IBM Capstone Project – Allegheny County / City of Pittsburgh 2017 Crash Data

Introduction

Allegheny County in the City of Pittsburgh government official wants to analyze the data collected for car accidents. They want to know by using data science methodologies and algorithms of the circumstances recorded, whether there is a way to predict and decrease the frequency of accidents. The decrease of accidents can help city officials better plan their personnel hours and resources where needed most. The goal is to plan resources more efficiently and effectively based on the predicted severity type of an accident. If an accident severity is level 1 (killed) or 2 (major injury), then more resources should be directed for accident assistance or to educate the drivers.

Data Source

The historical data of the crash data in 2017 contains locations and information about every crash incident reported to the police in Allegheny Country in 2017. This will be used to train and test the developed model.

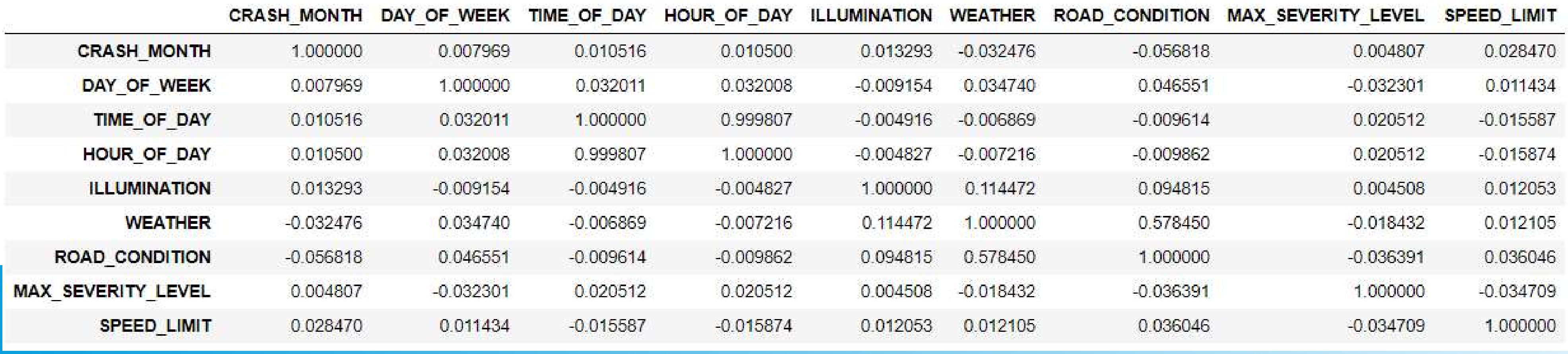
The link of the data can be found in: https://data.wprdc.org/datastore/dump/bf8b3c7e-8d60-40df-9134-21606a451c1a

Data Understanding

- This data will be used as an input to build a machine learning mechanism to predict the severity of the types of the accident based on key factors that might caused the accident or contribute to the accidents. The key factors can include “time of the day” “month” “day of the week,” “road conditions”, “illumination”, etc. The prediction is the “Max_Severity_Level”.
- There are 12537 rows with 190 columns in this data.

Data Processing

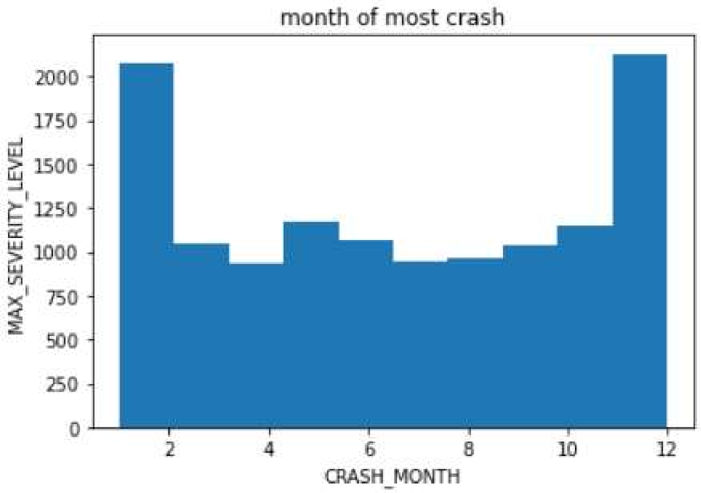
- Clean up the data
  - There are columns that are not needed information
  - There are missing data or mislabeled data in cell
- A broad view on the correlation between all the remaining data



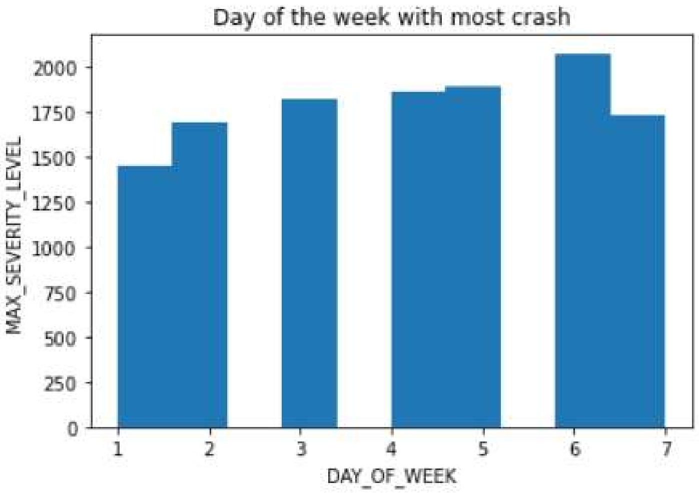
	CRASH_MONTH	DAY_OF_WEEK	TIME_OF_DAY	HOUR_OF_DAY	ILLUMINATION	WEATHER	ROAD_CONDITION	MAX_SEVERITY_LEVEL	SPEED_LIMIT
CRASH_MONTH	1.000000	0.007969	0.010516	0.010500	0.013293	-0.032476	-0.056818	0.004807	0.028470
DAY_OF_WEEK	0.007969	1.000000	0.032011	0.032008	-0.009154	0.034740	0.046551	-0.032301	0.011434
TIME_OF_DAY	0.010516	0.032011	1.000000	0.999807	-0.004916	-0.006869	-0.009614	0.020512	-0.015587
HOUR_OF_DAY	0.010500	0.032008	0.999807	1.000000	-0.004827	-0.007216	-0.009862	0.020512	-0.015874
ILLUMINATION	0.013293	-0.009154	-0.004916	-0.004827	1.000000	0.114472	0.094815	0.004508	0.012053
WEATHER	-0.032476	0.034740	-0.006869	-0.007216	0.114472	1.000000	0.578450	-0.018432	0.012105
ROAD_CONDITION	-0.056818	0.046551	-0.009614	-0.009862	0.094815	0.578450	1.000000	-0.036391	0.036046
MAX_SEVERITY_LEVEL	0.004807	-0.032301	0.020512	0.020512	0.004508	-0.018432	-0.036391	1.000000	-0.034709
SPEED_LIMIT	0.028470	0.011434	-0.015587	-0.015874	0.012053	0.012105	0.036046	-0.034709	1.000000

