Predicting the Severity of vehicle accidents in the city of Seattle Capstone project - Applied Data Science Capstone in Coursera

Presented by:
Nicolas Achury

20/09/2020

Introduction/Business Problem

Problem

How to reduce the number of vehicular accidents in Seattle using the best information available. For this, it is possible to predict the severity of an accident given certain factors such as the weather, road and light conditions and speeding among several other factors. In this way, drivers can take preventive actions such as more careful driving behaviours or even change their routes, which would be directly reflected in a reduction in the accident rate in the city.

Target Audience

The solution to this problem would allow the local government of Seattle to develop informative campaigns to reduce vehicle accidents, reducing the social and economic costs of attending future accidents. Likewise, drivers and even pedestrians would also benefit, since they are the ones who are directly involved in the vehicle accidents.

Data

- 37 attributes
- 194,673 vehicle accident records
- The records are updated weekly since 2004
- Each record includes information on the severity of the accident (fatality or prop damage), collision type, the total number of people involved in the collision, the date of the accident, the number or injuries and fatalities in the accident, light conditions during the collision, the road conditions, weather, and some other attributes related to the accident itself.

Methodology

Removing unnecessary data

The dataset provided by the local government of Seattle has 37 attributes that can be used to identify some patterns in the severity of the car accidents within the city. However, not all attributes have relevant information for the model development. For instance, those attributes related to unique key identifiers used by the SDOT Traffic Management Division do not provide any useful information.

Removing Missing Data

All records with at least one missing data at any of the attributes were deleted. This guarantees the dataset is complete. On top of that, only records with a Matched status were considered as it was assumed that the description of the vehicle accident was well represented and matched by the record itself.

After cleaning the provided data, a total of 182.894 complete and verified records were obtained. This dataset will be used for the development of a supervised learning model

Methodology

Processing Data

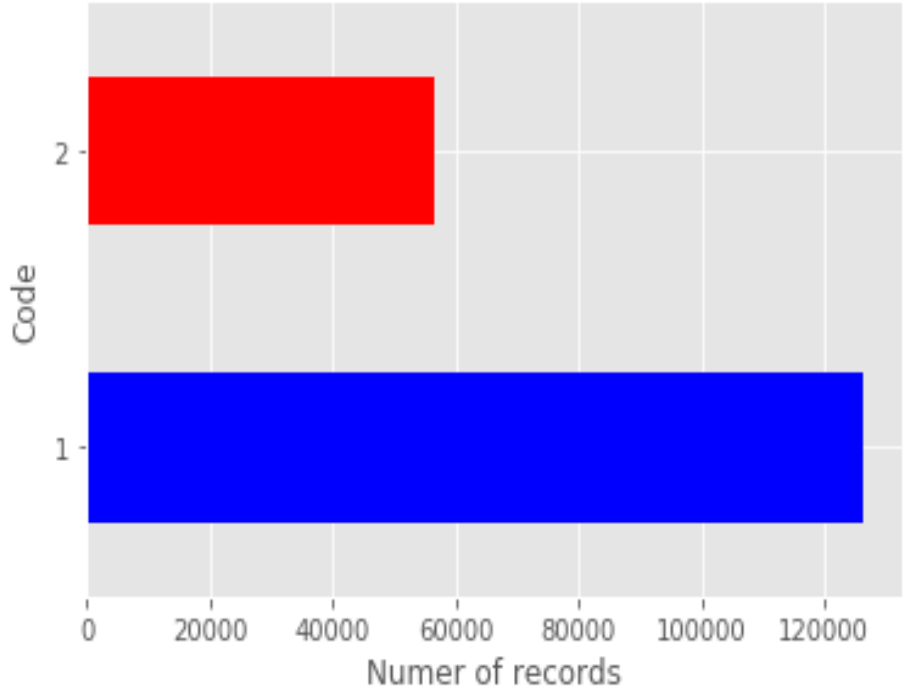
Attribute	Processing Description
‘INNATENTIONIND’	This attribute has the format of Y/N, buy only Y data was recorded. It was assumed that missing data, in the form of NaN, could be replaced with N. The attribute was reformatted to 1/0.
‘SPEEDING’	This attribute has the format of Y/N, buy only Y data was recorded. It was assumed that missing data, in the form of NaN, could be replaced with N. The attribute was reformatted to 1/0.
'UNDERINFL'	This attribute has different data formats: N, Y, 0 and 1. All records were reformatted to 1/0 and missing data, in the form of NaN, was deleted.
'HITPARKEDCAR'	Data in this attribute was reformatted to 1/0. Missing data, in the form of NaN, was deleted.
‘STATUS’	This attribute has the format of ‘Matched’ and ‘Unmatched’. The data was reformatted to 1/0 for Matched and Unmatched respectively. Missing data, in the form of NaN, was deleted.

