Predictions for failures in dyeing process

Dr. Sapana Tripathi

Contents

- Situation: Analysis of given dataset
- Task: Identification of issues and complications
- Action: Implementation of models
- Results and recommendations

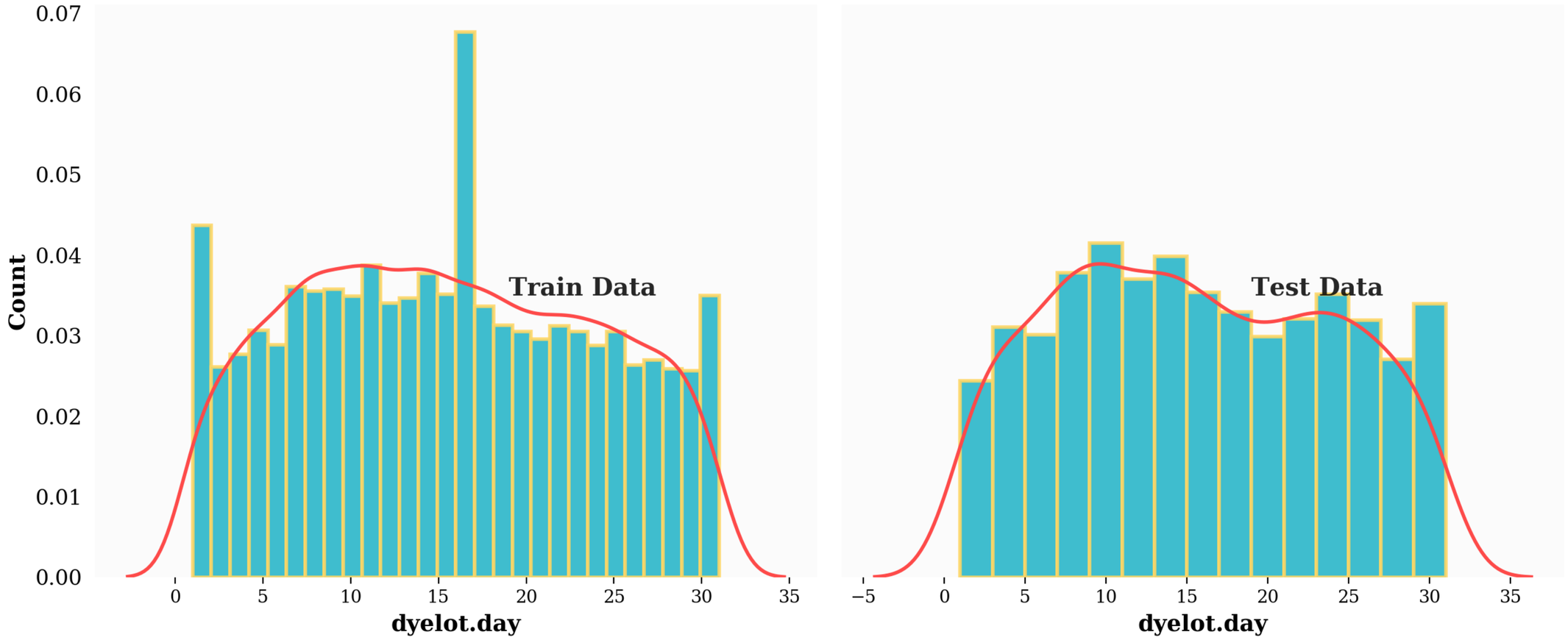
Training dataset : Features

- Number of features : 68
- Unique identification : Parent.batch.ID., Batch.ID
- Substrate related features
- Dye related features
- Recipe related features
- Yarn weight features
- Thread features
- Colour related features

