IBM Data Science Capstone Project
Car Accident Severity

Chantal Beatrice Bamigbele, PhD

9th November 2020

Abstract

Globally, road traffic accidents are an important public health concern which needs to be tackled. The City of Seattle is not an exception. The number of road traffic accidents has increased every year and the damages whether on proprieties such as cars, infrastructures or on humans injury, death are quite alarming. The aim of the present study is to develop a model to predict the severity risks of road accidents in the City of Seattle. To do so a dataset consisting of 194673 recorded accidents was analysed and machine learning models were built by applying 4 different algorithms in a binary classification setting (SEVERITYCODE 1 and SEVERITYCODE 2). The effect of parameters such as weather, road and light conditions and the type of road (ADDRTYPES) on accident severity risks was identified. This model will guide road planners in their future road construction, drivers as well as pedestrians in their behaviour on the road in a particular condition.

1. Background

1.1 Introduction/Business Problem

I have been hired by the city council of Seattle to build a model that will enable them to predict the severity of road accidents in order to protect lives, properties and infrastructures damages. In fact, the number of road traffic accidents increases every year and the damages whether on proprieties such as cars, infrastructures or on humans injury, death are quite alarming. The pressure is now on the council as the communities accuse the council not doing much to alleviate the problem, the budget allocates every year for road infrastructures repairs increases and always almost runs on deficit. Not to talk about the cost of other damages and the loss of a loved one. Therefore, I have the task to build a reliable and appropriate model to predict accidents severity and allow people to take informed decision on their driving.

The question is to build a model that will predict the severity of road accidents Therefore what are the conditions that affect the severity of road accidents? Are the data already available within the council or am I going to fetch them somewhere else? These are some of the issues at hands before talking about any modelling.

While discussing the issues with the council as I tried to understand the problem, I was given access to their database where is stored years of the Seattle Collision dataset in a csv format.

Next, I have to define measurable conditions that could affect the severity of the road accidents.

In the literature, it has been reported that many studies aim at predicting the severity of an accident using various information from the accident in order to understand what causes an accident to be fatal. Specifically, Chong et al. [1] identified that the seat belt usage, the light conditions and the alcohol usage of the driver are the most important features whereas Abellan´ et al. [2] found that the type and cause of the accident, the light condition, the sex of the driver and the weather were the most important features. Using BoxPlots, a way of statistically representing the distribution of the data through five main dimensions in addition to the aforementioned information, I identified the weather conditions, road conditions and light conditions as important measurable attributes to weigh the road accidents severity. Interestingly, less than 3% data of these attributes were missing.

Furthermore, it has also been reported that road accidents prediction are seen whether as (1) a regression problem, predicting the risk of accidents translated into different ways, or (2) a binary classification problem, predicting whether an accident will occur. I choose to approach it as a classification problem because the dataset only contains 2 severity codes implying a classification issue. Nevertheless I will be using different algorithm models in order to deliver a more accurate solution.

1.2 Interest

The main target audiences for this work will be road/city planners, drivers and pedestrians. Identifying and understanding the key factors influencing the severity of accidents, it may be possible for local authorities to alleviate the severity of accidents.

2. Data Understanding

Regarding the data, it was extracted from the Seattle City Council database as I mentioned before in a csv format. A close look at it showed that it needed a thorough data cleaning. In fact, there were many columns I didn't need for the machine learning model including empty columns. I got rid of missing data as well.

Only 2 severity codes (1, 2) were available with SEVERITYCODE 1 cases been almost 2 times SEVERITYCODE 2. This was not surprising as vehicle collision prediction has always suffered from data imbalance issue. And, knowing that when dealing with severe data imbalance, most machine learning algorithms do not perform well. I downsampled SEVERITYCODE 1 to match the severity code 2 samples size.

Moreover, most of the features were of type object, unfortunately, Sklearn Decision Trees which I was planning to use, do not handle categorical variables. I then converted these features to numerical values using pandas.get_dummies(). Since I had more than two categories, I might create a new problem in the way. As I kept assigning different integers to different categories, it might create a confusion. This strategy might as well defeat its own purpose. So instead of having one column with n number of categories, I used n number of columns with only 1s and 0s to represent whether the category occurred or not. To accomplish this task, I imported OneHotEncoder library.

3. Data Exploration

I then visualised these categorical variables by using Boxplots to determine the relationship between the different attributes and the severity code (Fig 1A, 1B, 1C, 1D). Furthermore, I performed group with multiple variables and looked at the correlation of these variables with the severity code wandering about the dependency of the severity code on these variables (Fig 2). I narrowed them down to the following variables: 'Raining', 'Clear', 'Overcast', 'Snowing', 'Dry', 'Ice', 'Snow/Slush', 'Wet', 'Dark - Street Lights Off', 'Dark - No Street Lights', 'Dawn', 'Daylight', 'Dusk', 'Alley', ‘Block' and ‘Intersection’ in order to feed the model with variables that meaningfully affect our target variable. Noteworthy, the correlation between the above variables and SEVERITYCODE was statistically significant, however there was almost no linear relationship (close to 0) meaning that they can't affect the SEVERITYCODE, probably in combination.

