L&T Loan Default Prediction

Aleksa Radojićić

Overview

- **Goal**: Predict the probability of Indian borrower defaulting on a vehicle loan in the first EMI (Equated Monthly Installments) on the due date (competition)

- **Problem type**: Supervised binary classification (target = default on first EMI)

- **Evaluation metric**: AUC (Area under the Curve)

Data Description

- Train set rows: 233,154
- Test (submission) set rows: 112,392

- 39 features:
  `PERFORM_CNS.SCORE.DESCRIPTION`,
  `asset_cost`,
  `disbursed_amount`,
  `State_ID`,
  `PRI.NO.OF.ACCTS`,
  `NO.OF_INQUIRIES`, ...

- Label: `loan_default`

Split 10% of original train set for testing:

- Train set: 209,835
- Test set: 23,316

EDA

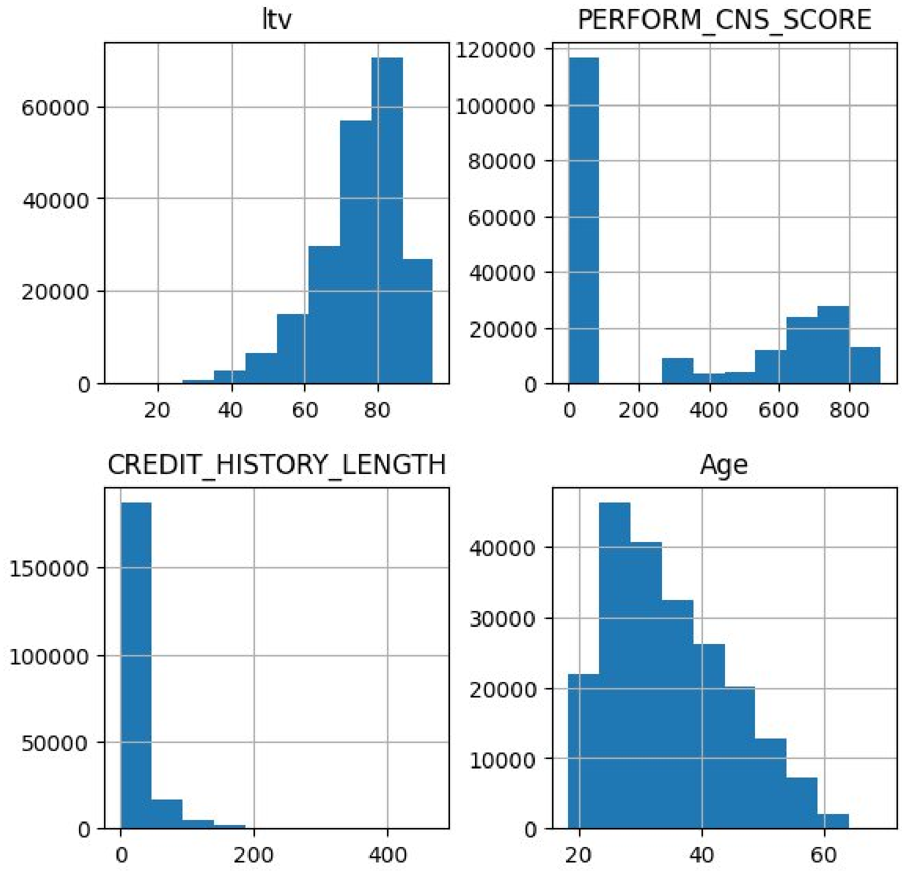
- Entire EDA performed on training data subset.

- Mean LTV (loan to value): 76.81.

- ~50.15% do not have bureau history available, meaning no CNS Score either.

