HACKATHON REPORT

Sri Hari Malla - CS19BTECH11039
Sarat Chandra Sai - CS19BTECH11003

November 15, 2021

1 Info:

- This is a Classification problem
- The Train data set initially has 42 columns and 51490 rows
- The test data set has 41 columns excluding Id column and 77235 rows
- The classification column is **Fault**

2 Flow:

- Briefly examine the data
- Need to properly pre process data before training the model
- Use multiple models to increase the accuracy.

3 Pre Processing Data:

3.1 Observations:

- Train data has many rows with all Nan values. They are to be dropped.
- Train data has also columns with the value of Nan exceeding 50%.
- Need to handpick columns by examining carefully and selecting required columns and leaving the useless columns(experimentation).
- Nan values are present in between the columns and needs to be replaced accordingly.

3.2 Actions Taken:

- Identified and removed the columns which has more than 70% of NaN values.
- Handpicked and removed the columns which are irrelevant to the classification.
- Dropped rows with only Nan values in Train data.
- Replaced all the Nan values in between with the mode of the data present using categorical imputer.
- The rows with only Nan Values in Test data is manipulated by the imputer.
- Used Label encoder to label encode all the string values present in the data to int and float.

4 Model Selection:

4.1 Used models:

- Gradient Boosting Classifier (GBC)
- Random Forest Classifier (RFC)
- XG Boosting Classifier (XGB)
- Cat Boosting Classifier(with 5 fold validation) (CBC)
- **Clearly Gradient Boosting Classifier and Cat Boosting Classifier out performed the other models while experimenting.**
- **Let's see Gradient Boosting classifier and Cat Boosting classifier where the best accuracy ais measured.**

4.2 Gradient Boosting Classifier:

- We know that GBC uses Random Forest and Decision Tree intuition but starts the process of combining in the start instead of end.
- We've also seen that GBC yields better performance than RFC and XGB given the noise is low.
- So we've reduced the noise in the train data as much as we can to achieve most from GBC.
- Hyperparameter tuning is comparatively harder in case of GBC than RFC and XGB.

- After trying out various parameters, the best accuracy is obtained with the **hyperparameters : nestimators : 351, Learning rate : 0.06, Max Depth : 7**
- **The Accuracy obtained is 87.231**
- By using K-Cross Validation, accuracy can be improved slightly as it uses multiple random training sets to train the model.
- By using the 5 Split Validation, **Accuracy increased to 87.254**
- Using 10 splits decreased the accuracy.

4.3 Cat Boosting Classifier:

- On seeing Various other websites and exploring the models, we came through the Cat boosting model which we never encountered before. Fortunately, it yielded the best of all.
- It is found to be the most innovative algorithm for processing data having categorical features.
- For small datasets, GBC overfits easily whilst in CBC it is handled all by its own. Overfitting problems are relatively low in CBC when compared to Other algorithms.
- Hyperparameters used **od_wait:200, loss_function:Cross Entropy, Boost_rounds:1800**
- On using CBC we initially got accuracy of **87.347**
- By using the same trick of cross validation, we got accuracy of **87.464**
- Using 10 folds decreased the accuracy.

LaTeX generated document

THE END