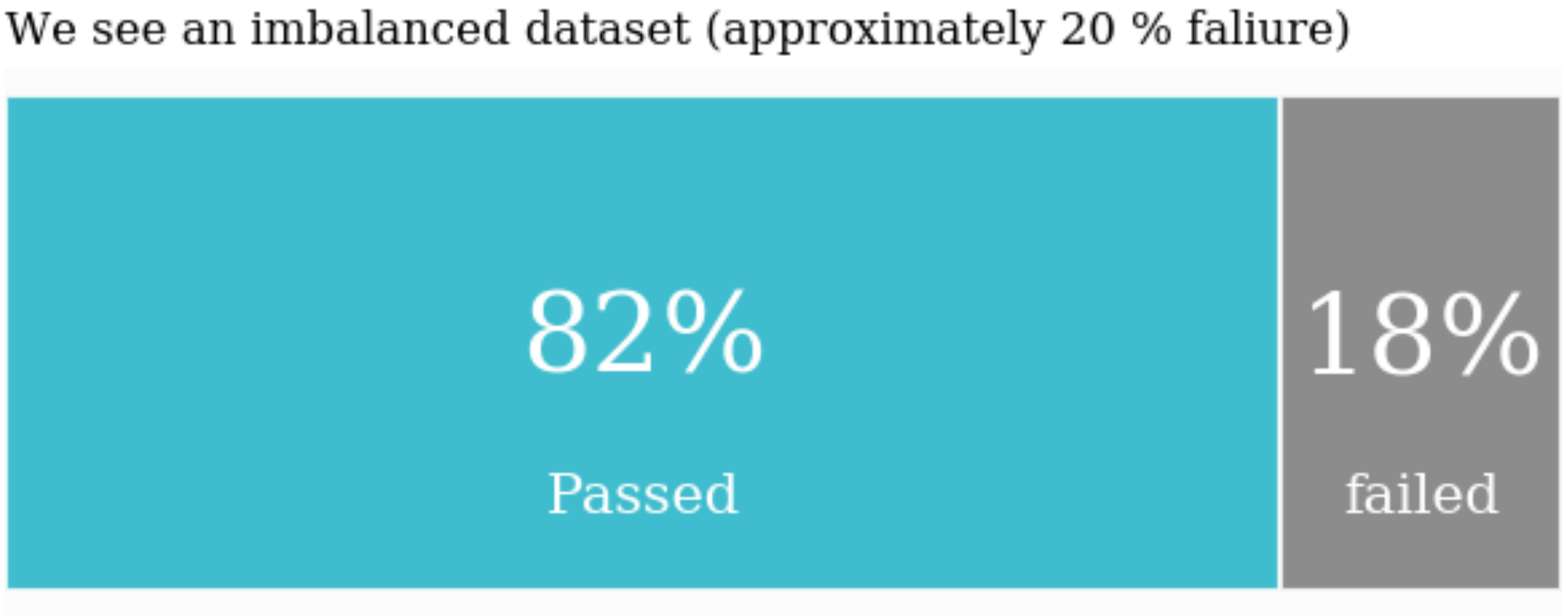
Situation

- Objective: Textile industry requires same coloured fibres in a batch
- Issue: Failure in dyeing batches with same colour
- Given: Dataset of dyed batches of 2017
- Actions:
  - Identify features responsible for failures
  - Create a model

Dataset (2017): How the data is recorded?



Distribution of target in training dataset



Task

- Treatment of raw and imbalanced dataset
- Classification of failed and passed batches :
  - Successful batch : class = 1
  - Failed batch : class = 0
- Estimation of probability of failed batches for the given dataset

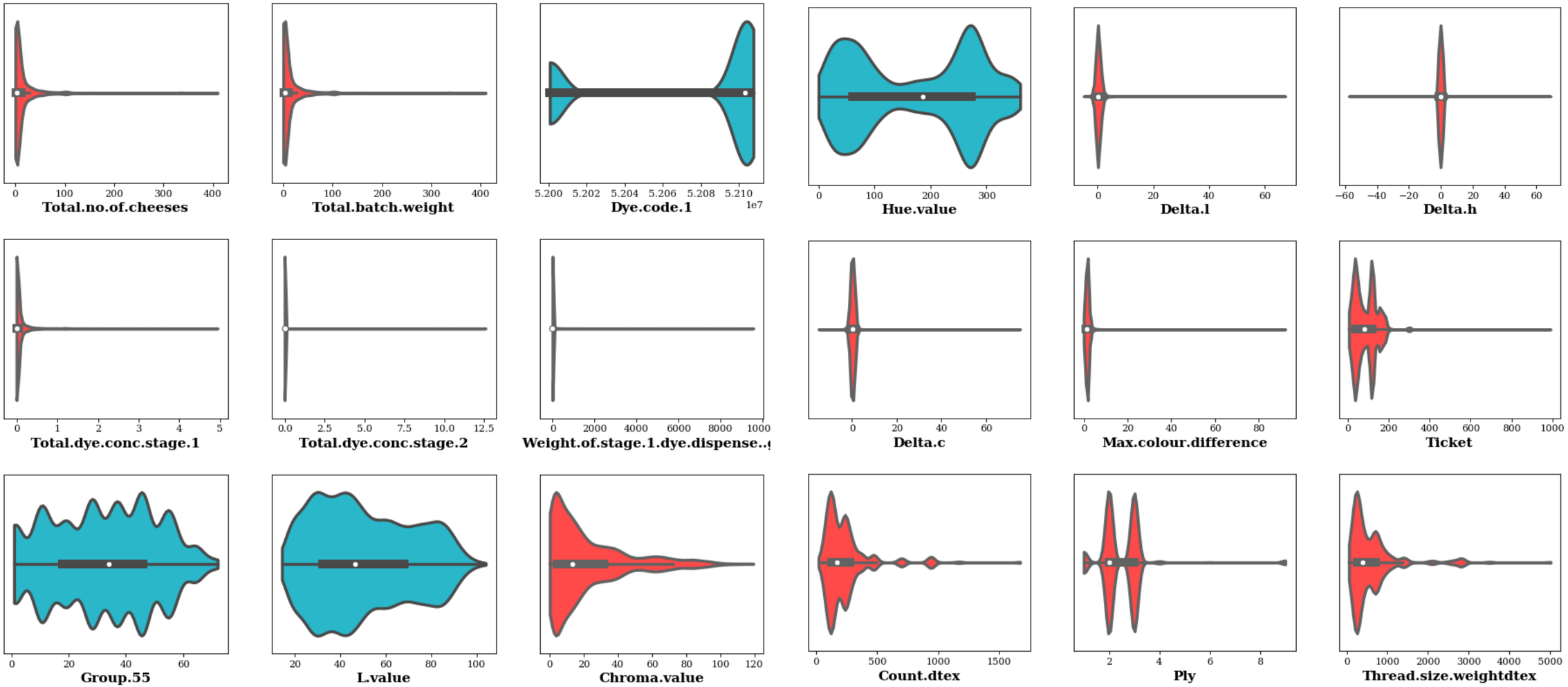
How to solve the imbalanced dataset problem?

