IBM and Coursera Applied Data Science Capstone Project

REPORT: Predicting Car Collision Severity in Seattle

By Tom K. Walter (04 September 2020)

Inhalt

1. Introduction: Business Understanding

1.1 Background

For any traffic participant, pedestrian, cyclist, or motorist, an accident is an unexpected and undesired thing to experience. Property damage, personal injury or death as consequences of collisions not only impact the lives of individuals but also society and economy at large. When cars were first introduced in the US at the beginning of the 20th century, they were few numbers but the fatalities they caused were many. Neither comprehensive traffic laws nor significant safety features in cars or on roads existed. Since then, landmarks in auto safety (produced by commercial, technological, legal, or moral forces) have reduced both motor vehicles accident deaths and pedestrian deaths to an all-time low (see Figure 1 and Figure 2).[1] To name a few such landmarks: car manufacturers have introduced seatbelts and airbags; civil engineers have paved roads and added reflecting guard-rails to highways; moral campaigners have lobbied against drunk driving; and last but not least government has promulgated and enforced many traffic (safety) laws. 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 (see Figure 2). Although motor vehicle accident deaths and pedestrian deaths in the US are at an all-time low, each death or injury that still occurs is a tragedy and thus remain a public health concern.[2]

Figure 1: Motor Vehicle Accident Deaths, US, 1921-2015

Source: Pinker, Steven. *Enlightenment now*. Penguin, 2018.
http://www.informedforlife.org/demos/FCKeditor/UserFiles/File/TRAFFICFATALITIES(1899-2005).pdf.
http://www-fars.nhtsa.dot.gov/Main/index.aspx.
https://crashstats.nhtsa.dot.gov/Api/Public/ViewPublication/812384.

[1] Both figures are taken from Pinker, Steven. *Enlightenment now*. Penguin, 2018, p. 222-225.

[2] The WHO defines it as public health issue that causes approximately 1.35 million deaths around the world each year and leave between 20 and 50 million people with non-fatal injuries. https://www.who.int/health-topics/road-safety.

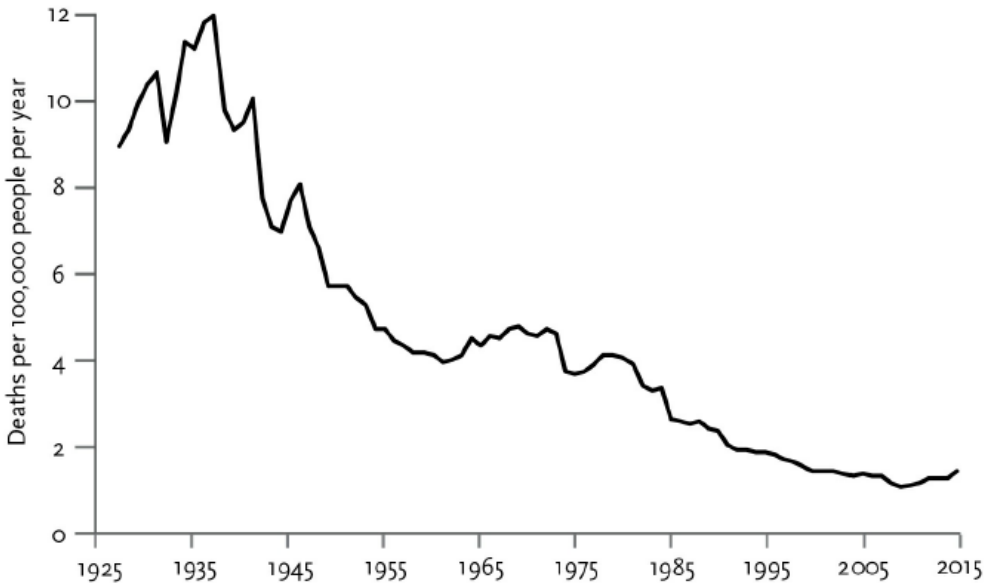

Figure 2: Pedestrian Deaths, US, 1927-2015

Sources: Pinker, Steven. *Enlightenment now*. Penguin, 2018.
For 1927–1984: Federal Highway Administration 2003.
For 1985–1995: National Center for Statistics and Analysis 1995.
For 1995–2005: National Center for Statistics and Analysis 2006.
For 2005–2014: National Center for Statistics and Analysis 2016.
For 2015: National Center for Statistics and Analysis 2017.

1.2 The Objective

Before AI-driven cars will become ubiquitous and hopefully reduce traffic accidents to near zero, this report argues that Machine Learning Algorithms are the next important step to reduce traffic accidents. Machine Learning Algorithms can analyze historical data on traffic collisions and determine what features can best predict the occurrence and severity of accidents. The immediate application is that traffic authorities can use the algorithm's output in combination with electronic traffic signs to warn motorists (and other traffic participants) about potentially dangerous traffic conditions. The future application is that such an algorithm can update AI-driven cars on traffic condition in the area (in addition to the data that the car collects and analyzes itself) to improve its driving and the safety for all traffic participants.

As a case study, this report will use the Collision Data collected by SDOT Traffic Management Division, Traffic Records Group (from 2004 to present) for the city of Seattle to build a classification model (a supervised machine learning algorithm) for predicting the severity of collisions.[3] Following best practice and to ensure reproducible results, this report uses the Cross-Industry Standard for Data Mining methodology (CRISP-DM, see Figure 3).[4] In the second chapter, this report will investigate the Collision Dataset to determine the potential features for a machine learning model by performing an exploratory analysis on the attributes. The third chapter uses the insights from the second to pre-process all relevant feature variables and the target variable putting them into final dataset ready for machine

[3] See https://data-seattlecitygis.opendata.arcgis.com/datasets/5b5c745e0f1f48e7a53acec63a0022ab_0.
[4] See https://en.wikipedia.org/wiki/Cross-industry_standard_process_for_data_mining.

learning. Chapters four and five will serve to build various type of classification machine learning models and to evaluate their performance in predicting collision severity. Given the limitations of this report, the final chapter will discuss the meaning of the results.