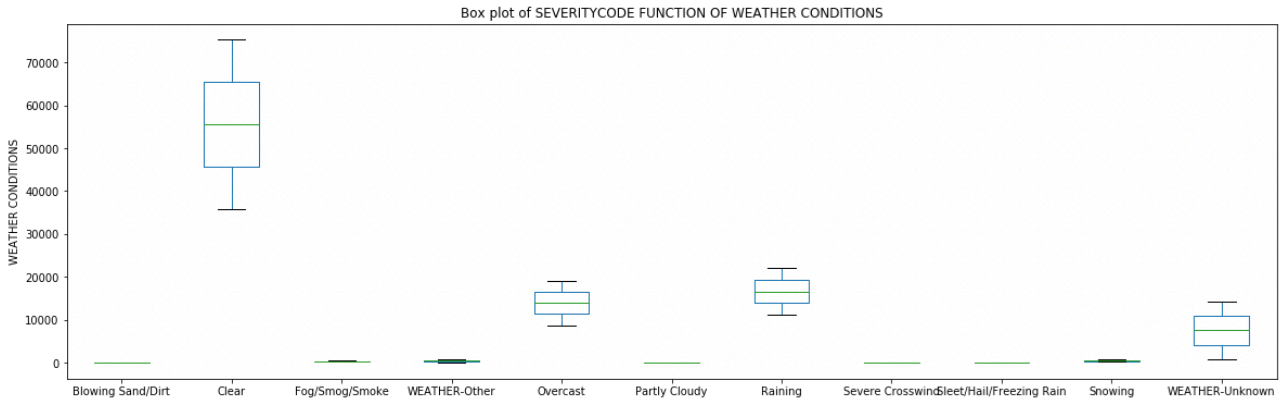

Fig 1A. Box plot of Severity code based on weather conditions

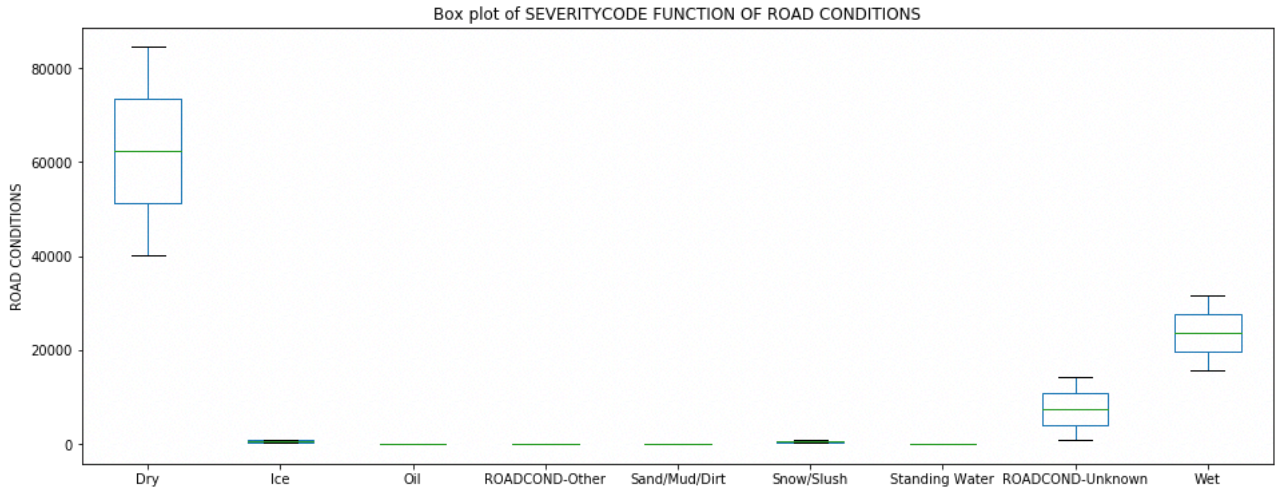

Fig 1B. Box plot of Severity code based on road conditions

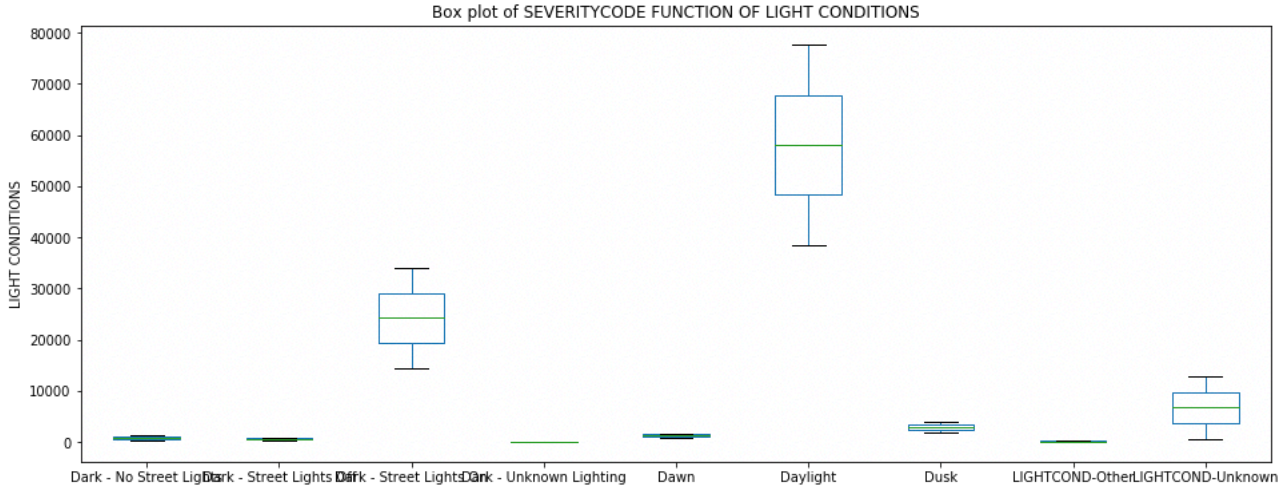

Fig 1C. Box plot of Severity code based on light conditions

