CS5063: Foundations of Machine Learning

Hackathon

Submitted By:
Vishal Singh Yadav – CS20MTECH01001
Kaggle Username: raspyduke694, Public/leaderboard name: CS20MTECH01001

EDA and Data Cleaning:

I first check the number of null values in each column. Every column that had more than 50% null values was removed. This reduced the number of columns by 6.

Then I checked the data by going through it and found some columns which should not affect the prediction, such as *ReportID*, *Person ID*, etc. After removing these columns, I parsed the *date/time of crash*.

I converted the *date/time of crash* to 5 columns containing day, month, year, day of the week, the hour of the day. Then I removed the original *Date/Time of crash* column.

Additionally, a few more columns, such as vehicle type and model number, were removed. After that, all null values were filled with the median value of the column and all categorical columns converted to numerical where one number denoted a class.

Modelling:

I first submitted a random submission to get the baseline which gave me a score of 0.5. Then I used Linear Classifier, but that resulted in a poor score of 0.6. Similarly, SVM also provided a poor score of 0.5.

Then I tried Random Forest Classifier, and it gave an excellent score of 0.85. After this, I decided that tree-based algorithms are suited for this data.

I then decided to implement boosting further to improve the model as well as classification accuracy. I tried XGBoost, which provided a slightly better score. Also, knowing that LightGBM and CatBoost delivers an even better score, I decided to try them as well. This also increased my score slightly.

Since I had four models (RandomForestClassifier, XGBoost, LightGBM, and CatBoost) that had similar scores of around 0.85-0.86, I decided to ensemble these three methods. I went with the majority voting for final predictions based on these three models. I used two ensemble weighing methods which have been submitted for final scoring. One of them equally weights each model and make predictions, whereas the other considers each model according to the score they provided on the public leaderboard. I noticed that weighted ensemble provide a bit higher score but still lower than CatBoost score.

The files *LGBM.csv*, *RandomForest.csv*, *CATBoost.csv* and *XGB.csv* have the predictions for these individual classifiers.

The files *ensemble_equal_weights.csv* and *ensemble_score_weigths.csv* are contain ensembled predictions.

Submitted Files:

The following files are submitted with the code

- LGBM.csv
- RandomForest.csv
- CATBoost.csv
- XGB.csv
- ensemble_equal_weights.csv
- ensemble_score_weigths.csv

Of these *CATBoost.csv* and *ensemble_score_weights.csv* is to be considered for final scoring