IBM Machin Learning Capstone Project

Author: Atena Vahedian

Introduction

In this project, a public data set is used to demonstrate Data Analysis and Machine Learning skills obtained in the IBM Data Science Program. Coursera has provided the data set which includes records of accidents and their severity level. The objective of this project is to: understand the data, define relevant attributes that cause the accident and build a predictive model. This analysis will help in potential driving hazard identification and can help drivers choose the best rout based on weather and road condition.

Data

The collision data set provided by Coursera includes all collision reports in Seattle since 2004. The data set has 194673 records and 37 attributes, each record is labelled by the accident's severity level. After proper data evaluation and plotting, 10 features were finalized for modeling purposes. Therefore, the effect of Address Type, Collision Type, Weather Condition, Road Condition, Light Condition, Speed Violation, Number of Pedestrians, Vehicles, Cyclists and Total Number of People Involved in the Accident were used to estimate severity of accidents.

Methodology

Exploratory Data Analysis

Before the modeling process, it is important to investigate the data. This means finding a better understanding of the data as well as cleaning the data set based on the results.

By looking at the size of data set, number of missing data at each column, and data types, 10 attributes were chosen for further evaluation. **Table 1** shows first five rows of the data set after filtering.

Table 1 - first five rows of the data set after filtering.

	SEVERITYCODE	ADDRTYPE	COLLISIONTYPE	PERSONCOUNT	PEDCOUNT	PEDCYLCOUNT	VEHCOUNT	WEATHER	ROADCOND	LIGHTCOND	SPEEDING
0	2	Intersection	Angles	2	0	0	2	Overcast	Wet	Daylight	NaN
1	1	Block	Sideswipe	2	0	0	2	Raining	Wet	Dark - Street Lights On	NaN
2	1	Block	Parked Car	4	0	0	3	Overcast	Dry	Daylight	NaN
3	1	Block	Other	3	0	0	3	Clear	Dry	Daylight	NaN
4	2	Intersection	Angles	2	0	0	2	Raining	Wet	Daylight	NaN

Now that dataframe has gone through preliminary filtering, more in-depth evaluations can be performed.

Table 2 - Statistical description of numerical data in the filtered dataframe.

	SEVERITYCODE	PERSONCOUNT	PEDCOUNT	PEDCYLCOUNT	VEHCOUNT
count	194673.000000	194673.000000	194673.000000	194673.000000	194673.000000
mean	1.298901	2.444427	0.037139	0.028391	1.920780
std	0.457778	1.345929	0.198150	0.167413	0.631047
min	1.000000	0.000000	0.000000	0.000000	0.000000
25%	1.000000	2.000000	0.000000	0.000000	2.000000
50%	1.000000	2.000000	0.000000	0.000000	2.000000
75%	2.000000	3.000000	0.000000	0.000000	2.000000
max	2.000000	81.000000	6.000000	2.000000	12.000000

Table 3 - Statistical description of the object type data in the dataframe.

	ADDRTYPE	COLLISIONTYPE	WEATHER	ROADCOND	LIGHTCOND	SPEEDING
count	192747	189769	189592	189661	189503	9333
unique	3	10	11	9	9	1
top	Block	Parked Car	Clear	Dry	Daylight	Y
freq	126926	47987	111135	124510	116137	9333

Table 4 - Distribution of accidents based on the severity code at each address type.

ADDRTYPE	SEVERITYCODE	
Alley	1	0.890812
	2	0.109188
Block	1	0.762885
	2	0.237115
Intersection	1	0.572476
	2	0.427524

Table 5 - Distribution of accidents based on the severity code at different weather condition.

WEATHER	SEVERITYCODE	
Blowing Sand/Dirt	1	0.732143
	2	0.267857
Clear	1	0.677509
	2	0.322491
Fog/Smog/Smoke	1	0.671353
	2	0.328647
Other	1	0.860577
	2	0.139423
Overcast	1	0.684456
	2	0.315544
Partly Cloudy	2	0.600000
	1	0.400000
Raining	1	0.662815
	2	0.337185
Severe Crosswind	1	0.720000
	2	0.280000
Sleet/Hail/Freezing Rain	1	0.752212
	2	0.247788
Snowing	1	0.811466
	2	0.188534
Unknown	1	0.945928
	2	0.054072

Table 6 - Distribution of accidents based on the severity code whether or not speeding was a factor.

SPEEDING	SEVERITYCODE	
N	1	0.705099
	2	0.294901
Y	1	0.621665
	2	0.378335

Table 7 - Distribution of accidents based on the severity code and road condition

ROADCOND	SEVERITYCODE	
Dry	1	0.678227
	2	0.321773
Ice	1	0.774194
	2	0.225806
Oil	1	0.625000
	2	0.375000
Other	1	0.674242
	2	0.325758
Sand/Mud/Dirt	1	0.693333
	2	0.306667
Snow/Slush	1	0.833665
	2	0.166335
Standing Water	1	0.739130
	2	0.260870
Unknown	1	0.950325
	2	0.049675
Wet	1	0.668134
	2	0.331866

Table 8 - Distribution of accidents based on the severity code and light condition.