Time factor for accident rates

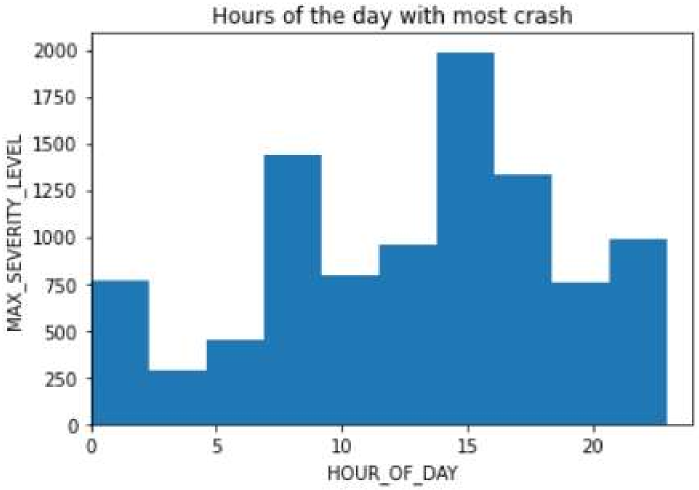
The historical data of the crash data in 2017 contains times, hour and month or days of the week for each and every crash. Therefore we plot them to get a look at whether there are obviously correlation or they can be explained by logic.



month of most crash



Day of the week with most crash



Hours of the day with most crash

Methodology

After looking at the time correlation to the accident rate, now we will build and compare using the following Machine Learning techniques:

Logistic Regression

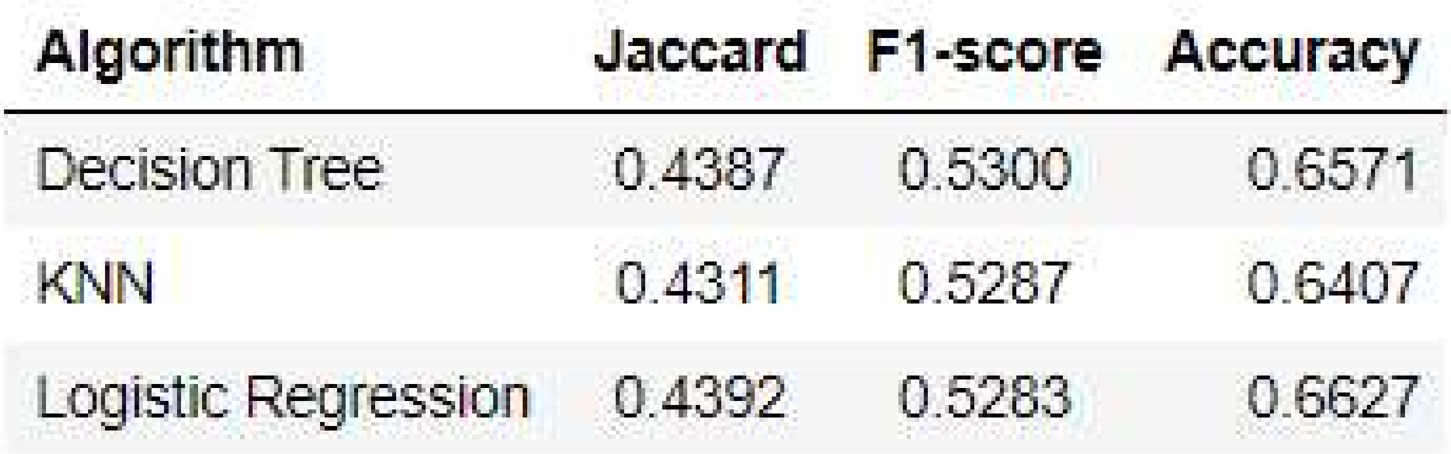
K-Nearest Neighbor (KNN)

Decision Tree

Testing and Training Set

- The data is also split between the Training set and the testing set.
- Decision Tree - use 6837 as training set and 2931 as testing set.
- KNN - use 5860 as training set and 3908 as testing set.
- Logistic regression - use 5860 as training set and 3908 as testing set.

Result



Algorithm	Jaccard	F1-score	Accuracy
Decision Tree	0.4387	0.5300	0.6571
KNN	0.4311	0.5287	0.6407
Logistic Regression	0.4392	0.5283	0.6627

Conclusion

- The comparison of the results indicated that the Decision Tree was the better prediction tool in terms of the higher result in accuracy.