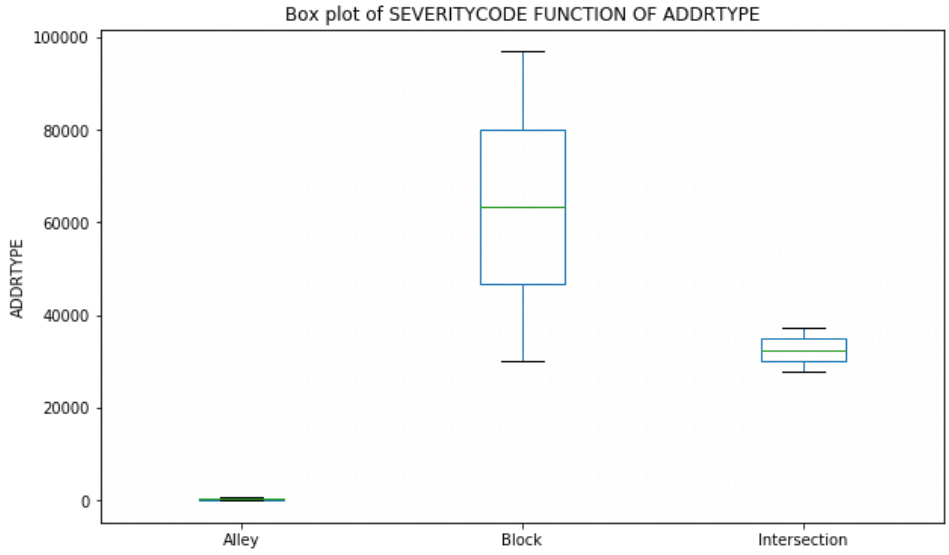

Fig 1D. Box plot of Severity code based on road types

The correlation matrix corresponding to the above variables is shown in Fig 2. I could see some of the correlations defined earlier such as clear weather, dry and wet road and block and intersection.

