CAPSTONE PROJECT -CAR ACCIDENT SEVERITY

Theodoros P. Kantas

September 2020, IBM Data Science Certificate

Table of Contents

1.Introduction/Business Problem

Road accidents are always a serious and frequent issue.

In this project we aim to inform the drivers for the possibility of an accident occurring when there are specific weather, road and visibility conditions, in order to be more prepared.

2.Data

We use shared data for Seattle city for our project. Data are from 2004 until now (05/2020).

Our dataset has 38 columns and 194.673 rows

The dataset contains data related to the severity of accidents, our aim to predict. In our dataset the column related to severity of an accident, the SEVERITYCODE has two values, value 1 refers to “Property Damage Only Collision” and value 2 refers to “Injury Collision”.

Some data needs to be balanced. The main feature, SEVERITYCODE , is imbalance (number of value1=136.485 vs value2=58.188)

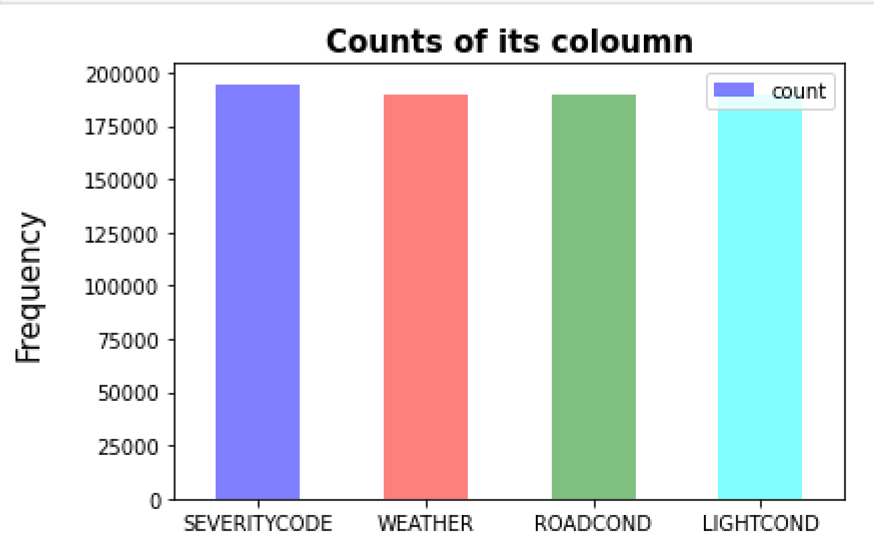
Some columns, such as SPEEDING(only 9333 values in 194.673 rows), PEDROWNOTGRNT(only 4667 values in 194.673 rows) etc,   contain small number of values or are not related to our model, so we are going to exclude them.