LIGHTCOND	SEVERITYCODE	
Dark - No Street Lights	1	0.782694
	2	0.217306
Dark - Street Lights Off	1	0.736447
	2	0.263553
Dark - Street Lights On	1	0.701589
	2	0.298411
Dark - Unknown Lighting	1	0.636364
	2	0.363636
Dawn	1	0.670663
	2	0.329337
Daylight	1	0.668116
	2	0.331884
Dusk	1	0.670620
	2	0.329380
Other	1	0.778723
	2	0.221277
Unknown	1	0.955095
	2	0.044905

Table 9 - Distribution of accidents based on the severity code and collision type.

		SEVERITYCODE
COLLISIONTYPE	SEVERITYCODE	
Angles	1	0.607083
	2	0.392917
Cycles	2	0.876085
	1	0.123915
Head On	1	0.569170
	2	0.430830
Left Turn	1	0.605123
	2	0.394877
Other	1	0.742142
	2	0.257858
Parked Car	1	0.944527
	2	0.055473
Pedestrian	2	0.898305
	1	0.101695
Rear Ended	1	0.569639
	2	0.430361
Right Turn	1	0.793978
	2	0.206022
Sideswipe	1	0.865334
	2	0.134666

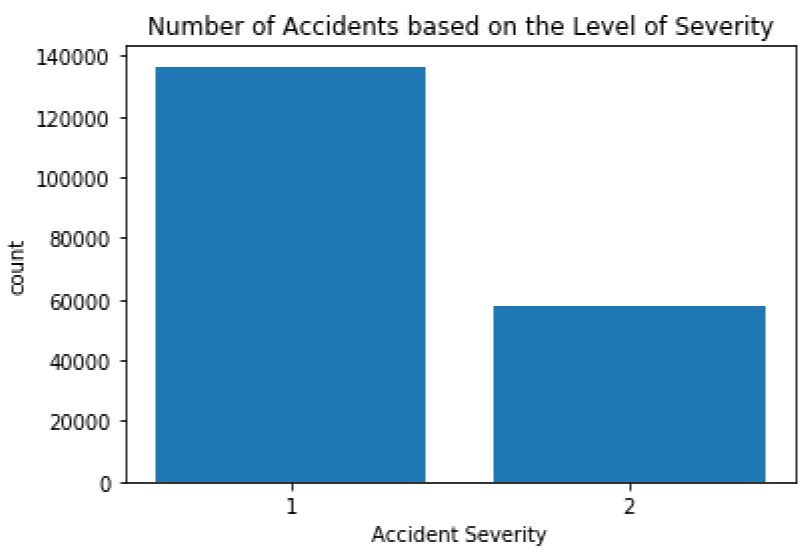

Figure 1 - Distribution of accidents based on the severity code.

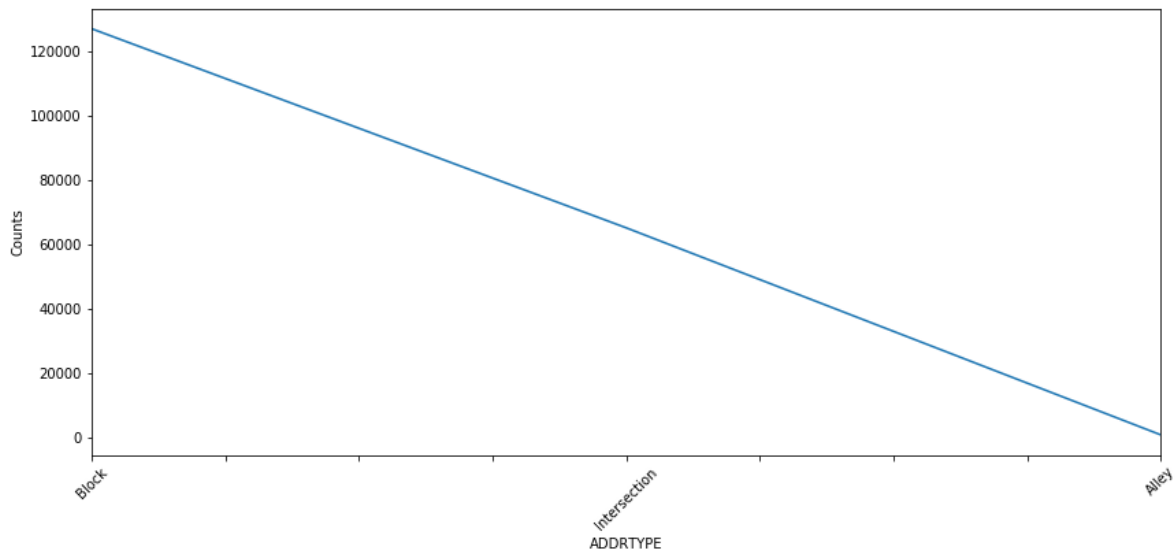

Figure 2 – Number of accidents at various address types.