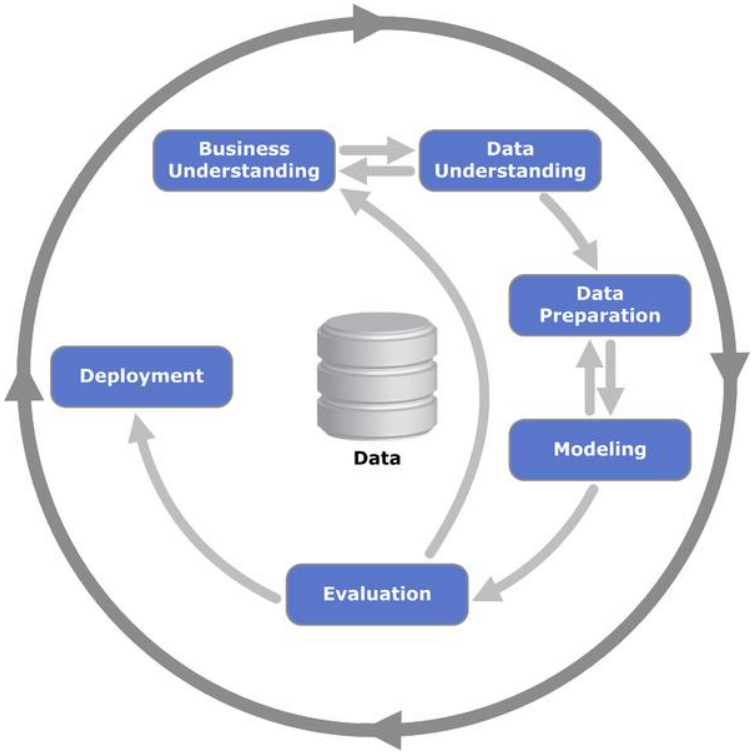

Figure 3: Cross-industry standard process for data mining, known as CRISP-DM

Sources: https://en.wikipedia.org/wiki/Cross-industry_standard_process_for_data_mining

2. Data Understanding

2.1 Understanding the Dataset

The purpose of this chapter is to gain a preliminary understanding of the Collision Dataset for the reason of selecting the label variable and potential features to predict it. In the next chapter, the selected label and feature variables will undergo pre-processing, which includes all measures of transforming and cleaning the data so that it can be read by the machine learning algorithm.

As stated before, this report uses the Collision Dataset from the city of Seattle.[5] The given dataset has 194,673 rows of observations, i.e. reported collisions, and 38 columns, which correspond to various attributes about and around the collisions. The dependent variable, also known as label, is the class that the classification algorithm should predict. In the Collision Dataset, the label is SEVERITYCODE, which can take two values:

- 1 = property damage only,

[5] The dataset was obtained here: https://s3.us.cloud-object-storage.appdomain.cloud/cf-courses-data/CognitiveClass/DP0701EN/version-2/Data-Collisions.csv.
An updated version of the dataset is available on the Seattle City website: https://data-seattlecitygis.opendata.arcgis.com/datasets/5b5c745e0f1f48e7a53acec63a0022ab_0?geometry=-123.374%2C47.452%2C-121.288%2C47.776.

- 2 = injury collision.

Of those collisions, 136,485 belong to class 1, property damage, and 58,188 belong to class 2, injury collision. An imbalance in the labels is natural in with real-world phenomena but will bias the machine learning algorithm. The issue of imbalance will be addressed in the next chapter. The information about the label can be taken either from the metadata or from the column SEVERITYDESC.[6] Thus, SEVERITYCODE.1 being a duplicate of SEVERITYCODE can be dropped along with SEVERITYDESC and COLLISIONTYPE.

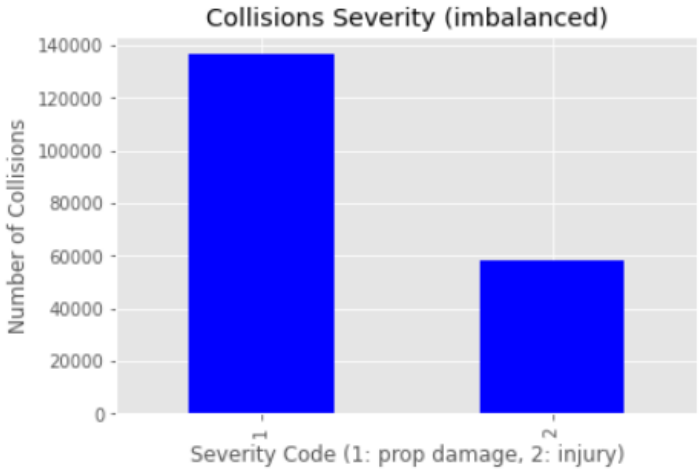

Figure 4: Collision Severity (imbalanced)

The metadata also reveals that there are several columns that contain unique identification keys for each collision, but which hold no value for a machine learning algorithm.[7] They serve the authority or organization that created them for cross-referencing purposes. Among those columns are OBJECTID, INCKEY, COLDETKEY, REPORTNO, STATUS, INTKEY, EXCEPTRSNCODE, EXCEPTRSNDESC, SDOT_COLCODE, SDOT_COLDESC, SDOTCOLNUM, ST_COLCODE, ST_COLDESC, SEGLANEKEY, and CROSSWALKKEY. As they hold no value for machine learning, they can also be dropped to narrow the feature set.

The remaining columns describe either the outcome of the collision or the conditions surrounding the accident. Among those that describe the outcome of the collision are PERSONCOUNT, PEDCOUNT, PEDCYLCOUNT, and VEHCOUNT (continuous) as well as PEDROWNOTGRNT and HITPARKEDCAR (categorical). They describe the number of various traffic participants and vehicle affected by the accident. Although continuous variables are generally a good source of information for supervised machine learning algorithm, the causality (cause-effect direction) is important. Especially, PERSONCOUNT or VEHCOUNT would make excellent target variables for prediction in a different

[6] The metadata was obtained here: https://www.seattle.gov/Documents/Departments/SDOT/GIS/Collisions_OD.pdf.
[7] Ibidem.

type of supervised machine learning algorithm, namely regression. They are not truly independent of the collision or its severity but rather the result of it. Therefore, these columns are also be dropped.

2.2 Feature Selection

Now, the rest of the features only describe the conditions before or surrounding the collision. The variables given as X and Y refer to longitude and latitude respectively (and are renamed as such), and they also make a written address in the variables LOCATION and ADDRTYPE redundant. The categorical variables JUNCTIONTYPE, WEATHER, ROADCOND, and LIGHTCOND describe the environmental factors and can be considered independent variables. They will be kept as features whose variation may explain the variation in collision severity. Similarly, the variation in the INCDATE and INCDTTM may help to explain the variation in the collision severity. Likewise, the variation in human behavior as observed by the columns INATTENTIONIND, UNDERINFL, and SPEEDING may help explain the variation in collision severity. As they precede the accident, they can also be considered independent variables and kept as features.