Balancing Dataset

The dataset must be balanced as the number of non-severe vehicle accidents surpasses the severe ones by 69,644 records. An imbalanced dataset may result in a biased model. For the model development, non-severe records were down sampled, to avoid randomly duplicating severe records.

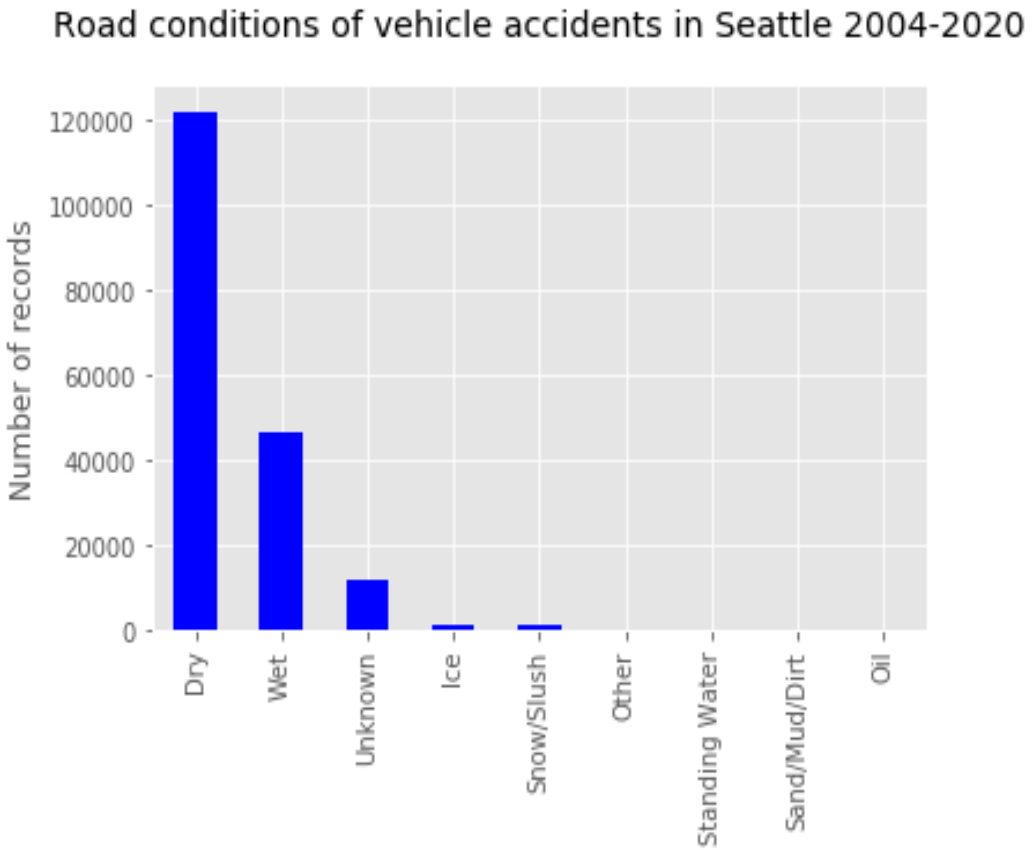
Data Visualization

