Predicting car accident severity

1. Introduction

1.1 Background

Car accidents are one of the most frequent cause of serious injuries and casualties in modern societies. There are many potential approaches which can be considered in order to reduce their occurrence. However, the main objective should be to reduce the ones which are more serious.

There are several variables which might have some influence when trying to predict the probability of an accident. But if the objective is to predict its severity, it might not be intuitive to determine which ones are the most relevant.

1.2 Problem

Data with a list of accidents which have occurred in the city of Seattle is available, labelled in terms of their severity. Severity is expressed as a binary variable: 2 when the accident caused some injure and 1 when there is only property damage. This project aims to predict severity of the accident in terms of the different variables which can be found in the dataset.

1.3 Interest

As it was said before, reducing the number of severe accidents is very important. Building a model which can predict the severity of an accident would be valuable for that purpose. Authorities could use this knowledge to take measures in order to mitigate this problem.

In addition, this model could be used in navigation applications. This would improve the journey which these applications recommend, based on the probability that a severe accident might occur in any potential journey.

2. Data acquisition and cleaning

2.1 Data Sources

The data which will be used to elaborate this report is the .csv file provided by IBM. Some databases on the website Kaggle were acquired as well as a complement.

2.2 Data cleaning

The dataset has a total of 194674 rows. Some of them have empty values.

As a first approach *dropna()* function was applied to the whole dataframe. Only 29943 rows were deleted. This is considered reasonable, so this was the chosen method to delete empty values.

Another problem is that the dataset is not balanced: severity 2 accidents (which caused injuries) are undersampled in the dataset. In order to solve this, the method *sample()* was used. Two dataframes were created from the original (severity-1 and severity-2), a sample with fraction 0.5 was taken from severity-1 dataframe and finally the two dataframes were concatenated.

2.3 Feature selection

There are some features which are not valuable for our problem. For instance, there are several features which are codes, just useful to identify the accident. OBJECTID, INCKEY, SHAPE, COLDKEY are some of these.

Other variables might be valuable to describe the characteristics of the accident, but not to predict its severity. For example, COLLISIONTYPE, SDOT_COLCODE. In fact, if these features were introduced in the model, they might improve the accuracy. However, this would be misleading. The only reason this might happen is that these features *describe* the severity of the accident, they do not predict it.

The features which were used for the analysis can be classified as the following: time variable (date), coordinates and location (spatial), weather and road conditions and finally features which describe the location of the accident (ADDRTYPE for instance).

There are some features which are categorical variables with various categories. In order to train the machine learning models with success these features were transformed to binary features using one hot encoding. For instance, weather and road conditions are categorical variables with various categories.

The feature *DATE* was transformed into five new features: hour, day, month, year, weekday. For the model weekday and hour were transformed into binary features: weekend/workday and night/day.

For the feature *LOCATION* there are too many categories to do one hot encoding. Because of this, it is necessary to do some feature engineering. Some new features will be created related to the frequency of accidents in each location. This will be explained more in detail in the following sections.

3. Exploratory data analysis

3.1 Relationship between weather and severity

It is quite reasonable to claim that bad weather such as raining would increase the probability of suffering an accident. However, when analyzing its influence on accident severity (Figure 1) the influence is quite small. For the usual weather conditions (overcast, clear and raining) there is almost no difference when referring to severity.