Table 1: Pre-Selected Features

	Variable	**Description**
1	SEVERITYCODE	code that corresponds to severity of the collision: • 1 = property damage • 2 = injury
2	LONGITUDE	longitude
3	LATITUDE	latitude
4	JUNCTIONTYPE	category of junction at which collision took place
5	WEATHER	description of the weather conditions during the collision
6	ROADCOND	condition of the road during the collision
7	LIGHTCOND	light conditions during the collision
8	INCDATE	date of the incident
9	INDTTME	date and time of the incident
10	INATTENTIONIND	whether or not collision was due to inattention
11	UNDERINFL	whether or not driver was involved under the influence of drugs or alcohol
12	SPEEDING	whether or not the speeding was a factor in the collision

3. Data Preparation

The ultimate purpose of this chapter is to ready the pre-selected variable set for machine learning. Data preparation, also called pre-processing, includes steps such as exploratory data analysis (EDA) to confirm the selection of features, dealing with missing values, and converting variables into machine-legible data types. These steps will be applied to the environmental variables, human behavior variables and date-time variables. At the end, this chapter will also address the issue of label imbalance and employ a resampling method for balancing.

3.1 Environmental Variables

LONGITUDE and LATITUDE are continuous, numerical variables that together pinpoint the location of each recorded collision. Their minimum and maximum value map a square-shaped area of Seattle in which all collisions have been recorded. LONGITUDE and LATITUDE have 5,334 missing values. Common strategies of dealing with missing values in continuous variables are by replacing them either with their median or their mean. In this case, the median would just the point to the center of the square-shaped map of Seattle. But the mean would generally point to the area were collisions happen most frequently (see Figure 5). LONGITUDE has a mean of approximately -122.33052 and LATITUDE has a mean of approximately 47.61954. Their missing values will be replaced with their mean values respectively.

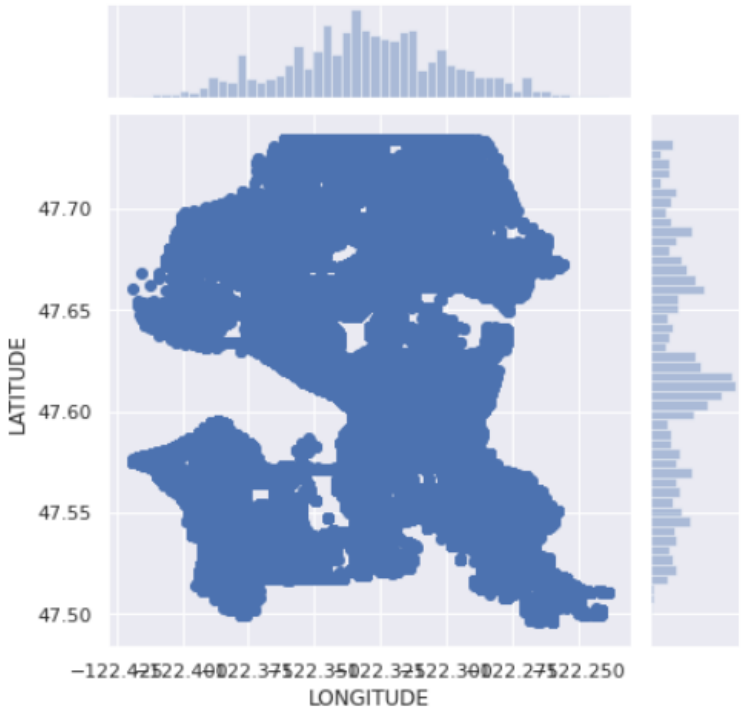

Figure 5: Location and Distribution of Collisions (by LONGITUDE and LATITUDE)

Next, JUNCTIONTYPE, WEATHER, ROADCOND, and LIGHTCOND are all categorical variables that describe the place and surrounding conditions of the accident. A common strategy of replacing missing values is to replace with the mode, the most frequent value.

For JUNCTIONTYPE, the mode is “Mid-Block (not related to intersection)” and it replaces 6,329 missing variables. “Unknown” values are dropped because their weight is insignificant. Afterwards, the different categories are converted into dummy variables for the machine learning algorithm, and the original variable JUNCTIONTYPE is dropped.

For WEATHER, the mode “Clear” will replace all 5,081 missing values. Since there is no indication what “Other” weather is, it is merged with the category “Unknown”. “Unknown” is renamed as “Unknown Weather” to distinguish it from the variable “Unknown Roadcond” to be created in the

next step. Additionally, as “Partly Cloudy” is insignificant to it is merged with “Overcast”. WEATHER is also dropped in place for its dummies.

“Dry” is the mode of ROADCOND and will replace all the 5,012 missing values. Again, “Other” is merged with “Unknown”. Moreover, “Standing Water” is added to “Wet” and “Snow/Slush” is added to “Ice” as the former in those pairs are very similar to the latter. “Unknown” is renamed “Unknown Roadcond”. It should not be dropped as it is the third largest category.

Lastly, LIGHTCOND is encoded as a single dummy, where 0 stands for “Daylight” and 1 for all other conditions. 5,170 missing values are also replaced with “Daylight,” the most frequent category.

3.2 Human Behavior Variables

The variables INATTENTIONIND, UNDERINFL, and SPEEDING describe human behaviors that are assumed to increase the chances of an accident.

For INATTENTIONIND, 29,805 observations are given as “Y”, which means the missing 164,868 will assumed to be “N” and filled in accordingly. They are encoded as 1 and 0 respectively. It is also very counter-intuitive as this suggests that most accidents happen even though people are paying attention to traffic. Since attention span may be highly depended on the time of the day, the variable INCDTTM is dropped.

The column UNDERINFL has four values (“Y”, “N”, 1, and 0). Since no further indication is given in the metadata, it is assumed that “Y” equates to 1 and “N” equates to 0. Here a highly counter-intuitive picture emerges, most accidents happen when people are not under the influence of drugs or alcohol (see Figure 6).