- K-Fold cross-validation (Stratified)
- Hyper- parameter tuning
- Evaluation matrix like recall, precision, confusion matrix
- AUC_ROC curve
- Under-sampling
- Over-sampling
- SMOTE

Actions

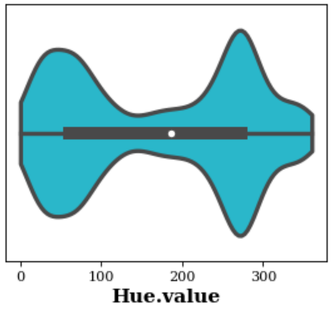
- Data cleaning: Missing values, Duplicates, Outliers (Z-Score)
- Decoding coded features
- Feature engineering: Label Encoding, Feature Importance
- Primary Modelling: Logistic regression, Random Forest, Adaboost
- Secondary Modelling: Selection of one model
- Results and Evaluation of model
- Application of model on test dataset

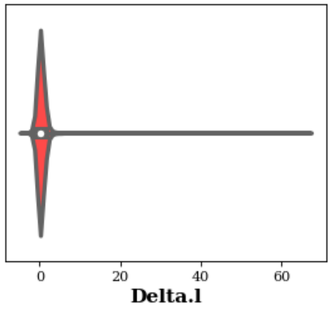
Univariant analysis of features

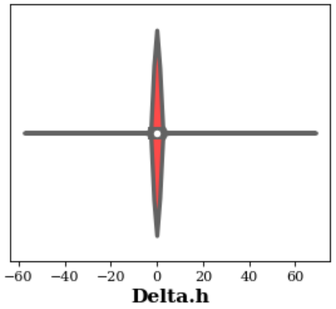


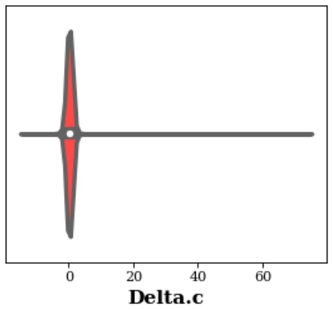
Univariant analysis of features

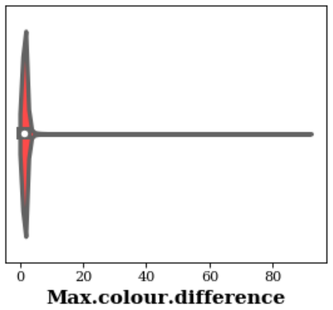
Before outlier removal

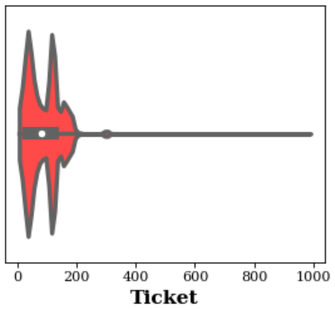


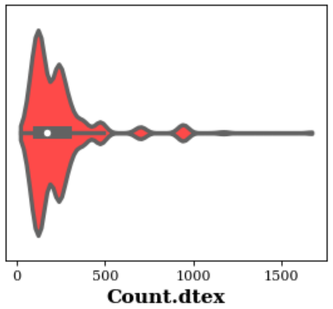


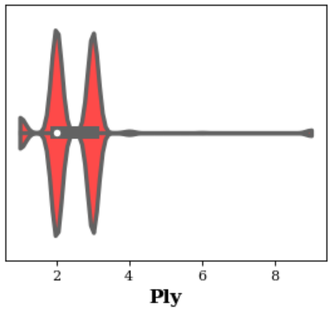


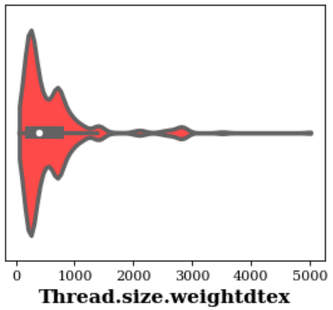




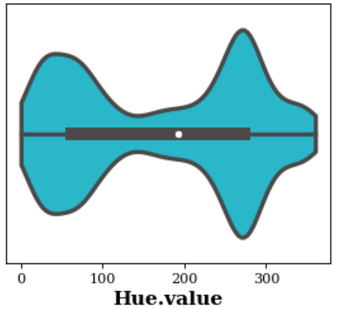


