Leveraging Data Science to Predict Car Collision Severity

Theo Hayeck

September 2020

Contents

1 Introduction

1.1 Background

A traffic collision, also called a motor vehicle collision, car accident, or car crash, occurs when a vehicle collides with another vehicle, pedestrian, animal, road debris, or other stationary obstruction, such as a tree, pole or building. Traffic collisions often result in injury, disability, death, and property damage as well as financial costs to both society and the individuals involved.[0] When motor-vehicles were first introduced to the US in the early 20th century, they were few in numbers but resulted in a disproportionate number of roadside casualties

compared to the traditional horse-drawn carriages at the time. Neither comprehensive traffic laws nor significant safety features in cars or on roads existed. Since then numerous improvements in safety features, manufacturing regulations, and traffic laws have been introduced. For comparison, in 2015 almost 5,000 pedestrians died in traffic accidents, whereas in 1937, 15,000 pedestrians were killed, when the US had far fewer cars and two-fifths of its current population.[1]

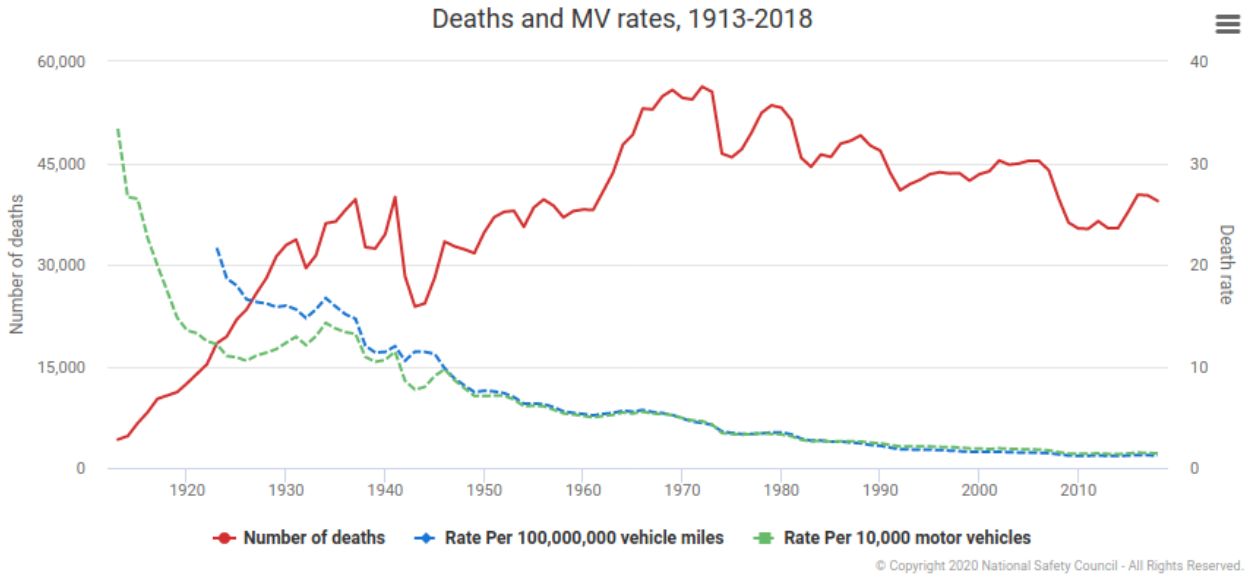

Figure 1: Deaths and MV rates 1913-2018

In 2015, the Seattle Department of Transportation (SDOT) released a 10-year plan for providing safe and sustainable transportation infrastructure for the city. A "Vision Zero" goal was adopted to eliminate serious and fatal crashes by the year 2030. However, since the release of the plan, the annual total number of collisions has only decreased by 27.6%, with the total number of severe collisions (where a collision results in an injury or a fatality) has decreased by only 18.4% [2], and making up a larger proportion of all recorded incidents annually. These findings suggest that although the initiatives implemented by the SDOT to reduce collisions were successful to a point, there are still improvements that could be made by more accurately targeting the causes of such collisions; such as: weather conditions, speeding regulations and light conditions.

In this report, we will be answering the question "Can we build a Machine Learning Model that than predict the collision severity of a crash given a set of characteristics?". This will be vital to the Seattle Department of Transportation in order to better allocate resources to limit these collisions, and to other road users who may choose to use this model in order to be alerted when they are entering potentially dangerous locations/conditions.

2 Data Understanding

2.1 Label Selection

As a case study, this report will use the collision data collected by the SDOT Traffic Management Division, Traffic Records Group (from 2004 to present) for the city of Seattle [3]. In this section, we will be undertaking an exploratory analysis on the dataset to determine the potential features that should be used in the machine learning model to predict the car collision severity.

The dataset consists of 194,673 incidents recorded by the SDOT, with 38 columns corresponding to various attributes about each collision. A SEVERITYCODE is assigned to each incident, classifying each as:

- 1 = property damage only
- 2 = injury collision

This will be used as our dependent variable and label for each collision. The data shows that the majority of all collisions (136,485) are of type 1, and the remaining 58,188 are of type 2. This is important to note because this imbalance of class labels will result in a bias to the machine learning model if it is not addressed later.

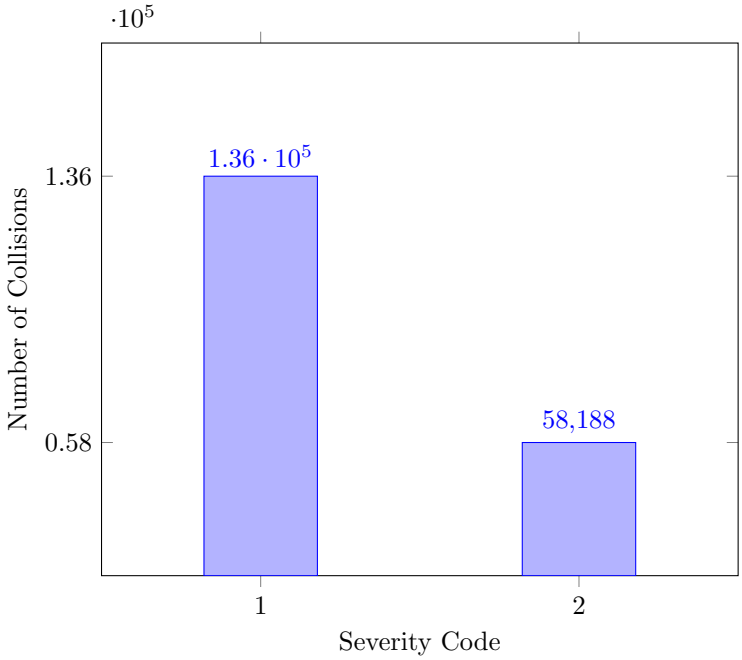

Figure 2: Collision Severity (Inbalanced)

The information about the label can be taken either from the metadata or from the column SEVERITYDESC.2. Thus, SEVERITYCODE.1 being a duplicate of SEVERITYCODE can be dropped along with SEVERITYDESC and COLLISIONTYPE.

2.2 Feature Selection

In this subsection we will explore what features should be used in our machine learning model that contribute to predicting the severity of a car collision.

Firstly, the dataset contains numerous other unique identification keys that could be used to identify each collision or it's location. These include:

- OBJECTID
- INCKEY
- COLDETKEY
- INTKEY
- SEGLANEKEY
- CROSSWALKKEY
- REPORTNO

However, none of these can uniquely identify the *severity* of the collision and therefore cannot act as the dependent variable. Furthermore, none contribute relevant information that could be used find the collision severity. Therefore, all of these columns can be dropped from the feature set as they hold no value to the machine learning model.

The remaining columns provide information regarding the conditions relating to the collision (such as weather and road conditions), or the resulting impact of the collision to it's surroundings (such as the number and type of person or vehicle affected). Such columns include:

- ADDRTYPE
- LOCATION
- JUNCTIONTYPE
- COLLISIONTYPE
- WEATHER
- ROADCOND
- LIGHTCOND
- SPEEDING
- UNDERINFL
- PERSONCOUNT
- PEDCOUNT
- PEDCYLCOUNT
- VEHCOUNT
- PEDROWNOTGRNT
- HITPARKEDCAR
- STATUS
- EXCEPTRSNCODE
- EXCEPTRSNDESC
- SDOT_COLCODE
- SDOT_COLDDESC
- SDOTCOLNUM
- ST_COLDDESC
- X & Y Coordinates of the collision

ADDRTYPE, LOCATION, and X & Y coordinates all provide information relating to the location of the collision. Since X & Y coordinates provide an exact location of the crash, ADDRTYPE and LOCATION in comparison to X & Y coordinates provide less specific information, and can be considered dispensable. Therefore, these shall be dropped from the feature set.