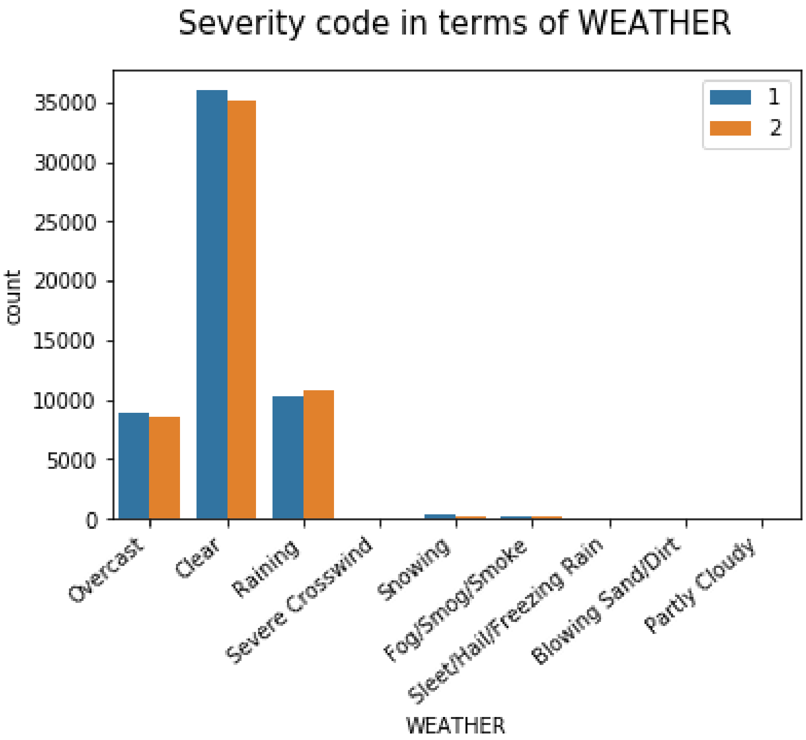

Figure 1: Countplot of accidents in Seattle based on the weather conditions.

In order to analyze the rest of weather conditions and be able to visualize them, it is necessary to plot them separately (Figure 2). The results are quite counterintuitive: extreme weather conditions such as snow might increases the probability of suffering an accident, but frequently it will be a less severe one. This could make some sense, as many small incidents would happen caused by snow that would not cause injures (severity 1).

However, this sample (extreme weather conditions) is very small compared to the whole dataset. It will not have a large effect when elaborating the predictive model.