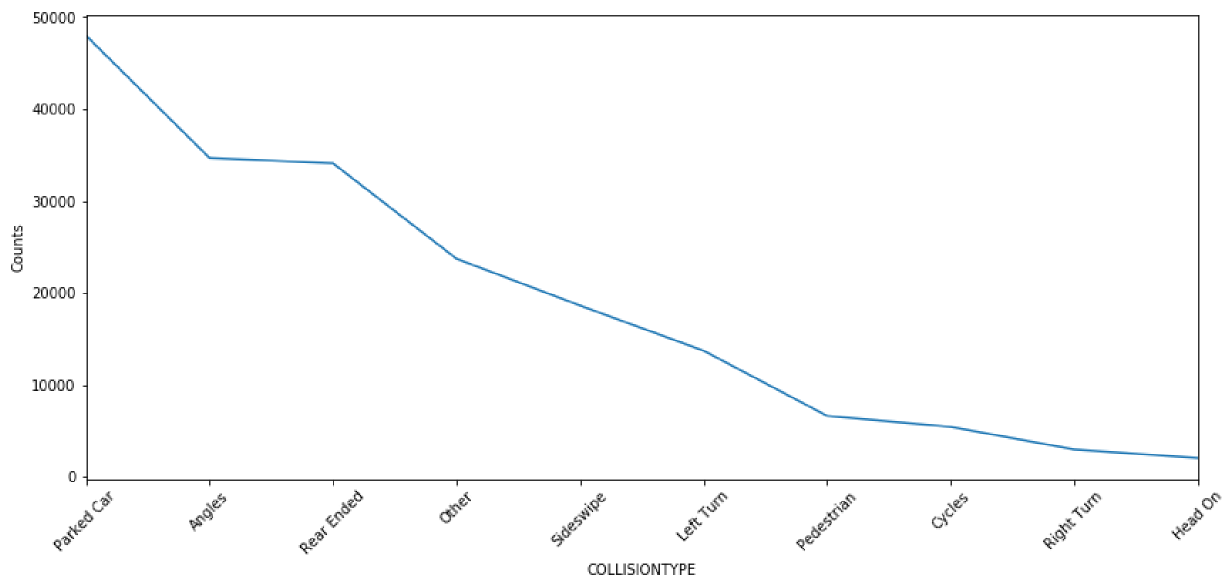

Figure 3 – Number of accidents for various collision types.

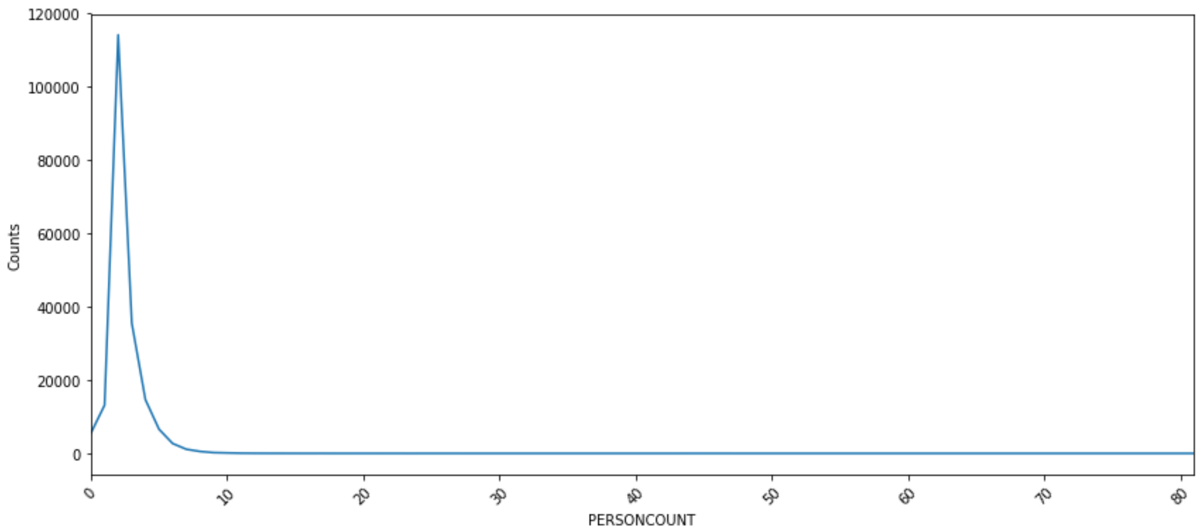

Figure 4 – Number of accidents versus number of people involved in the accident.

