Capstone Project - Car accident severity

BUSSINESS UNDERSTAND

In order to reduce the frequency of car collisions in a city, the model to be developed to predict severity of an accident given certain conditions such as whether, road and visibility, this algorithm will alert drivers to be more careful. It will establish sites where the probability of accident is too high.

DATA UNDERSTANDING

'SEVERITYCODE' is the target variable, it is used measure the severity of an accident from 0 to 5 within the dataset. Attributes used to weigh the severity of an accident are 'WEATHER', 'ROADCOND' and 'LIGHTCOND'. 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.

In it's original form, this data is not fit for analysis. For one, there are many columns that we will not use for this model. Also, most of the features are of type object, when they should be numerical type.

We must use label encoding to covert the features to our desired data type.

METODOLOGY

The most important features to predict the severity of accidents: Weather, Roadcond and Lightcond.

After balancing SEVERITYCODE feature, and standardizing the input feature, the data has been ready for building machine learning models.

We will use the following models:

K-Nearest Neighbor (KNN)

KNN will help us 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

Because 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.

RESULTS AND EVALUATION

The final results of the model evaluations are summarized in the following table:

KNN
Jaccard Score: 0.68
F1 Score: 0.60

Decision Tree
Jaccard Score: 0.70
F1 Score: 0.58

Linear Regression
Jaccard Score: 0.70
F1 Score: 0.58

Based on the above table, KNN is the best model to predict car accident severity.

DISCUSSION

Once we analyzed and cleaned the 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 most sense because of its binary nature.

Evaluation metrics used to test the accuracy of our models were jaccard index, f-1 score. Choosing different k, max depth and hyparameter C values helped to improve our accuracy to be the best possible.

CONCLUSION

Based on the dataset from weather, road, and light conditions pointing to certain classes, we can conclude that particular conditions have a somewhat impact on whether or not travel could result in property damage (class 1) or injury (class 2).