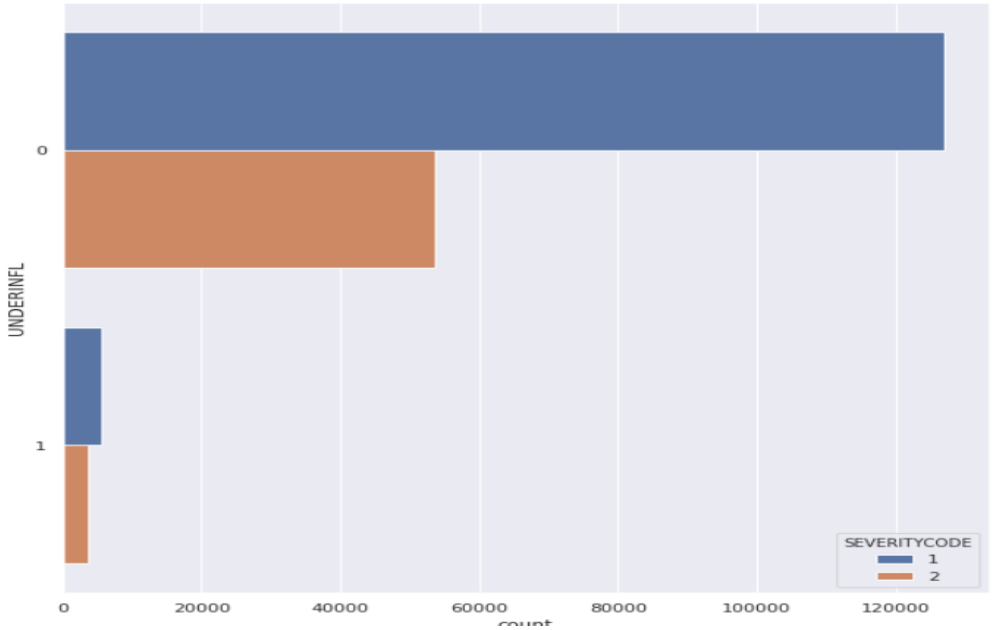

Figure 6: Influence of Drugs or Alcohol on the Frequency of Collisions

For SPEEDING, 9,333 observations are given as “Y”, which means the missing 185,340 will assumed to be “N”. They are also encoded as 1 and 0. Counter-intuitively, most collisions occurred without speeding.

3.3 Date-Time Variables

The assumption about the date of the collision is closely related to those about the human behavior variables, i.e. during certain days of the week, people are less attentive, for instance, due to exhaustion from work. A cursory analysis confirms that most collisions happen on Fridays while the fewest happen on Sunday (see Figure 7). The variable INCDATE is converted into dummy variable, where 1 stands for weekend and 0 weekdays.

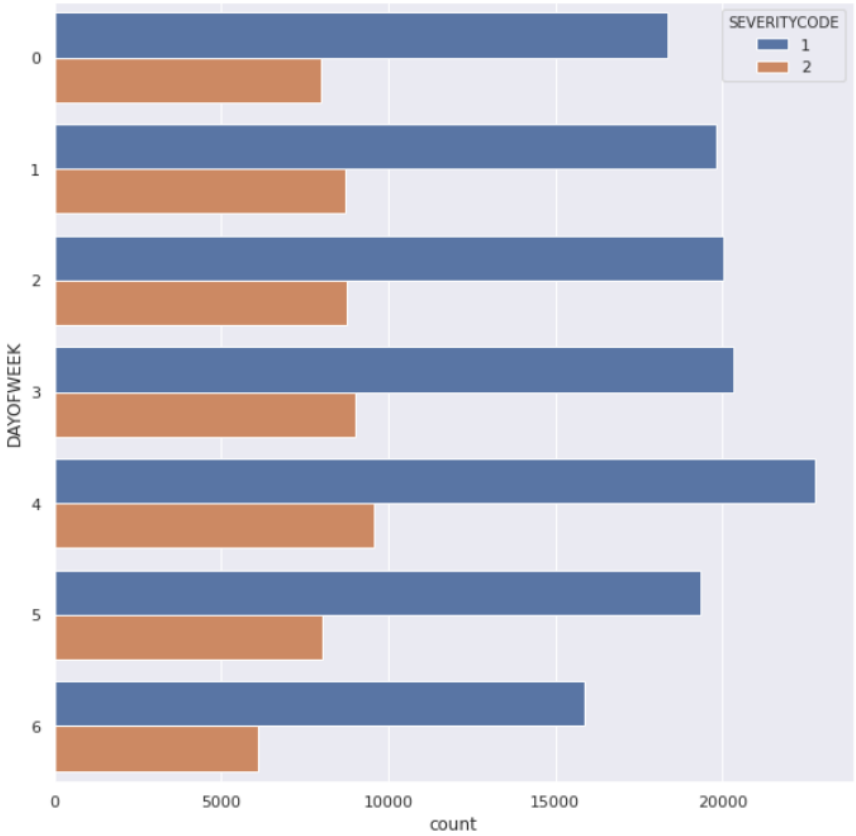

Figure 7: Collision Frequency by Day of the Week

(0 = Monday, 1 = Tuesday, 2 = Wednesday, 3 = Thursday, 4 = Friday, 5 = Saturday, 6 = Sunday)

3.4 Balancing the Dataset

In this dataset, 58,188 injury collision (class 2) and 136,485 property-damage-only collision (class 1) are recorded. Many real-world classification problems have an imbalanced class distribution such as fraud detection, churn prediction, or extreme event prediction. Slight imbalances between classes are considered to have ratio around 4:6, while severe imbalances are considered to have a ratio of 1:100 or more. Thus, the Collision Dataset with an approximate ratio of 58:137 has a severe imbalance. The abundant class, in this case property damage collision (class), is called majority class, whereas the one with fewer samples is called minority class. By definition, a minority class is difficult to predict because of its few examples. This means it is more challenging for a machine learning model

to learn the characteristics of examples from this class and to distinguish it from the majority class. In fact, most classification algorithms are designed and demonstrated on problems that assume an equal distribution of classes.

The two common strategies of rectifying imbalanced classes in machine learning 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 und 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 employs the method random under-sampling to rectify the imbalance and thereby reducing class 1 collisions to 58,188 samples.

3.5 Expected Outcomes and Hypotheses

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 28 features for analysis, 26 of which are dummy 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.

4. 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 Neighbors,
- Decision Tree,
- Random Forrest,
- Logistic Regression,
- Artificial Neural Networks.

For modelling and evaluation, 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 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 on Seattle's streets.

4.1 K-Nearest Neighbors (KNN)

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).[8] 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 neighbors for comparison. A too small k-value may capture too much noise, while increasing it endlessly will run into diminishing marginal gains for mean accuracy. Given the large feature set of the training data, this challenged is approached by iterating through integers 15 to 24 and compare them against their respective mean accuracy (see Figure 8). The best mean accuracy is achieved at approximately 0.6060 with k equaling 23. A full evaluation of model performances will be given in the next chapter.

