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  | SplitVar | SplitcodePred | LeftNode | Rightnode | MissingNode | ErrorReduction | Weight | Prediction |
| **0** | 10 | 5.000000e-01 | 1 | 26 | 30 | 8.098481 | 168821 | -4.662897e-04 |
| **1** | 8 | 5.000000e-01 | 2 | 18 | 25 | 7.580047 | 163425 | -2.889976e-04 |
| **2** | 12 | 5.000000e-01 | 3 | 16 | 17 | 8.143335 | 157580 | -1.008566e-04 |
| **3** | 1 | 5.000000e-01 | 4 | 11 | 15 | 8.988196 | 152108 | 8.518385e-05 |
| **4** | 11 | 5.000000e-01 | 5 | 9 | 10 | 11.309334 | 131429 | 2.440941e-04 |
| **5** | 9 | 5.000000e-01 | 6 | 7 | 8 | 11.052571 | 126725 | 4.655694e-04 |
| **6** | -1 | 6.495039e-04 | -1 | -1 | -1 | 0.000000 | 122044 | 6.495039e-04 |
| **7** | -1 | -4.330009e-03 | -1 | -1 | -1 | 0.000000 | 4681 | -4.330009e-03 |
| **8** | -1 | 4.655694e-04 | -1 | -1 | -1 | 0.000000 | 126725 | 4.655694e-04 |
| **9** | -1 | -5.722415e-03 | -1 | -1 | -1 | 0.000000 | 4704 | -5.722415e-03 |
| **10** | -1 | 2.440941e-04 | -1 | -1 | -1 | 0.000000 | 131429 | 2.440941e-04 |
| **11** | 15 | 5.000000e-01 | 12 | 13 | 14 | 13.834721 | 20679 | -9.247977e-04 |
| **12** | -1 | -6.085714e-04 | -1 | -1 | -1 | 0.000000 | 19841 | -6.085714e-04 |
| **13** | -1 | -8.411965e-03 | -1 | -1 | -1 | 0.000000 | 838 | -8.411965e-03 |
| **14** | -1 | -9.247977e-04 | -1 | -1 | -1 | 0.000000 | 20679 | -9.247977e-04 |
| **15** | -1 | 8.518385e-05 | -1 | -1 | -1 | 0.000000 | 152108 | 8.518385e-05 |
| **16** | -1 | -5.272318e-03 | -1 | -1 | -1 | 0.000000 | 5472 | -5.272318e-03 |
| **17** | -1 | -1.008566e-04 | -1 | -1 | -1 | 0.000000 | 157580 | -1.008566e-04 |
| **18** | 42 | 5.000000e-01 | 19 | 23 | 24 | 19.762937 | 5845 | -5.361242e-03 |
| **19** | 66 | 5.000000e-01 | 20 | 21 | 22 | 8.875621 | 2380 | 1.977931e-03 |
| **20** | -1 | -1.721073e-03 | -1 | -1 | -1 | 0.000000 | 1234 | -1.721073e-03 |
| **21** | -1 | 5.960978e-03 | -1 | -1 | -1 | 0.000000 | 1146 | 5.960978e-03 |
| **22** | -1 | 1.977931e-03 | -1 | -1 | -1 | 0.000000 | 2380 | 1.977931e-03 |
| **23** | -1 | -1.040229e-02 | -1 | -1 | -1 | 0.000000 | 3465 | -1.040229e-02 |
| **24** | -1 | -5.361242e-03 | -1 | -1 | -1 | 0.000000 | 5845 | -5.361242e-03 |
| **25** | -1 | -2.889976e-04 | -1 | -1 | -1 | 0.000000 | 163425 | -2.889976e-04 |
| **26** | 42 | 5.000000e-01 | 27 | 28 | 29 | 10.558629 | 5396 | -5.835815e-03 |
| **27** | -1 | 2.106461e-03 | -1 | -1 | -1 | 0.000000 | 1991 | 2.106461e-03 |
| **28** | -1 | -1.047989e-02 | -1 | -1 | -1 | 0.000000 | 3405 | -1.047989e-02 |
| **29** | -1 | -5.835815e-03 | -1 | -1 | -1 | 0.000000 | 5396 | -5.835815e-03 |
| **30** | -1 | -4.662897e-04 | -1 | -1 | -1 | 0.000000 | 168821 | -4.662897e-04 |

