Predicting the severity of car crashes with Machine Learning

Introduction

Car crashes impact every industrialized economy massively. Nearly 1.25 million people die in road crashes each year, on average, 3,287 deaths a day. Moreover, 20–50 million people are injured or disabled annually. Road traffic crashes rank as the 9th leading cause of death and accounts for 2.2% of all deaths globally. Road crashes cost USD 518 billion globally, costing individual countries from 1–2% of their annual GDP.

In the USA, over 37,000 people die in road crashes each year, and 2.35 million are injured or disabled. Road crashes cost the U.S. $230.6 billion per year or an average of $820 per person. Road crashes are the single greatest annual cause of death of healthy U.S. citizens travelling abroad
Looking at the severity of this problem, I decided to analyse the accidents' data to discover something useful. And here I am, sharing my findings.

The question that I am trying to address here is around determining the severity of an accident (1 or 2) based on factors like *road and weather conditions, lighting, address type, etc.*
The Traffic Police Department does take strict measures with aim of making driving the safest possible and ensure minimum collisions. If we can help extract some meaningful insight out by analysing real time collisions data, it might come in handy to facilitate the traffic department to devise better strategies and alerts around safe driving rules. This data, post pre-processing, will be combined, with detailed predictive models using machine learning techniques to improvise accuracy and classify future likeability of accidents as well as help label them into one of the two severity codes.

Data

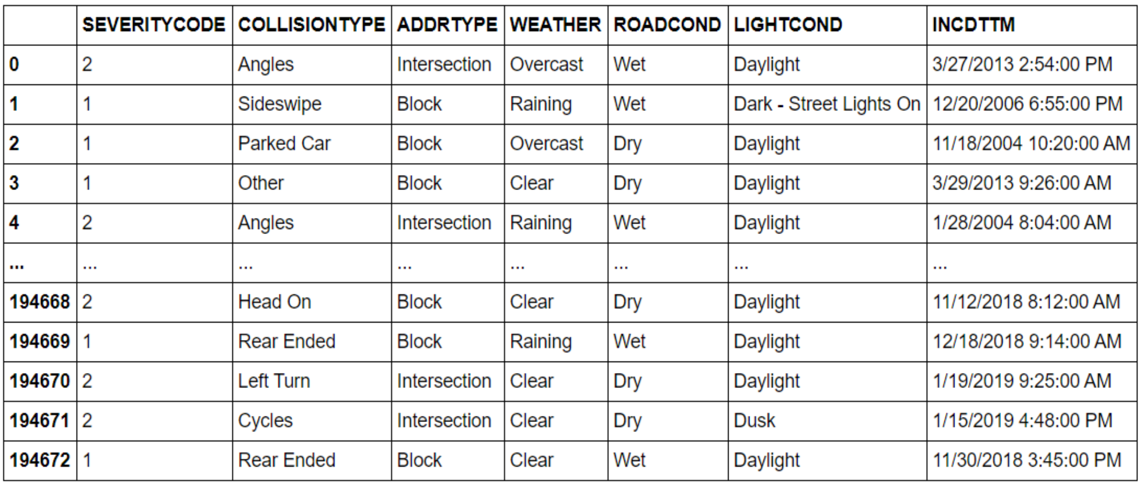
The data used for this analysis is provided by the Seattle Department of Transportation, which includes 194,673 real-cases of accidents reported from 2004 to 2020. There are 37 attributes, including but not limited to, address type, junction, location, incident timings, road, weather and lighting conditions. Each row contains a primary key of interest i.e. the severity code where 1 depicts low and 2 depicts high leading to possible fatality of the driver.

In order to execute my problem, I will (1) clean the data to remove excess columns, replace NaN data values, or clear empty columns. Then, I will (2) ensure all data types of each data is correct and (3) create multiple regression to see which variables most impact the accidence severity code and then finally use or do (4) KNN or Decision Tree or Logistic Regression to classify new conditions as either 1 or 2 in severity code.

As an introductory example, ROADCOND is likely to be important. When road conditions are wet, the risk of accident increases to 2. So, I will analyze it as a part of classification process while also doing accuracy evaluations depending on the machine learning technique deployed. If it is a decision tree, I will attempt to use entropy and gain to get the best ordering of different variables in classification and finally use accuracy metrics to provide evaluation on the test data.

Methodology

The first step taken for implementing this ML based solution was to download the dataset from Week1 of the course under Cognitive Class and saved it as a *Pandas* dataframe in a notebook. It can be seen clearly from retrieving the head of the data that the first column is severity code which is binary in terms of 1 or 2. Since it clearly defines the severity of an accident, this will be the main variable of study i.e. X.