Furthermore, PEDROWNOTGRNT and HITPARKEDCAR provide information relating to whether or not the pedestrian right of way was not granted and whether or not the collision involved hitting a parked car, respectively. Although these provide information relating to a potential cause and effect of the collision, they both do not affect the *severity* of the collision. Therefore, these too will be dropped from the feature set.

EXCEPTRSNDESC, COLLISIONTYPE, SDOT_COLCODE, SDOT_COLDDESC SDOTCOLNUM and ST_COLDDESC provide information describing the collision, such as the angle of the collision. Individually, these can be relevant to predicting the severity of the crash. However, as mentioned later in this report, I have chosen to use information provided by the state rather than by the SDOT due to the state's ST_COLCODE providing more reliable and reproducible descriptions. Therefore, these columns will also be dropped from the feature set.

Finally, PERSONCOUNT, PEDCOUNT, PEDCYLCOUNT, and VEHCOUNT all provide a continuous measurement of the number of people, pedestrians, cyclists and vehicles involved in the crash. However, the number of participants involved in the collision does not directly affect the severity of the crash. These therefore do not contribute any relevant information needed by the machine learning model and therefore will also be dropped from the feature set.

This leaves the remaining columns as features for the machine learning model to be developed later on in this report. The table below shows each feature to be used along with a description of what information it provides. These features provide information that is relevant to predicting the severity of the collision.

Table 1: Selected features and their descriptions

Feature	Description
SEVERITYCODE	Code that corresponds to severity of the collision
X	The longitudinal coodinate of the collision:
Y	The latitudinal coodinate of the collision
JUNCTIONTYPE	category of junction at which the collision took place
WEATHER	Description of the weather conditions during the collision
ROADCOND	Condition of the road during the collision
LIGHTCOND	Light conditions during the collision
INDTTME	Date and time of the incident
INATTENTIONIND	Whether or not the collision was due to inattention
UNDERINFL	Whether or not driver involved was under the influence of drugs or alcohol
SPEEDING	Whether or not speeding was a factor in the collision
ST_COLCODE	A code issued by the state corresponding to a description of the collision.

- **X & Y:** The location of the collision is important because there could be a possible relation between the location of the collision and its severity. For example, a collision on a motorway is more likely to be a type 1 than on a quite road.
- **JUNCTIONTYPE & ROADCOND:** The type and condition of the road at the collision can also be a factor in determining the severity of the collision. It is commonly known that some junctions are more dangerous than others, and road conditions can affect a car's handling and stopping distance. These therefore may determine how severe a collision is.
- **WEATHER & LIGHTCOND:** These environmental factors may impede a driver's ability to see, this could therefore affect the driver's response time and result in a more severe collision is vision was impaired.
- **INATTENTIONIND, UNDERINFL & SPEEDING:** These human factors can result in the driver being unable sufficiently control their vehicle, leading to a greater probability of a more harmful crash due to a greater speeds and lack of control.
- **INDTTME:** The date and time of a collision is also significant when predicting the severity of a collision. For example, despite 60% less traffic on the roads, more than 40% of all fatal car accidents occur at night. [4]

- **ST_COLCODE:** The type of crash can have a significant affect on the severity of the collision. For example, a head on collision with another vehicle will have a higher probability of being more severe than a collision by two vehicles both moving in the same direction.

All other columns in the dataset not aforementioned can be assumed dropped. This is because of information pertaining to these columns already being provided (duplicate features) and/or are more accurately represented by other selected features.

3 Data Preparation

The purpose of this section is to clean and format the selected features to be used in the machine learning model. Examples of data preparation, often called pre-processing, include the handling of missing or NaN values (cleaning) and converting variables into machine-legible data types (formatting). Data preparation also involves balancing labels to ensure an unbiased model, this will also be discussed at the end of this section.

3.1 Cleaning

The features X & Y provide information relating to the longitude and latitude of the collision location. In the dataset, there are 5,334 missing values between them. Since longitude and latitude are both continuous variables, a common strategy of dealing with missing values would be to replace them with either the median or mean of X & Y. In this context, using a median value of X & Y would simply just point to the center of the X*Y map of Seattle (See Figure 3 below). However, using the mean would point to an area of the map where collisions happen most frequently. X has a mean value of -122.330518439041, and Y having a mean value of 47.6195425176886. These values shall be used to fill in X & Y's missing values, respectively.

JUNCTIONTYPE, WEATHER, ROADCOND and LIGHTCOND are all categorical variables that describe the surrounding conditions present at the time of the collision. To handle missing values for these type of variables, a most frequent value for each variable called the mode is chosen to replace the missing values.

- **JUNCTIONTYPE**, has 6,329 missing values. The mode, "Mid-Block (not related to intersection)", is chosen to replace these values. The JUNCTIONTYPE "Unknown" is also replaced with the mode, since this alone does not contribute any meaningful weight to the machine learning model. Afterwards,the different categories are converted into dummy variables for the model, and the original variable JUNCTIONTYPE is dropped.
- **WEATHER**, has 5,081 missing values. The mode, "Clear", is chosen to replace these. Furthermore, since there is no explanation of what "Other" weather is, it is merged with the category "Unknown". "Unknown" is renamed to "Unknown Weather", this is to distinguish it from other "Unknown" values given in other features. Additionally, "Partly Cloudy" is synonymous to "Overcast", these are therefore merged together as "Overcast". Again, the feature header WEATHER is also dropped in place for its dummy.
- **ROADCOND**, has 5,012 missing values. The mode, "Dry", is chosen to replace these. Similar to WEATHER, "Other" is merged with "Unknown" and renamed as "Unknown Roadcond".
- **LIGHTCOND**, has 5,170 missing values. These are replaced by its mode, "Daylight".

3.2 Formatting

- **INATTENTIONIND** has 29,805 observations given as "Y", we will therefore assume that the missing 164,868 observations are "N". To use these observations in our machine learning model, we need to format them in such a way that the model can easily interpret. We will therefore encode them as 1 and 0, respectively.
- **UNDERINFL** observations are given as either "Y", "N", "1", or "0". Following the logic we have used for INATTENTIONIND, we can safely assume that "1" and "Y" are synonymous. Likewise, for "0" and "N". We will therefore convert any occurrence of "Y" and "N" to their respective number.
- **SPEEDING** has 9,333 "Y" observations. We can again assume that the remaining missing 185,340 values are to be "N". These values will also be encoded as 1 and 0 respectively.
- **INCDTTM** We shall be obtaining the day of the week, and the hour of the day of each collision using the date-time string found in this column. The INCDTTM feature will then be discarded.

One-Hot Encoding: Where any feature is expressed as a categorical variable (such as WEATHER, LIGHTCOND and WEATHERCOND) we will perform one-hot encoding, such that the feature is expressed as a matrix of 1's and 0's. With 1 signifying that the feature column was present, and 0 signifying that the feature was not. For example, if the weather at a collision was "Dry", a 1 will be present in the newly created, "Dry", feature column.

3.3 Balancing the Dataset

As mentioned previously, the dataset contains 58,188 collisions of class 2 and 136,485 collisions of class 1. If left as it is, this could result in a bias to the machine learning model. This is because it is more challenging for a machine learning model to learn the characteristics of examples from the minority class due to the comparatively less data it has for it. This makes it harder for the model to distinguish members of the minority class from members of the majority class.

The two commonly used strategies to balance classes in a dataset are random over-sampling (ROS) and random under-sampling (RUS). ROS is the process of supplementing the dataset with multiple, randomly chosen copies of cases from the minority class, until the number of samples match the majority class. RUS randomly deletes samples from the majority class until the number of samples matches the minority class.Both methods come with advantages and disadvantages. While ROS may inflate or exaggerate underlying patterns in the minority class, RUS may potentially discard important samples of majority class and distort its underlying patterns. A rule of thumb is to use ROS when the given dataset is small and RUS when the given dataset is large. As the collision dataset is sufficiently large, this report will employ the random under-sampling method to balance the dataset and thereby reduce class 1 collisions to 58,188 samples.