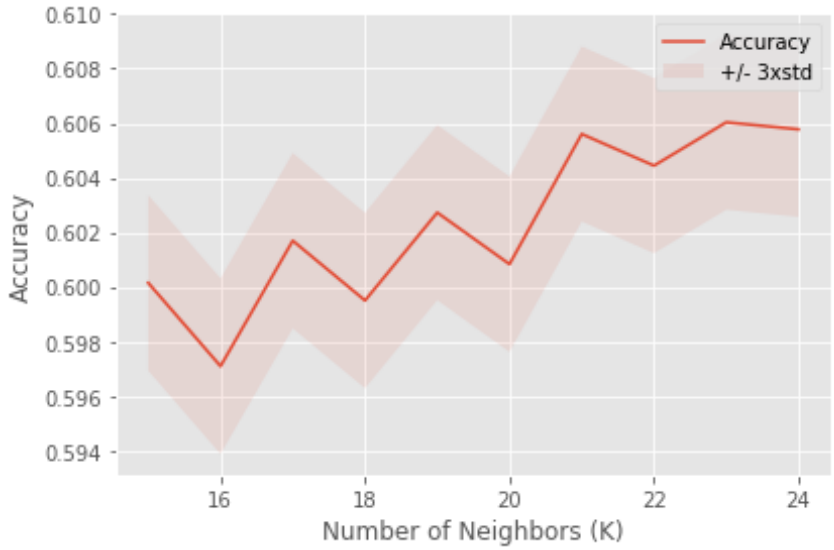

Figure 8: K-Values with Corresponding General Accuracy

4.2 Decision Tree

Decision Tree classifier is another of 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, leaves represent class labels and branches represent conjunctions of features that lead to those class labels. The criterion by which branches eventually lead to pure leaves, where 100% of its cases fall into a single class is called 'entropy'.[9] 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 Decision Tree that has been given a maximum depth of 4 (Figure 9).

Without a maximum depth, the Decision Tree algorithm will run until all leaves have become are pure. This has led to a general accuracy score of 0.5719 on the given feature set for predicting severity classes of collisions. The advantages of Decision Trees are that they work well with categorical

[8] See https://scikit-learn.org/stable/modules/neighbors.html#classification; and https://www.saedsayad.com/k_nearest_neighbors.htm.
[9] See https://scikit-learn.org/stable/modules/generated/sklearn.tree.DecisionTreeClassifier.html.

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.

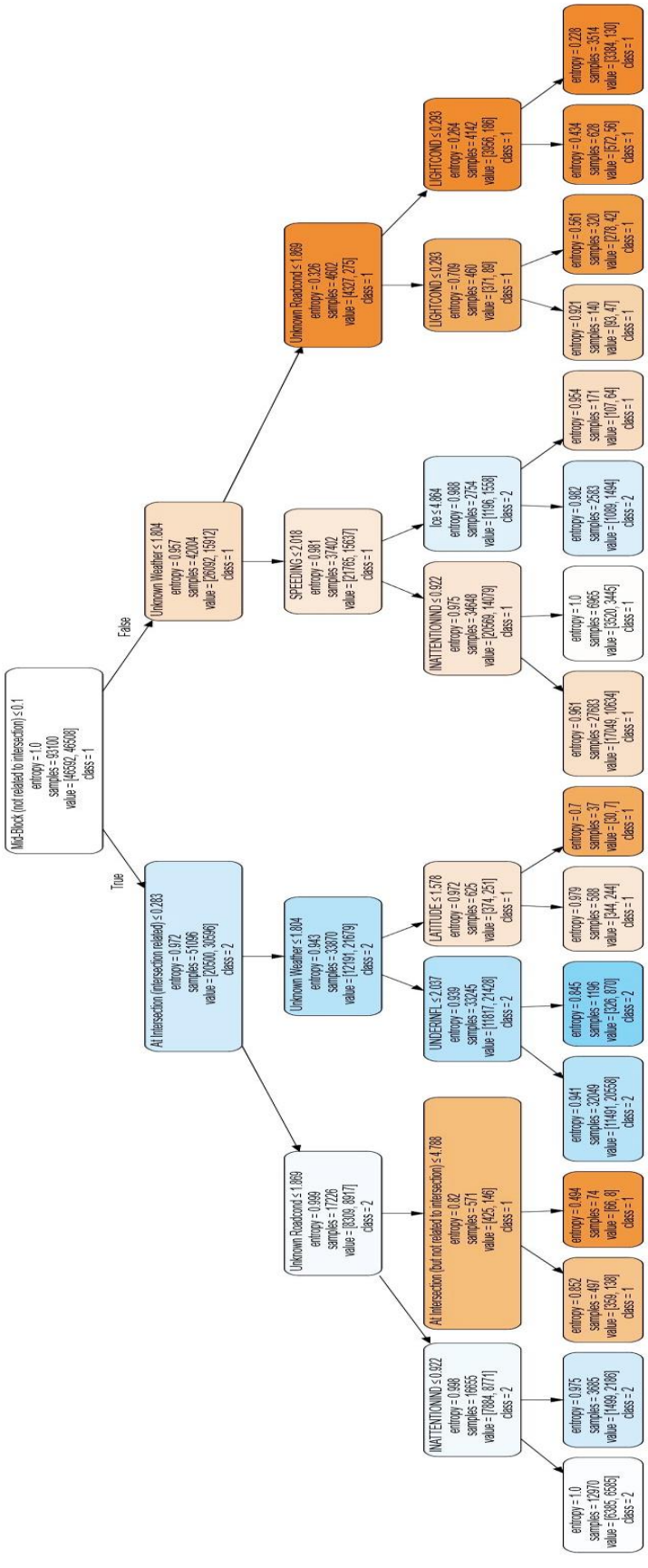

Figure 9: Decision Tree Diagram (maximum depth = 4)

Key: If the condition on top of the node is true the tree branches to the left, if false to the right.
The ultimate goal of decision tree is to iterate through all features and find the condition that lead to pure leaves, where entropy = 0.

4.3 Random Forest

The Random Forrest improves on the advantages and disadvantages of Decision Trees. It is called Random Forest because it operates by constructing a multitude of Decision Trees on various, random sub-samples of the dataset and then outputs the class that is the mode of the classes.[10] 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 0.5870, 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.

