Bureau ID Assignment Report

Contents

Harsh Nandwani
+91 9867280369
harsh0nandwani@gmail.com

Approach Taken

The approach taken was the one that could deliver best results in minimal code and time as the assignment was assigned during the festival Ganesh Chaturthi. The following steps were done:

- Exploratory Data Analysis using YProfile
- Feature Selection, with features having correlation coefficient $\eta \geq 0.05$
- Model Selection and Hyperparameter tuning using MLJAR

Exploratory Data Analysis

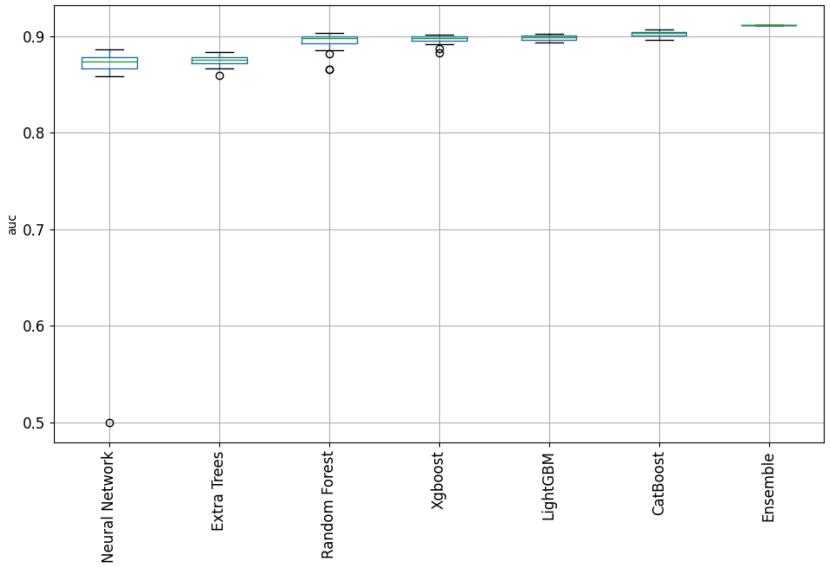
YProfile was used to generate the EDA report based on which feature selection was done. The EDA report is submitted under the filename ‘EDA_Report.html’

Model Training

MLJAR was used to train the models for model selection and tuning.
The models considered were:

- Random Forest
- Extra Trees
- LightGBM
- Xgboost
- CatBoost
- Feed Forward Neural Network
- Ensemble

Metric for model selection used: AUC



Insights

Overview | Alerts 61 | Reproduction

Dataset statistics

Number of variables	55
Number of observations	10000
Missing cells	148108
Missing cells (%)	26.9%
Duplicate rows	9
Duplicate rows (%)	0.1%
Total size in memory	4.1 MiB
Average record size in memory	433.0 B

Variable types

Numeric	9
DateTime	1
Text	12
Categorical	31
Boolean	2

Important features with $\eta \geq 0.05$

- ADDRESS TYPE
- AGE
- ASSET CTG
- ASSET MODEL NO
- EMPLOYER TYPE
- HDB BRANCH STATE
- PRIMARY ASSET MAKE
- TOTAL ASSET COST