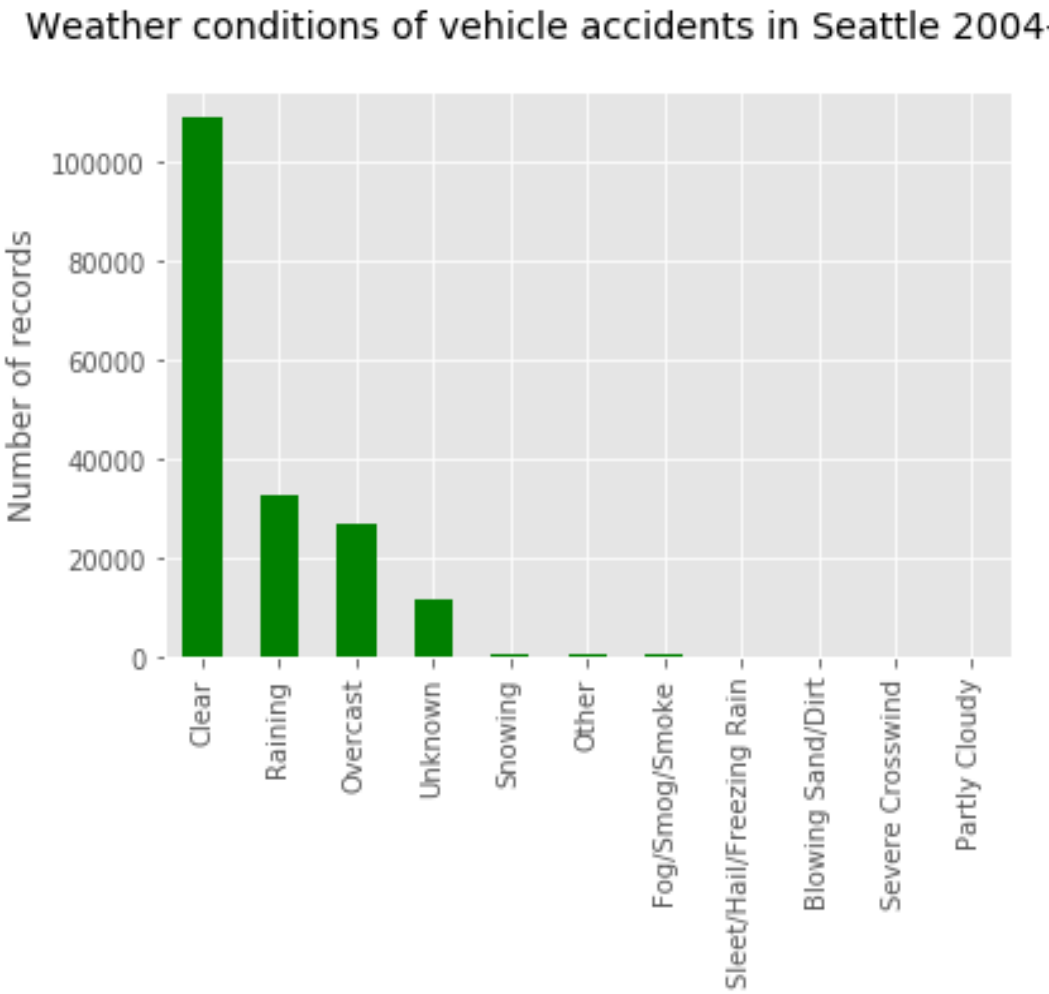
Number of Severe (code 2) and Non-severe (code 1) vehicle accidents in Seattle 2004-2020



the number of non-severe vehicle accidents surpasses the number of severe ones. There have been 126.269 non-severe accidents in Seattle since 2004, compared to 56.625 severe cases in the same period. It is possible to infer that the dataset is unbalanced.