3.Methodology

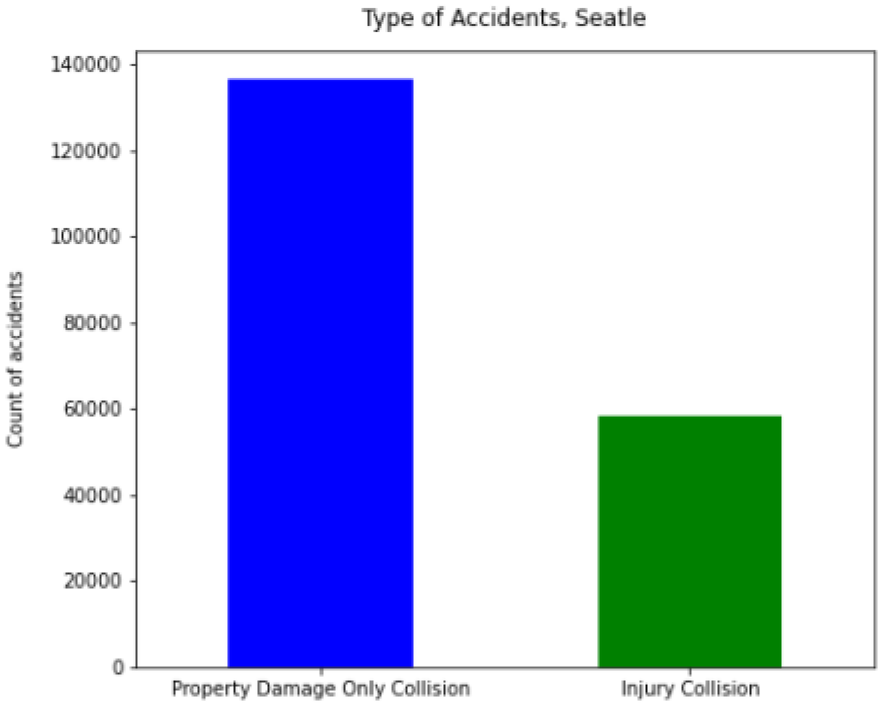
3.1 Exploratory Analysis

In order to understand better our data we create some basic graphs, and analyze the values of the selected columns(features).

Firstly a graph of type bar was created for all of ours columns.



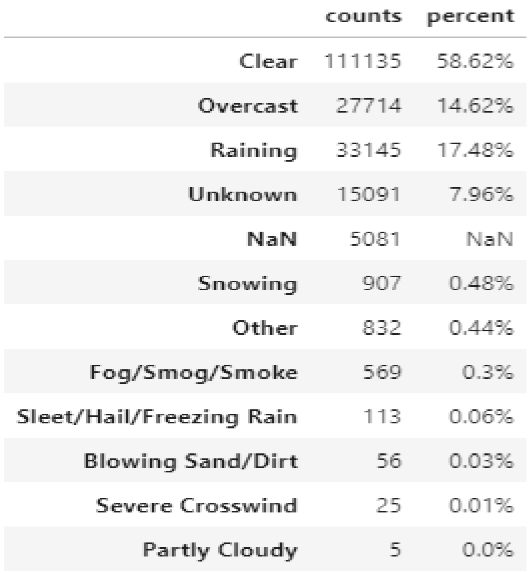
A graph of type bar was created for our main feature , the severity of an accident.



A short analysis of the values of the features that affect the severity of an accident is depicted below :

Weather

The weather conditions (WEATHER Column) that prevailed when the accident happened.



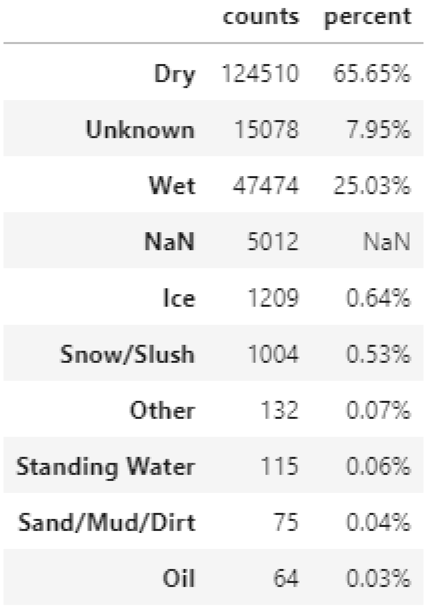
	counts	percent
Clear	111135	58.62%
Overcast	27714	14.62%
Raining	33145	17.48%
Unknown	15091	7.96%
NaN	5081	NaN
Snowing	907	0.48%
Other	832	0.44%
Fog/Smog/Smoke	569	0.3%
Sleet/Hail/Freezing Rain	113	0.06%
Blowing Sand/Dirt	56	0.03%
Severe Crosswind	25	0.01%
Partly Cloudy	5	0.0%

A check for the missing values is also performed. NaN =5081 and there are going to be deleted.

The above values were encoded according to their order of appearance, except the “Unknown” and “Other” values which encoded together.

Road Conditions

The Road Conditions (ROADCOND Column) that prevailed when the accident happened.