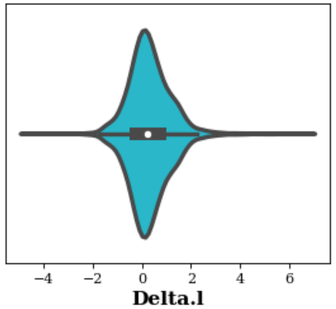


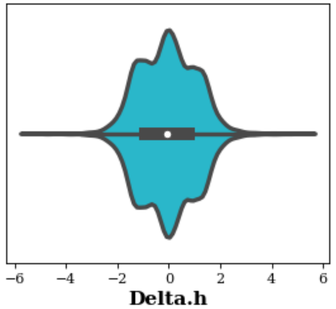


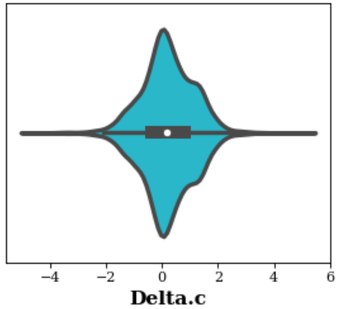


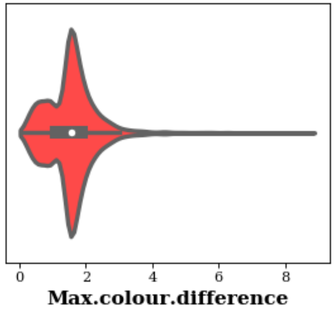
After outlier removal

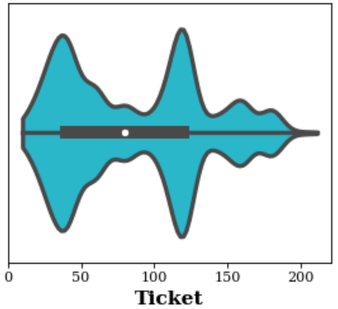


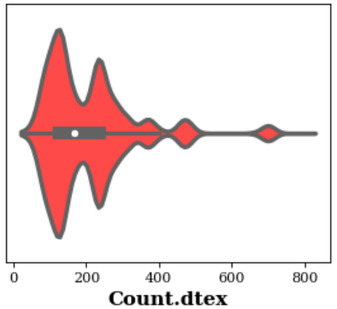


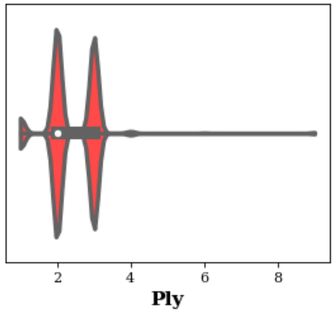


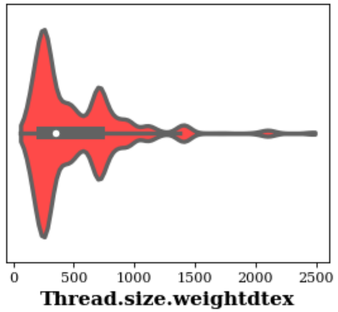




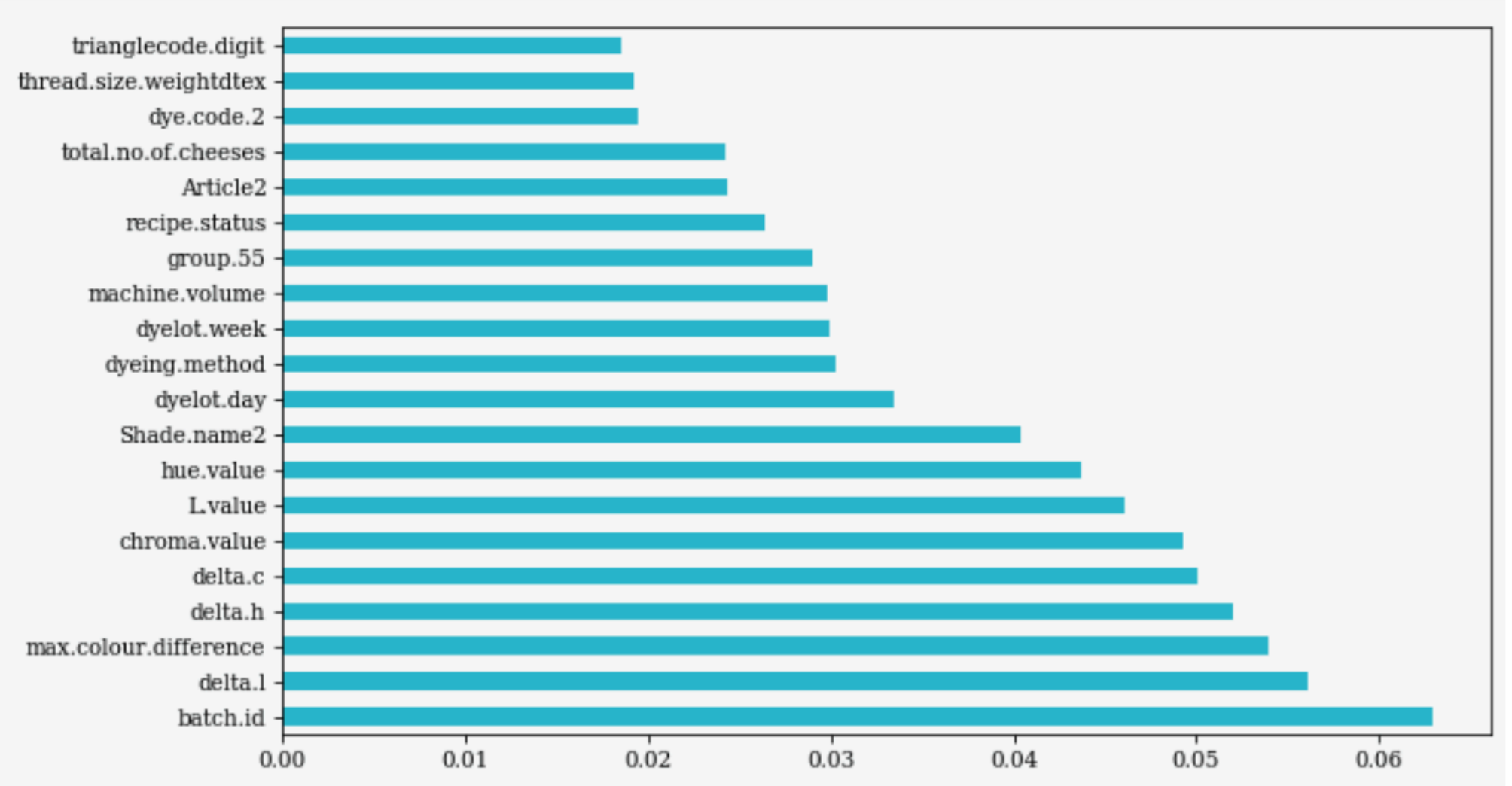






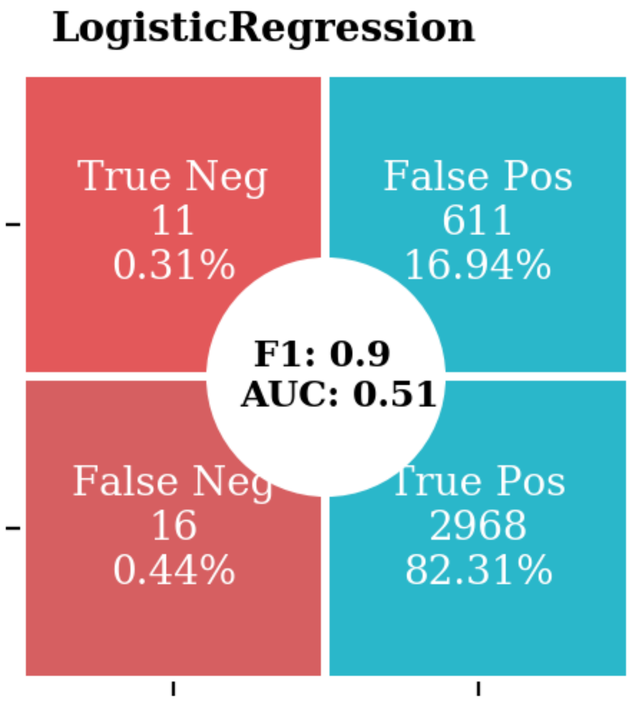


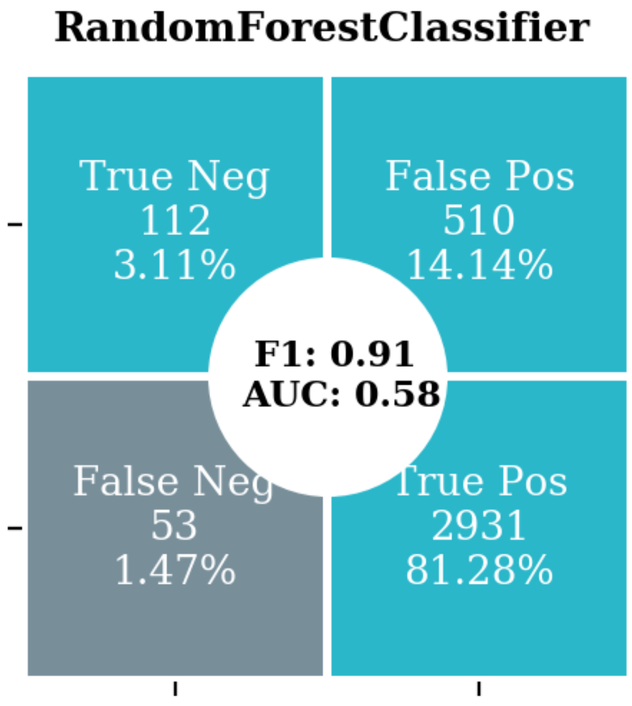
Feature selection using Feature Importance