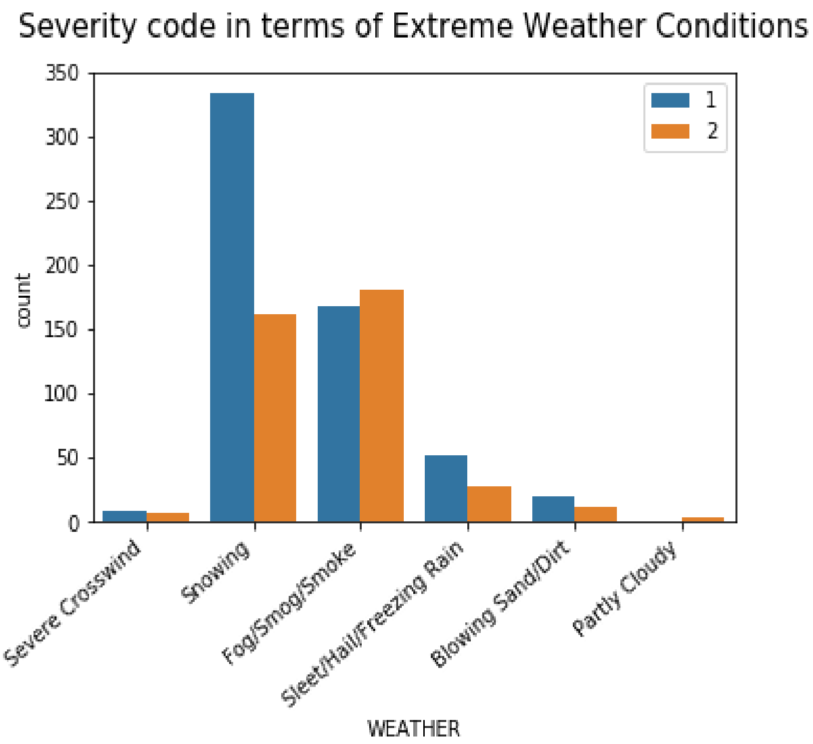

Figure 2: Countplot of accidents in Seattle based on extreme weather conditions

3.2 Relationship between road condition and severity

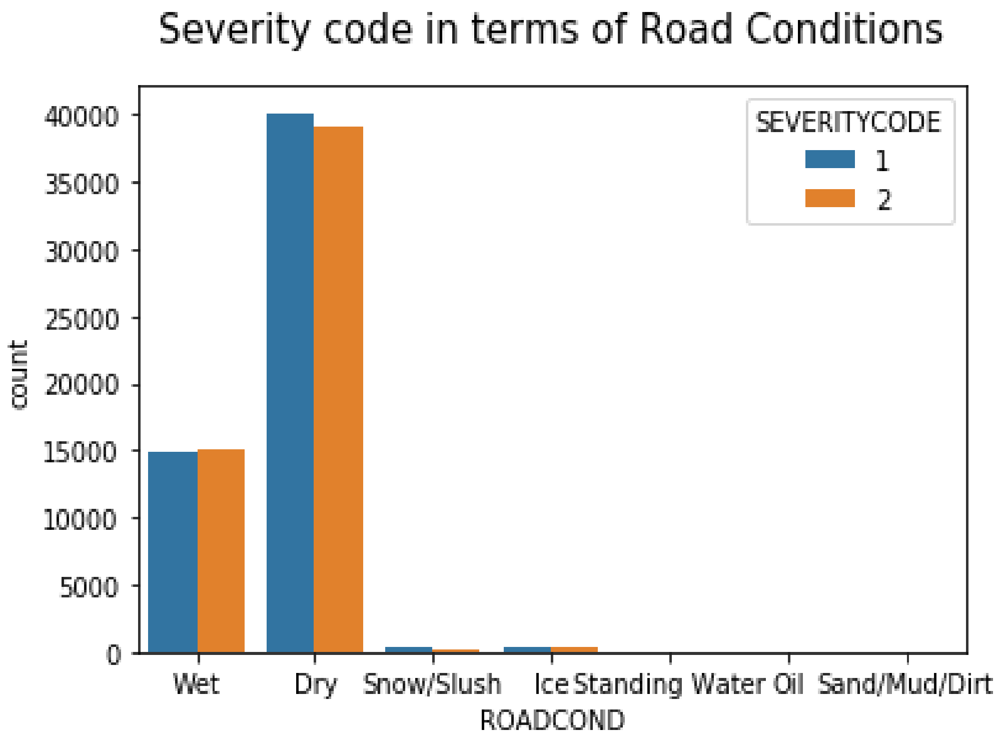

Figure 3: Countplot of accidents in Seattle based on road conditions

Road conditions affect accident severity in a similar way as the weather, in fact, the main road conditions are caused by weather.

It can be noticed that oil increases severity. However, the sample is very small. Further investigation would be needed.

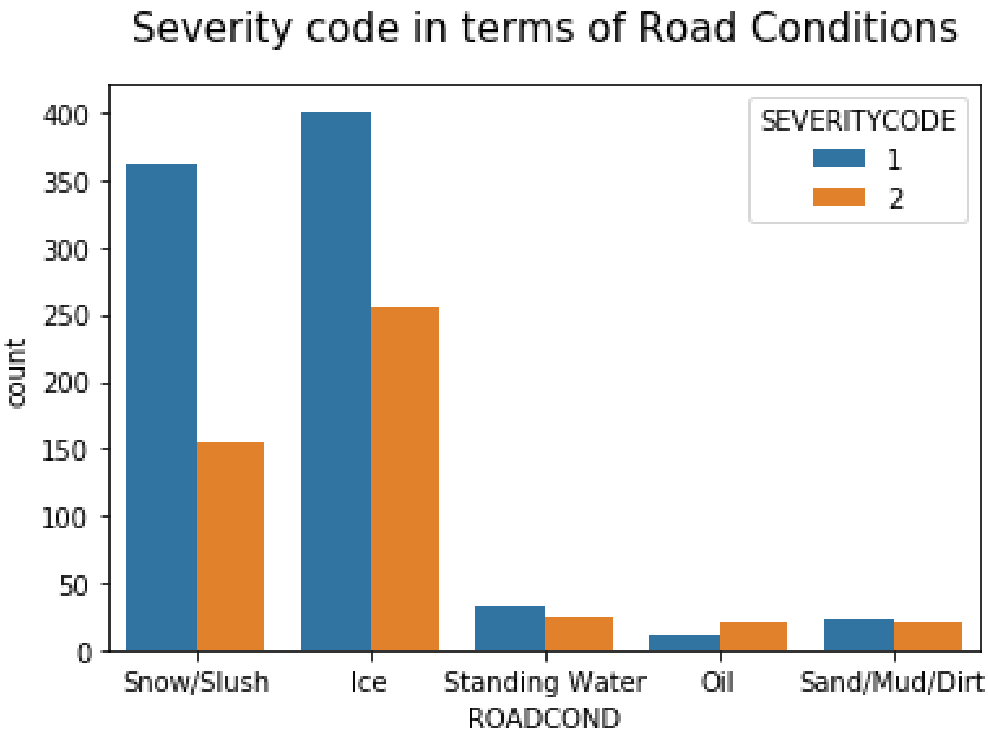

Figure 4: Countplot of accidents in Seattle based on extreme road conditions

3.3 Relationship between light condition and severity

At night accidents seem to be less severe. However, the effect is small.