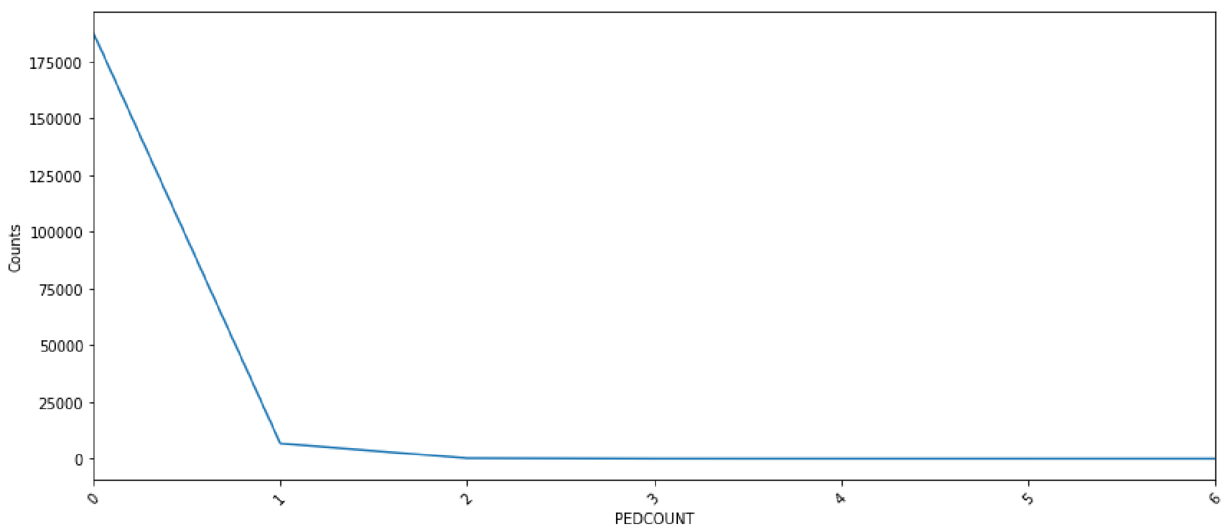

Figure 5 – Number of accidents versus number of pedestrians involved in the accident.

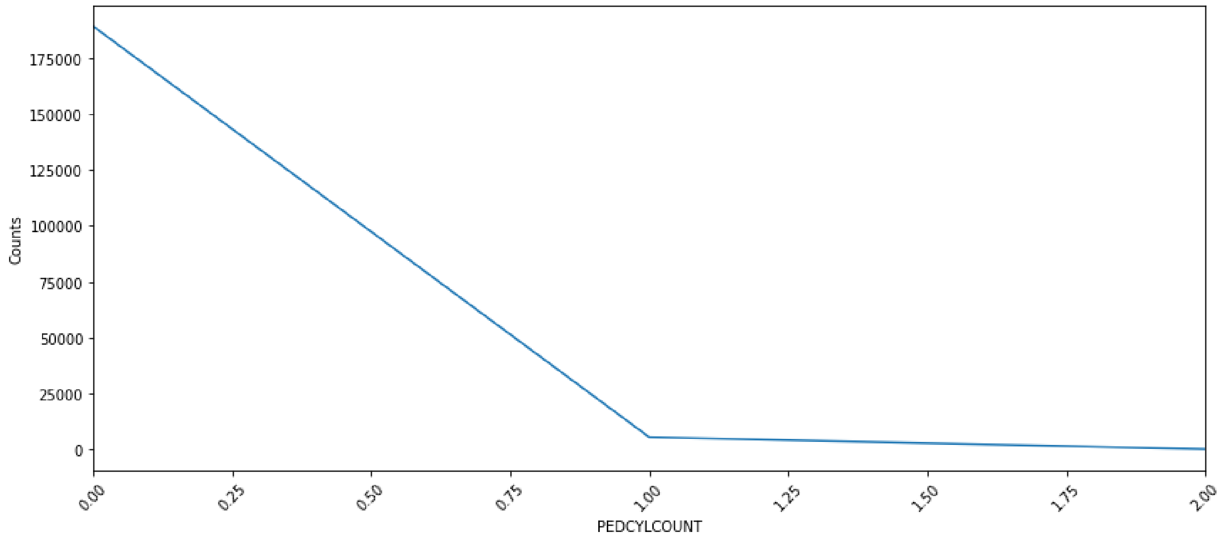

Figure 6 – Number of accidents versus number of cyclists involved in the accident.