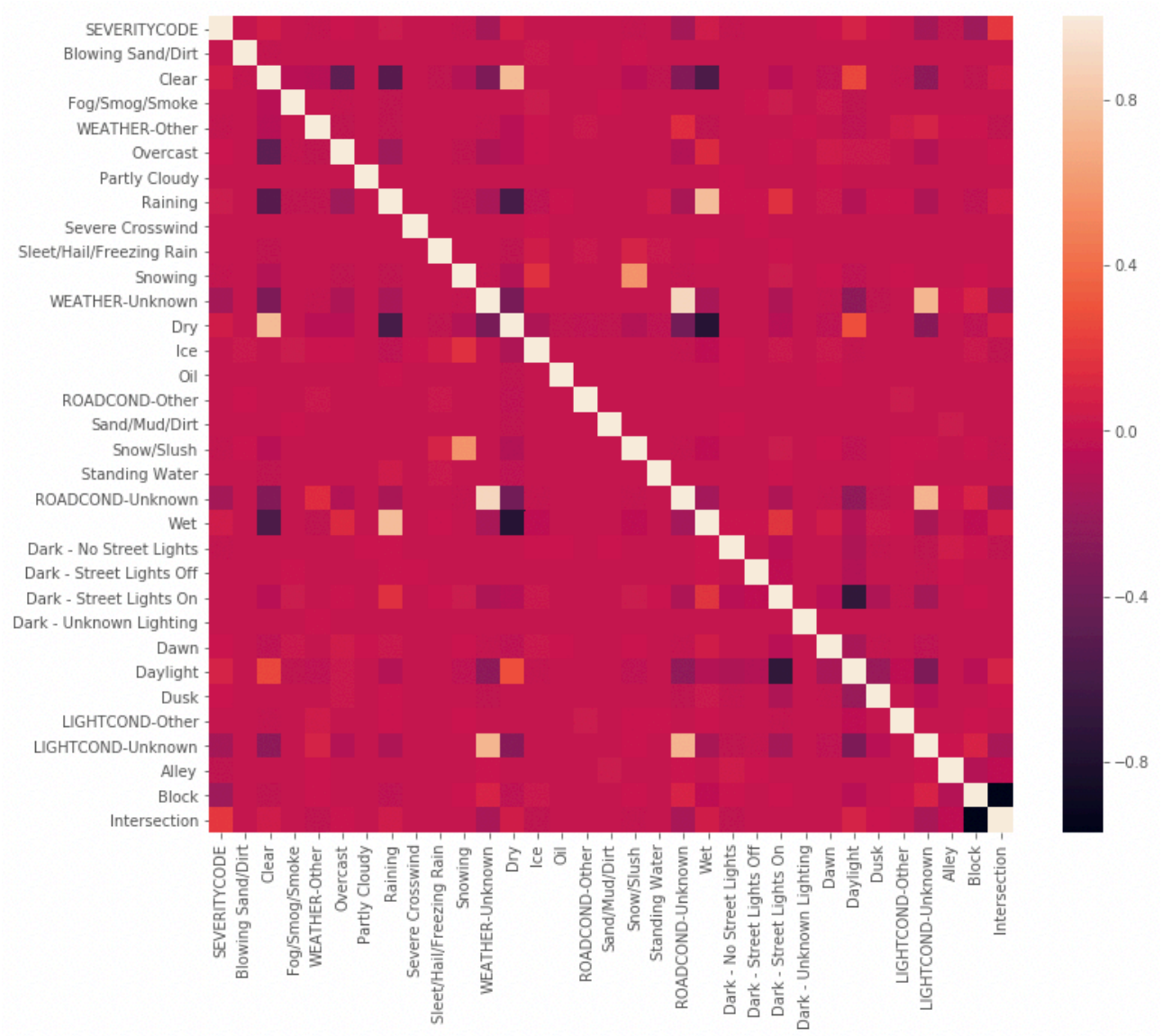

Fig 2. Correlation matrix of selected variables with severity code.

4. Predictive modelling

Finally, I built a supervised machine learning algorithm that will accurately predict road accidents severity. As I mentioned before, road accidents prediction is defined as a classification problem or a regression problem. Most of the studies that performed classification only reported the accuracy metric which is not well suited for problems with data imbalance such as road accident prediction [3]. The studies that performed regression used different definitions for the risk of accidents, which makes comparisons difficult. Therefore, in order to build an accurate model, I built several prediction models using various machine learning algorithms such as K Nearest Neighbour (KNN), Decision Tree, Support Vector Machine and Logistic Regression.

K-Nearest Neighbour (KNN) which will help me predict the severity code of an outcome by finding the most similar to data point within k distance.

Decision Tree which will give me a layout of all possible outcomes to fully analyse the consequences of a decision. It context, the decision tree observes all possible outcomes of different weather, road and light conditions.

Support Vector Machine which is used for classification, regression and outliers detection. Logistic Regression which is perfect for binary data such as the present dataset containing only two severity codes.

The balanced dataset were then split in to training and testing subsets using the train_test_split function from scikit learn. The parameter test_size was set to 0.3, meaning that 70% of the balanced data were used for training the model and 30% of the data were reserved for testing.

4.1 K-Nearest Neighbour model

KNN is a pattern-recognition algorithm which maps an input dataset to a multi-dimensional hyperspace and then attempts to classify a data point of unknown classification based on the classifications of is k-nearest neighbours in this hyperspace. The optimum choice of k is highly dependent on the dataset in question, and in practice it is usually necessary to train and test kNN models using a range of k, measuring the accuracy of each. k ~kNN model was built for k = 1–50 using the kneighborsclassifier in scikit learn. The resulting model accuracy as a function of k is shown in Fig 3. The best accuracy was 0.59 with k= 48. Accuracy is the most intuitive performance measure, and defined as the ratio of the number of correctly classified objects to the total number of objects evaluated.