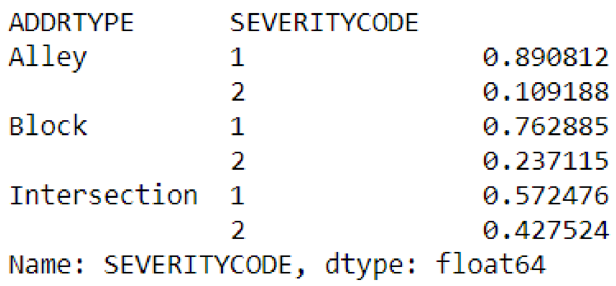


	SEVERITYCODE	COLLISIONTYPE	ADDRTYPE	WEATHER	ROADCOND	LIGHTCOND	INCDTTM
0	2	Angles	Intersection	Overcast	Wet	Daylight	3/27/2013 2:54:00 PM
1	1	Sideswipe	Block	Raining	Wet	Dark - Street Lights On	12/20/2006 6:55:00 PM
2	1	Parked Car	Block	Overcast	Dry	Daylight	11/18/2004 10:20:00 AM
3	1	Other	Block	Clear	Dry	Daylight	3/29/2013 9:26:00 AM
4	2	Angles	Intersection	Raining	Wet	Daylight	1/28/2004 8:04:00 AM
...	...	...	...	...	...	...	...
194668	2	Head On	Block	Clear	Dry	Daylight	11/12/2018 8:12:00 AM
194669	1	Rear Ended	Block	Raining	Wet	Daylight	12/18/2018 9:14:00 AM
194670	2	Left Turn	Intersection	Clear	Dry	Daylight	1/19/2019 9:25:00 AM
194671	2	Cycles	Intersection	Clear	Dry	Dusk	1/15/2019 4:48:00 PM
194672	1	Rear Ended	Block	Clear	Wet	Daylight	11/30/2018 3:45:00 PM

In the process, there were certain variables that were ***omitted***:
Fields that were descriptive but were either hard to analyze/categorize and not descriptive of the event were omitted from this study.
These include namely attributes such as X or Y location, status, intkey, objectid, coldetkey, crosswalkkey, etc. majorly, these were keys to help link the dataset with other datasets which were not used and hence were not relevant for this study. On the other hand, I decided to exclude junction type since it was like address type and to avoid similar data columns. Moreover, While the location might be relevant, it is not a good indicator because it might lead people to avoid a certain area, however, the goal of this study is maximizing public safety regardless of the location.

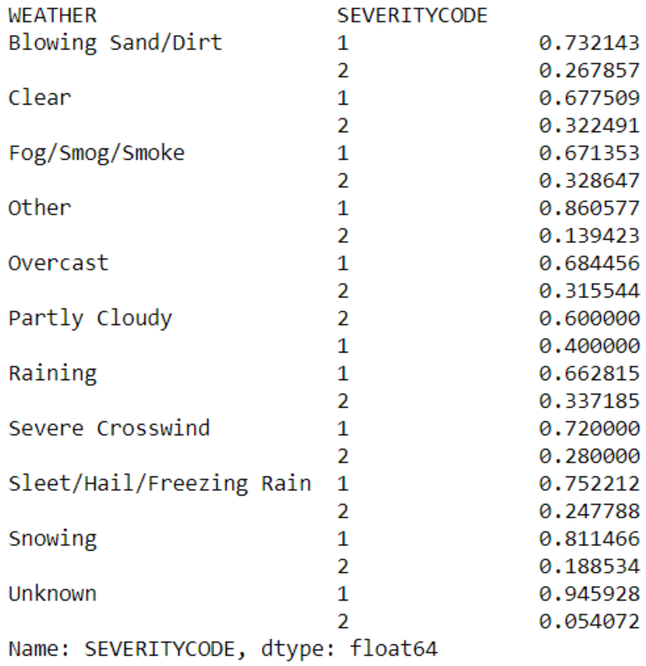
The independent feature that have been shortlisted to train the model to predict accident severity are as mentioned below

- ADDRTYPE – Alley, Block, or Intersection. This could potentially show which types of common regions are harder to drive by safely



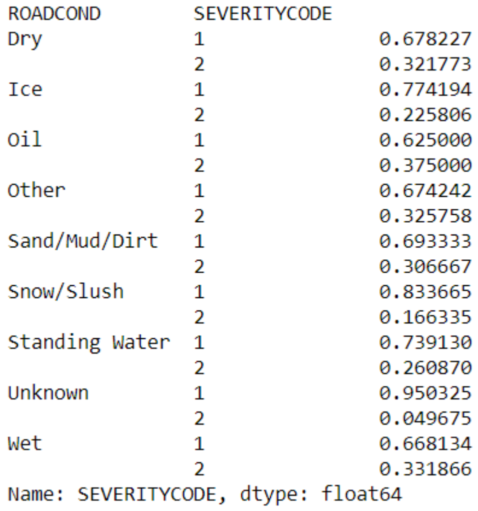
```
ADDRTYPE       SEVERITYCODE
Alley          1               0.890812
               2               0.109188
Block          1               0.762885
               2               0.237115
Intersection   1               0.572476
               2               0.427524
Name: SEVERITYCODE, dtype: float64
```

- WEATHER- Cloudy, Rainy, Sunny, Windy, etc. it has multi-type classification, making it remarkably simple to use and implement. As well as that, weather is in general an important indicator of how smooth one drives.