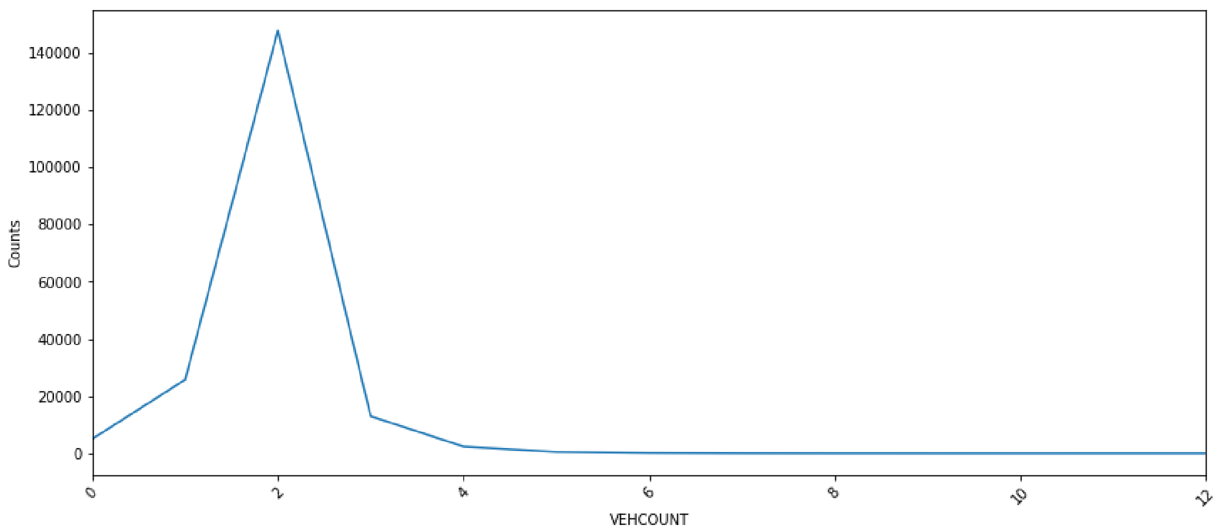

Figure 7 – Number of accidents versus number of vehicles involved in the accident.

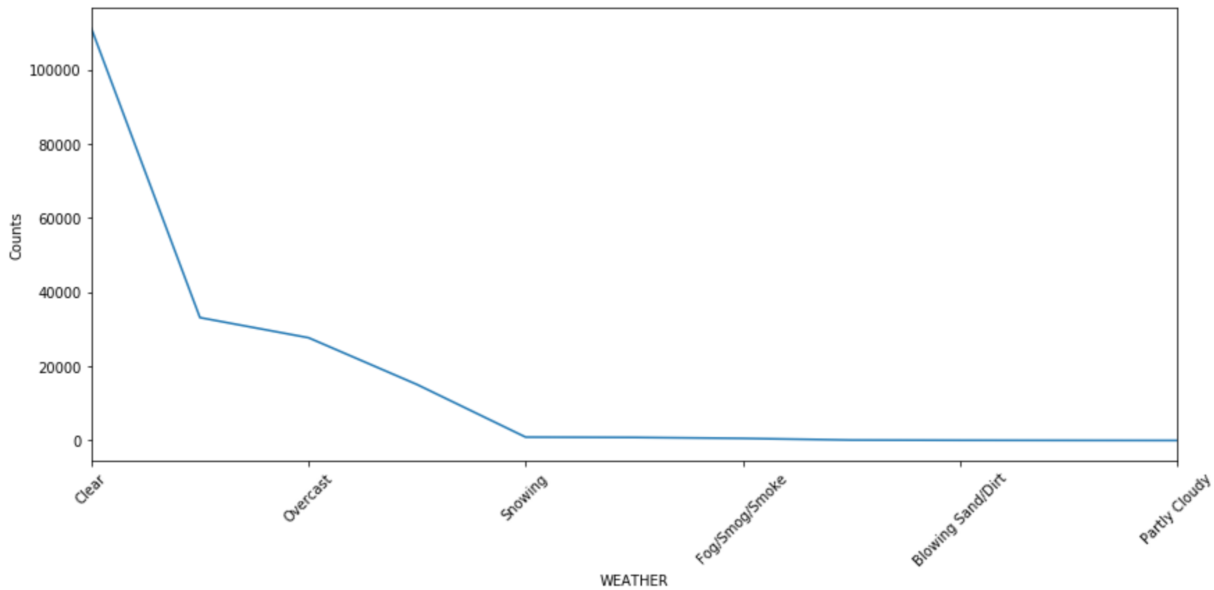

Figure 8 – Number of accidents at weather conditions.