- phone_nameMatchScore

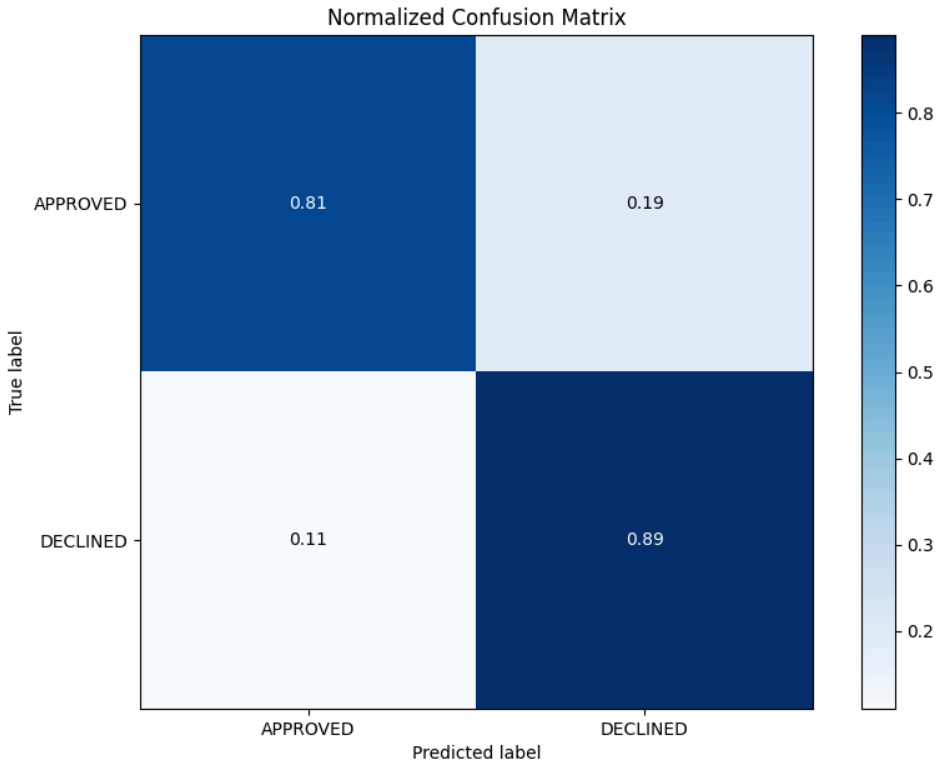
Constant Values:

- AADHAR VERIFIED
- MOBILE VERIFICATION

ADDRESS TYPE has very high Correlation with Application Status.

Conclusions

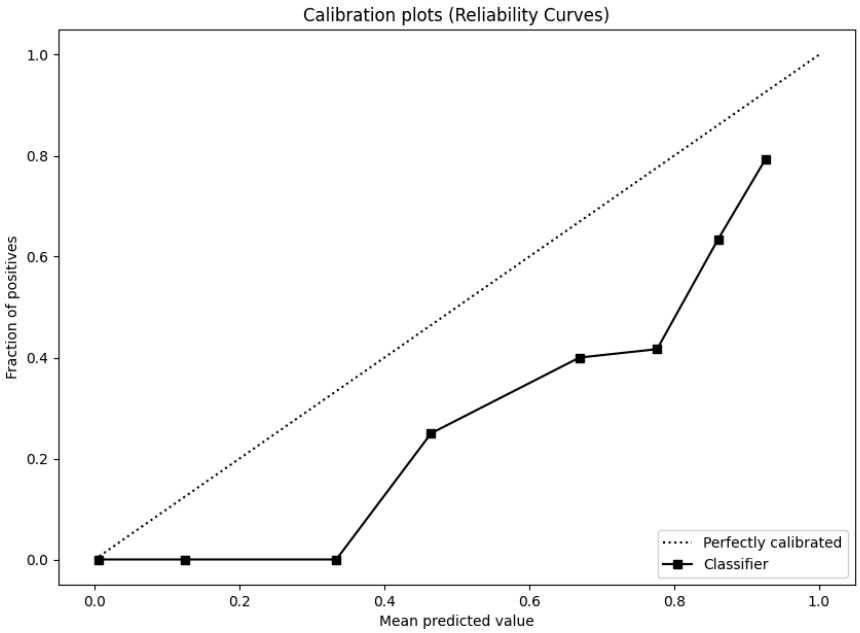
It was found that the Ensemble gave the best results, with an AUC of .91.
Further model exploration can be done along with more tuning with grid search.
Additionally different encoding techniques can be tested with.