3.4 Exploratory Data Analysis

3.4.1 Relationship between location and collision severity

It is well known that some locations result in a greater number of crashes than others. But does this mean that there is a relationship between the *severity* of a collision and the location in which it occurred?

Figure 3 shown below, displays the location of each major hotspot where at least 100 collisions (class 1 and 2) have taken place. Where those of just class 1 are depicted with yellow and blue circles, and hotspots containing at least 1, class 2 collision, as red circles. Since there are zero yellow and blue circles on the figure, it is clear that out of each hotspot where there have been more than 100 collisions, every spot has had at least 1, class 2 (serious injury) collision. Furthermore, the locations of the hotspots are often on major motorways in Seattle, such as Rainier Ave S. We can therefore conclude that these locations are particularly dangerous, and that there is a relationship between location and collision severity.

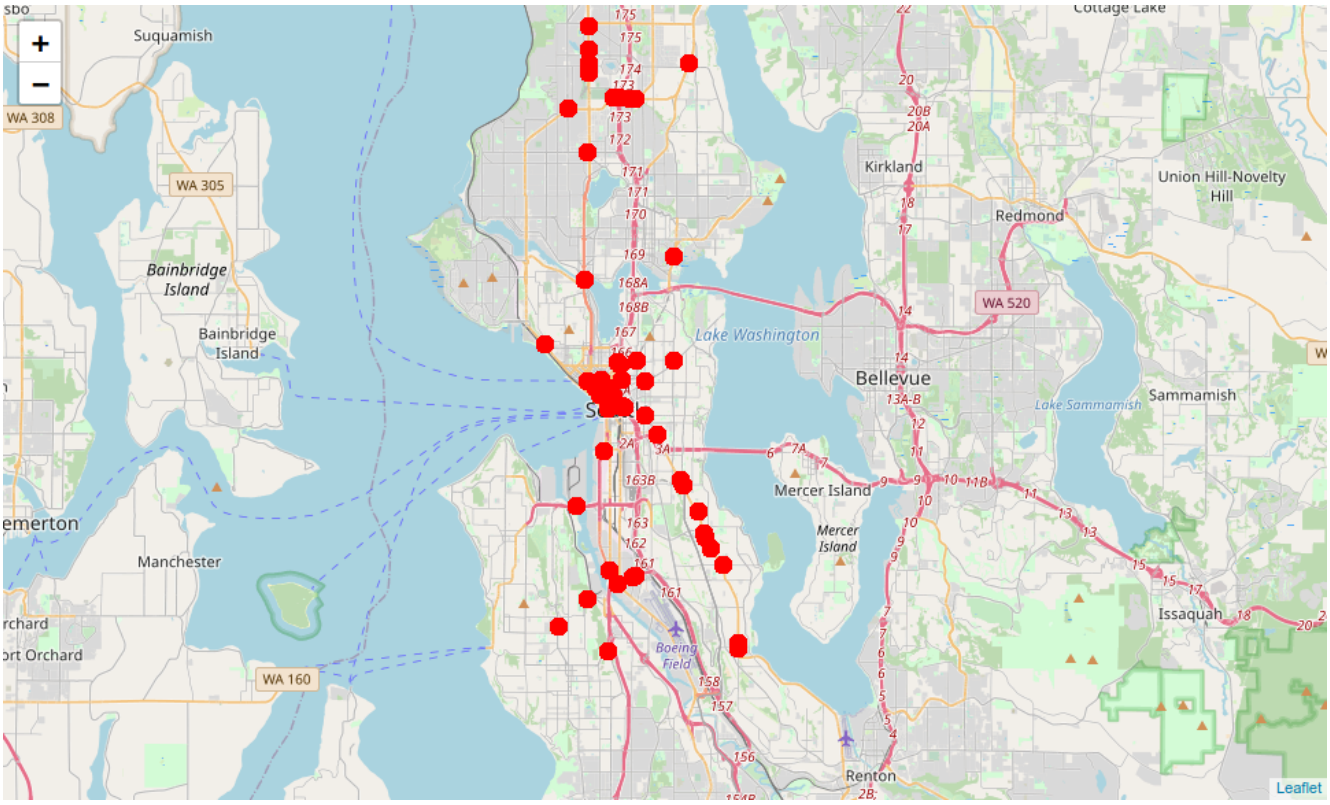

Figure 3: Locations in which more than 100 collisions (Class 1 & 2) have ocurred

3.4.2 Relationship between junction type and collision severity

It is often recognised by drivers that there are harder types of junctions to navigate than others. However, does this mean that these junctions result in a higher proportion of collisions than others?

It is shown in Figure 4 below that the most number of collisions occur at a Mid-Block. There is also a higher ratio of class 2 collisions to class 1 at "At Intersection (intersection related)" than any other type of junction. This means that if a collision was to happen at this junction type, it will result in the highest chance of the collision being of class 2 than at any other junction. However, we do not know how many junctions of each type are in Seattle, we would therefore need to normalize the total number of each type of junction in Seattle before we are able to draw any conclusions about which type is more likely to result in a Class 1 or 2 type collision.

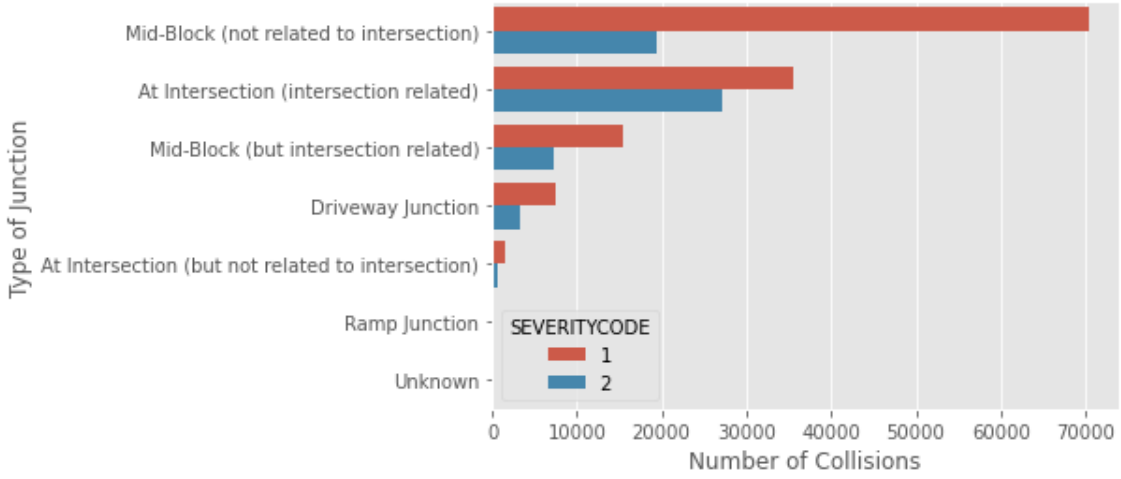

Figure 4: Bar chart to show relationship between junction type and the number of collisions

3.4.3 Relationship between weather and collision severity

Rain and snow often cause hazardous driving conditions by reducing the handling and increasing the stopping distances of vehicles on the road. It is therefore wise to investigate any relationship between weather conditions and the severity of a collision.

As you can see from Figure 5 below, the majority of collisions occur during times when the weather is "Clear". This is to be expected, since the weather is clear in Seattle most of the time. This pattern is expressed for the rest of the dataset, where the greatest number of collisions occur during the most common weather conditions.

However, it is also clear that Raining and Fog/Smog/Smoke weather conditions have the highest class 2:1 ratio out of all the weather conditions, at 0.5087 and 0.4895 respectively. Again, meaning that if a collision was to happen during this weather condition, it will result in the highest chance of the collision being of class 2 than during any other weather condition.