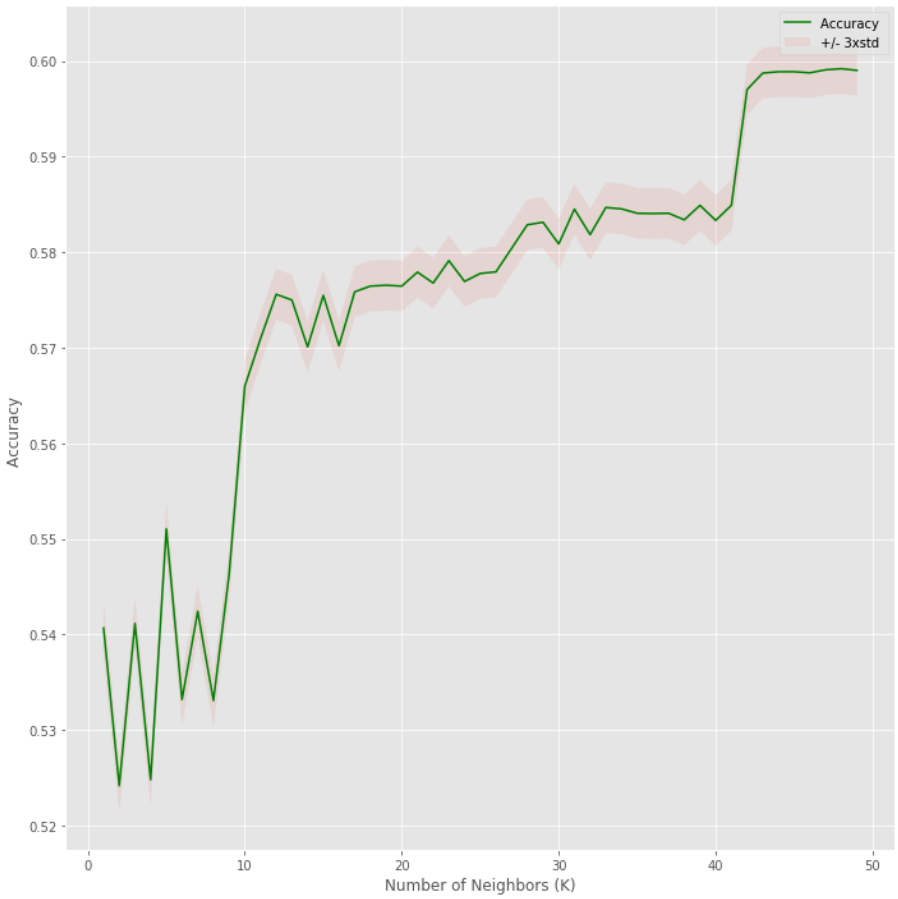

Fig 3. Model accuracy for different number of neighbours (K).

4.2 Performance metrics for classification

The most widely used technique for summarising the performance of a classification algorithm is the Confusion Matrix. Figure 4 shows the KNN confusion matrix with useful information about how well the model does. Its elements were used to calculate many performance metrics to get even more information as shown on table 1.