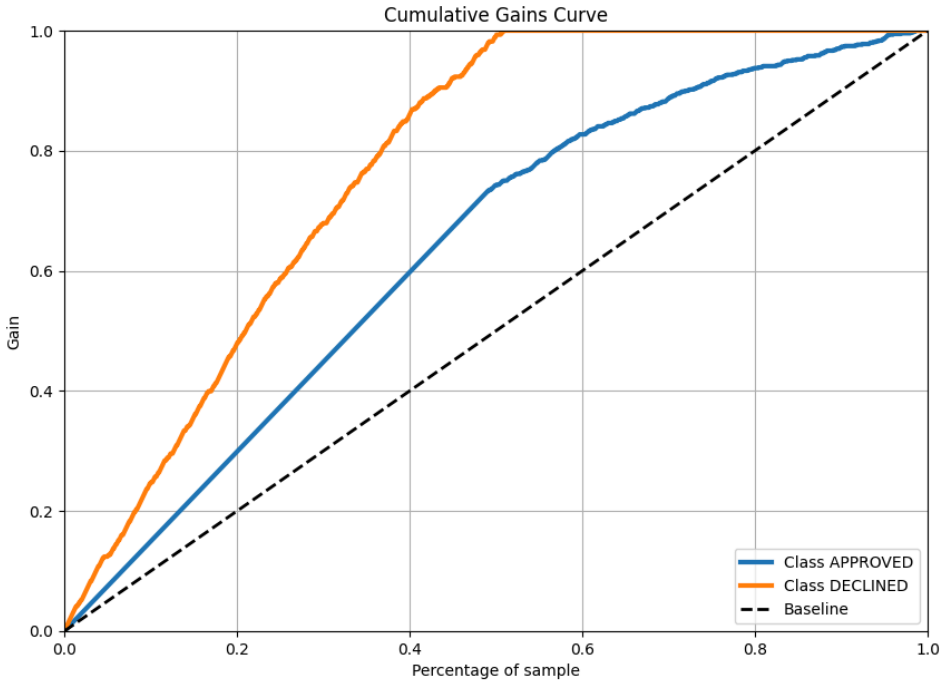


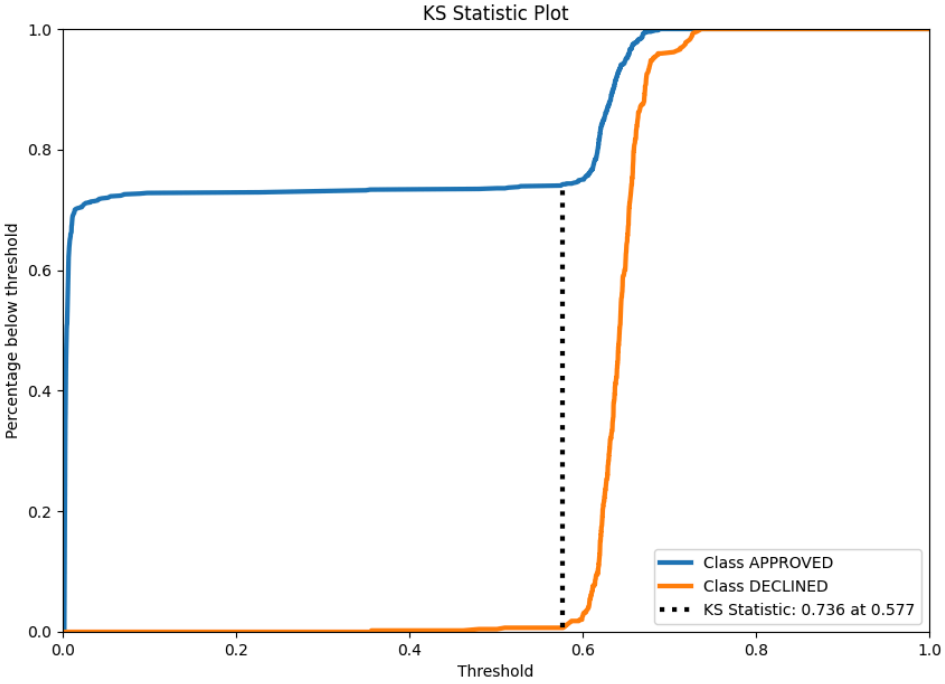
Training Results

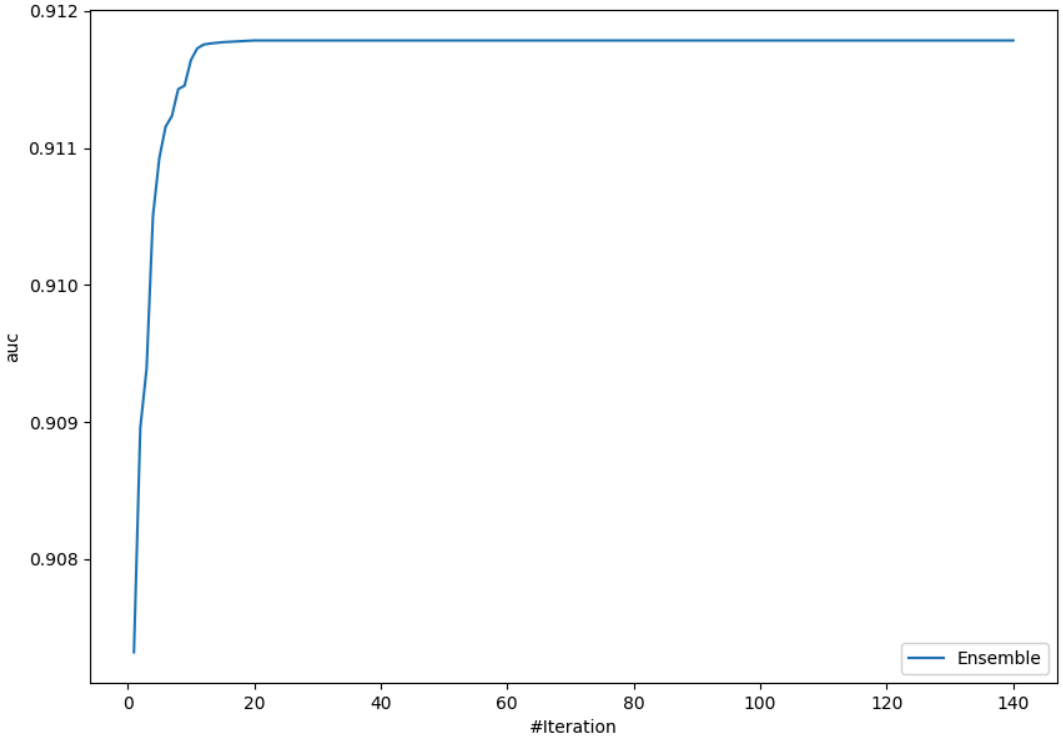
- Accuracy: 0.8423880597014926
- Precision: 0.9417525773195876
- Recall: 0.8148974130240857
- Specificity: 0.8148974130240857
- F1_score: 0.8737446197991393
- AUC: 0.911784
- ROC: 0.91