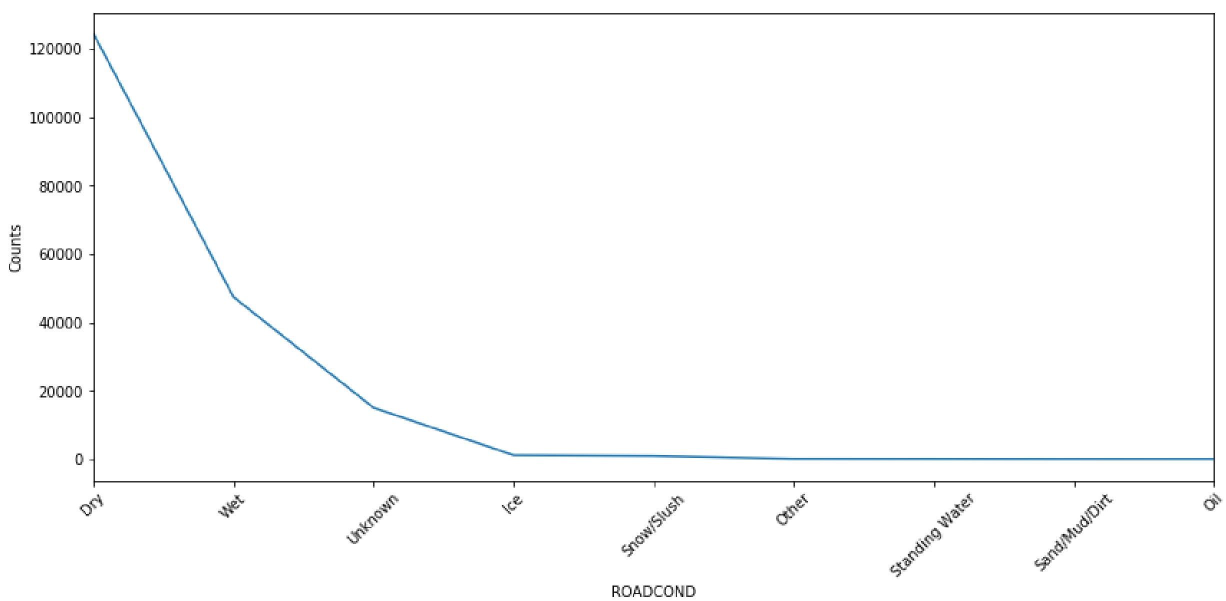

Figure 9 – Number of accidents at road conditions.

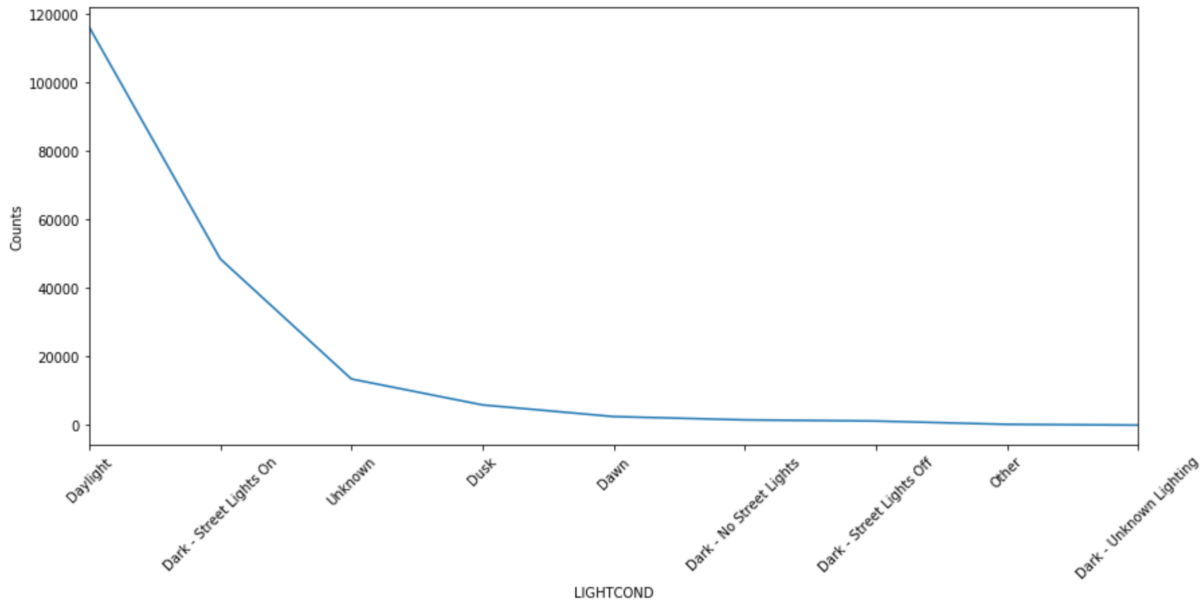

Figure 10 – Number of accidents at light conditions.

