Car Severity Prediction in Seattle

Coursera Capstone Project

Ankit Kapoor

Background

- Car accidents contributes to biggest cause of injuries and deaths
- These impact:

1. Delivery times
2. Pollution
3. Commuting time

Business Problem

- So this project aims to reduce the collisions in a community and for that an algorithm has to developed .
- This is done to predict the severity of accidents depending upon the factors such as current weather, road and visibility conditions which are already given to us.
- An intimation will be given to driver when these conditions are bad by the model.
- So here the target audience are the drivers in the region(Seattle) which is mentioned in dataset and it is important to solve since it aims to reduce the collisions or accidents

Stakeholders

- Individuals
- Work commuters
- Taxi drivers ,truck drivers ,bus drivers
- Businesses logistic companies
- public/private passengers
- Bus & Taxi companies

Data Sources

- Collision data from Seattle Department of Transportation Open Data Program in CSV format
- https://s3.us.cloud-object-storage.appdomain.cloud/cf-courses-data/CognitiveClass/DP0701EN/version-2/Data-Collisions.csv

Data Understanding

- So in this case we are using 'SEVERITYCODE' as predictor or target variable since we can measure or depict the severity of an accident from 0 to 5 within the dataset.
- Here attributes used to weigh the severity of an accident are 'WEATHER', 'ROADCOND' and 'LIGHTCOND'.

- The codes for severity are as follows:
- 0 : Little to no Probability (Clear Conditions)
- 1 : Very Low Probablility - Chance or Property Damage
- 2 : Low Probability - Chance of Injury
- 3 : Mild Probability - Chance of Serious Injury
- 4 : High Probability - Chance of Fatality

Models

- K-Nearest Neighbor (KNN): This will help us to predict the severity code of an outcome by finding the most similar to data point within k distance.
- Decision Tree : A decision tree model gives us a layout of all possible outcomes so we can fully analyze the consequences of a decision. It context, the decision tree observes all possible outcomes of different weather conditions.
- Logistic Regression: Since our dataset only provides us with two severity code outcomes, our model will only predict one of those two classes. This makes our data binary, which is perfect to use with logistic regression.

Evaluation

- Earlier we had categorical data that was of type 'object'. This is not a data type that could be used to fed to algorithm, so we used label coding & created new classes that were of type int8; a numerical data type.
- After solving that issue we were presented with another - imbalanced data. As mentioned earlier, class 1 was nearly three times larger than class 2. The solution to this was downsampling the majority class with sklearn's resample tool. We downsampled to match the minority class exactly with 58188 values each.
- After analysing and cleaning data, it was then fed through three ML models; K-Nearest Neighbor, Decision Tree and Logistic Regression. Although the first two are ideal for this project, logistic regression made the most sense because of its binary nature.
- Evaluation metrics used to test the accuracy of our models were jaccard index, f-1 score and logloss for logistic regression. Choosing different k, max depth and hyperamater C values helped to improve our accuracy to be the best possible.

Conclusion

- Conclusion Based on historical data from weather conditions pointing to certain classes, we can conclude that particular weather conditions have a somewhat impact on whether or not travel could result in property damage (class 1) or injury (class 2).