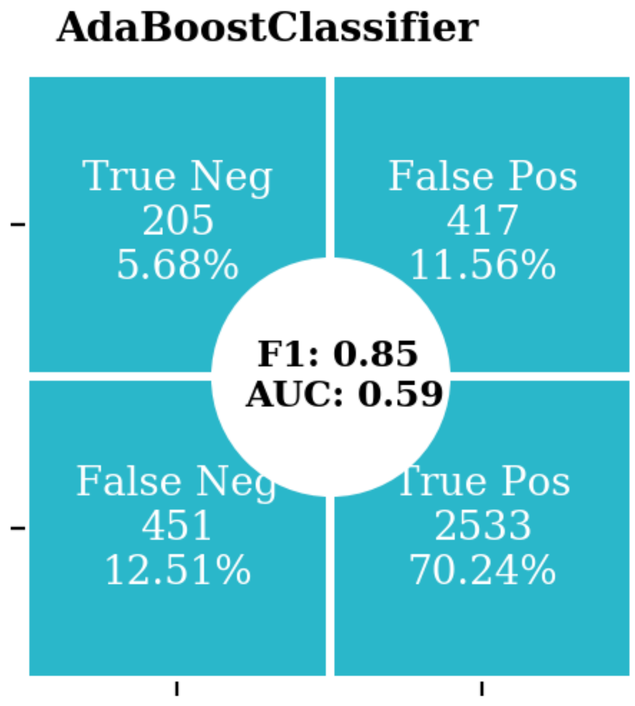


Results

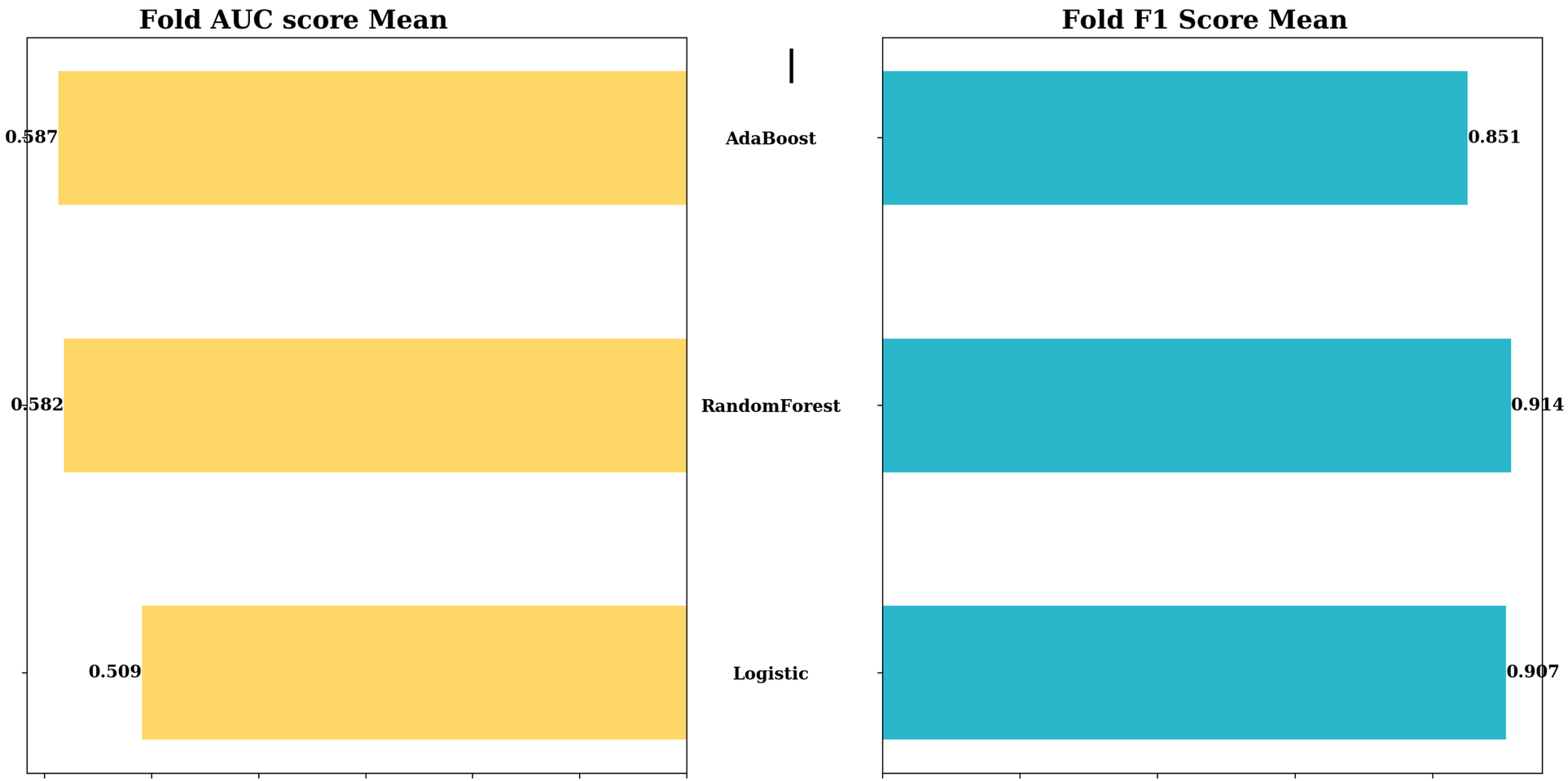
Baseline Modelling results : Confusion Matrix