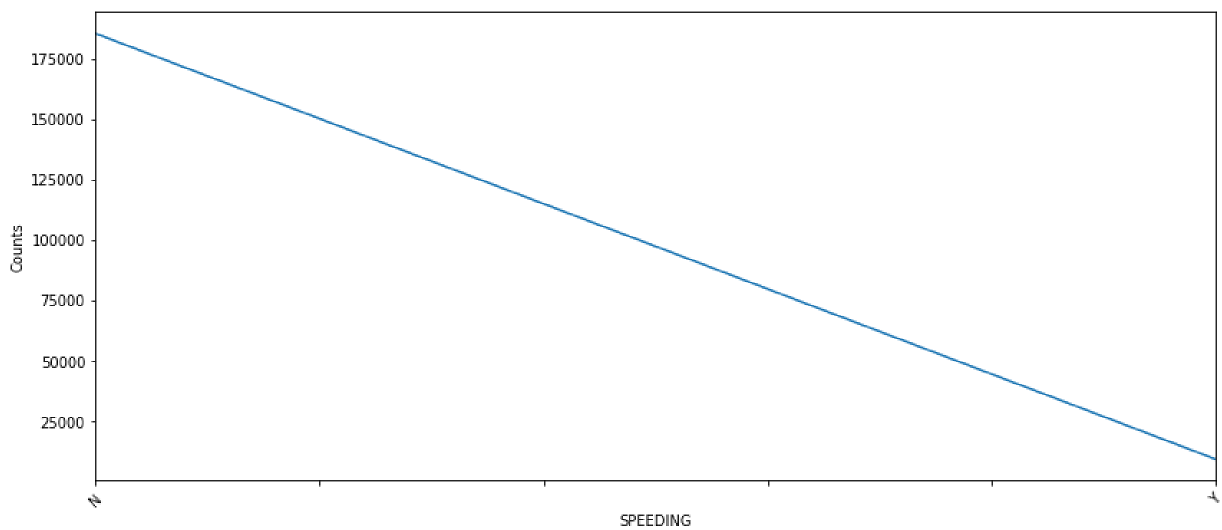

Figure 11 – Number of accidents whether or not involved speeding.

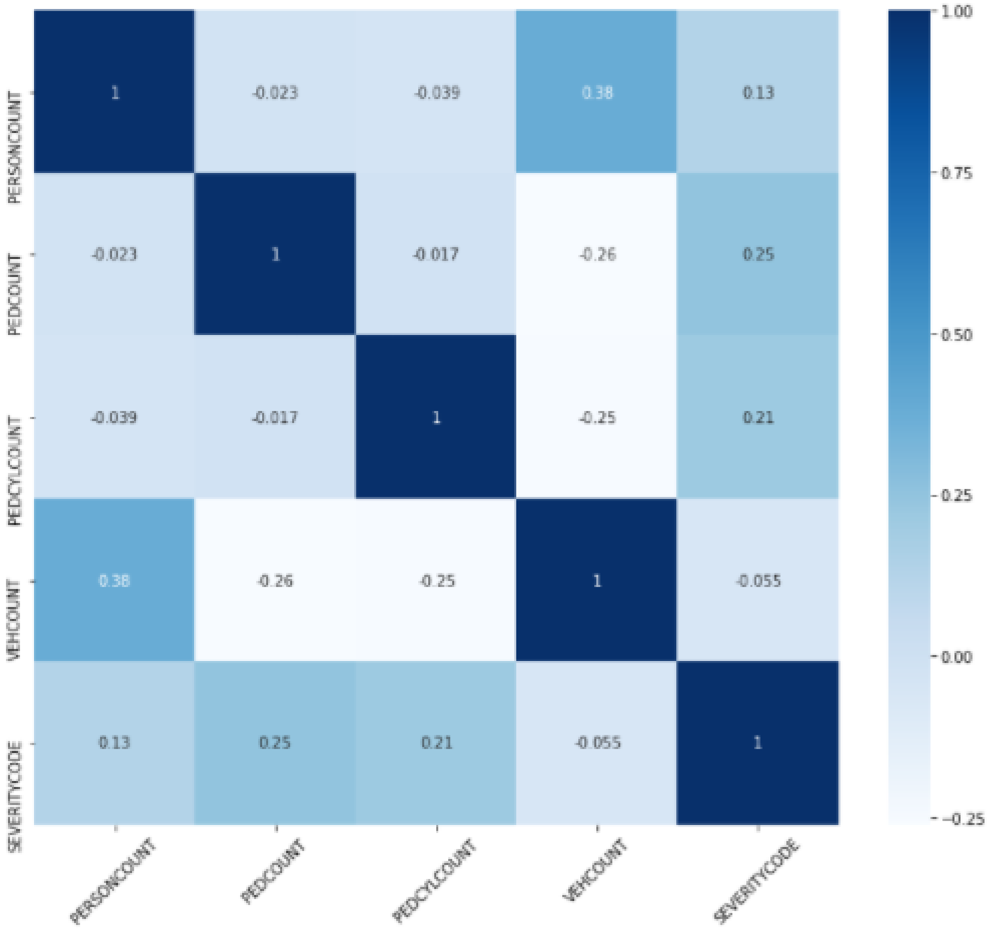

Figure 12 – Heatmap showing the correlation between numerical values in the dataframe.