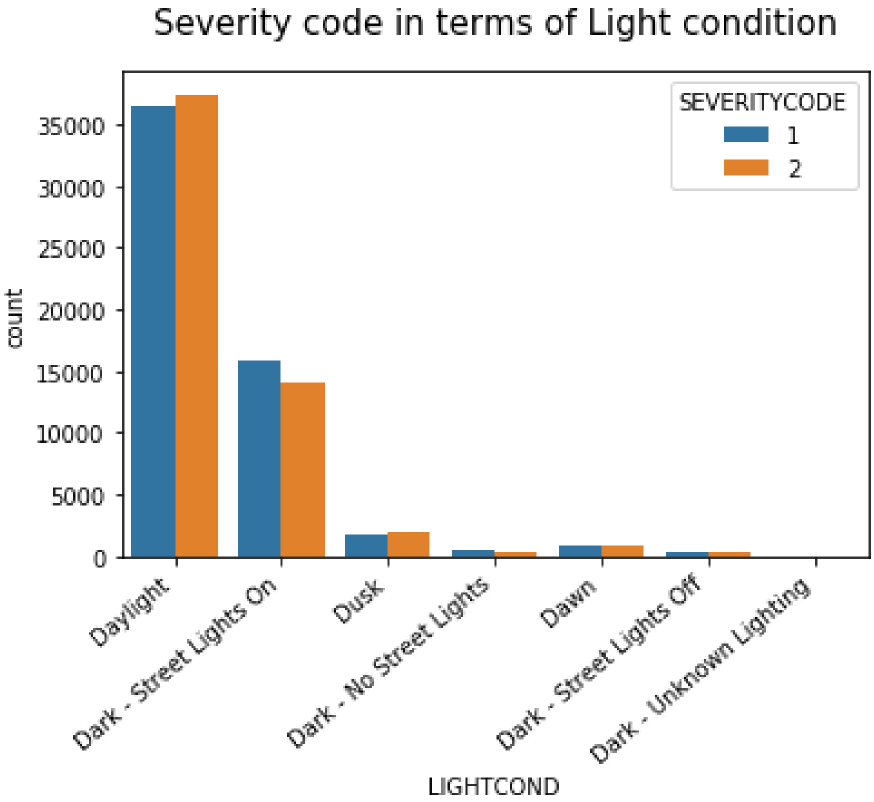

Figure 5: Countplot of accidents in Seattle based on light conditions

3.4 Relationship between address type and severity

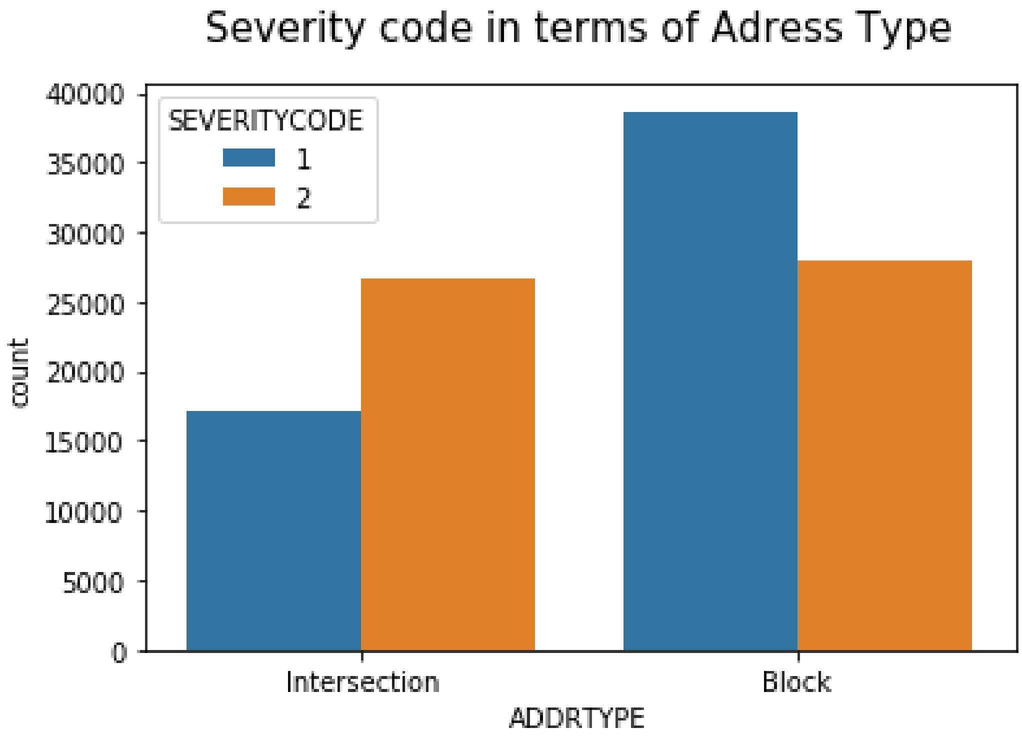

Figure 6: Countplot of accidents in Seattle based on the address type

It can be noticed that address type is very relevant as variable which affects accident severity. Accidents in intersections tend to be more severe than the ones in blocks.

The location of the accident seems to be important in our analysis. Other characteristics of the accident location might affect accident severity as well.