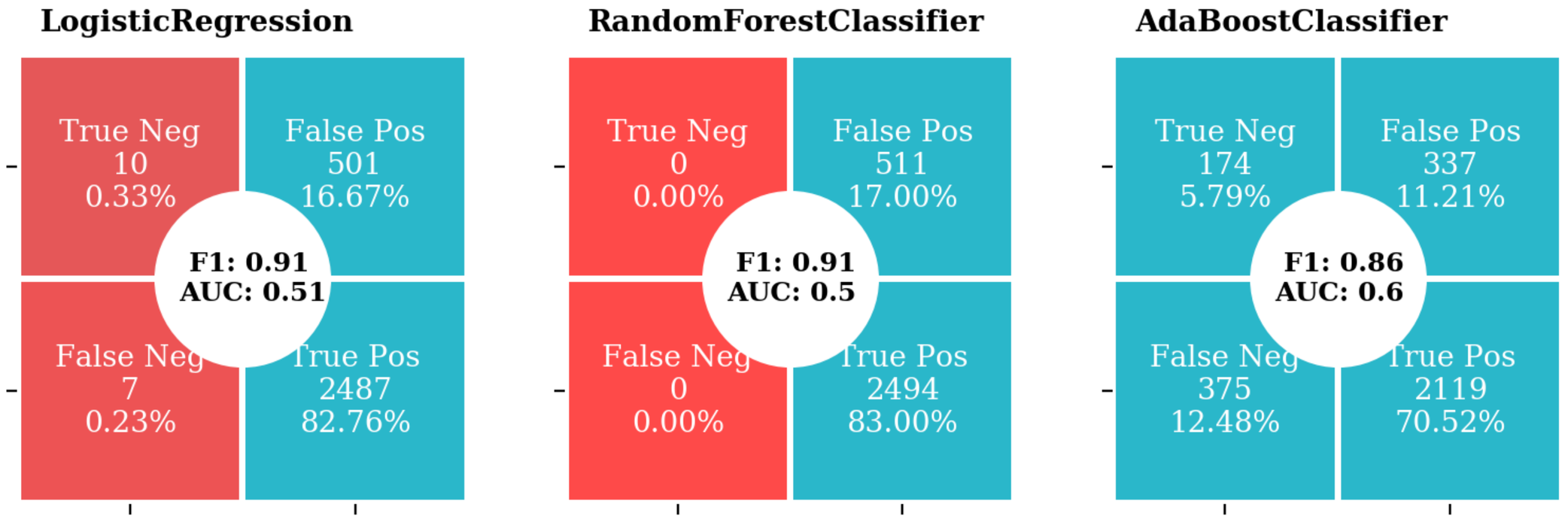




Stratified CrossValidation Results

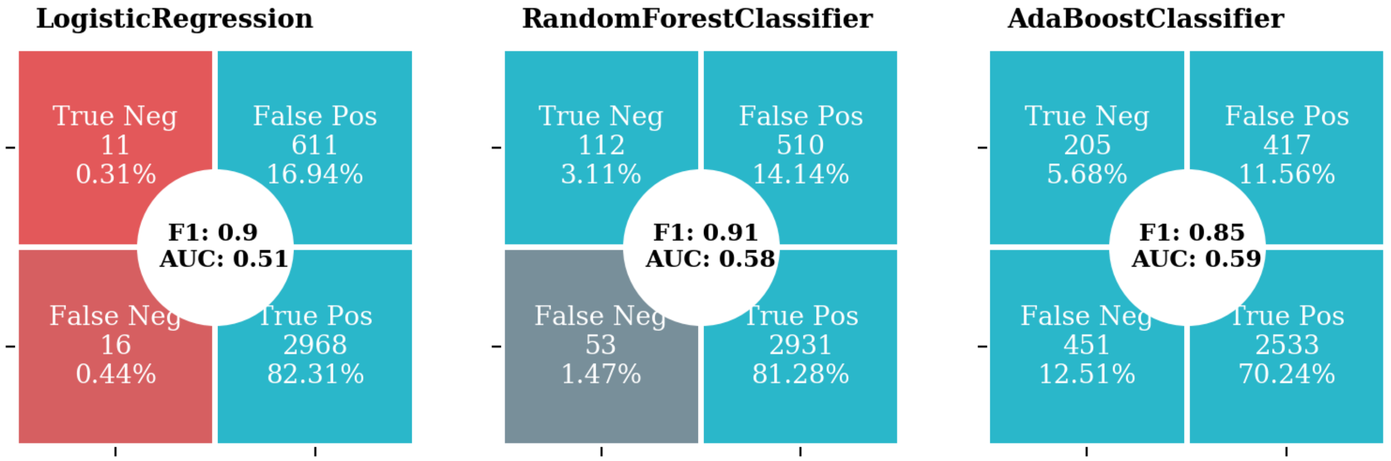


CrossValidation + Hyper-parameter Tuning

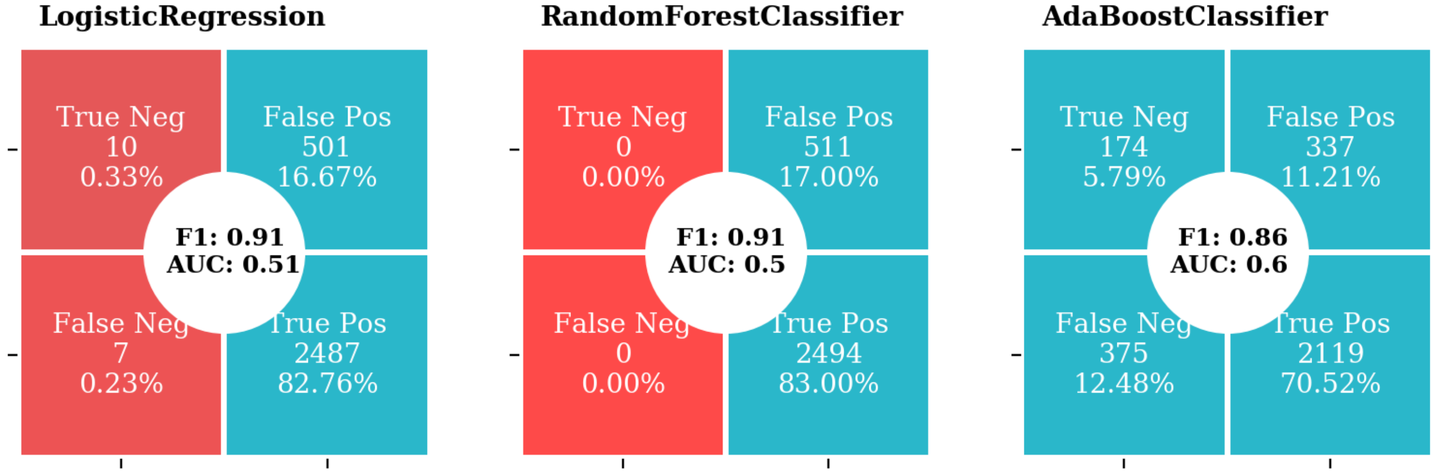


CrossValidation + Hyper-parameter Tuning Comparison

Baseline model

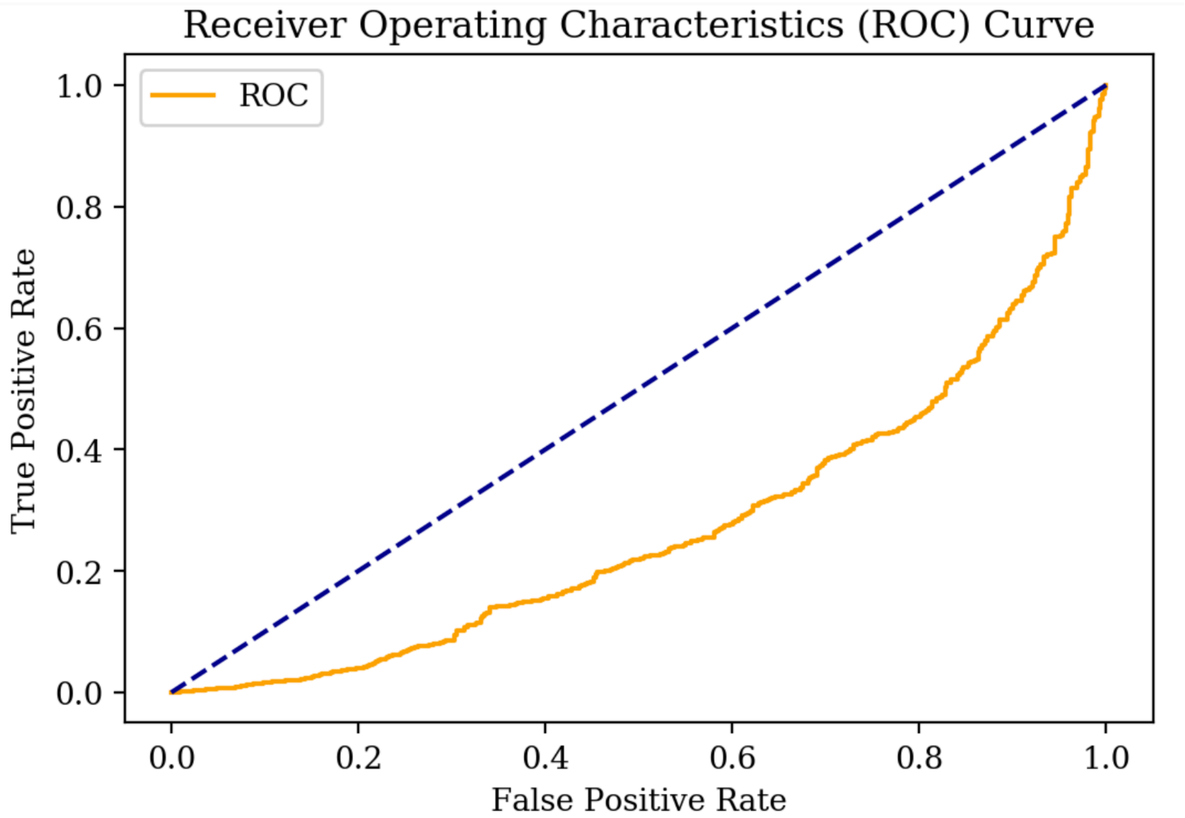


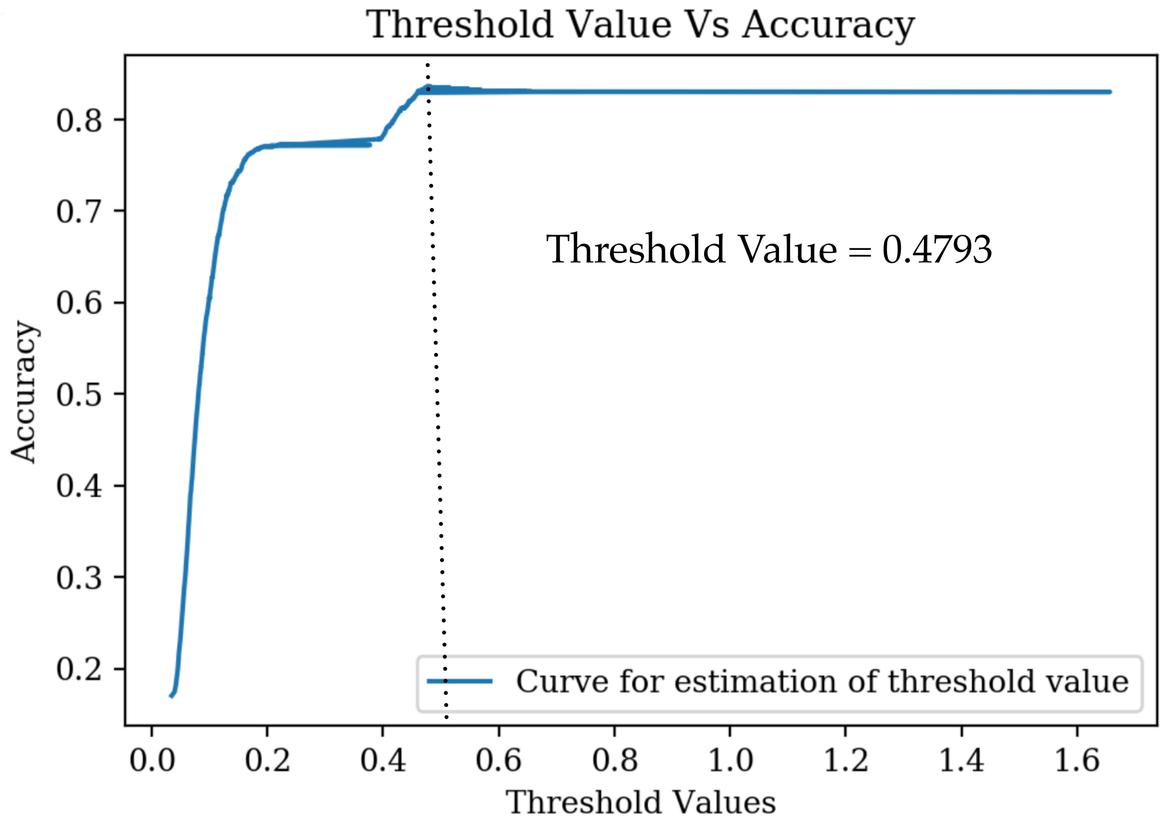
CrossValidation + Hyper-parameter tuning



How to find threshold for the binary classification problem?

AUC-ROC and Threshold value curve





Conclusions

- Presence of **skewness** and **kurtosis**.
- **Lowest** False positive values for **Adaboost.**
- **Highest** roc_auc_score (0.6) for **Adaboost.**
- Threshold value for the binary classification is **0.4793.**

Future-Prospects

- Need for further refinement of data
- Scaling and transformations
- One Hot encoding
- Dimensionality reduction : PCA, UMAP
- Testing of data with additional models
- SMOTE