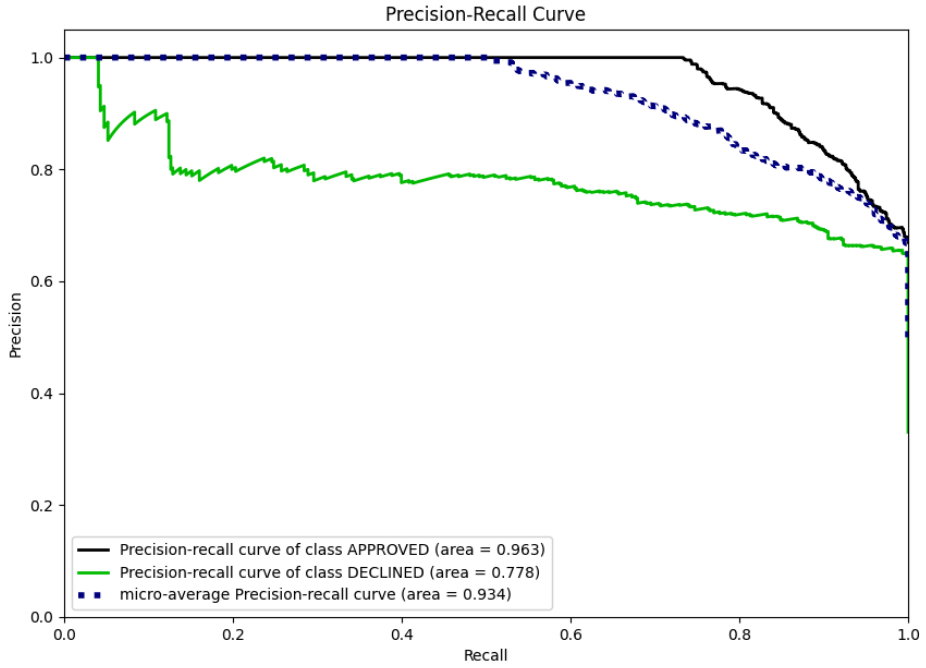
Plots

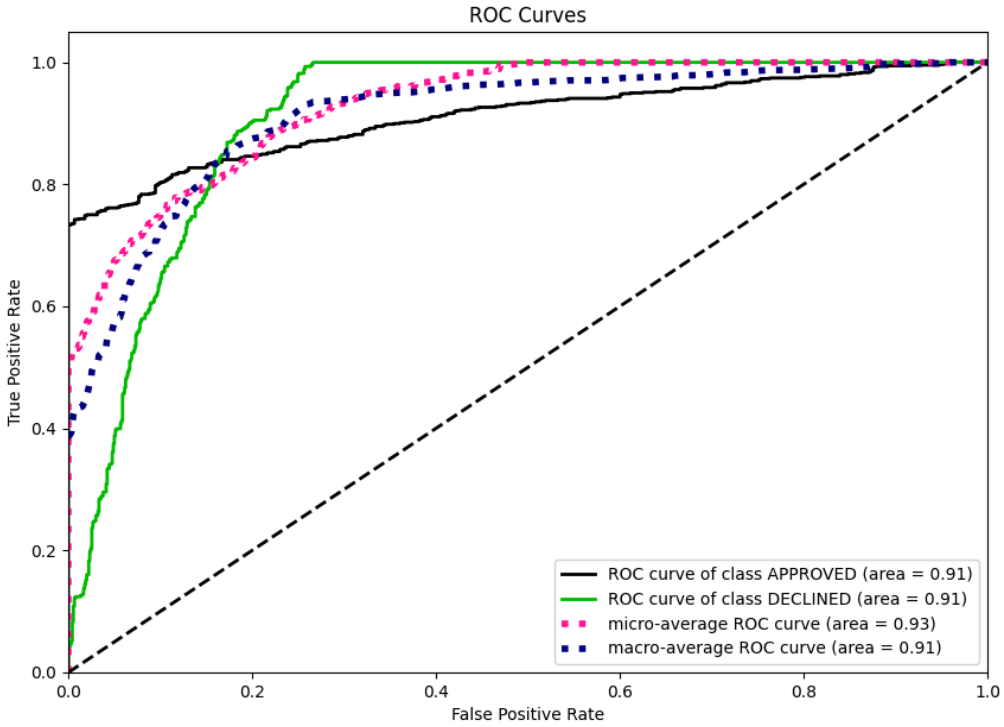



Cumulative Gains Curve
Gain
Percentage of sample
Class APPROVED
Class DECLINED
Baseline
0.0
0.2
0.4
0.6
0.8
1.0


KS Statistic Plot
Percentage below threshold
Threshold
Class APPROVED
Class DECLINED
KS Statistic: 0.736 at 0.577
0.0
0.2
0.4
0.6
0.8
1.0


0.912
0.911
0.910
0.909
0.908
auc
0
20
40
60
80
100
120
140
#Iteration
Ensemble


Precision-Recall Curve
1.0
0.8
0.6
0.4
0.2
0.0
Precision
0.0
0.2
0.4
0.6
0.8
1.0
Recall
Precision-recall curve of class APPROVED (area = 0.963)
Precision-recall curve of class DECLINED (area = 0.778)
micro-average Precision-recall curve (area = 0.934)


ROC Curves
True Positive Rate
False Positive Rate
0.0
0.2
0.4
0.6
0.8
1.0
ROC curve of class APPROVED (area = 0.91)
ROC curve of class DECLINED (area = 0.91)
micro-average ROC curve (area = 0.93)
macro-average ROC curve (area = 0.91)