3.5 Relationship between frequency of accidents in a location and severity

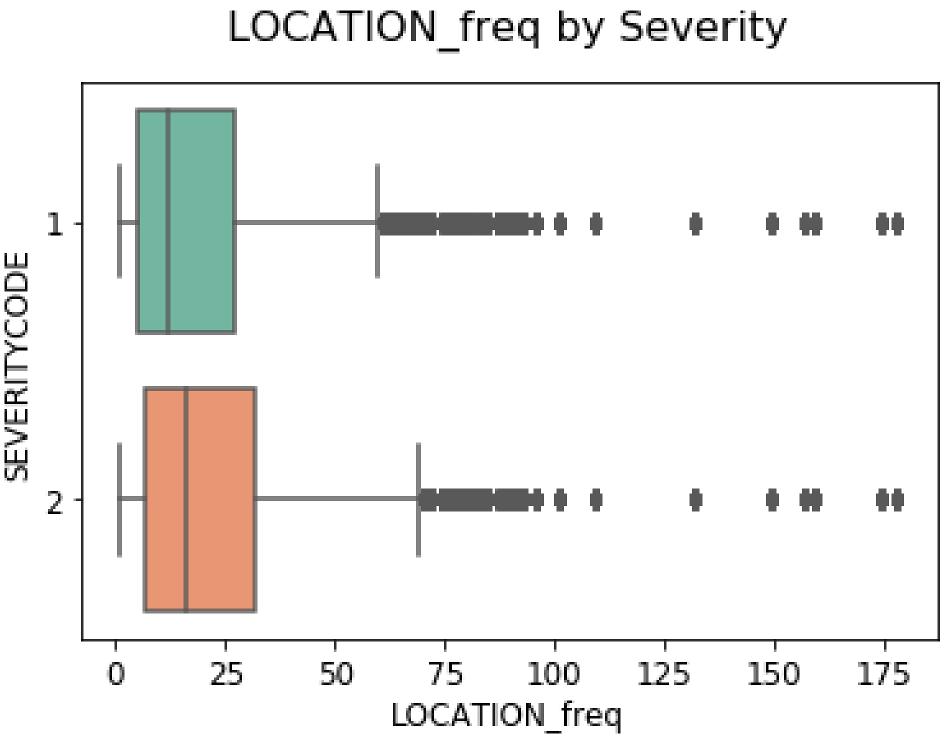

Figure 7: Boxplot where frequency of accidents by location is exposed both for accidents of severity 1 and 2

As it was said before, a new variable was created called *LOCATION_FREQ.* This variable measures the number of accidents (of any severity) which happened in each location. In Figure 7 accidents which are more severe (Severity 2) tend to be more concentrated in certain locations. In other words, there are *blackspots* where accidents tend to occur, but this effect is even larger with more severe accidents.