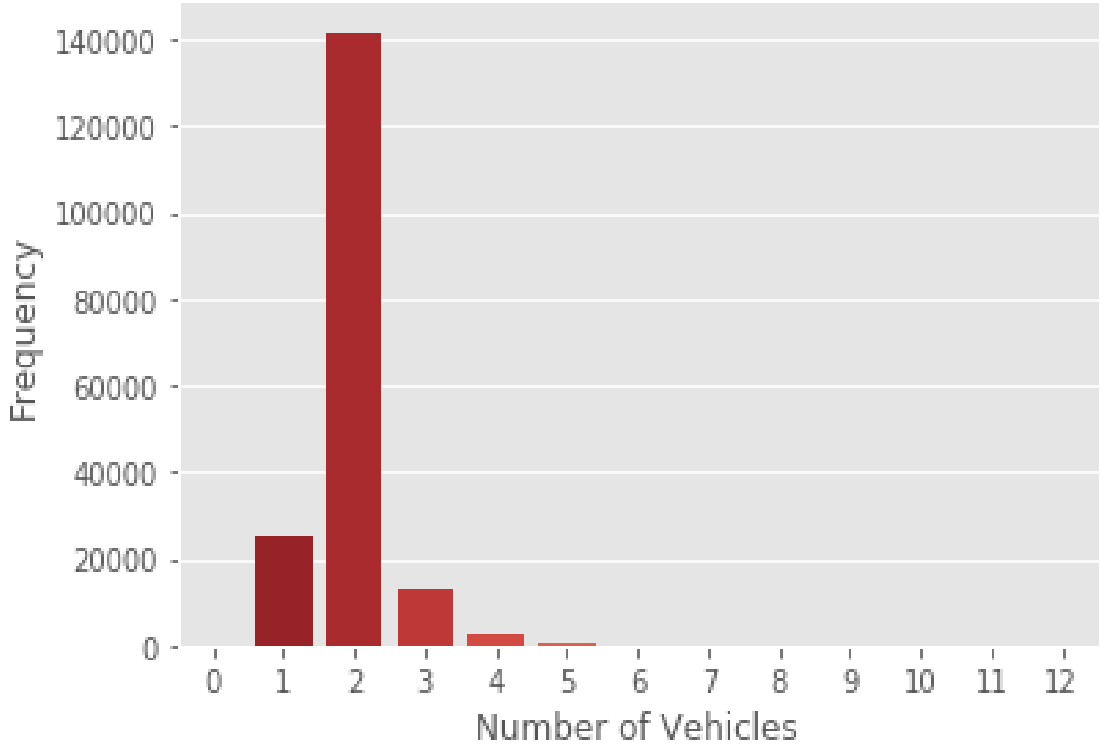
Data Visualization





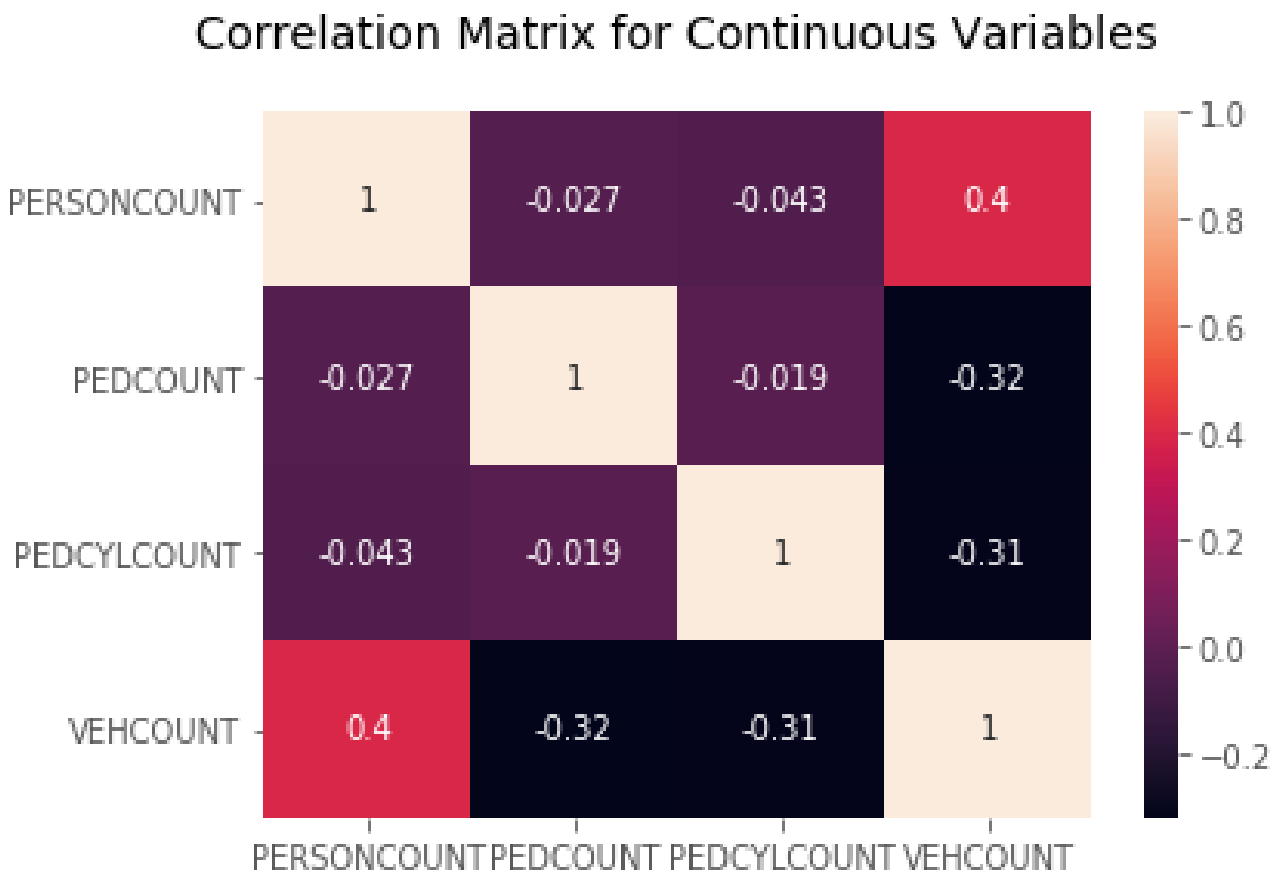
Data Visualization

Histogram of number of vehicles involved in accidents in Seattle 2004-2020



Most of the vehicle accidents in Seattle occur in dry road conditions, clear weather and involve usually 2 vehicles per accident (including bicycles). The second most common conditions in Seattle's vehicle accidents occur in wet road conditions possibly due a raining day.

Data Analysis



The ones' diagonal in the matrix indicates a perfect correlation of one variable with itself, or in other words, is the same variable. For all the other correlation data, it is possible to see low or near-zero values. This implies that the variables tend to be linearly independent and therefore, no collinearity issues are expected.

Data Analysis

One-way Analysis of Variance (ANOVA)

Categorical Variable	F-Value	P-Value
'ADDRTYPE'	3530.70	0
'COLLISIONTYPE'	5369.06	0
'JUNCTIONTYPE'	1403.6	0
'WEATHER'	374.03	0
'ROADCOND'	476.04	0
'LIGHTCOND'	463.64	0
'SPEEDING'	223.74	1.47e-50
'HITPARKEDCAR'	1577.65	0

All categorical variables considered have a significantly high F-value and corresponding zero or near-zero P-Values. This implies indeed, that there is a statistical difference between the means of the categorical variables and therefore they have a significant impact of the target variable SEVERITYCODE. In this sense, all variables will be used in the model development.

Results

Model	Accuracy Score	F1 Score	Jaccard Score	Log Loss
Logistic Regression	0.7007	0.6977	0.5004	0.5464
Decision Tree	0.7034	0.6989	0.4938	-
K-Nearest Neighbour	0.6867	0.6860	0.5039	-
Support Vector Machine (SVM)	-	-	-	-

Model	Computation Time [seconds]
Logistic Regression	1.39062
Decision Tree	0.96875
K-Nearest Neighbour	190.42187
Support Vector Machine (SVM)	-

Discussion

