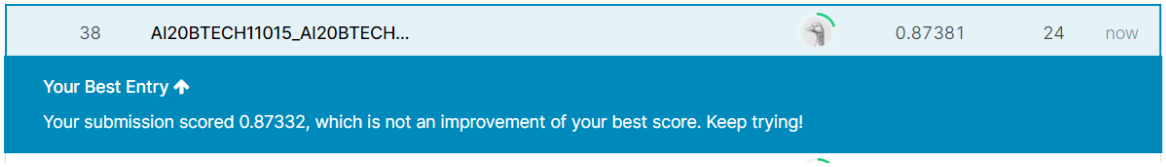
FOUNDATIONS OF MACHINE LEARNING (AI2000) HACKATHON REPORT

MURARISETTY ADHVIK MANI SAI
(AI20BTECH11015)
KARUMANCHI SACHIN
(AI20BTECH11013)

November 15, 2021

Kaggle Details

- Kaggle ID: User ID 8008719
- Kaggle User Name: AI20BTECH11015_AI20BTECH11013
- Best Accuracy: **0.87381**



1 PREPROCESSING

- After analyzing the data there are 42 features including the label in the train data set there are 51490 train samples in the test data set there are 77235 test samples. We have dropped some of the unnecessary columns that are not required to predict is it driver fault. And we removed the columns which are having more then 20 percent of data as NaN values or null values. We dropped some of the object data type columns which are gaving nearly thousands of categories.

  Here is the list of categories we dropped

'Agency Name','Vehicle Make','Vehicle Model','Equipment Problems','Location','Report Number','Local Case Number','Road Name','Cross-Street Name','Off-Road Description','Person ID','Vehicle ID','Crash Date/Time','Drivers License State'

- We replaced 'Vehicle Year' column with age of the vehicle till that incident.
- we replaced some nan values with mode of those column.

1.1 ENCODING

We tried many encoding tyechniques such as dummies encoding, LabelEncoder, CatBoostEncoder, polynomial encoder, one hot encoder. Of which best results was produces by using get_dummies().

2 MODEL USED

We used CatBoostClassifier, Lgbm, gradient boosting, xgb classifier and used those results to build a prediction with majority voting. We also tried using random forest. All the above mentioned classifiers performing descent on the validation set with an accuracy of 0.86%+. And all parameters are hypertuned to maximise accuracy. Since there are more categorical features we used the above models with encoded data and optimised the prediction.