Modeling

After careful evaluation of the dataset and better understanding the relationship between different features, it is the time to find the best predictive model. For this purpose, four different model would be applied on the data set and will be evaluated using various metrics to find the best option. First, it is required to use one hot encoding to convert categorical data to numerical values for modeling. Then, the train/test split is used on data which returns 4 different parameters. We will name them: X_trainset, X_testset, y_trainset, y_testset.

The train_test_split will need the parameters: X, y, test_size=0.2, and random_state=4.

The X and y are the arrays required before the split, the test_size represents the ratio of the testing dataset, and the random_state ensures that we obtain the same splits.

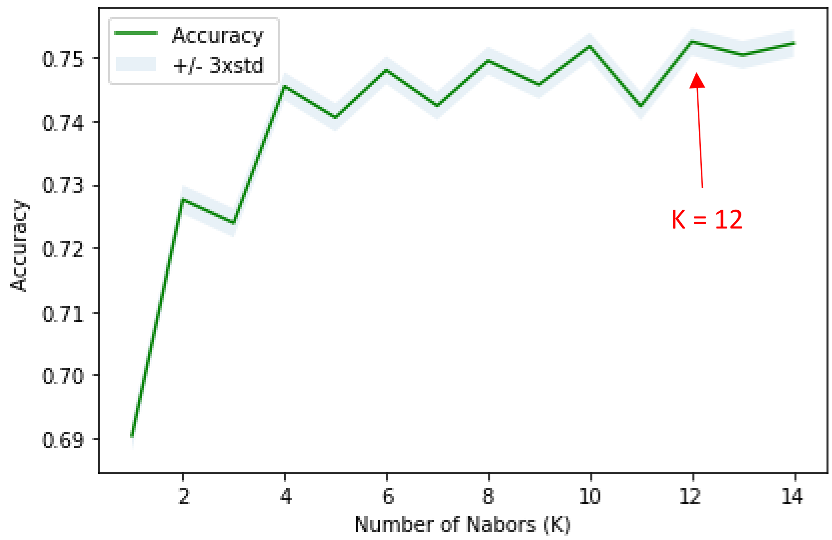

Figure 13 – Accuracy of the KNN algorithm for k values between 0 and 14.

Algorithms used in this project are:

1. K Nearest Neighbor (KNN)

2. Decision Tree

3. Support Vector Machine

4. Logistic Regression

Model Evaluation

The models built can be evaluated using various metrics such as Jacard Similarity Score, f1 Score and Log Loss.

Algorithm	Jaccard	F1-Score	Log Loss
KNN	0.75	0.72	NA
Decision Tree	0.75	0.69	NA
SVM	0.76	0.71	NA
Logistic Regression	0.76	0.72	0.48

Results

By looking at the evaluation results, we can see that most accidents are labeled as category “1” or less sever accidents. Results also show that most accidents happen at the block type address, however, they are less severe than the accidents happening at intersections. Most accidents reported here did not involve speed violation. Also, number of pedestrian, cyclist, and persons involved in the accident have positive relationship with the severity level of the accident and all accidents have less than 10 people involved and mostly no pedestrian. Most collisions are labeled as parked car type and lastly, most accidents happen at clear weather and dry road condition during daylight hours.

All four models predicted the severity code for the test set with relatively high accuracy score. KNN and SVM algorithms required longer computation time compared to the other two methods.

Discussion

Four different algorithms were applied on the filtered data set to find the most optimized predictive model for this data set. Data set was split into training and testing sets (80%, 20%) to compare the predicted data labels (severity code) with the recorded values. As shown here, all models have similar Jaccard accuracy values, but Decision Tree and Logistic Regression have comparatively less computation times and would be the optimum models for future predictions. Although some data columns were eliminated from the data set mostly due to high number of missing data, but the final attributes selected for the modeling seem to be the most relevant and reliable data. It would be recommended to collect additional data with less missing information as this could help improve the model accuracy.

Conclusion

According to this data set on collisions in Seattle from 2004 to the present, it is concluded that there is no particular relationship between bad weather, light and road conditions that affect collisions. It is observed that there were a lot more collisions that happened on dry roads and clear weather conditions during day time compared to when conditions were not ideal. This could mean that drivers tend to be more careful in driving in adverse weather, road and light conditions. The data shows that drivers are more likely to have a collision when weather conditions are good and roads are dry.

Author suggests using this model to predict a possibility of sever car accidents based on the time and location of driving and provide this information to the drivers. Having this information could help drivers choose the safest route for their travel and reduce the number of dangerous car accidents.