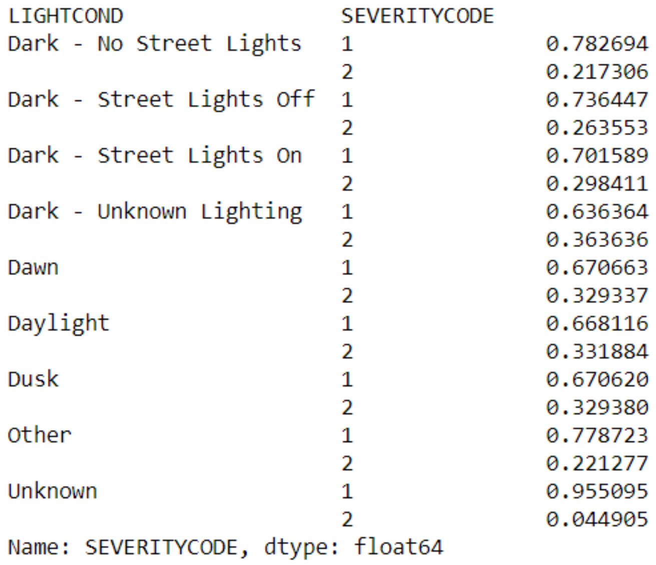
```
WEATHER                   SEVERITYCODE
Blowing Sand/Dirt         1               0.732143
                          2               0.267857
Clear                     1               0.677509
                          2               0.322491
Fog/Smog/Smoke            1               0.671353
                          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
Name: SEVERITYCODE, dtype: float64
```

- ROADCOND: this is again multi-class classification and helpful in cases where road conditions affect driving.



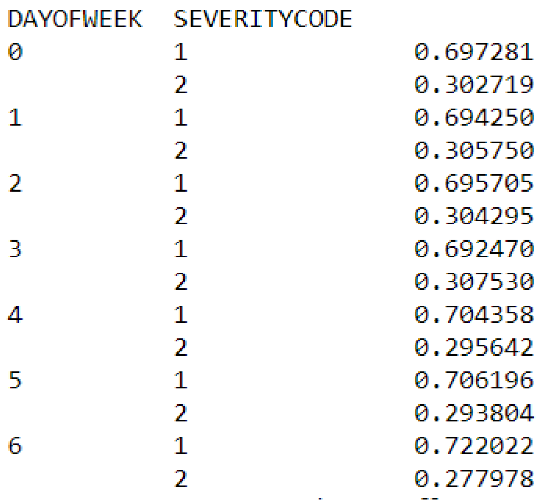
```
ROADCOND        SEVERITYCODE
Dry             1               0.678227
                2               0.321773
Ice             1               0.774194
                2               0.225806
Oil             1               0.625000
                2               0.375000
Other           1               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
Name: SEVERITYCODE, dtype: float64
```

- LIGHTCOND – The light conditions during the collision. This can help determine a correlation between how severe the collision could depend on the lighting conditions of the road.



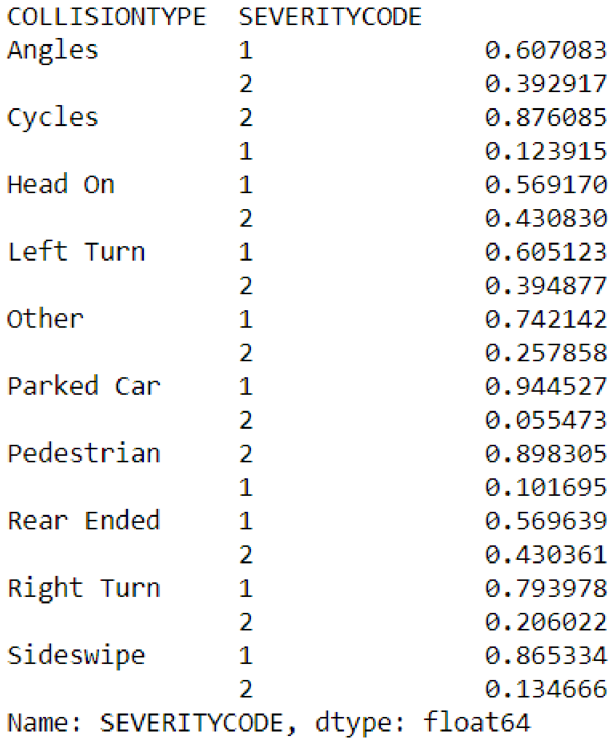
```
LIGHTCOND                 SEVERITYCODE
Dark - No Street Lights   1               0.782694
                          2               0.217306
Dark - Street Lights Off  1               0.736447
                          2               0.263553
Dark - Street Lights On   1               0.701589
                          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
Name: SEVERITYCODE, dtype: float64
```

