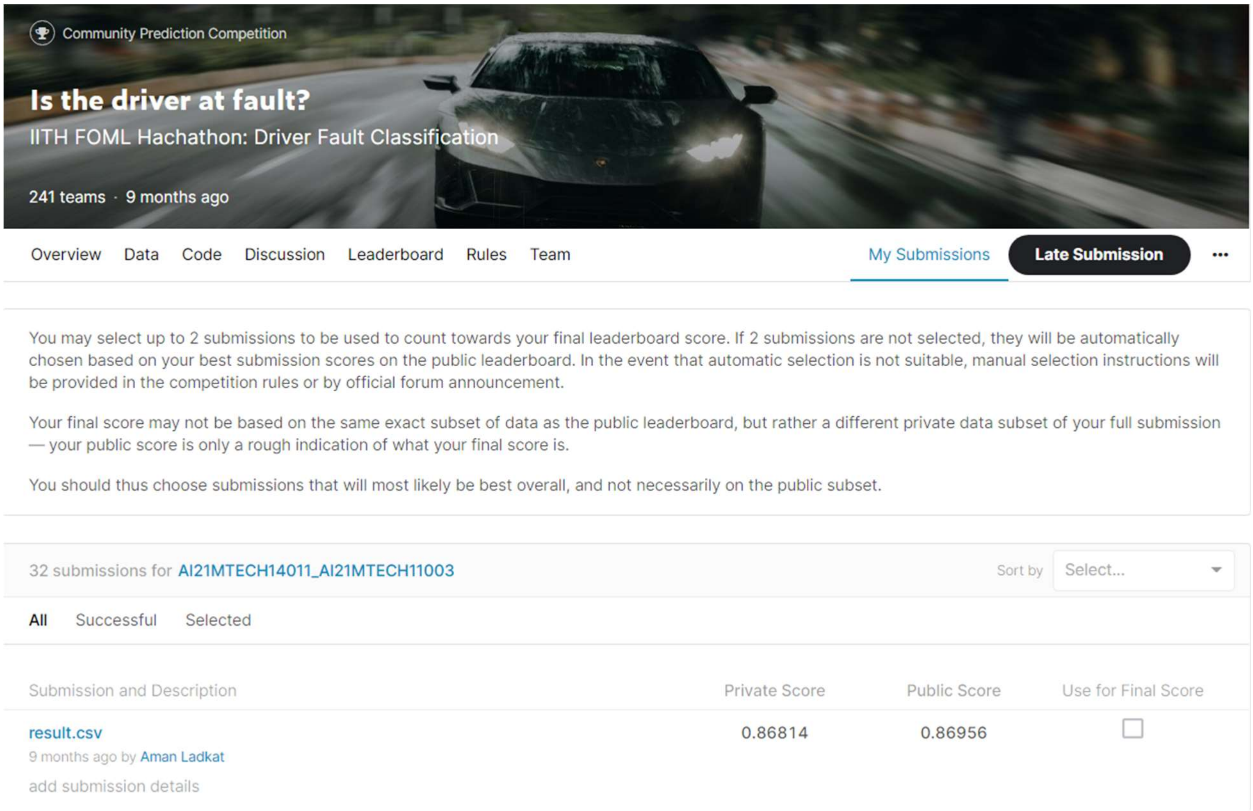
Hackathon Report

Team members: J.Shashank (AI21MTECH11003) and Aman Ladkat (AI21MTECH14011)

1. Pre-processing the dataset:
   a. Dropped columns with excessive null values.
   b. Filled columns having few missing values with mode of that column.
   c. Decomposed the ‘Crash Date/Time’ column into hour, minutes, date and day of week.
   d. Filled missing values of ‘Light’ column with mode of that column based on the hour of the day.
   e. Converted categorical data to numeric data using scikit-learn’s LabelEncoder.
   f. Normalized the dataset using scikit-learn’s MinMaxScaler.
   g. Replicated the above steps for pre-processing the test dataset.

2. Feature selection:
   a. Ranked the best features from the training dataset using scikit-learn’s feature selection tool ExtraTreesClassifier.
   b. Selected ‘k’ features from the ranked features for training our model.

3. Training the model:
   a. Filtered the dataset based on the number of columns we chose from the ranking of features.
   b. Divided the training data into training set and validation set with validation set consisting of 20% of rows of original dataset.
   c. Trained the dataset using RandomForestClassifier, GradientBoostingClassifier and XGBoostClassifier.
   d. The best accuracy score was achieved with the XGBoostClassifier.
   e. In order to find the best score, we fine-tuned the hyperparameters of each classifier.
   f. We also fine-tuned the number of features to select for our model by checking accuracy score for various values of ‘k’.



Is the driver at fault?

IITH FOML Hachathon: Driver Fault Classification

241 teams · 9 months ago

You may select up to 2 submissions to be used to count towards your final leaderboard score. If 2 submissions are not selected, they will be automatically chosen based on your best submission scores on the public leaderboard. In the event that automatic selection is not suitable, manual selection instructions will be provided in the competition rules or by official forum announcement.

Your final score may not be based on the same exact subset of data as the public leaderboard, but rather a different private data subset of your full submission — your public score is only a rough indication of what your final score is.

You should thus choose submissions that will most likely be best overall, and not necessarily on the public subset.

32 submissions for AI21MTECH14011_AI21MTECH11003

Submission and Description	Private Score	Public Score	Use for Final Score
result.csv 9 months ago by Aman Ladkat add submission details	0.86814	0.86956	□