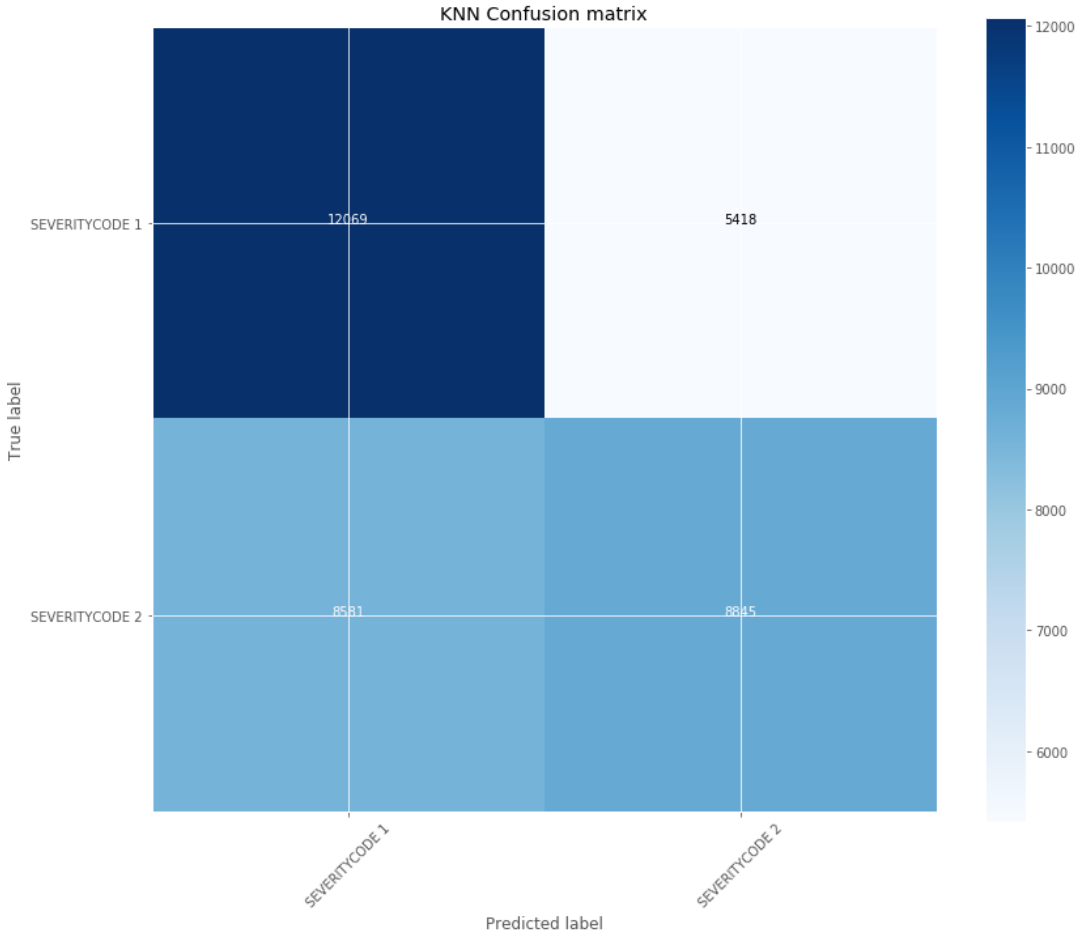

Fig 4. KNN confusion matrix

	precision	recall	f1-score
1	0.58	0.69	0.63
2	0.62	0.51	0.56

Table 1. KNN performance metrics

1. Precision it is simply a ratio of correctly predicted positive data objects to the total predicted positive data objects.
2. Recall it is defined by the number of correct positive results divided by the total number of relevant samples (all samples that should have been identified as positive).

3. F-score it can be defined as a weighted average of the precision and recall. An F-score is considered perfect when reaches its best value at 1, while the model is a total failure when it reaches the 0 value.

4.2 Decision Tree model

The objective was to branch the tree until every “leaf” contains only SEVERITYCODE 1 or SEVERITYCODE 2. One of the key attractions of building a decision tree model is that it provides easy insight into which features in the feature set most cleanly separate the target variable between categories. A decision tree model was built for the Seattle accident dataset using the decision- treeclassifier in scikit learn, using the “entropy” criterion (Fig 5). The Decision Trees's Accuracy was 0.60, a little bit better than the KNN’s which was 0.59. The Decision Tree Confusion Matrix and the performance metrics were determined (Fig 6, Table 2).