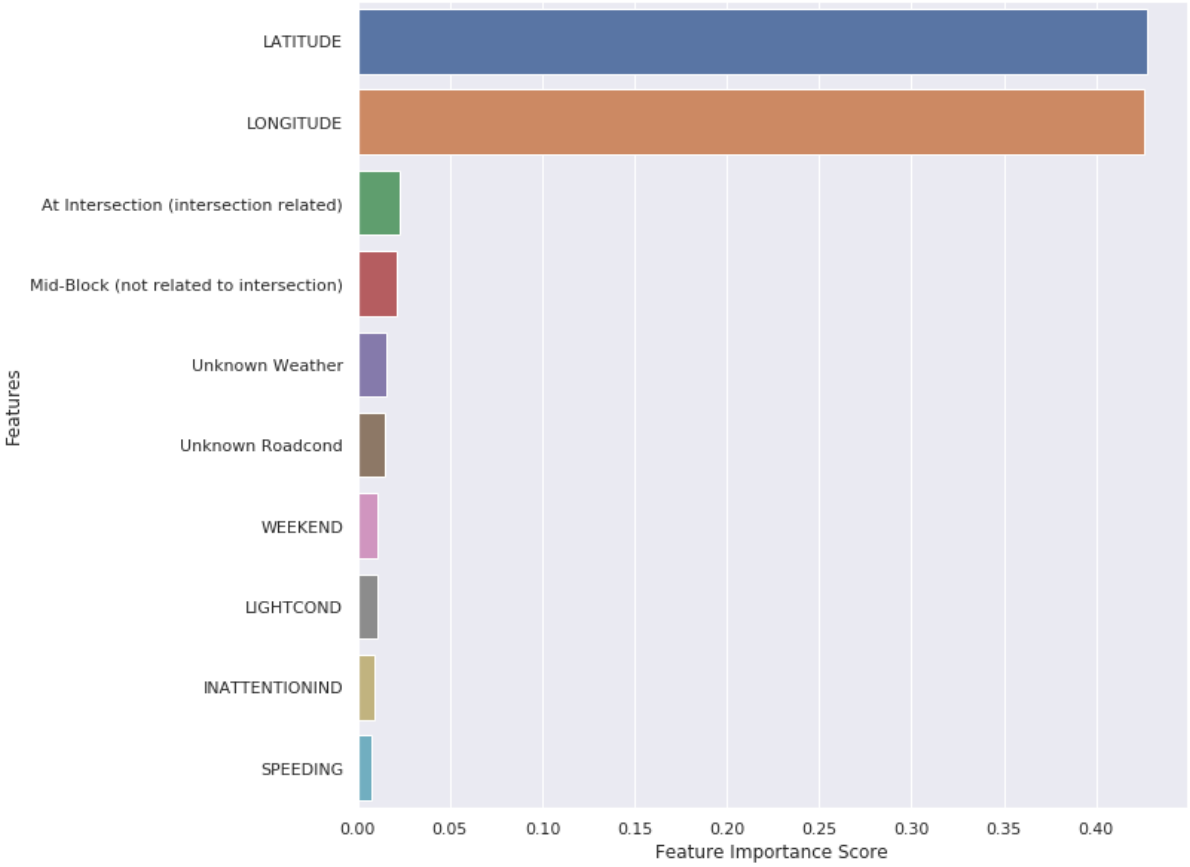

Figure 10: Random Forest's Feature Importance Scoring

Another crucial ability of Random Forest is 'feature importance scoring' (see Figure 10).[11] The higher a feature is scored, the more important the feature is in the total reduction of entropy and thus finding pure leaves (the total score must equal 1). By far, the two most important features in predicting collision severity on Seattle's streets are the continuous variables LATITUDE and LONGUTIDE. At much lower importance, they are followed by whether the location of a collision happens either 'At Interaction (Intersection related)' or happens 'Mid-Block (not related to intersection)'. On place four

[10] See https://scikit-learn.org/stable/modules/generated/sklearn.ensemble.RandomForestClassifier.html.
[11] Ibid.

and five of importance scoring are ‘Unknown Roadcond’ and ‘Unknown Weather’ (features that cannot be translated into actions for improving traffic safety). Other features such as WEEKEND, LIGHTCOND, INATTENTION, SPEEDING, and others that could be turned into actionable insights for improving traffic safety, have almost insignificant importance score in predicting collision severity. This confirms the expectations about the large size of the feature set and the large number of dummy variables as postulated in section 3.5. The implications of this ranking in feature importance will also be discussed in the next chapters.

4.4 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’.[12]

Liblinear is a library for large linear classification that support logistic regression and linear support vector machines.[13] 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 dataset, 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.[14] The ‘SAGA’ solver is a variant of SAG that also supports the non-smooth error minimization. It is recommended for multi-class classification.[15] 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 around 0.62. As the Liblinear solver seems the most appropriate for the type of dataset that is given here, it will be used for

[12] See https://scikit-learn.org/stable/modules/generated/sklearn.linear_model.LogisticRegression.html.
[13] Fan, Rong-En, Kai-Wei Chang, Cho-Jui Hsieh, Xiang-Rui Wang, and Chih-Jen Lin. "LIBLINEAR: A library for large linear classification." *Journal of Machine Learning Research* 9, no. Aug (2008): 1871-1874.
[14] Schmidt, Mark, Nicolas Le Roux, and Francis Bach. "Minimizing finite sums with the stochastic average gradient." *Mathematical Programming* 162, no. 1-2 (2017): 83-112.
[15] Defazio, Aaron, Francis Bach, and Simon Lacoste-Julien. "SAGA: A fast incremental gradient method with support for non-strongly convex composite objectives." In *Advances in neural information processing systems* (2014), pp. 1646-1654.

the full evaluation later. Additionally, since Logistic Regression can also estimate probability of class, it will also be included in the full evaluation.

4.5 Artificial Neural Networks

An Artificial Neural Network 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 Neural Networks (NNs), 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. To transfer their signal outputs to other neurons, neurons have connections named edges and neurons are aggregated into layers. Signals can traverse from the first layer to the last layer multiple times and thus also recreate the feedback mechanism of biological neurons.