- INCIDENT TIME – It is a critical parameter as it can help determine the frequency pattern between the time of the day and the severity of the collision. In general, driving at night-time during rush hours makes one more prone to accidents.



```
DAYOFWEEK  SEVERITYCODE
0          1               0.697281
           2               0.302719
1          1               0.694250
           2               0.305750
2          1               0.695705
           2               0.304295
3          1               0.692470
           2               0.307530
4          1               0.704358
           2               0.295642
5          1               0.706196
           2               0.293804
6          1               0.722022
           2               0.277978
```

- COLLISION TYPE: This variable can help develop a pattern between the severity of the collision and the collision type



```
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
Name: SEVERITYCODE, dtype: float64
```

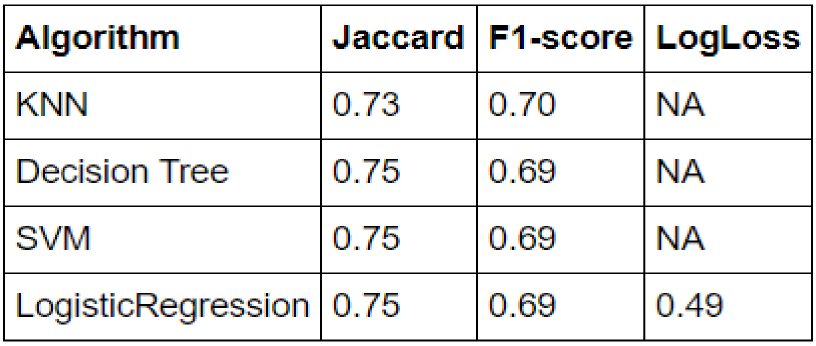
Data Preparation

The first and foremost step was to adjust the data types. Next, post initial skimming through dataset was to shortlist the into the above given chosen fields and severity code. Then, testing and training dataset split was done to train the model. Since the dataset is huge, for memory efficiency, I assigned 20% for test data. This is to make sure that the data model is trained comprehensively on training data but is not over sensitive to test data.

Post applying data transformations and refining the data, the training data was trained against 4 Machine learning techniques i.e., Support Vector Machine, K-Nearest Neigbors, Logistic Regression, and Decision Tree. Post training the model, evaluation was done for each technique using Jaccard Index, F1 score and logloss (Only for logistic regression)

Results

Using the less-time taking test and training split, the following accuracy metrics were received.



Algorithm	Jaccard	F1-score	LogLoss
KNN	0.73	0.70	NA
Decision Tree	0.75	0.69	NA
SVM	0.75	0.69	NA
LogisticRegression	0.75	0.69	0.49

Based on these observations, we can clearly infer that all of them yield remarkably similar accuracy. It was a matter of what technique runs the fastest and it turned out that logistic regression ran the quickest with accuracy of ~75%

Discussion

The results as shown above are remarkably interesting. Given the huge dataset and quite a lot of exceptions to fast time and memory space, the accuracy is extremely high and surprising.

The logistic regression seems to fit the best, and based on these results, these recommendations can be assumed:
1) Environmental factors such as road conditions, weather conditions are indeed remarkably close predictor of an accident severity. Hence, monitoring and controlling these can help to reduce fatalities significantly.
2) Incorporating these predictors and a safe guide as part of a marketing campaign can be extremely powerful in maximizing public safety
3) Further powerful geospatial data and cholropeth maps can be used to include even the location of the data
4) Stronger and faster database softwares and more parameters can be helpful to analyze and make larger, generalized inferences.

Conclusion

As initiated and purposed, the study is effective in predicting or classifying future records and parameters to determine efficiently the accident severity. These results, if not extensive, can be helpful, to further develop and implement in the form of a basic marketing campaign for the Seattle public to reduce fatalities and severe injuries. Moreover, the data of 200,000 collisions is a large database with 75% accuracy to incorporate in GPS tracking (using geo spatial data), or automobile functioning to alarm the driver when the road conditions or any paramters become risky.