PREDICTING THE PROBABILITY OF PROPERTY OR PHYSICAL DAMAGE DURING A ROAD TRIP

By: Darnel Lloyd

INTRODUCTION

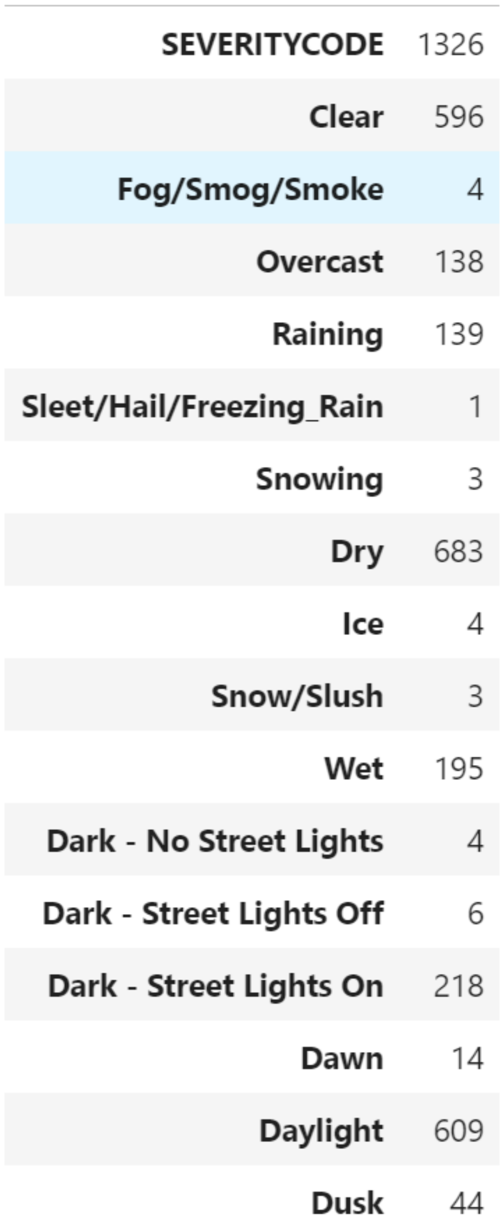
Predicting the probability of property or physical damage during a road trip.

- It is estimated that yearly there are more than 6 million road accidents, with 2 million plus injuries and over 37 thousand deaths.
- Numerous factors influence this statistics including road condition, vehicle condition, weather, driver’s condition etc.
- Calculation the likelihood of an accident can provide drivers with an awareness of what to expect.

DATA SELECTION AND CLEANING

- The data used for my research contains vehicle collision in Seattle from the year 2004 to present and can be found at : https://s3.us.cloud-object-storage.appdomain.cloud/cf-courses-data/CognitiveClass/DP0701EN/version-2/Data-Collisions.csv

- It contains roughly 194673 rows and 38 columns.
- A subset of the data set containing 1000 rows were chosen.
- The initial columns *SEVERITYCODE, LIGHTCOND, WEATHER* and *ROADCOND* where selected as relevant columns.
- The subentries where concated and the final list of columns were *'SEVERITYCODE', 'Clear', 'Fog/Smog/Smoke', 'Overcast', 'Raining', 'Sleet/Hail/Freezing_Rain', 'Snowing', 'Dry', 'Ice', 'Snow/Slush', 'Wet', 'Dark - No Street Lights', 'Dark - Street Lights Off', 'Dark - Street Lights On', 'Dawn', 'Daylight', 'Dusk']*
- The dependent variable "SEVERYCODE" represent values 1 (property damage) and 2 (injury).