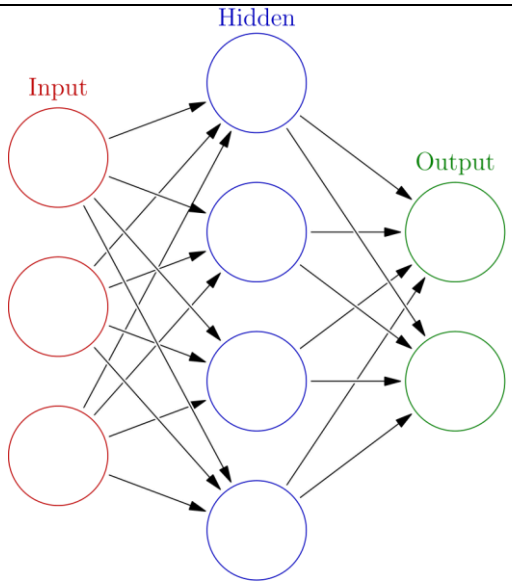

Figure 11: Artificial Neural Network Structure

Source: https://commons.wikimedia.org/w/index.php?curid=24913461.

An NN classifier trains itself by iterating through the original input data and comparing the final output (i.e. the prediction) with a given target label.[16] As for any other ML algorithm, the difference between predicted and true label is the error. With successive iterations through the data, the NN adjusts the functions of neurons and layers, which will cause the NN 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, NN can also estimate the probability of a case falling into a dependent class. The general accuracy score attained by artificial NN is 0.6243.

[16] Glorot, Xavier, and Yoshua Bengio. "Understanding the difficulty of training deep feedforward neural networks." In *Proceedings of the 13th International Conference on Artificial Intelligence and Statistics* (2010), pp. 249-256.

5. Evaluation

5.1 Formal Evaluation Metrics

Table 2: Classification Algorithms and Accuracy Scores

	Gen. Accuracy	Jaccard-Score	F1-score	Log-Loss
K-Nearest Neighbor	0.606032	0.414656	0.605196	NaN
Decision Tree	0.571920	0.411111	0.571595	NaN
Random Forest	0.587085	0.410946	0.587052	NaN
Logistic Regression	0.619995	0.440933	0.619862	0.643860
Neural Network	0.624377	0.447345	0.624300	0.640896

Table 3: Variation in Accuracy Scores

	Gen. Accuracy	Jaccard-Score	F1-score	Log-Loss
mean	0.602	0.425	0.602	0.642
std	0.022	0.018	0.022	0.002

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 2 and 3 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 60.2% and they all fall within one standard deviation of 2.2%. Except for the Decision Tree, which is within 2 standard deviations below the mean, 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 42.5% and fall within one standard deviation of 1.8%, except for the NN. The NN is within 2 standard deviations above the mean, showing much better 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. 3 out of 5 classification models fall within one standard deviation of the mean of 0.602. Again, the decision tree performed worse (within 2 standard deviations below the mean), while the artificial NN performed best (within 2 standard deviations above the mean).

Logistic Regressions and artificial NNs 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 NN model did slightly better than the Logistic Regression model by a difference of 0.002 in terms of Log-Loss.

5.2 Best Classifier and Best Determinants of Collision Severity?

In conclusion, the artificial NN performs best in all evaluation categories. However, it is not a winner by a magnitude only by a margin. 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. This confirms the negative rather than the positive apprehensions formulated in section 3.5 about the lack of numerical features and abundance of dummy ones. To recall, the classification model was hypothesized to be a function where negligent human behavior and adverse driving conditions increase the severity of collisions. The abundance of dummy features has led to only moderate accuracy for all models. Moreover, the feature importance scoring has demonstrated that most conditions -that would make good policy-levers for improving traffic safety- had almost no significance in determining the classification of severity (see Figure 10). Ultimately, the top three best models for predicting collision severity on the streets of Seattle have been: been:

1. Artificial Neural Networks
2. Logistic Regression
3. K-Nearest Neighbors

6. Conclusion

6.1 Summary of Findings

This section summarizes the findings on determinants of collision severity obtained from both the exploratory data analysis (EDA) and the data mining section of this report. It was initially hypothesized that collision severity is a function of adverse driving conditions and human behavior. Already the EDA delivered counterintuitive answers to this hypothesis. Figures 12a, b, and c demonstrate that the top categories in ROADCOND, WEATHER, and LIGHTCOND show that the best driving conditions, namely dry streets, clear weather, and daylight, are also the ones with the most

frequent accidents in both classes. The second most frequent amount of accidents happen in categories that only present a slight deterioration in those conditions: wet streets, rain, and dark but with streetlight on. Much more adverse driving conditions are listed but the amount of accidents they list are is significantly less compared to the top-ranking categories. This poses the question whether good driving conditions make traffic participants more negligent than they should be? However, the observations in INATTENTIONIND, UNDERINFL, and SPEEDING compound the problem even further. Negligent or reckless human behavior is also not a driver of the amount and severity of accidents as Figures 6 and 13 a and b show.

There may be two answers to this conundrum. First, it can be assumed that, when possible, traffic participant such as drivers and pedestrians avoid adverse weather conditions and most of them commute to work during daylight hours. In the same vein, it can be assumed that most drivers will try to avoid speeding or driving under the influence (regardless whether for their own safety or because of fear of being caught and punished). The second answer concerns the severity of collisions. As collision class 1 dominates, it can be assumed that most safety measures in cars and on roads have already reduced the severity of collisions from death or injury down to property damage only.

Finally, the classification models employed have not delivered an entirely satisfactory answer as to what combination of factors will increase the severity class of collisions. The feature importance scoring has illustrated that combination of location (LATITUDE and LONGITUDE), JUNCTIONTYPE: At Intersection (intersection related) or Mid-Block (not intersection related), unknown weather and unknown road conditions, as well as bad light conditions, inattention, and speeding may increase the severity of collisions.

