Predicting the Car Accident Severity

Applied Data Science Capstone by IBM/Coursera

Business Problem

Seattle, a city on Puget Sound in the Pacific Northwest, is surrounded by water, mountains and evergreen forests, and contains thousands of acres of parkland. Washington State's largest city, it's home to a large tech industry, with Microsoft and Amazon headquartered in its metropolitan area. The traffic is also huge as you guess and accidents are very common.

It is a challenge to the government to control the accidents. The data of previous data were taken and now the task is to make better use of the available data. Many of the accidents in the city are because of the negligence of the people driving the vehicles. But also there are cases of some uncontrollable factors like light, weather, roads, etc. So the accidents because of these uncontrollable factors can be controlled by using the previous data and making an efficient solution out of it. For example, an alert can be sent to the drivers predicting the chances of accidents to take place based on the factors previously mentioned.

The target audience of this project are Government of Seattle, local police and rescue teams, also for car financing corporations. They can gain a lot of profit from implementing this thing.

We will use our Data Science technology to make out an absolute working solution for it now.

Data

The data collected here was huge and was collected by the Seattle Police Department and Accident Traffic Records Department from 2004 to present.
The data consists of 37 independent variables and 194,673 rows.

Depending on the definition of our problem, factors that will influence our decision are:
- Road condition
- Weather condition
- Light condition

After studying the data, as per our requirements mentioned above I have decided to pick up three independent variables - light condition, weather condition and road condition and severity code as the target variable.

Our target variable is "SEVERITY CODE", contains numbers that correspond to different levels of severity caused by an accident from 0 to 4.

Severity codes are as follows:

0. Little to no Probability (Clear Conditions)
1. Very Low Probability — Chance or Property Damage
2. Low Probability — Chance of Injury
3. Mild Probability — Chance of Serious Injury
4. High Probability — Chance of Fatality

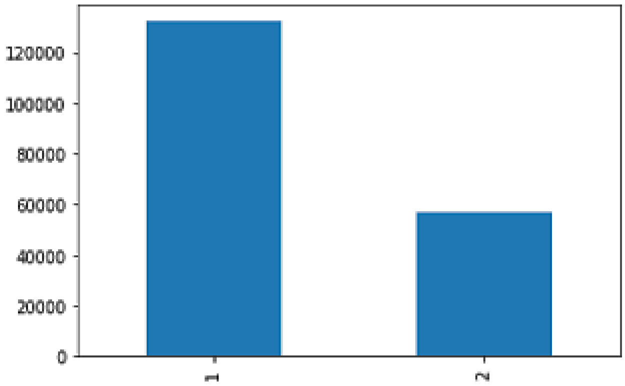
Methodology

As I mentioned earlier WEATHER, ROAD CONDITION, LIGHT CONDITION are the factors predicting the results more accurately.
Our data was just prepared enough to get the target variable SEVERITY CODE.

SEVERITY CODE:

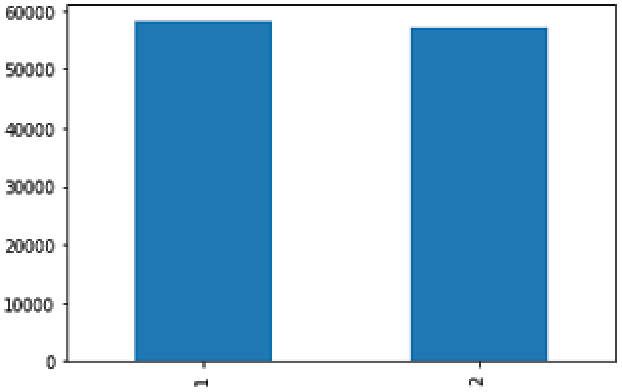
Before :

The data is biased towards the first class in the data set which may not give us the expected outcomes.

