Predicting Automobile Accident Severity

- *Shreya Goyal*

1. Introduction

As countries and cities develop and industrialize, traffic has become one of the key components of our lives. While the hustle-bustle denote the fast-pace of our lives and progress, it can in many cases lead to car collisions and fatalities.

Today, analyzing what variables impact these collisions, I will be investigating:

Question: How to determine automobile accident severity (1 or 2) based on factors such as road and weather conditions, lighting, address type, etc.

As such, given this goal, traffic police of different regions take rigorous steps to make drivers aware of safe driving rules to avoid such collisions. Given this interest, this report is intended to analyze and interpret this very question using ample real-data. 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.

These models, will target audiences such as existing and potential drivers in terms of safety mechanisms and things-to-keep-in-mind before driving to ensure maximum public and personal safety.

2. 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, indient timings, and road weather or lighting conditions.

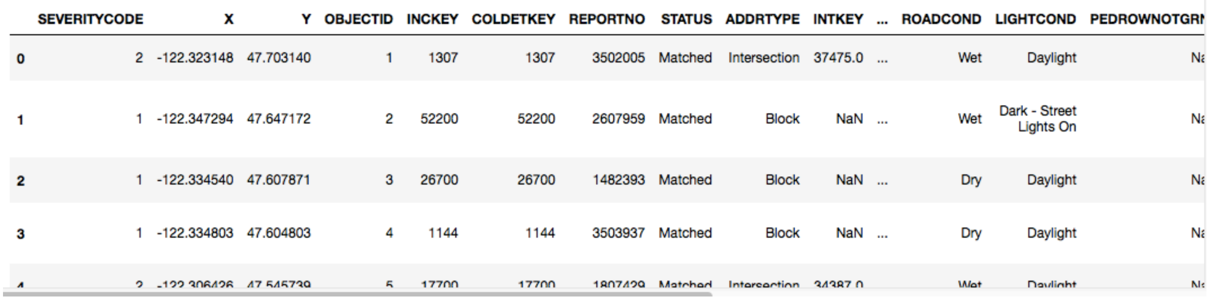
Each row contains a primary key of interest i.e. the severity code where 1 depicts low and 2 depicts high leading to possible fatality of the driver.

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.

3. Methodology

To execute this task, I first downloaded 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 my main variable of study i.e. X.



	SEVERITYCODE	X	Y	OBJECTID	INCKEY	COLDETKEY	REPORTNO	STATUS	ADDRTYPE	INTKEY	...	ROADCOND	LIGHTCOND	PEDROWNOTGR
0	2	-122.323148	47.703140	1	1307	1307	3502005	Matched	Intersection	37475.0	...	Wet	Daylight	Na
1	1	-122.347294	47.647172	2	52200	52200	2607959	Matched	Block	NaN	...	Wet	Dark - Street Lights On	Na
2	1	-122.334540	47.607871	3	26700	26700	1482393	Matched	Block	NaN	...	Dry	Daylight	Na
3	1	-122.334803	47.604803	4	1144	1144	3503937	Matched	Block	NaN	...	Dry	Daylight	Na
4	2	-122.306426	47.545739	5	17700	17700	1807429	Matched	Intersection	34387.0		Wet	Daylight	Na

In the process, I omitted as well as chose to include certain variables, such as:

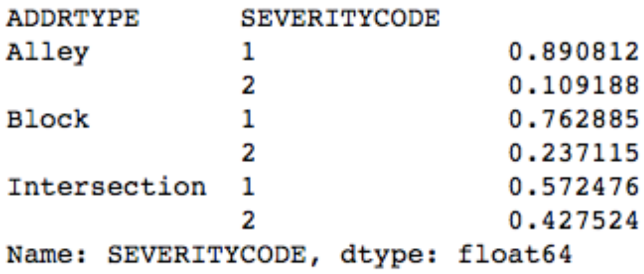
OMITTED

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 similar to 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 maximize public safety regardless of the location.