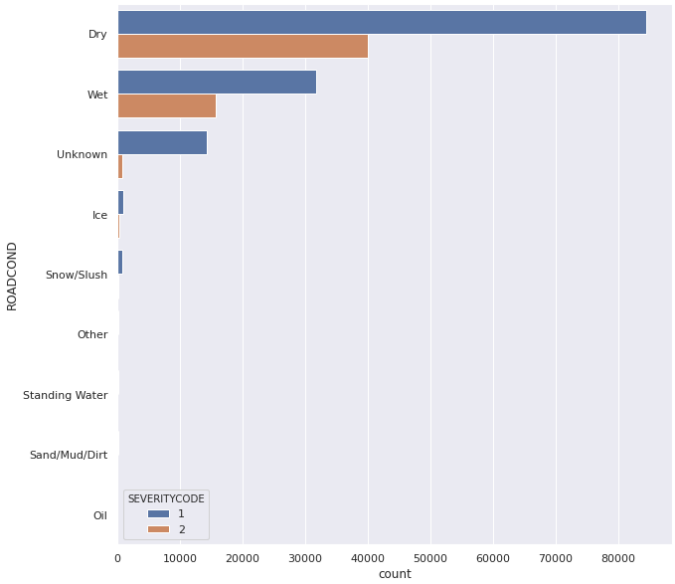

Figure 12a: ROADCOND Frequency and Severity of Collision

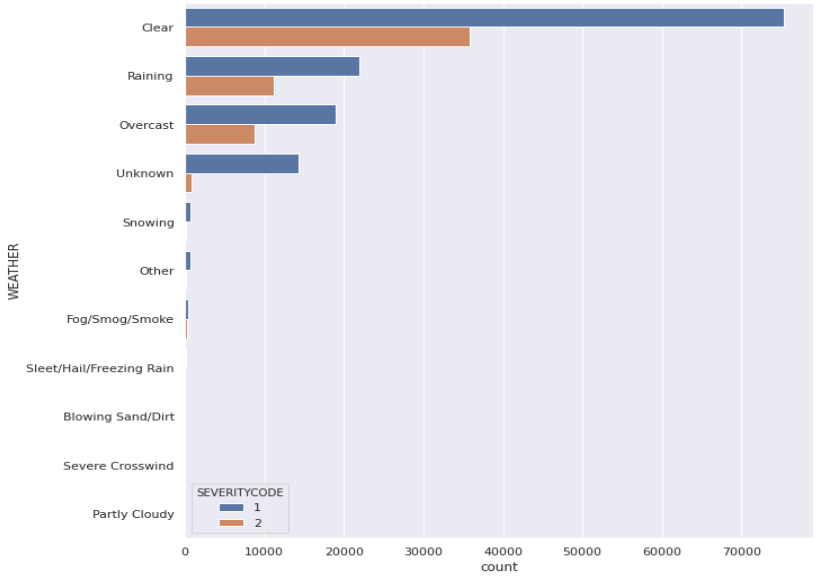

Figure 12b: WEATHER Frequency and Severity of Collision

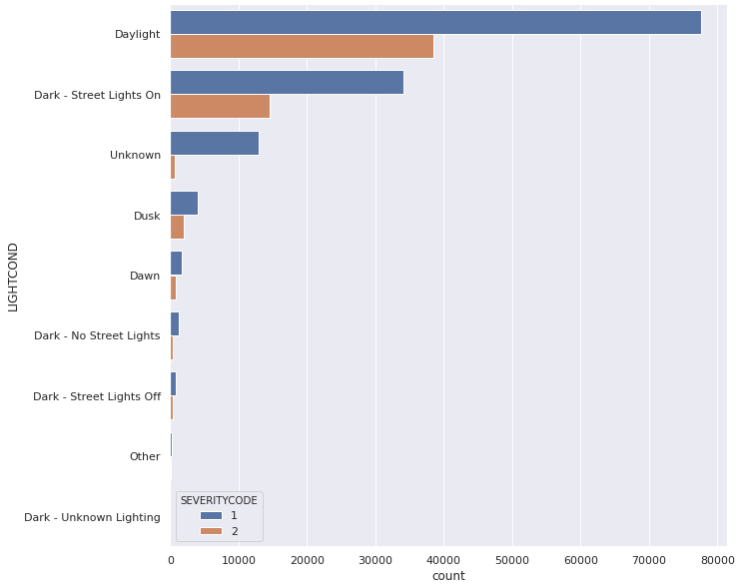

Figure 12c: LIGHTCOND Frequency and Collision Severity

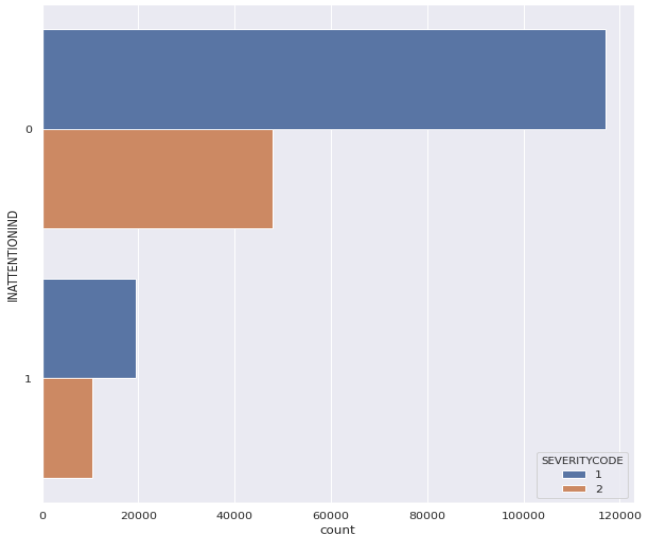

Figure 13a: INATTENTIONIND Frequency and Collision Severity

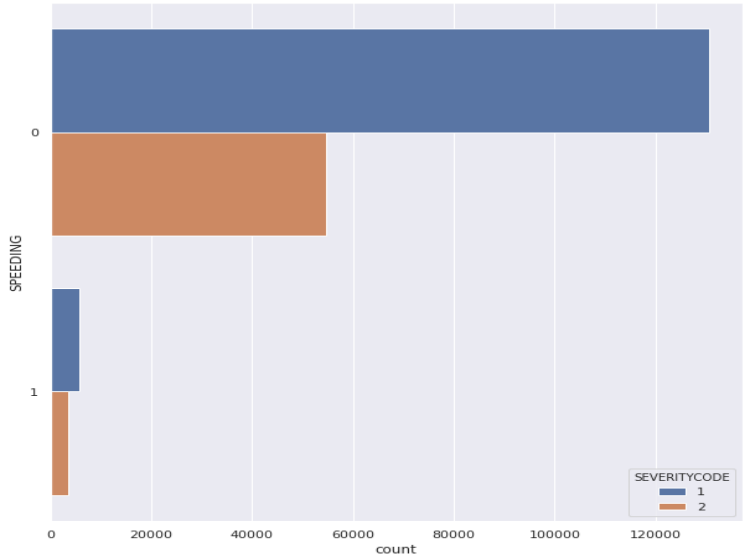

Figure 13b: SPEEDING Frequency and Collision Severity

6.2 Future Research and Deployment

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 dummies. 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 repost was not entirely futile as it revealed that Seattle has generally safe drivers and traffic participants. 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.