DATA SCIENCE CAPSTONE PROJECT

CAR ACCIDENT SEVERITY REPORT

-HAMSAA S K

Introduction | Business Understanding

- In an effort to reduce the frequency of car collisions in a community, an algorithm must be developed to predict the severity of an accident given the current weather, road and visibility conditions.

- When conditions are bad, this model will alert drivers to remind them to be more careful.

Data Understanding

- Our predictor or target variable will be 'SEVERITYCODE' because 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 Probablility - Chance or Property Damage

  2 : Low Probability - Chance of Injury

  3 : Mild Probability - Chance of Serious Injury

  4 : High Probability - Chance of Fatality

Extract Dataset & Convert

- In its 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.

Methodology

- Our data is now ready to be fed into machine learning models.
- We will use the following models:
- **K-Nearest Neighbour (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 analyse 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.
- Let's get started!

Results & Evaluation

K-Nearest Neighbor

```
]: # Jaccard Similarity Score
   jaccard_similarity_score(y_test, Kyhat)
[197]: 0.564001947698565

]: # F1-SCORE
   f1_score(y_test, Kyhat, average='macro')
[198]: 0.5401775308974308
```

Model is most accurate when k is 25.

Decision Tree

```
]: # Jaccard Similarity Score
   jaccard_similarity_score(y_test, DTyhat)
[213]: 0.5664365709048206

]: # F1-SCORE
   f1_score(y_test, DTyhat, average='macro')
[214]: 0.5450597937389444
```

Model is most accurate with a max depth of 7.

Logistic Regression

```
]: # Jaccard Similarity Score
   jaccard_similarity_score(y_test, LRyhat)
[247]: 0.526021825680978 4

]: # F1-SCORE
   f1_score(y_test, LRyhat, average='macro')
[248]: 0.5116020939633 83

]: # LOGLOSS
   yhat_prob = LR.predict_proba(X_test)
   log_loss(y_test, yhat_prob)
[249]: 0.6849535383198887
```

Model is most accurate when hyperparameter C is 6.

DISCUSSION

- In the beginning of this notebook, we had categorical data that was of type 'object'. This is not a data type that we could have fed through an algorithm, so label encoding was used to 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.
- Once we analyzed and cleaned the data, it was then fed through three ML models; K-Nearest Neighbour, 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 and log loss for logistic regression. Choosing different k, max depth and hypermeter 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).