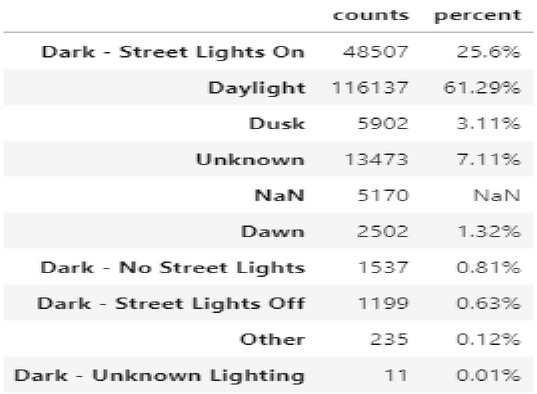
	counts	percent
Dry	124510	65.65%
Unknown	15078	7.95%
Wet	47474	25.03%
NaN	5012	NaN
Ice	1209	0.64%
Snow/Slush	1004	0.53%
Other	132	0.07%
Standing Water	115	0.06%
Sand/Mud/Dirt	75	0.04%
Oil	64	0.03%

A check for the missing values is also performed. NaN =5012 and there are going to be deleted.

The above values were encoded according to their order of appearance, except the “Unknown” and “Other” values which encoded together.

Light Conditions

The Light   conditions (ROADCOND Column) that prevailed when the accident happened.



	counts	percent
Dark - Street Lights On	48507	25.6%
Daylight	116137	61.29%
Dusk	5902	3.11%
Unknown	13473	7.11%
NaN	5170	NaN
Dawn	2502	1.32%
Dark - No Street Lights	1537	0.81%
Dark - Street Lights Off	1199	0.63%
Other	235	0.12%
Dark - Unknown Lighting	11	0.01%

A check for the missing values is also performed. NaN =5170 and there are going to be deleted.

The above values were encoded according to their order of appearance, except :

- “Dark - No Street Lights”, “Dark - Street Lights Off” , “Dark - Unknown Lighting” values which encoded together
- “Unknown” and “Other” values which encoded together.

3.2 Machine Learning Model

We practice with different classification algorithms, such as Decision Trees, Logistic Regression and KNN (k-Nearest Neighbor) in order to predict the severity of an accident based on the selected features. We did not use the Support Vector Machine (SVM) due to the size of dataset. We use various evaluations metrics such as Accuracy, F1 Score, Jaccard Index, Precision/Recall score, LogLoss score.

4.Results

The results per algorithm used are depicted below:

Decision Tree

Decision Tree Algorithm	Scores
Accuracy	0.69618147
F1	0.57150931
Jaccard Index	0.69617344
Precision/Recall	0.69615506

Logistic Regression

Logistic Regression Algorithm	Scores
LogLoss score	0.59545798
Accuracy	0.69615506
F1	0.57144760
Jaccard Index	0.69615506
Precision/Recall	0.69615506

KNN (k-Nearest Neighbor)

KNN (k-Nearest Neighbor)	
Accuracy with K=4	0.696023027
F1	0.574753254
Jaccard Index	0.691666004
Precision/Recall	0.696145824

Algorithm	Accuracy	F1	Jaccard Index	Precision/Recall	LogLoss
Decision Tree	0.69618147	0.57150931	0.69617344	0.69615506	
Logistic Regression	0.69615506	0.57144760	0.69615506	0.69615506	0.59545798
KNN (k-Nearest Neighbor)	0.696023027	0.574753254	0.691666004	0.696145824	

Comparing the results of the three algorithms used, it is understood that their performance where almost the same.

5.Conclusion

As it was mentioned in the data section the SEVERITYCODE had only two values , whereas the original dataset had 5 values (3—fatality,2b—serious injury,2—injury,1—prop damage,0—unknown).

The distribution of the two values was problematic and the dataset needed to be balanced.

The unknown values of the features that were used were between 7-8% of the total, a considerable percent.

The three algorithms that were used, probably would perform better if the above restrictions did not exist or existed to a lesser extent.