- All models have similar accuracy. However, the Decision Tree model has slightly better performance in terms of Accuracy Score and the F1 Score.
- Logistic Regression despite having a fairly similar score to the other models, it has a considerably high Log-loss. So, it won't be recommended.
- There is no performance metrics for the Support Vector Machine (SVM) as the model did not compute in more than 1 hour.
- In terms of computation times, the most efficient model is the Decision Tree, followed by the Logistic Regression and the K-Nearest Neighbour which took the longest computation time (more than 3 minutes in a Dell XPS 13 9370)
- The recommended model is a Decision Tree with an accuracy score of 70.34% and short computation times.

Conclusions

- The recommended model is a Decision Tree with max_depth =11 due to the relatively high accuracy and short computation times.
- The Logistic Regression and the K-Nearest Neighbour are not recommended due to a slightly lower accuracy, high Log-loss (Regression Model), and high computation times (K-Nearest Neighbour).
- The Support Vector Machine is not recommended due to computation times. This might be due the time required to map the large dataset to a high dimension space in order to the algorithm works.
- There is no evidence of collinearity issues in the continuous variables considered in the analysis.
- All categorical variables have significant impact on the Severity Code variable, this evidenced from high F-values and zero/near-zero P-values in the ANOVA analysis.
- In order to avoid biased models, the data had to be balanced. For this analysis a Down Sampling approach was taken.
- The number of non-severe vehicle accidents surpasses the number of severe ones. Since 2004, there have been 126.269 non-severe accidents in Seattle, compared to 56.625 severe cases in the same period