To understand this effect more clearly, two new variables were created: *LOCATION_FREQ_S1 and LOCATION_FREQ_S2.* The first one measures the number of accidents of severity 1 and the second one the number of accidents of severity 2 in each location. In Figure 8 this can visualized for the 20 locations with the highest number of accidents. For most locations severe accidents are oversampled. Only three locations are balanced and another two have the opposed situation (a much greater number of less severe accidents).

This clearly shows that these two variables have a big predictive power in order to predict the severity of car accidents, so they will be included in the model.

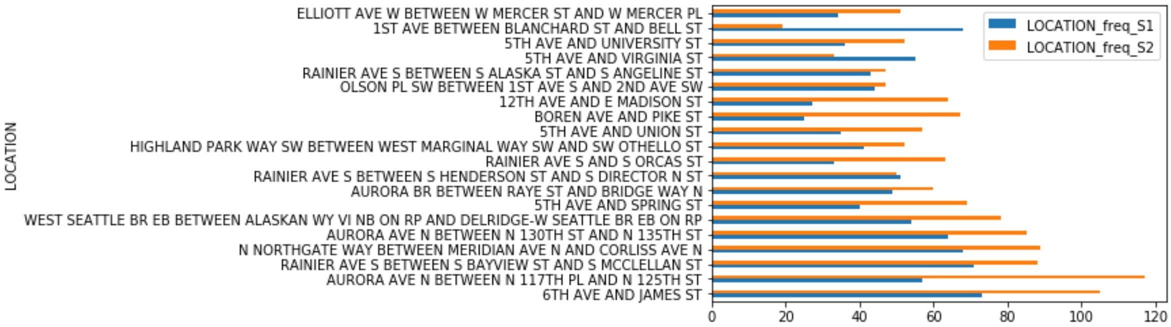

Figure 8: Horizontal boxplot where frequency of accidents by location is plotted for the 20 locations with the highest number of accidents

3.6 Relationship between hour of the accident and severity

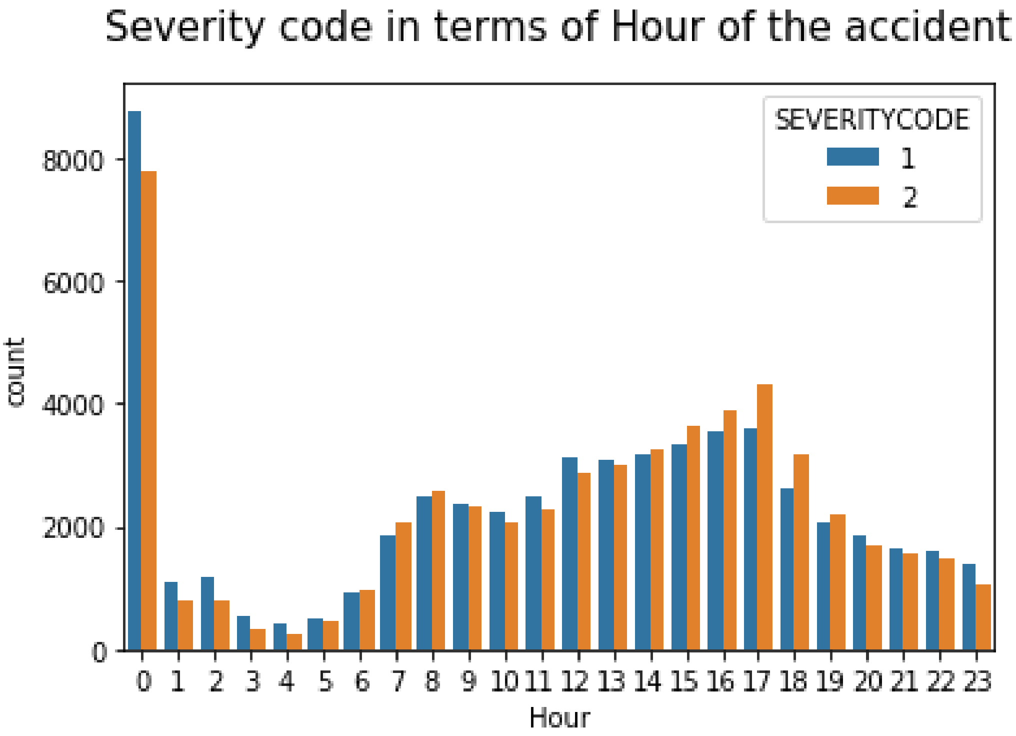

Figure 9: Countplot of accidents based on the hour of the accident

Figure 9 shows that the number of accidents greatly diminishes in the night, which is quite obvious because traffic flow is lower. Less severe accidents are a bit more frequent in the night. This was seen when light condition was analyzed.