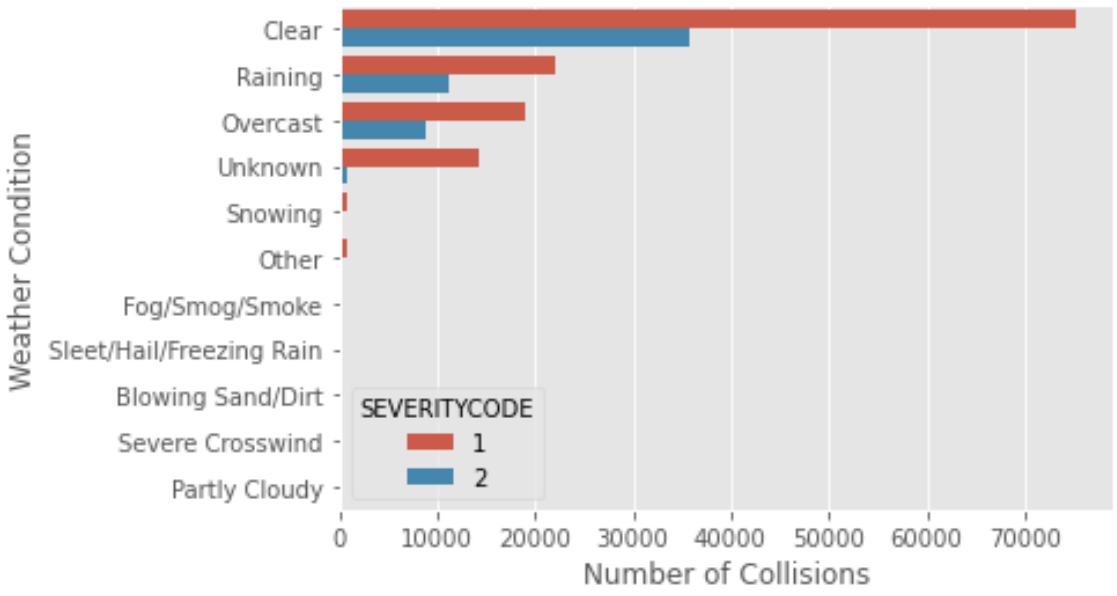

Figure 5: Bar chart to show relationship between weather conditions and the number of collisions

3.4.4 Relationship between road conditions and collision severity

It is well known that poor road conditions can reduce a driver's handling of their vehicle and increase the stopping distance should they need to react to a hazard. However, does this necessarily impact the severity of a collision?

It is shown by Figure 6 that oil on the road results in the greatest chance of a class 2 collision occurring out of all the road conditions (if the collision was to occur at all), with 40 resulting collisions being of class 1 and 24 being of class 2, the largest class ratio out of all road conditions at 0.6000. This is followed by wet road conditions having a ratio of 0.496. Again, this means that if a collision was to happen on a road with these conditions, it will result in the highest chance of the collision being of class 2 than if any other road conditions were present.

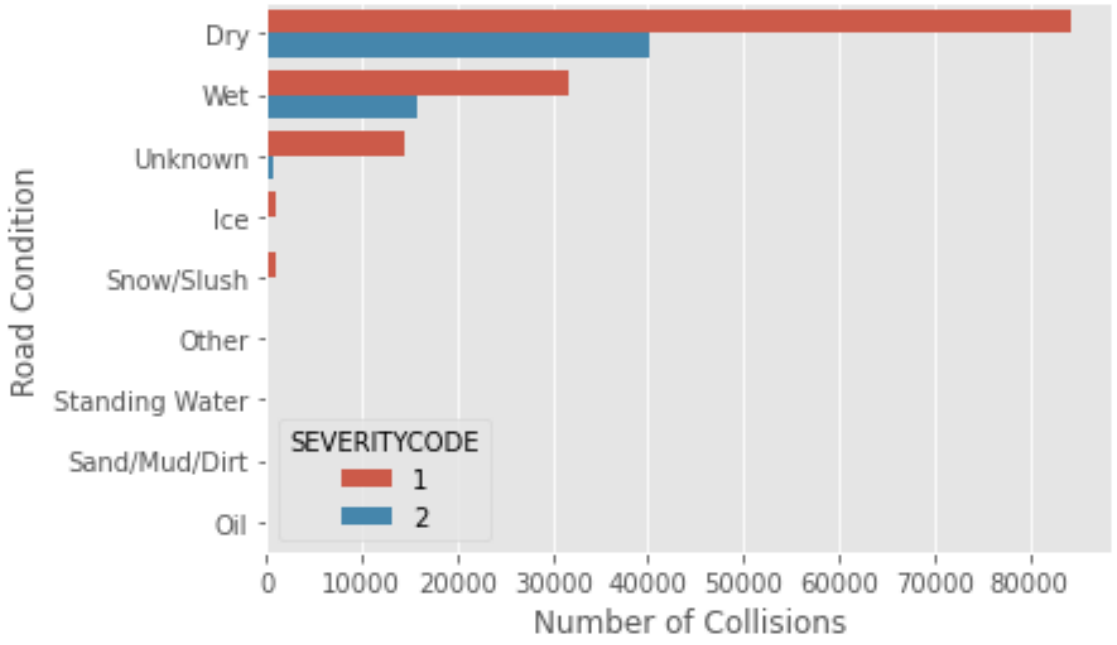

Figure 6: Bar chart to show relationship between road conditions and the number of collisions

3.4.5 Relationship between light conditions and collision severity

In lower light conditions it is often harder for drivers to see. However, does this necessarily relate to an affect on the collision severity of a crash?

In the figure below, you can see again that the most common lighting condition, "Daylight", results in the most collisions. However, "Dusk" results in the greatest chance of the collision being of class 2 if one was to occur out of all the light conditions, with a ratio of 0.491. It is to be noted that "Dark Unknown" did have a class ratio of 0.571. However, due to it being unknown, it is less reliable to use

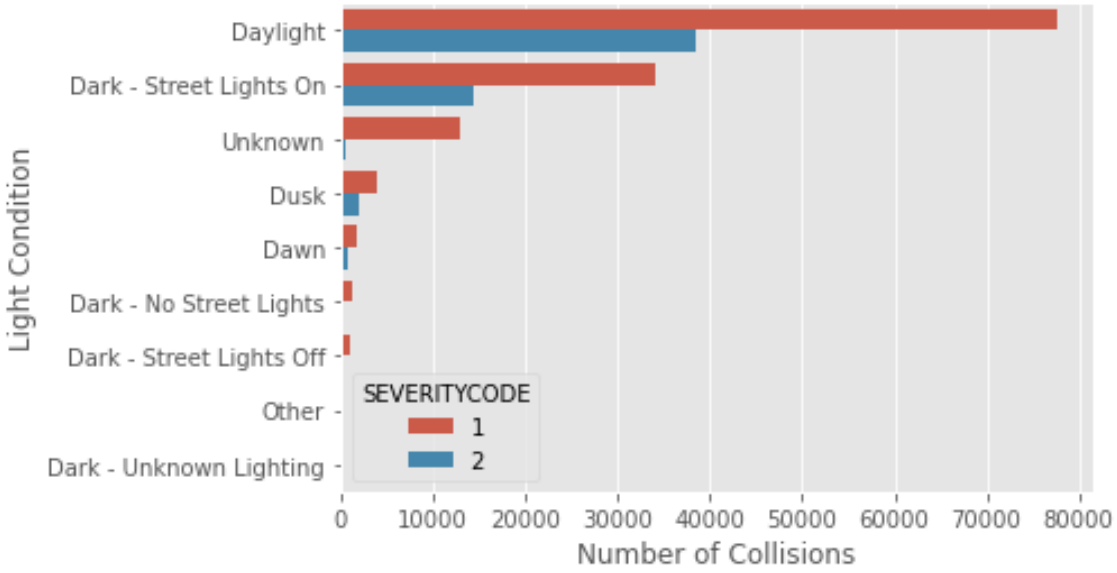

Figure 7: Bar chart to show relationship between light conditions and the number of collisions

3.4.6 Relationship between driver inattention and collision severity

A drivers lack of attention to their surroundings is often the cause of many collisions on the road. However, based on this data we can see that, the majority of the time, driver were paying attention leading up to the collision. Furthermore, we can see that out of those drivers who were not paying attention, a larger proportion of collisions were of class 2. This could suggest that there is a relationship between a lack of attention and an increased collision severity.