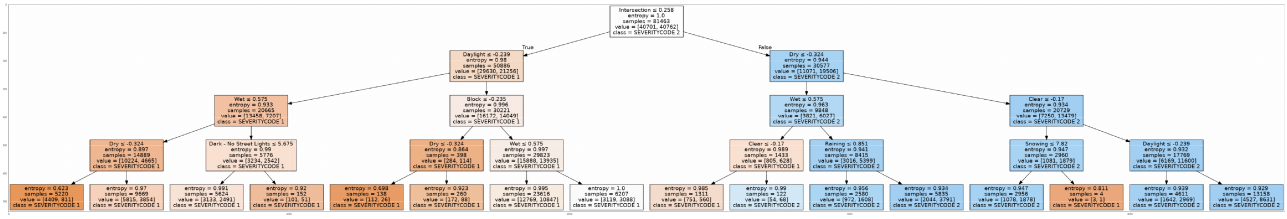

Fig 5. Decision Tree

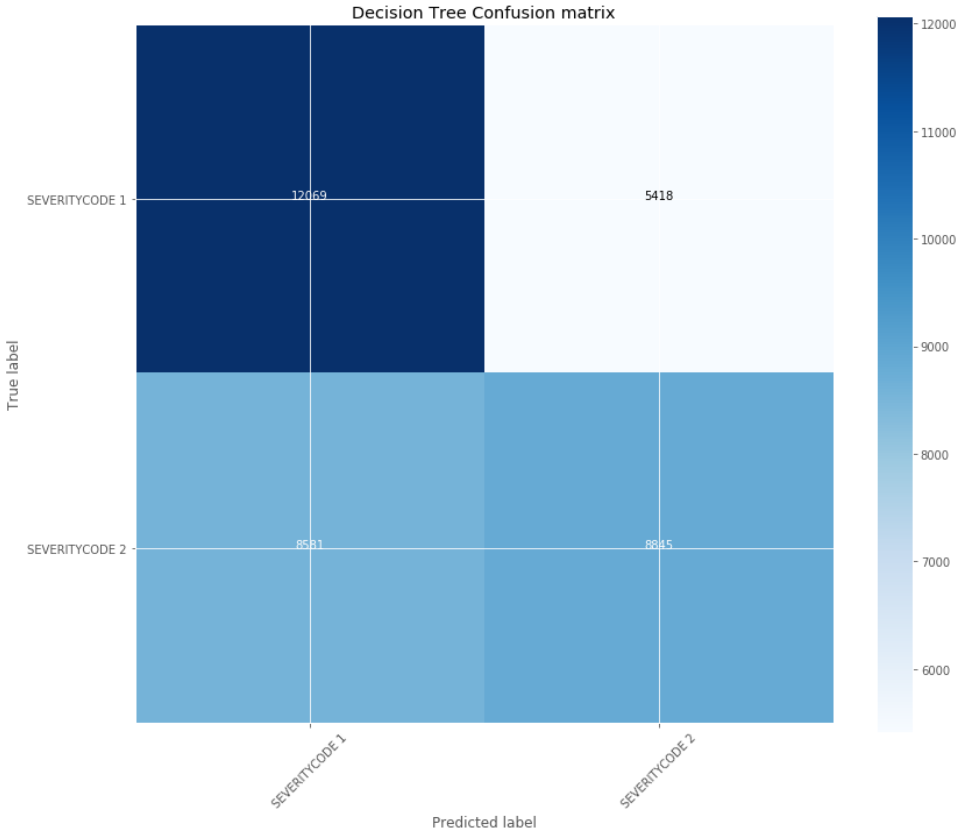

Fig 6. Decision Tree Confusion matrix

	precision	recall	f1-score
1	0.58	0.69	0.63
2	0.62	0.51	0.56

Table 2. Decision Tree performance metrics

4.3 Support Vector Machine model

Support Vector Machine (SVM) is used for classification, regression and outliers detection. This model was built using the sklearn.svm.svc, with a linear mapping kernel. The SVM Confusion Matrix and the performance metrics were determined (Fig 7, Table 3).

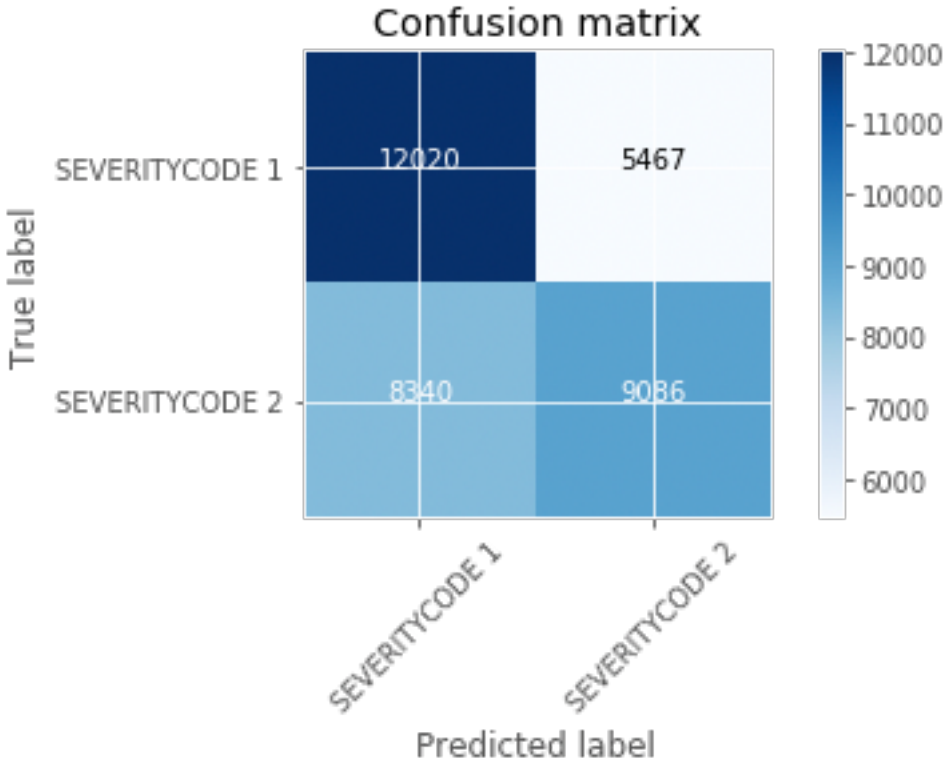

Fig 7. SVM Confusion matrix

	precision	recall	f1-score
1	0.59	0.69	0.64
2	0.62	0.52	0.57

Table 3. SVM performance metrics

4.4 Logistic Regression model

Logistic regression (LR) model is perfect for binary data such as the present dataset containing only two severity codes (1 & 2). This model was then built from the training set using scikit learn. The LR Confusion Matrix and the performance metrics were determined (Fig 8, Table 4).