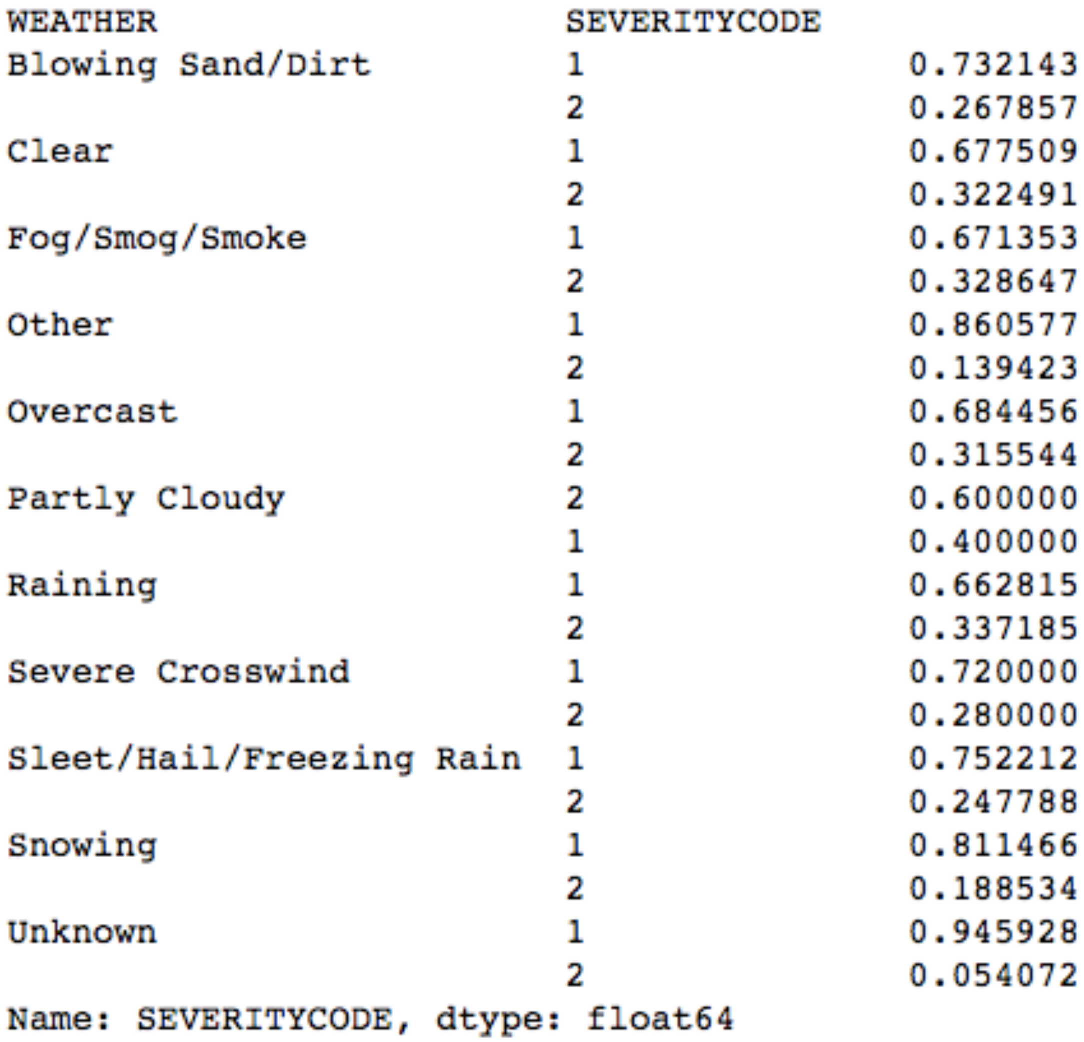
CHOSEN

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



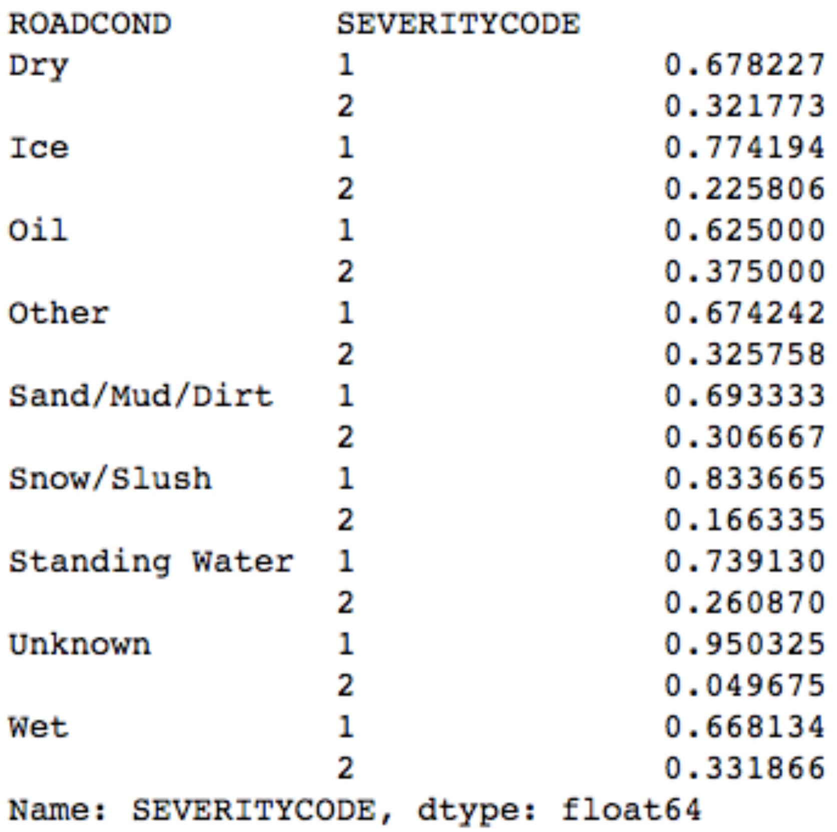
```
ADDRTYPE      SEVERITYCODE
Alley         1               0.890812
              2               0.109188
Block         1               0.762885
              2               0.237115
Intersection  1               0.572476
              2               0.427524
Name: SEVERITYCODE, dtype: float64
```