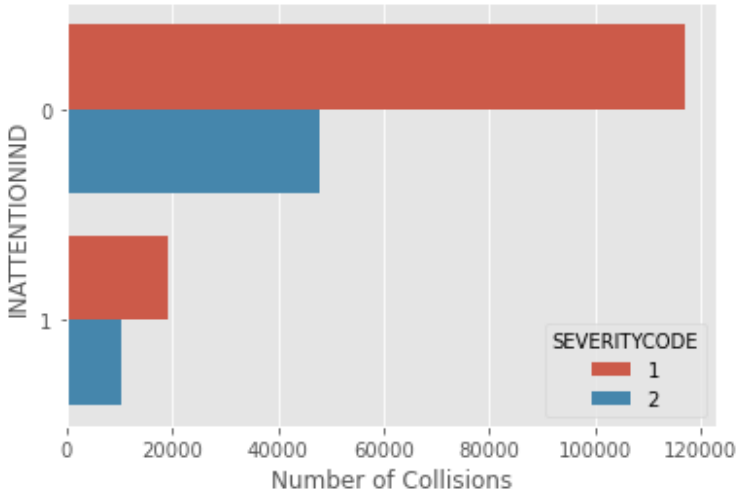

Figure 8: Bar chart to show relationship between a driver's attention and the number of collisions

3.4.7 Relationship between the influence of drugs or alcohol and collision severity

A similar trend is again seen when comparing drivers under the influence of drugs or alcohol during the time of the collision, where the majority of collisions occur when the driver is sober and not under the influence of anything; but with a higher proportion of severe, class 2, collisions occurring when the driver was under the influence.

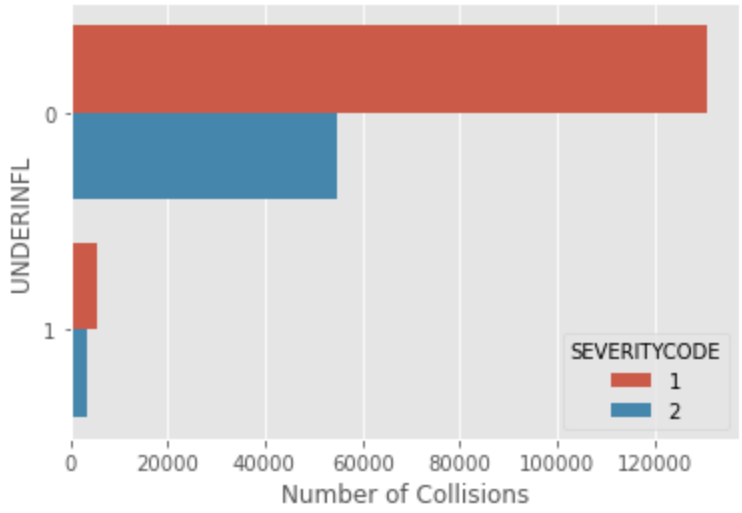

Figure 9: Bar chart to show relationship between drivers being under the influence and the number of collisions

3.4.8 Relationship between a driver speeding and collision severity

The trend again follows through when comparing drivers who were speeding during the time of the collision, where the majority of collisions occur when the driver was driving within the speed limit, but with a higher proportion of severe, class 2, collisions occurring when the driver was speeding.

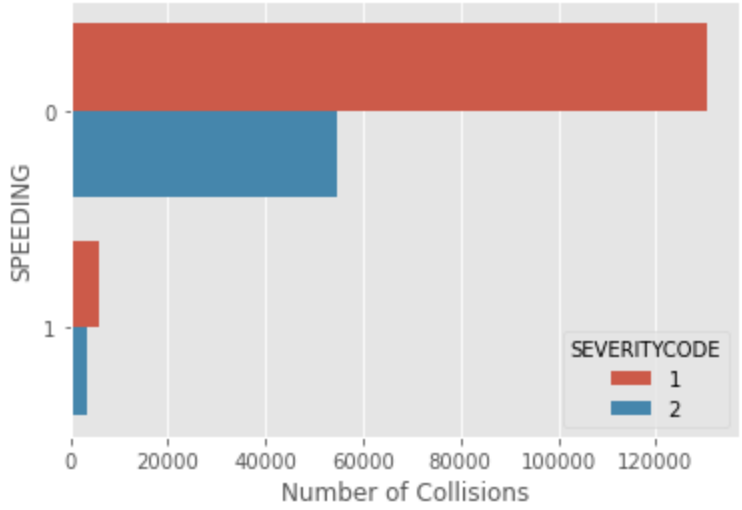

Figure 10: Bar chart to show relationship between drivers speeding and the number of collisions

3.4.9 Relationship between the time and date and collision severity

The assumption about the time and date of the collision is that during certain times, such as during week days or rush-hour, people are less attentive, for instance, due to exhaustion from work. As shown in figure(s) 11 and 12 below, this assumption holds. In figure 12 you can see that a bell shaped distribution is present, centered around 17:00 having the most number of collisions (of both class 1 and 2) in the day. Furthermore, in Figure 12 you can also see a gradual increase of collisions per day, from the start (Monday) to the end (Friday) of the working week. With a decrease in the number of collisions over the weekend. These findings reiterate that belief that when people get more tired during the day and week, more collisions occur. However, it is also to be noted that these patterns might arise as a result of there simply being more cars traveling on the roads during these times, where most people are likely to be travelling to and from work at around 8:00 and 17:00, respectively. Similarly, this is shown over the weekend, where a significant decrease in the number of collisions is seen; perhaps due to there being less cars on the road as people do not need to travel to work on the weekends. It is also important to note that in the event that a collision was recorded and the SDOT did not provide a time in which the collision occurred, just the date was inputted into the INCDTTM feature column. This has resulted in all of these events being assigned a time of 00:00, the start of the day; explaining the disproportionately large number number of collisions that have been shown to happen on the 0^{th} hour of the day in Figure 11.

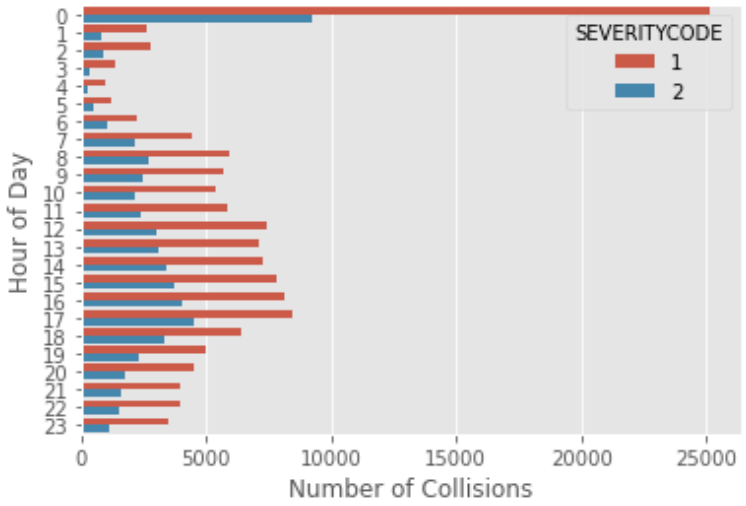

Figure 11: Bar chart to show relationship between the hour of the day and the number of collisions

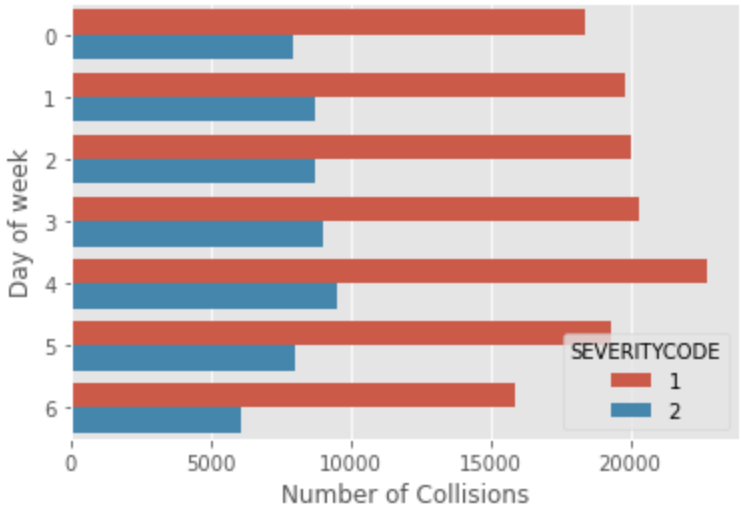

Figure 12: Bar chart to show relationship between the day of the week and the number of collisions

3.4.10 Relationship between ST_COLCODE and the number of collisions

Finally, it is has been proven that the type of collision a vehicle is in can affect its severity. As shown in Figure 13 given at the end of this report, you can see that there are specific ST_COLCODE's that result in more class 2 type collisions than class 1. These include ST_COLCODEs: 45, 0, 1 and 2. This again proves a relationship between the type of collision a vehicle is in and its severity.