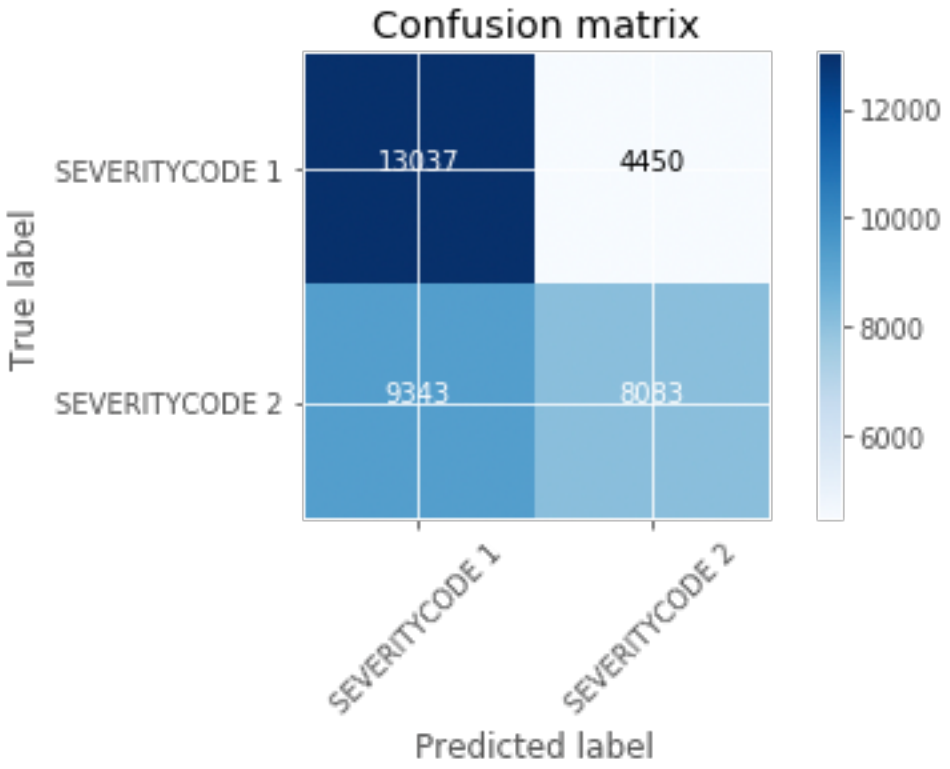

Fig 8. Logistic Regression Confusion matrix

	precision	recall	f1-score
1	0.58	0.75	0.65
2	0.64	0.46	0.54

Table 4. Logistic Regression performance metrics

4.5 Performance summary

The comparative performances of each of the four machine learning algorithms are summarised below in table 5.

	Algorithm	Jaccard	F1-score	LogLoss
1	KNN	0.599032	0.595649	NA
2	Decistion Tree	0.605562	0.597444	NA
3	SVM	0.604531	0.601777	NA
4	LogisticRegression	0.604932	0.596914	0.654374

Table 5. Report of models performance metrics

5. Conclusions

In this study, I built 4 algorithms models to predict the severity risk of accidents in the City of Seattle. I identified 'Raining', 'Clear', 'Overcast', 'Snowing', 'Dry', 'Ice', 'Snow/Slush', 'Wet', 'Dark - Street Lights Off', 'Dark - No Street Lights', 'Dawn', 'Daylight', 'Dusk', 'Alley', ‘Block' and ‘Intersection’ among the most important features that affect the severity of accidents. I built classification models to predict combination that would more likely lead to severe accidents. These models will be very useful in guiding road planners in their future road construction, drivers as well as pedestrians in their behaviour on the road in a particular condition.

6. Future directions

The accuracy in the classification problem of all models built was less than 65%. I think there is still room for models improvement. More data, especially data of different types, would help improve model performances significantly.

7. References

1. M.M.Chong,A.Abraham,andM.Paprzycki,“Traffic accident analysis using machine learning paradigms,” Informatica, vol. 29, pp. 89–98, 05 2005.

2. J.Abellan,G.L´opez,andJ.deO´na,“Analysisoftrafficaccident˜severityusing decision rules via decision trees,” Expert Systems with Applications, vol. 40, no. 15, pp. 6047 – 6054, 2013.

3. H.HeandE.A.Garcia,“Learningfromimbalanceddata,”IEEETransactionson Knowledge & Data Engineering (TKDE), no. 9, pp. 1263–1284, 2008.