Table 10: data representation of the 1000th decision tree in the model

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  | SplitVar | SplitcodePred | LeftNode | Rightnode | MissingNode | ErrorReduction | Weight | Prediction |
| **0** | 59 | 5.000000e-01 | 1 | 29 | 30 | 3.011652 | 168821 | -1.381588e-04 |
| **1** | 58 | 5.000000e-01 | 2 | 27 | 28 | 3.137466 | 118079 | -1.655025e-05 |
| **2** | 20 | 5.000000e-01 | 3 | 22 | 26 | 2.755307 | 115506 | 3.378129e-05 |
| **3** | 43 | 5.000000e-01 | 4 | 5 | 21 | 3.910891 | 110349 | 2.647489e-05 |
| **4** | -1 | -3.083270e-05 | -1 | -1 | -1 | 0.000000 | 88678 | -3.083270e-05 |
| **5** | 28 | 5.000000e-01 | 6 | 19 | 20 | 24.825976 | 21671 | 2.609782e-04 |
| **6** | 30 | 5.000000e-01 | 7 | 17 | 18 | 23.363185 | 20325 | -2.767289e-04 |
| **7** | 29 | 5.000000e-01 | 8 | 15 | 16 | 25.693055 | 19032 | -8.095272e-04 |
| **8** | 27 | 5.000000e-01 | 9 | 13 | 14 | 21.796444 | 17730 | -1.420370e-03 |
| **9** | 31 | 5.000000e-01 | 10 | 11 | 12 | 15.023243 | 16435 | -2.030422e-03 |
| **10** | -1 | -2.520285e-03 | -1 | -1 | -1 | 0.000000 | 15333 | -2.520285e-03 |
| **11** | -1 | 4.785427e-03 | -1 | -1 | -1 | 0.000000 | 1102 | 4.785427e-03 |
| **12** | -1 | -2.030422e-03 | -1 | -1 | -1 | 0.000000 | 16435 | -2.030422e-03 |
| **13** | -1 | 6.321870e-03 | -1 | -1 | -1 | 0.000000 | 1295 | 6.321870e-03 |
| **14** | -1 | -1.420370e-03 | -1 | -1 | -1 | 0.000000 | 17730 | -1.420370e-03 |
| **15** | -1 | 7.508638e-03 | -1 | -1 | -1 | 0.000000 | 1302 | 7.508638e-03 |
| **16** | -1 | -8.095272e-04 | -1 | -1 | -1 | 0.000000 | 19032 | -8.095272e-04 |
| **17** | -1 | 7.565667e-03 | -1 | -1 | -1 | 0.000000 | 1293 | 7.565667e-03 |
| **18** | -1 | -2.767289e-04 | -1 | -1 | -1 | 0.000000 | 20325 | -2.767289e-04 |
| **19** | -1 | 8.380515e-03 | -1 | -1 | -1 | 0.000000 | 1346 | 8.380515e-03 |
| **20** | -1 | 2.609782e-04 | -1 | -1 | -1 | 0.000000 | 21671 | 2.609782e-04 |
| **21** | -1 | 2.647489e-05 | -1 | -1 | -1 | 0.000000 | 110349 | 2.647489e-05 |
| **22** | 42 | 5.000000e-01 | 23 | 24 | 25 | 25.203266 | 5157 | 1.901230e-04 |
| **23** | -1 | -1.019879e-02 | -1 | -1 | -1 | 0.000000 | 1206 | -1.019879e-02 |
| **24** | -1 | 3.361227e-03 | -1 | -1 | -1 | 0.000000 | 3951 | 3.361227e-03 |
| **25** | -1 | 1.901230e-04 | -1 | -1 | -1 | 0.000000 | 5157 | 1.901230e-04 |
| **26** | -1 | 3.378129e-05 | -1 | -1 | -1 | 0.000000 | 115506 | 3.378129e-05 |
| **27** | -1 | -2.276012e-03 | -1 | -1 | -1 | 0.000000 | 2573 | -2.276012e-03 |
| **28** | -1 | -1.655025e-05 | -1 | -1 | -1 | 0.000000 | 118079 | -1.655025e-05 |
| **29** | -1 | -4.211477e-04 | -1 | -1 | -1 | 0.000000 | 50742 | -4.211477e-04 |
| **30** | -1 | -1.381588e-04 | -1 | -1 | -1 | 0.000000 | 168821 | -1.381588e-04 |