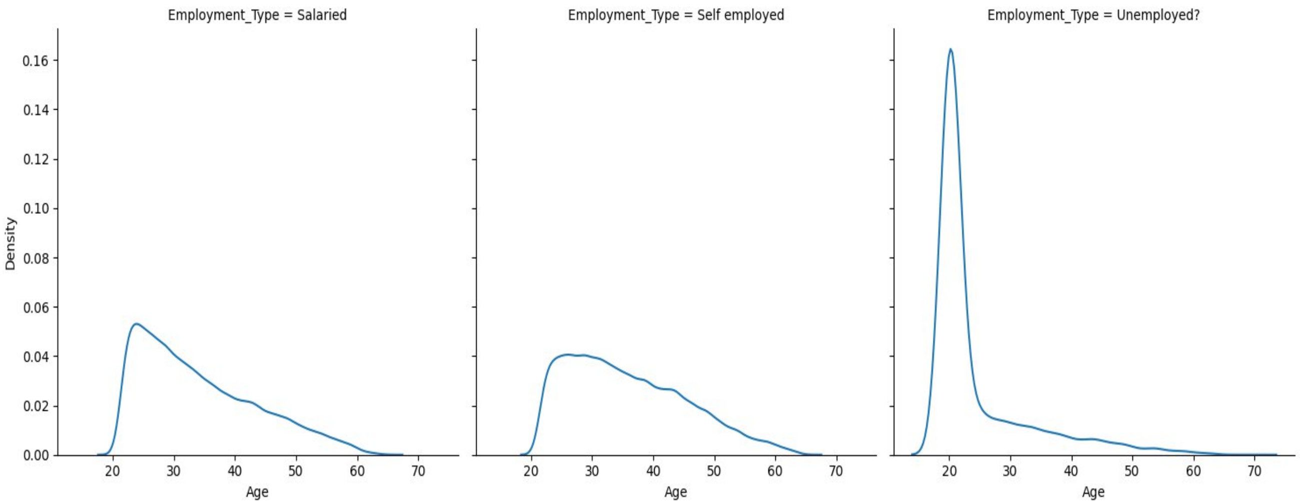
- Median age: 26.

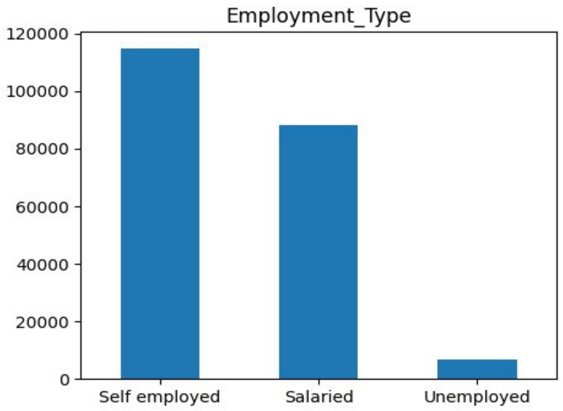
- For more than 50% of the population this was the first loan.



Preprocessing

- NA values were only in `Employment_Type` column. After examining relationship of `Age` to `Employment_Type`, there is a high number of young people in NaNs, considerably more than in other groups. Therefore, NaNs are filled with 'Unemployed' value.
- Removed rows where `PERFORM_CNS_SCORE_DESCRIPTION` = ‘Not Scored: More than 50 active Accounts found’ (3 entries).
- Removed columns for high correlation with others (>0.9): [`PRI_CURRENT_BALANCE`, `SEC_CURRENT_BALANCE`, `PRI_DISBURSED_AMOUNT`, `SEC_DISBURSED_AMOUNT`]





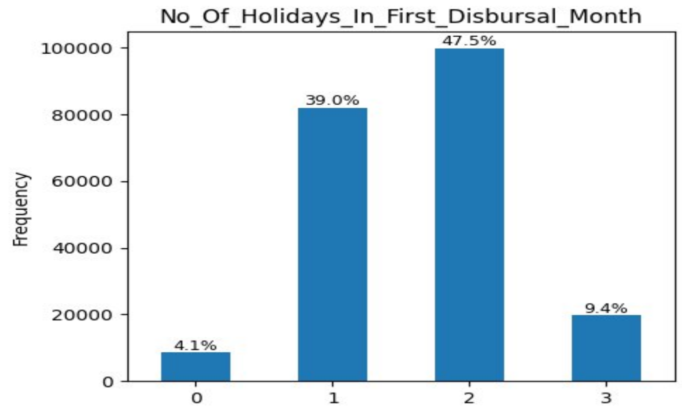
Feature Engineering – Numerical Features

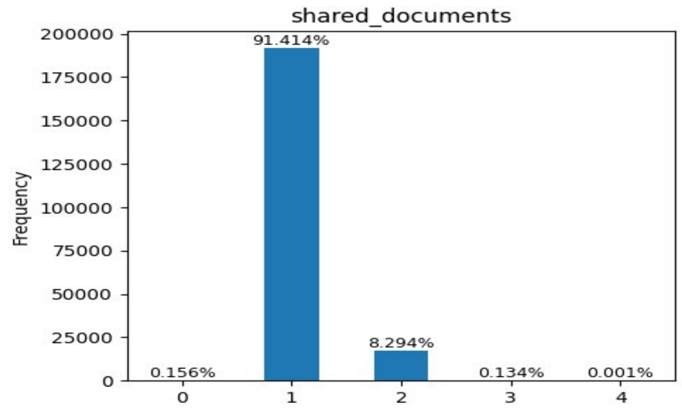
- Combined primary and secondary columns into total aggregates (prefix ‘TOT’).