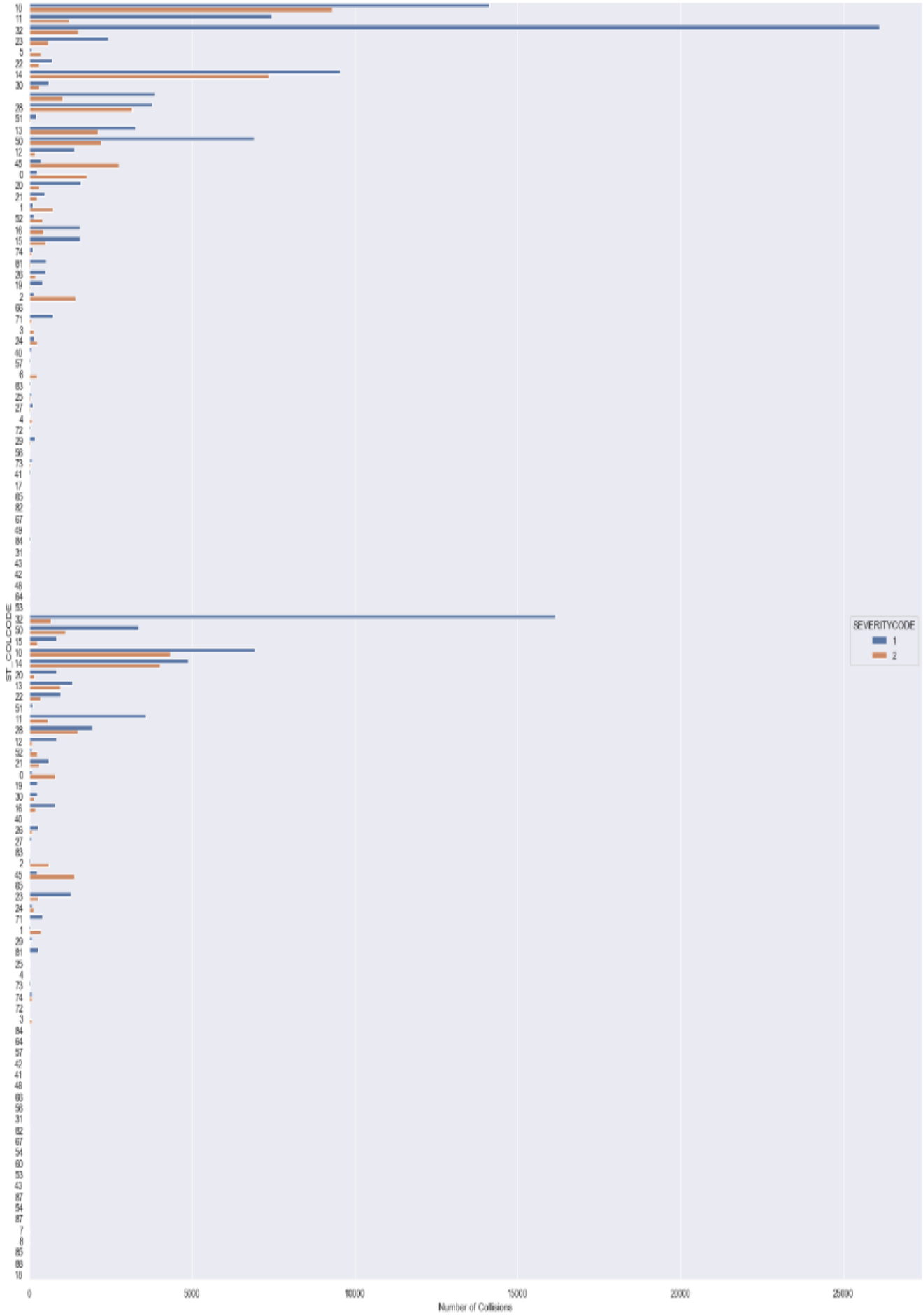

Figure 13: Bar chart to show the frequency of each type of state defined collision code.

4 Expected Outcomes

This section serves to formulate some hypotheses and anticipate potential problems with the final feature set that was created. To review, the final features set has 116,376 rows of observations equally balanced between property damage only (class 1) and injury collisions (class 2), which is still a good amount despite the reduction from RUS. However, since many columns in the original dataset were categorical, transforming them into dummies has resulted in a total of 41 features. While supervised ML algorithms, both regression and classification, should be able to handle big datasets with a large number of features, they perform better when most of those features are continuous variables. Moreover, some of these dummy features have a significant amount of observations listed under 'unknown' or 'other'. In case such categories turn out to have significant importance in the classification process, there is no means to alleviate these 'unknown' conditions to improve traffic safety. For future research on this subject, it is advised to quantify as many variables as possible during the survey. For instance, to replace WEATHER with the amount of precipitation and the windspeed; or to develop a scoring system for ROADCOND on scale from 1 to 10 (1 being the worst road conditions and 10 being the best). Ideally, to improve the traffic safety on the streets of Seattle, the classification models should highlight either features that can be influenced or be warned about by authorities. Features that can be influenced by authorities to reduce class 1 and class 2 collisions include human behavior variables. For instance, SPEEDING and UNDERINFL already represent violation of traffic laws and are already policed. They make for easy policy-levers. Furthermore, traffic authorities can employ a system of electronic traffic signs to warn commuters about sudden changes in ROADCOND and WEATHER as well as generally remind commuters to drive carefully (INATTENTIONIND or WEEKEND). Thus, the classification model is hypothesized to be a function of negligent human behavior and adverse driving conditions, for it to find significant options for road safety improvement measures. However, the opposite outcome would be that the classification models produce low accuracy scores on these features and reveal that adverse conditions do not lead to worse collisions. This could be cautiously interpreted as good news because existing measures to improve road safety have reduced the influence of adverse conditions on collision severity, but better data (as defined before) should be collected to produce more accurate results

5 Modelling

In machine learning, classification is considered supervised learning, which means learning where the class-labels are already given in the dataset. It can be binary or multi-class classification. Since there are only two classes of accident severity given, this report will develop binary classification models. The assumption is that the permutation or combination of all independent features in the dataset will have recurring patterns that 'predict' the classes. The classification algorithm will find the common pattern of combinations that correspond to either one of the dependent classes. This has many applications in various fields of business and science ranging from spam, fraud, or churn prediction over handwriting and face-recognition towards extreme event prediction and medical diagnosis. Common classification algorithms include:

- K-Nearest Neighbours
- Decision Tree
- Random Forrest
- Logistic Regression
- Artificial Neural Network

For modelling and evaluation, the dataset will be split into a training and a testing subset. The classification algorithm is trained to find the underlying pattern that predicts the classes only from the training subset, while the testing subset simulates out-of-sample accuracy testing. Since there are no means of anticipating how different ML algorithms perform on a given dataset, data scientists use this as form of controlled experiment in order to discover which algorithm and which hyperparameter values result in the most accurate classification model. Therefore, several classification algorithms will be modelled and tested to determine the most appropriate model for predicting collision severity in Seattle.

5.1 K-Nearest Neighbour (K-NN)

K-Nearest Neighbors is a comparatively simple classification algorithm that stores all available cases and classifies a new case based on a similarity to its 'nearest neighbors'. For instance, if an unknown case is compared to 5 neighboring cases and 3 out of 5 of those neighbors are class 2 while the rest are class 1, the unknown case will be classified as class 2. The distance between cases is usually measured by a standard mathematical formula for distances between points in multidimensional planes (e.g. Minkowski distance, which is also employed here). KNN models work best when the dataset is balanced and its features have been normalized. The challenge in

building an accurate KNN model lies in determining the value of K, namely the numbers of neighbours for comparison. A too small k-value may capture too much noise, while increasing it endlessly will run into diminishing marginal gains for accuracy. Given the large feature set of the training data, this challenge is approached by iterating through the integers 15 to 24 and comparing them against their respective mean accuracy (see graph below). The best mean accuracy is achieved at approximately 0.6987 with k equaling 23. A full evaluation of model performances will be given in the next chapter.

