CS5590: Foundations of Machine Learning

Hackathon Report

# Soumi Chakraborty

ES19BTECH11017

# Nandita Tunuguntla

CS19BTECH11051

# 1 Model: XGBoost

# We chose XGBoost for our model after preprocessing. We also tried the following models and after excessively tweaking the parameters of all of these, we decided to go with XGBoost as it gave the highest validation accuracy:

* Gradient Boosting
* Random Forests
* AdaBoost
* Logistic Regression

We got an accuracy of 85.5% from XGBoost which gave us an accuracy of 86.6% in the test set. The parameters of the model were the following:

* Learning\_rate: 0.1
* n\_estimators: 200
* max\_depth: 8
* colsample\_bytrees: 0.7
* subsample: 0.7
* alpha: 10

2 Preprocessing

2.1 NaN Columns

We removed all columns which had more than 40000 NaN values.

2.2 One Hot Encoding

We used one hot encoding for the features 'Non-Motorist Substance Abuse', 'Vehicle First Impact Location', 'Vehicle Second Impact Location', 'Light'

2.3 Label Encoding

We used label encoding for all the columns which have categorical object type data.

2.4 Missing Values

We replaced most missing values with the value “UNKNOWN” because we treated missing values to be unknown.

2.5 Large Number of Categories

Analysed each feature and removed the ones which seemed to have too many irrelevant categories. For example, IDs and names of places/vehicles/people.