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



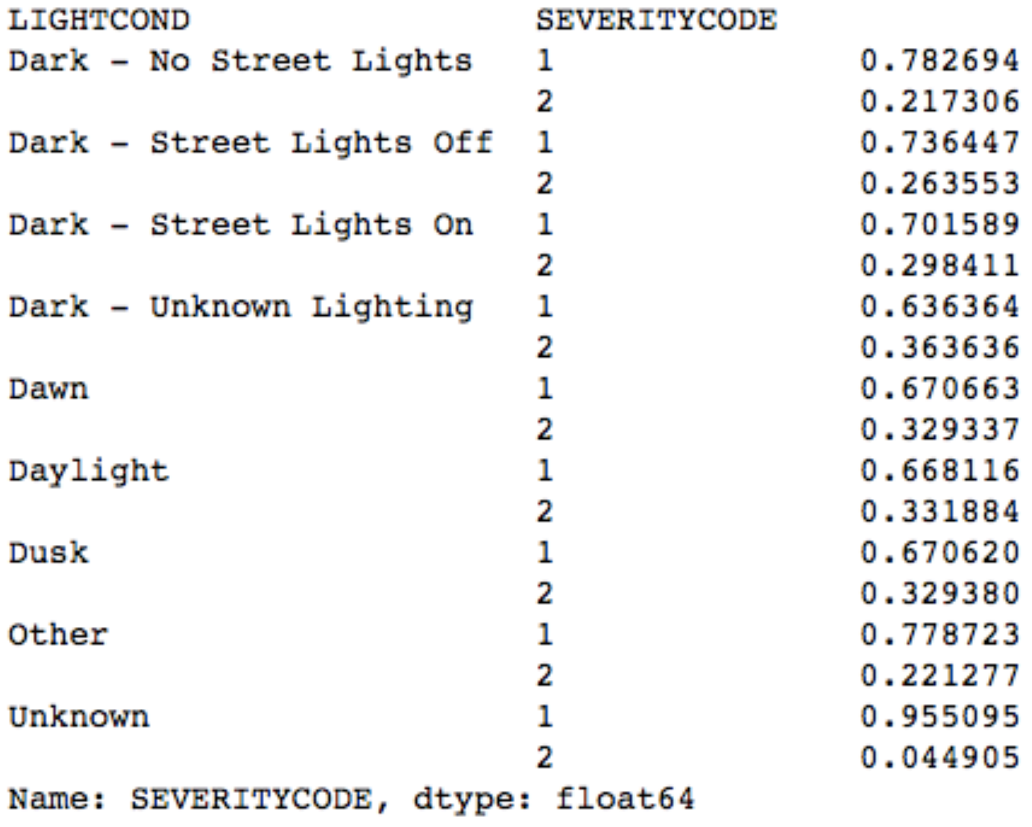
```
WEATHER                   SEVERITYCODE
Blowing Sand/Dirt         1               0.732143
                          2               0.267857
Clear                     1               0.677509
                          2               0.322491
Fog/Smog/Smoke            1               0.671353
                          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.



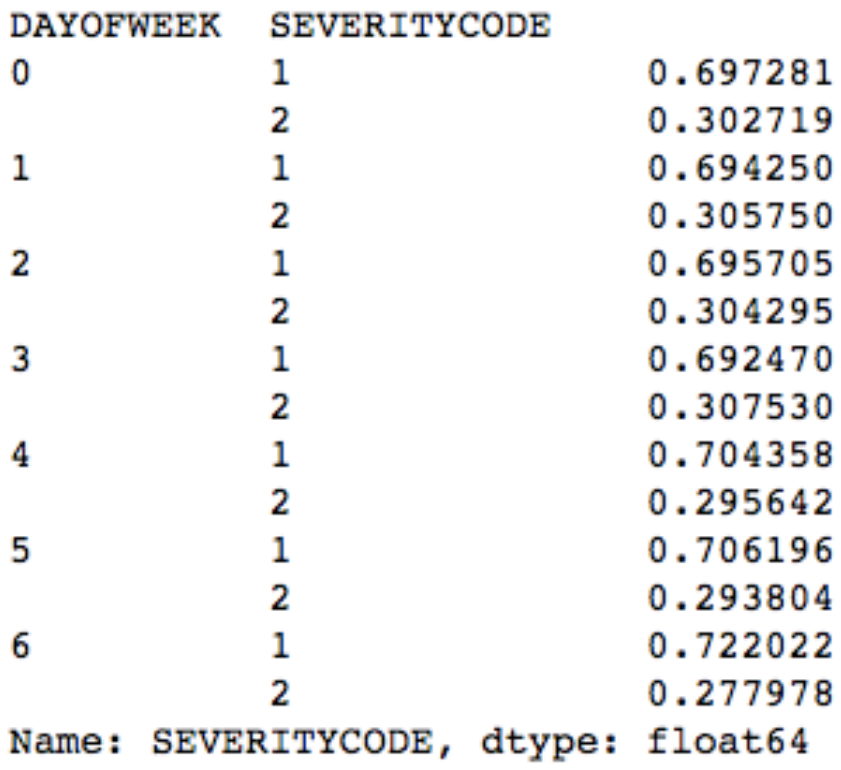
```
ROADCOND         SEVERITYCODE
Dry              1               0.678227
                 2               0.321773
Ice              1               0.774194
                 2               0.225806
Oil              1               0.625000
                 2               0.375000
Other            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 – similar to roadcond



```
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 – this is very important because driving at night in general is riskier or driving during rush hours too



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

Data Preparation

First of all, I adjusted the data types and shortlisted the dataset into the above given chosen fields and severity code. Then, I chose to randomly split data into training and test. Since the dataset is huge, for memory efficiency, I assigned 95% for test data. This also makes sense intuitively because automobile accident situations can be very unique and hence test data is better to achieve accuracy in that aspect.

Choosing and refining the data, I then chose to conduct all four types of ML techniques, i.e., Support Vector Machine, K-Nearest Neigbors, Logistic Regression, and Decision Tree. Then analysis of accuracy was done on the test data for each method.

4. Results

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

Machine Learning Technique	Accuracy Metrics
K-Nearest Neighbor	0.68
SVM	0.66
Decision Tree	0.69
Logistic Regression	0.70

Based on these observations, we can clearly infer that all of them yield very similar accuracy metrics with logistic regression at the highest of 70%.

5. Discussion

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

The logistic regression seems to fit the best, and based on these results, these recommendations can be assumed:

(+) Environmental factors such as road conditions, weather conditions are indeed very close predictor of an accident severity. Hence, monitoring and controlling these can help to reduce fatalities significantly.

(+) Incorporating these predictors and a safe guide as part of a marketing campaign can be extremely powerful in maximizing public safety

(-) further powerful geospatial data and cholropeth maps can be used to include even the location of the data

(-) stronger and faster database softwares and more parameters can be helpful to analyze and make larger, generalized inferences.

6. 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 70% accuracy to incorporate in GPS tracking, or automobile functioning to alarm the driver when the road conditions or any paramters become risky.