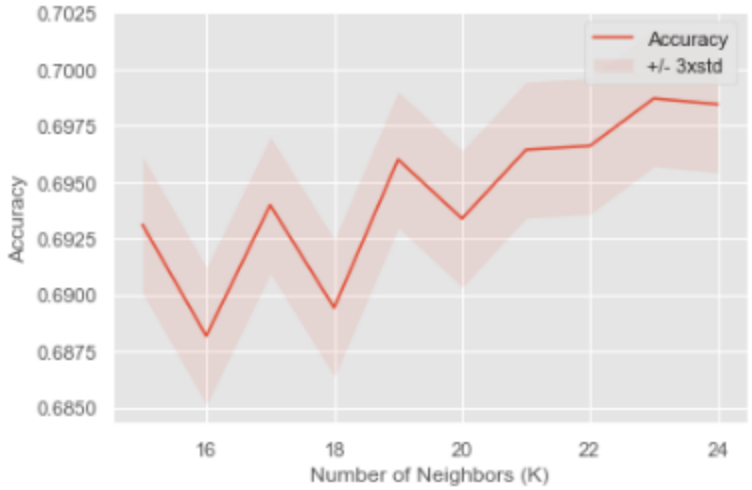

Figure 14: A line chart to show the K-NN classifier accuracy with respect to the hyper-parameter K

5.2 Decision Tree

Decision Tree classifier is another predictive modelling approach in machine learning. The Decision Tree iterates through the features about a case (represented in the branches) to conclusions about the case's target value (represented in the leaves). In Decision Trees, leafs represent class labels and branches represent conjunctions of features that lead to those class labels. The criterion by which branches eventually lead to pure leafs, where 100% of its cases fall into a single class is called 'entropy'. Entropy describes the amount of information disorder, if the sample of cases in a node is completely homogenous (i.e. 100% a single class), then the entropy equal 0. If the sample of cases in a node is completely heterogenous (i.e. split 50-50 between the 2 classes), then entropy equals 1. This process is illustrated below with a Decision Tree that has been given a maximum depth of 4 (Figure 15). Without a maximum depth, the Decision Tree algorithm will run until all leafs have become pure. This results in a general accuracy score of 0.6310 on the given feature set for predicting severity classes of collisions. The advantages of Decision Trees are that they work well with categorical data and are easy to interpret (as they mirror human decision making). There disadvantages are that they can become overly complex quickly and small changes in the training data can upset their whole structure.

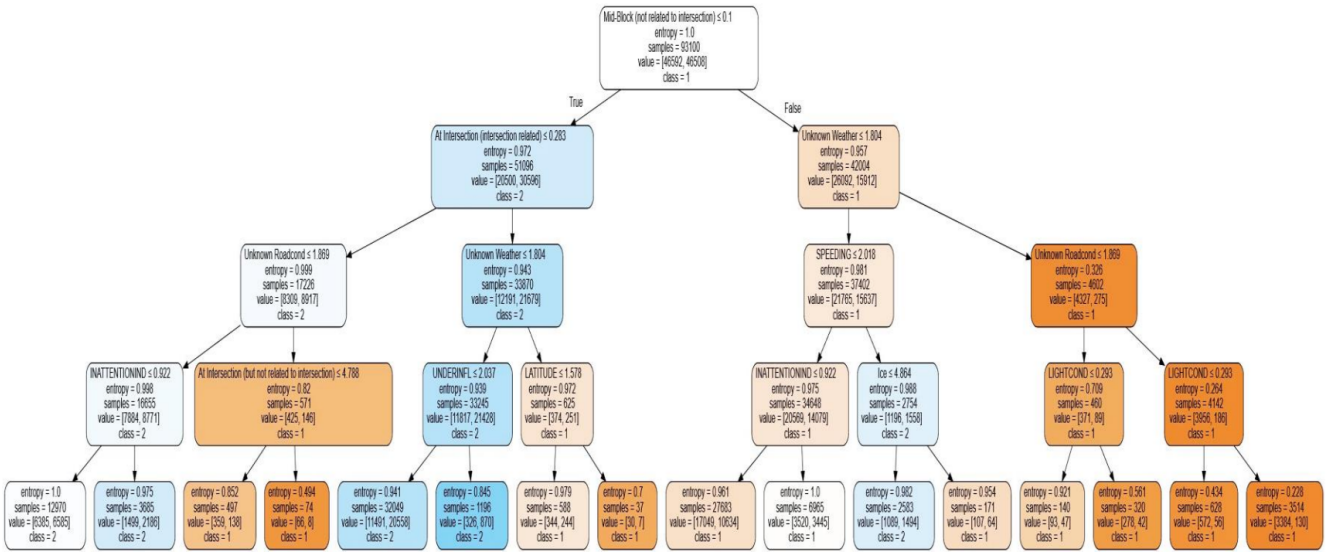

Figure 15: Decision Tree with a maximum depth = 4

5.3 Random Forest

The Random Forrest improves on the advantages and disadvantages of Decision Trees. It is called Random Forest because it operates by constructing numerous Decision Trees on various, random sub-samples of the dataset and then outputs the class that is the mode of the classes. The number of trees in the forest is set to 100 and the criterion 'entropy' ensure the same method of information gain is used as before. Overall, the Random Forrest produced attained a general accuracy of 0.6825, slightly higher than that of single Decision Tree. Fortunately, this means that the previous Decision Tree model was very close to the precision of the Random Forrest, but inversely this suggests that a much higher accuracy with both these models cannot be attained on the current training dataset.

We can see from below that by far the dominant features that predict the severity of the crash is the location of which it took place (X & Y coordinates), followed by the time of the day the crash occurred, then in what day of the week. The ST_COLCODEs given by the state, 32 and 45, also have a considerable impact on determining the severity of the collision.

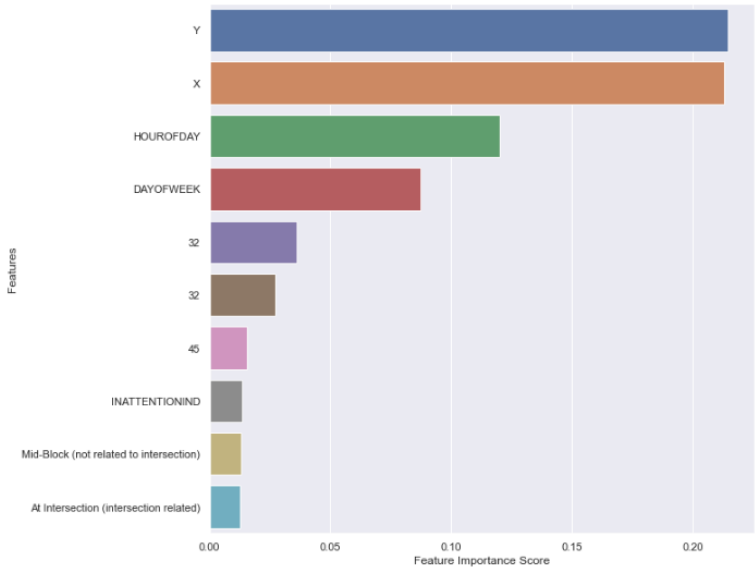

Figure 16: Random Forest feature importance scoring

5.4 Logistic Regression

In statistics and machine learning, Logistic Regression is employed not only to model binary classification as pass/fail, win/lose, or healthy/sick, but also the probability of a case falling into either class. Logistic Regression analysis is an alternative method to linear regression. Whereas linear regression tries to predict a continuous outcome by finding a linear equation, Logistic Regression does not look at the relationship between target and features as a straight line. Instead, Logistic Regression uses the natural logarithm function to find the relationship among the features that separate the targets into classes. In order to minimize the error of fitting this function to the training dataset, several solvers can be used such as liblinear, SAG, and SAGA.