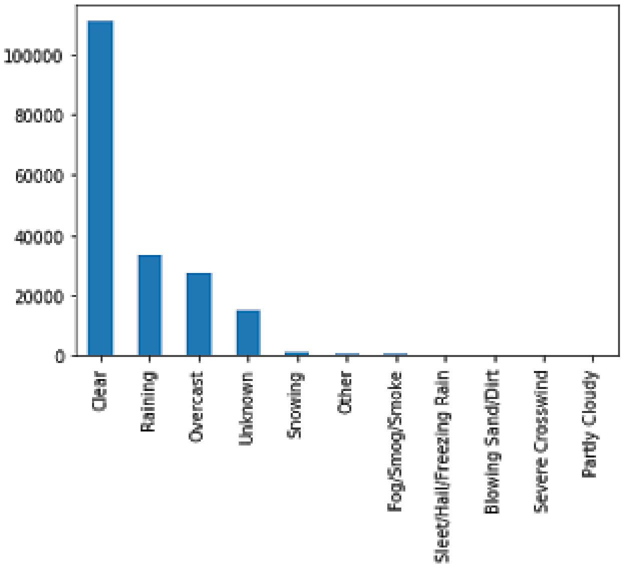


After Downsampling :

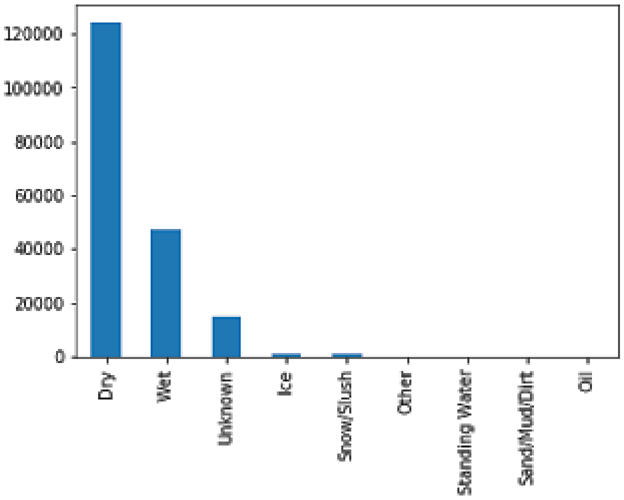
Now the data is fine to model the data and have the equal distribution.



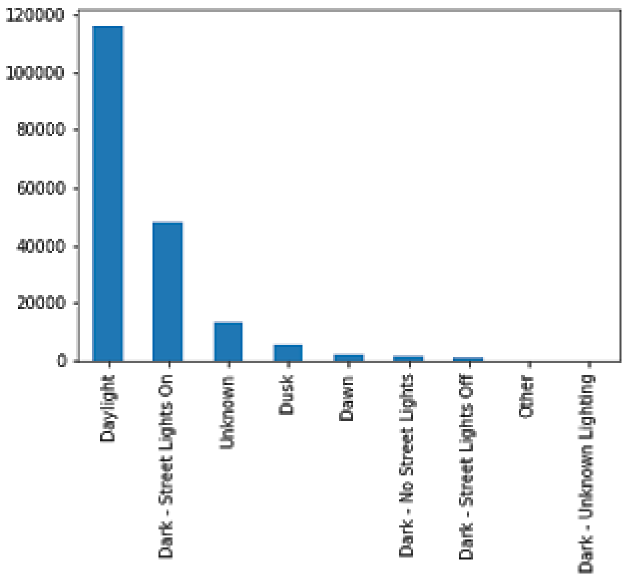
WEATHER CONDITION:



ROAD CONDITION:



LIGHT CONDITION:



Analysis

Decision Tree Model

```
Accident_Severity_Model = DecisionTreeClassifier(criterion='entropy',
max_depth=5)
Accident_Severity_Model.fit(X_train, y_train)
```

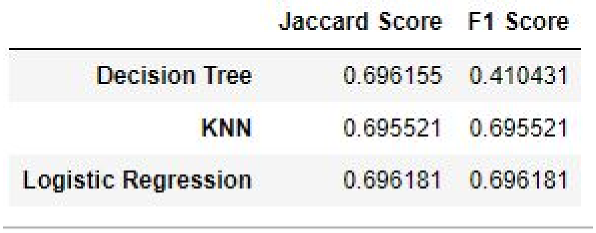
KNN Model

```
neigh = KNeighborsClassifier(n_neighbors=k).fit(X_train,y_train)
neigh
```

Logistic Regression

```
from sklearn.linear_model import LogisticRegression
from sklearn.metrics import confusion_matrix
LR = LogisticRegression(C=6,solver='liblinear').fit(X_train,y_train)
LR
```

Result



	Jaccard Score	F1 Score
Decision Tree	0.696155	0.410431
KNN	0.695521	0.695521
Logistic Regression	0.696181	0.696181

Conclusion

This model can be used to determine the probability of the accident taking place and upto what extent the damage would be.