Hour 0 is an outlier. The reason might be that 0 is the default hour and other accidents which happened at another hour appear with that value.

3.7 Relationship between weekday of the accident and severity

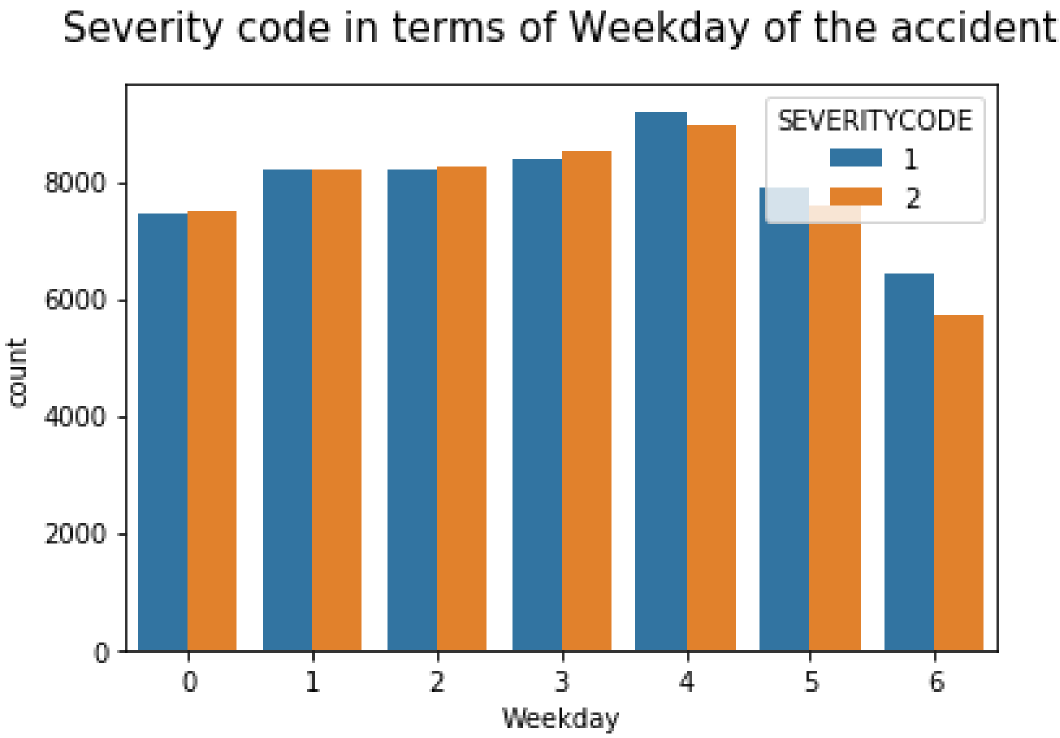

Figure 10: Countplot of accidents based on the day of the week of the accident

Accidents are less frequent in the weekend and a bit more frequent on Friday.

In addition, less severe accidents are a bit more frequent in the weekend. Because of this, this feature was transformed into a binary one: weekend/workday.

4. Predictive modelling

The models used to solve the problem are classification algorithms, as this is a classification problem. The chosen ones were KNN (nearest neighbors), logistic regression, SVM (support vector machine) and Decision Tree.

SVM was not finally included because it took too long to train, so results will be exposed for the other three algorithms.

The dataset was divided into a train/test split with test size of 0.3.

Preprocessing of the data was done with the *StandardScaler()* method.

5. Results

KNN was run from K=1 to K=15. Accuracy for each value can be seen in Figure 11. Value K=13 was chosen for the final model.