Liblinear is a library for large linear classification that support logistic regression and linear support vector machines. A linear classifier model works by making a classification based on the value of the linear combination of the feature values. Liblinear is recommended for large-scale and high-dimension datasets, same as is given here. It uses coordinate descent which successively approximates minimization. Stochastic Average Gradient or SAG minimizes the sum of error on a smooth convex line. It is also considered a fast solver for large datasets, when both the number of samples and the number of features are high. The SAGA solver is a variant of SAG that supports the non-smooth error minimization. It is recommended for multi-class classification. All solvers also need regularization parameter C (to prevent over-fitting). The value of C is inverse to its regularization strength; i.e. smaller values specify stronger regularization.

Despite applying different solvers to the training data in order to classify collision severity, all three Logistic Regression models have attained an identical general accuracy score of around 0.7106. As the Liblinear solver seems the most appropriate for the type of dataset that is given here, it will be used for the full evaluation later. Additionally, since Logistic Regression can also estimate probability of class, it will also be included in the full evaluation

5.5 Artificial Neural Networks

An Artificial Neural Network (ANN) is constituted of nodes that are called artificial neurons, which loosely mirror the neurons in a biological brain. Each connection between nodes, like the synapses in a biological brain, can transmit a signal to other neurons. In ANNs, neurons receive a number of signals or input data, they perform some mathematical function on that data, and then output a signal to the next neuron if the value of their activation function is larger than a given threshold value. To transfer their signal outputs to other neurons, neurons have connections named edges and neurons are aggregated into layers, with the middle layers called hidden layers. Signals can traverse from the first layer to the last layer multiple times and thus also recreate the feedback mechanism of biological neurons.

An Artificial Neural Network classifier trains itself by iterating through the original input data and comparing the final output (i.e. the prediction) with a given target label. As for any other ML algorithm, the difference between predicted and true label is the error. With successive iterations through the data, the ANN adjusts the functions of neurons and layers, which will cause the ANN to converge on predicted outputs with minimal difference to the true target class. Here, the number of iterations is set at 200 and solver is set for adam. Similar to Logistic Regression, ANN can also estimate the probability of a case falling into a dependent class. The general accuracy score attained by the Artificial Neural Network is 0.6933.

6 Evaluation

Table 2: Machine Learning models and their accuracy scores.

Classifier	Accuracy	Jaccard-Score	F1-score	Log-Loss
K-Nearest Neighbour	0.6987	0.5181	0.6956	n/a
Decision Tree	0.6310	0.4644	0.6351	n/a
Random Forest	0.6825	0.5138	0.6873	n/a
Logistic Regression	0.7106	0.5197	0.7084	0.5413
Artificial Neural Network	0.6933	0.5051	0.6955	0.5714

Table 3: Variation in accuracy scores.

	Accuracy	Jaccard-Score	F1-score	Log-Loss
count	5.0000	5.0000	5.0000	2.0000
mean	0.6854	0.5042	0.6844	0.5564
std	0.0293	0.0230	0.0286	0.0213
min	0.6351	0.4644	0.6351	0.5413
25%	0.6876	0.5051	0.6873	0.5488
50%	0.6962	0.5138	0.6955	0.5564
75%	0.6977	0.5181	0.6956	0.5639
max	0.7106	0.5197	0.7084	0.5714

Various evaluation metrics can be used to assess and compare the performance of the ML classification models developed in the previous chapter. All accuracy scores in tables below are obtained by testing the models on the previously created testing subset to simulate their performance on out-of-sample cases. The simplest accuracy score, titled General Accuracy before, is the number of all correct predictions over the total number of samples. It also can be read as the percentage of how many predictions were correct of all prediction that were made. All classification models tested have an average General Accuracy score 68.54% and they all fall within one standard deviation of 2.93%. Except for the Decision Tree, which is within 2 standard deviations below the mean accuracy for all the classifiers, illustrating the worst performing model in terms of General Accuracy.

The Jaccard Score, also known as Jaccard similarity coefficient is a metric for calculating the dissimilarity between two sample sets, i.e. the predicted classes and the actual classes. The Jaccard Score is defined as the size of the intersection divided by the size of the union of the two sample sets. It can also be read as percentage; the higher the values, the higher the overlap between the samples. The classification models developed have an average Jaccard score of 50.42% and fall within one standard deviation of 0.5042. Except for the Decision Tree which is within 2 standard deviations above the mean, showing lower performance than the other models in terms of Jaccard Score

While the General Accuracy and Jaccard-Score are relatively simple metrics of hit or miss, the F1-Score is more sophisticated accuracy measure because it measures the balance between true positive and false positives. The F1-score is the harmonic mean of the precision and recall (ranging between 0 for worst and 1 for best). Precision is the number of true positive results divided by the number of all positive results, including false positives.

Recall is the number of true positive results divided by the number of all samples that should have been identified as positive. 4 out of 5 classification models fall within one standard deviation of the mean of 0.6844. Again, the decision tree performed worse (within 2 standard deviations below the mean), while the Logistic Regression performed best, with the highest F1-score of 0.7084.

Logistic Regressions and artificial Neural Networks can estimate the probability of a case falling in either of the dependent classes on top of their binary classification capabilities. A common metric the assess the uncertainty of the predicted probabilities is logistic loss (also called cross-entropy loss), abbreviated as Log-Loss. Log-Loss scores read in the opposite direction as the previous accuracy metrics. For any given problem, a lower Log-Loss value means better predictions as it means lower uncertainty. The ANN model did slightly better than the Logistic Regression model by a difference of 0.0301 in terms of Log-Loss.

7 Conclusion

7.1 Summary of Findings

Finally, as mentioned above, the feature importance scoring has illustrated that the combination of location (X & Y) have the biggest impact on determining the severity of the collision. Followed by the time of the day the crash occurred (TIMEOFDAY), then in what day of the week (DAYOFWEEK). The ST_COLCODEs, 32 and 45, being present and JUNCTIONTYPE: At Intersection (intersection related) or Mid-Block (not intersection related) also contribute to determining the severity of the collision, but to a lesser extent.

It is surprising to see that based on the findings made in this report, road and weather conditions have little impact in determining the severity of a collision in Seattle. Furthermore, inattention, a driver being under the influence, and speeding also do not heavily influence the severity of a collision - despite the fact that they increase the likelihood of a collision, of any class, occurring.

In conclusion, Logistic Regression and Artificial Neural Network performs best in all evaluation categories. However, there is very little overall variation in the performance of all models and no single model significantly standouts as better, despite their very different approaches. Ultimately, the top three best models for predicting collision severity on the streets of Seattle have been:

- Artificial Neural Network
- Logistic Regression
- K-Nearest Neighbours

7.2 Further Research and Recommendations

While the intentions of this report were to develop a ML classification model to predict accident severity and that answers from those models may help implement improvements to traffic safety in Seattle. The way data about the collisions is currently collected is insufficient to make satisfactory predictions about the severity of collisions and to reveal what combination of features lead to increased severity. As stated before, supervised ML algorithms, both regression and classification, perform much better when most of those features are continuous variables rather than categorical. For future research on this subject, it is advised to quantify as many variables as possible during the data collection. For instance, WEATHER can be replaced with the amount of precipitation and the windspeed; ROADCOND could be measured in by a scoring system on scale from 1 to 10 (1 being the worst road conditions and 10 being the best). Then the remaining indicator features such as INATTENTION, SPEEDING, or UNDERINFL may also have a bigger weight on the predictions. However, the research done in this report did accurately identify the key possible causes that result in differing collision severity. Most of which, the Seattle Department of Transport can easily implement improvements on. Based on my results, I would recommend improvements to traffic calming measures in collision hotpots outlined in section 3.4.1, or introducing stricter traffic regulations in key times of the day and week outlined in section 3.4.9. Finally, I would research further into possible ways to improve Mid-Blocks and Intersections outlined in section 3.4.2. In its current form there are no deployment options for the models developed here. When better data is available, this project should be repeated in order to determine if more can be done to alleviate severe collisions that still happen.

8 References

- [0] https://en.wikipedia.org/wiki/Traffic_collisions
- [1] https://injuryfacts.nsc.org/motor-vehicle/historical-fatality-trends/deaths-and-rates/

- [2] https://www.seattle.gov/visionzero
- [3] https://s3.us.cloud-object-storage.appdomain.cloud/cf-courses-data/CognitiveClass/DP0701EN/version-2/Data-Collisions.csv
- [4] https://seriousaccidents.com/legal-advice/top-causes-of-car-accidents/nighttime-driving/