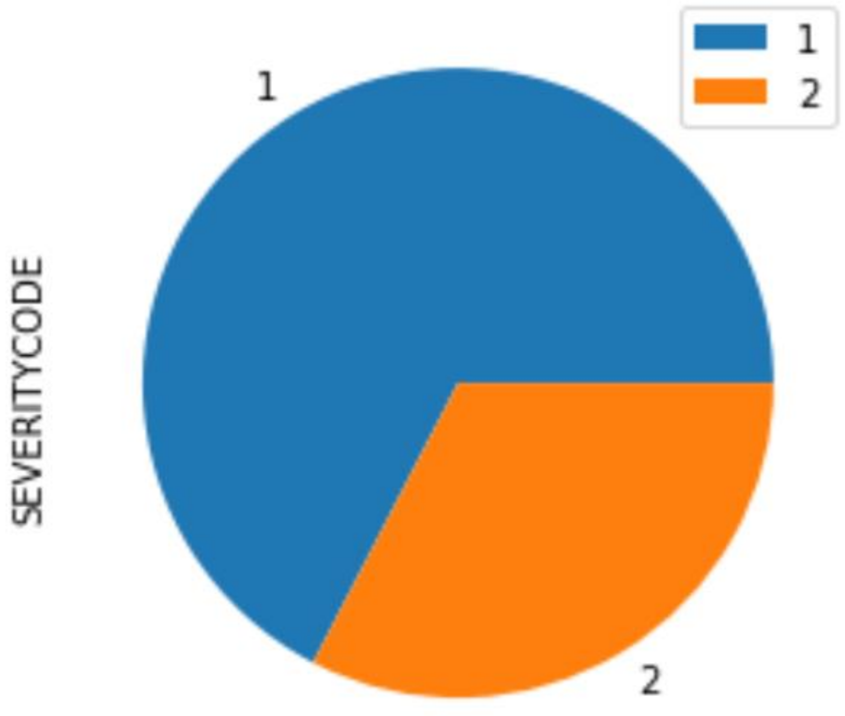


SEVERITYCODE	1326
Clear	596
Fog/Smog/Smoke	4
Overcast	138
Raining	139
Sleet/Hail/Freezing_Rain	1
Snowing	3
Dry	683
Ice	4
Snow/Slush	3
Wet	195
Dark - No Street Lights	4
Dark - Street Lights Off	6
Dark - Street Lights On	218
Dawn	14
Daylight	609
Dusk	44

Analysis and visualizations

- We notice that are few features in the data set appear less frequently in our data sample. These features where kept in order to take into account actual, real life road conditions even if they occur rearly.

- When it comes to the severity code, we notice a fairly distributed amount between injury and property damage. Therefore the model will have sufficient data from each variable to train on.



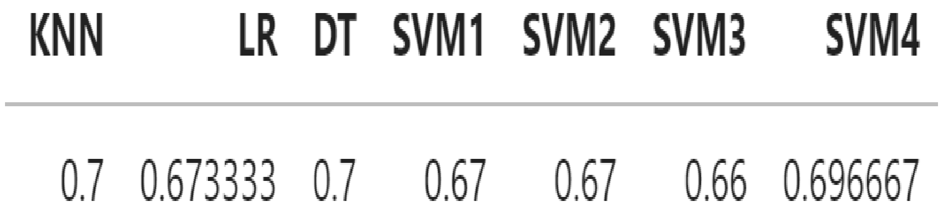
Modelling and Accuracy

Given that we are trying to predict a categorical variable (injury or property damage), several machine learning clustering algorithms were tested. We utilized:

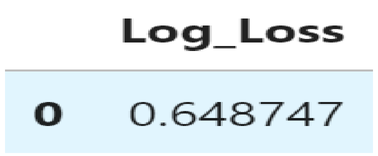
- K Nearest Neighbour (KNN); with an optimum value of k=22 used.
- Decision Tree (DT); with 1 branch.
- Logistic regression (L_R); with a solver=liblinear.
- Support Vector machine (SVM); using different kernels including rbf, linear, ploy and sigmoid.

- The following accuracies from the various models were obtained. KNN and DT produced the best results.

- A log loss of 0.648747 was also obtained.



KNN	LR	DT	SVM1	SVM2	SVM3	SVM4
0.7	0.673333	0.7	0.67	0.67	0.66	0.696667



	Log_Loss
0	0.648747

Conclusion

The initial aim of this project was to find an adequate model in order to predict the likelihood of physical or property damage while on a road trip. Form the research we can conclude that a classification model using K Nearest Neighbour (K=22) or a Decision Tree will achieve the best predictive results.