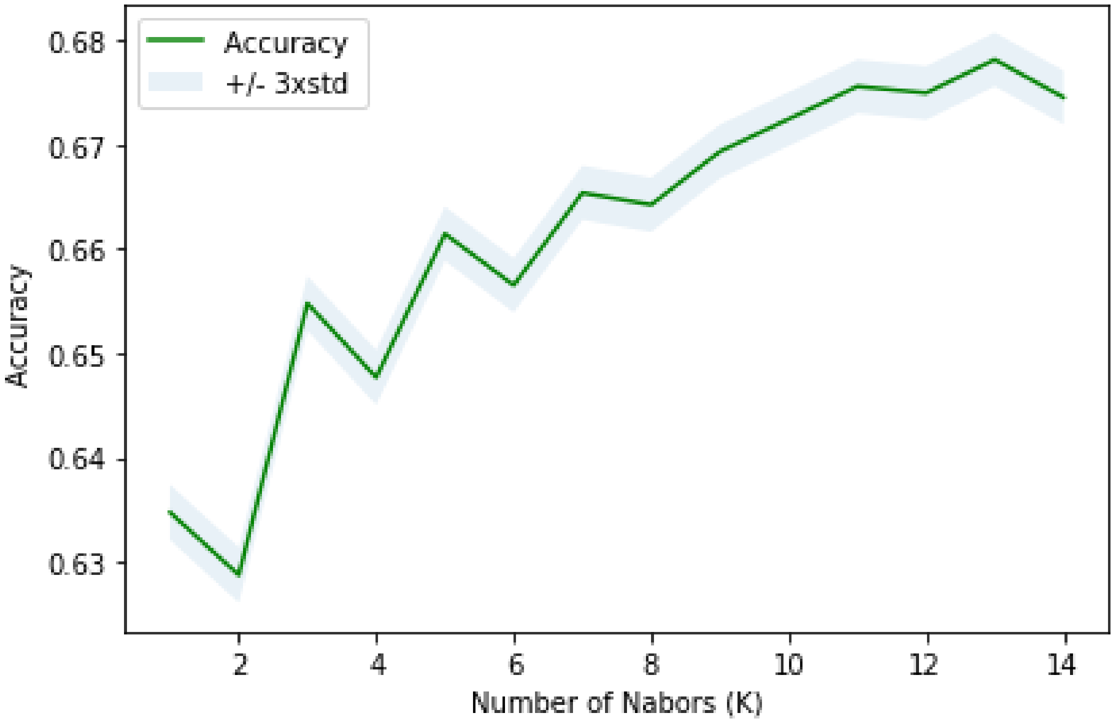

Figure 11: Evolution of accuracy with the number of neighbors for KNN algorithm

Performance of the three models was quite similar. Decision tree was the one which obtained the highest accuracy. In addition, it was the one which needed less time to be trained.

About the false positives/negatives it can be considered that predicting less severity is worse than predicting more severity. In this case logistic regression is the one which predicts incorrectly less severity 1 when true severity is 2.

	KNN	*Logistic Regression*	*Decision Tree*
Accuracy	0.68	0.66	0.69

Confusion matrix	`[[11203, 5572]`
	`[ 5095, 11270]]`
		`[[11671, 5104]`
		`[ 6308, 10057]]`
			`[[10477, 6298]`
			`[ 4451, 11914]]`

Table 1: Accuracies and confusion matrixes of the different models

The tree class in sklearn makes it possible to check feature importance for Decision Tree model. As expected, *LOCATION_FREQ_S1 and LOCATION_FREQ_S2* were the best predictors by far. Address type was also important, but the use of location frequency makes it a bit redundant. Weather and road conditions can be useful when there are extreme conditions like snow.

6. Discussion

As exposed in the results section, information about the location of the accident is very valuable when trying to predict accident severity. In the dataframe the only feature which describes the location of the accident is *ADRTYPE*. Because of this it was demonstrated that accidents in intersections tend to be more severe.

It would be nice to have more features which describe the location of the accident. Based on the importance of *LOCATION_FREQ_S1 and LOCATION_FREQ_S2* it is expected that other characteristics of the location are also important when trying to predict severity. For instance, the speed limit in each location could provide additional information.

7. Conclusions

In this report the relationship between car accident severity in Seattle and different features such as weather, hour of the day and location of the accident was analyzed. I identified the information about the location as the most valuable in order to predict severity. Based on this, I was able to add some variables which made it possible to build a model that predicts accident severity. This model can be valuable for authorities in order to reduce severe accidents or for traffic application in order to recommend better journeys.