- Add other numerical columns:
  [`PRI_OVERDUE_TO_ACTIVE_ACCTS_RATIO`,
  `SEC_OVERDUE_TO_ACTIVE_ACCTS_RATIO`,
  `TOT_OVERDUE_TO_ACTIVE_ACCTS_RATIO`,
  `PRI_ACTIVE_ACCTS_RATIO`,
  `SEC_ACTIVE_ACCTS_RATIO`,
  `TOT_ACTIVE_ACCTS_RATIO`,
  `DELINQUENT_TO_NEW_ACCTS_RATIO`,
  `DELINQUENT_TO_ALL_ACCTS_RATIO`,
  `PERFORM_CNS_SCORE_NORMALIZED_BY_ltv`,
  `Invalid_Age_First_Loan`].

Feature Engineering – Categorical Features

- Added categorical columns:
  `Invalid_Age_First_Loan`
  `No_Of_Holidays_In_First_Disbursal_Month`,
  `shared_documents`,
  `Month_of_Birth`



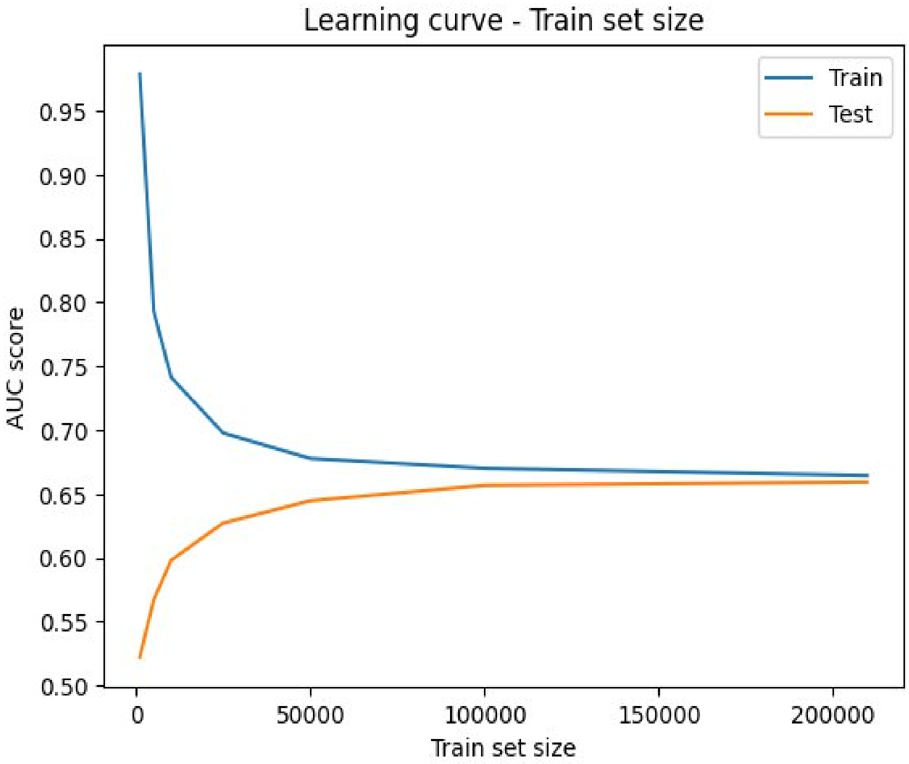


Modelling

Learning Curve – Train Set Size

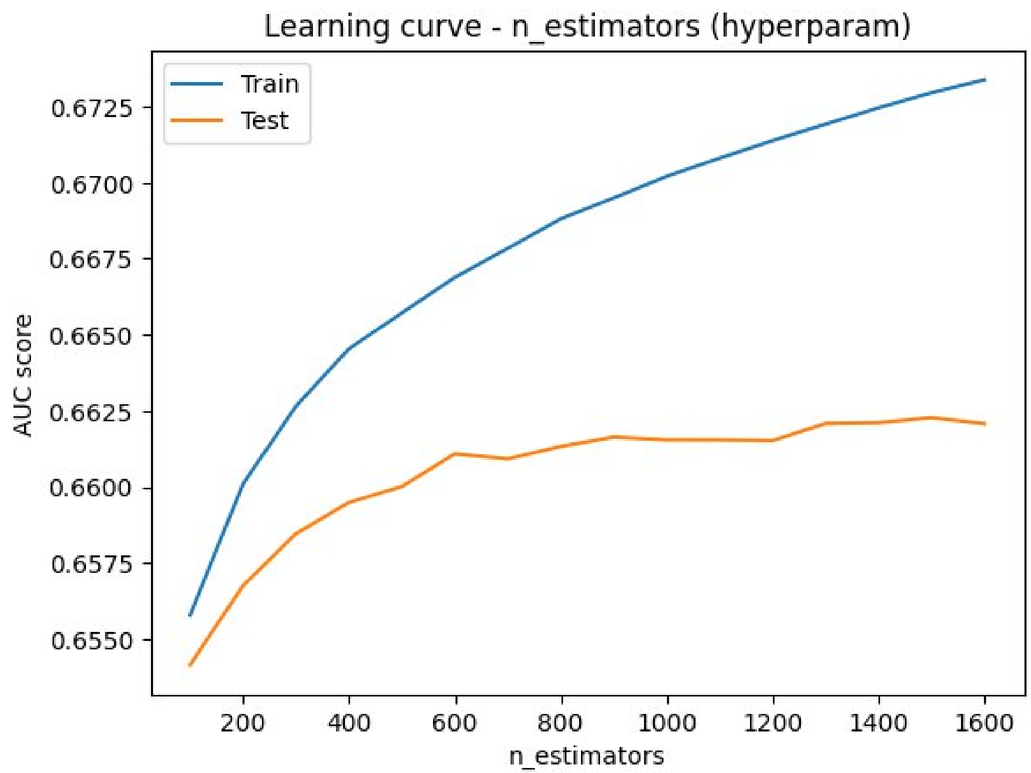
Train set sizes:
[1,000; 5,000; 10,000; 25,000; 50,000; 100,000; 209,835]

Conclusion:
Adding more data pass 200,000 set size won’t improve test AUC (high bias), implying stronger model is required for better performance.



Learning Curve – `n_estimators`

Train AdaBoost model with varying values of hyperparameter `n_estimators` from 100 to 1600 to estimate gain.



Conclusion:
Increasing the number of estimators beyond 600 yields only marginal improvements in test AUC, suggesting that further gains may require tuning other hyperparameters, such as the base tree depth or enhancing feature engineering.

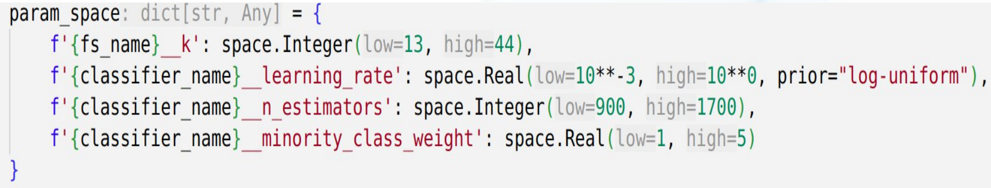
Hyperparameters Tuning pt. 1 – BayesSearchCV

- Total features after feature engineering: 44

- Pipeline:

1. Preprocessing pipeline
2. Feature selection using `mutual_info_classif`
3. AdaBoost model

- BayesSearch stratified 5-fold cross validation with 30 iterations, with hyperparameter space:



```
param_space: dict[str, Any] = {
    f'{fs_name}__k': space.Integer(low=13, high=44),
    f'{classifier_name}__learning_rate': space.Real(low=10**-3, high=10**0, prior="log-uniform"),
    f'{classifier_name}__n_estimators': space.Integer(low=900, high=1700),
    f'{classifier_name}__minority_class_weight': space.Real(low=1, high=5)
}
```

Hyperparameters Tuning pt. 2 – GridSearchCV

- GridSearch stratified 5-fold cross validation for the split criterion hyperparameter: ["gini", "entropy"]

- Refit the best model on the entire train data.

Results

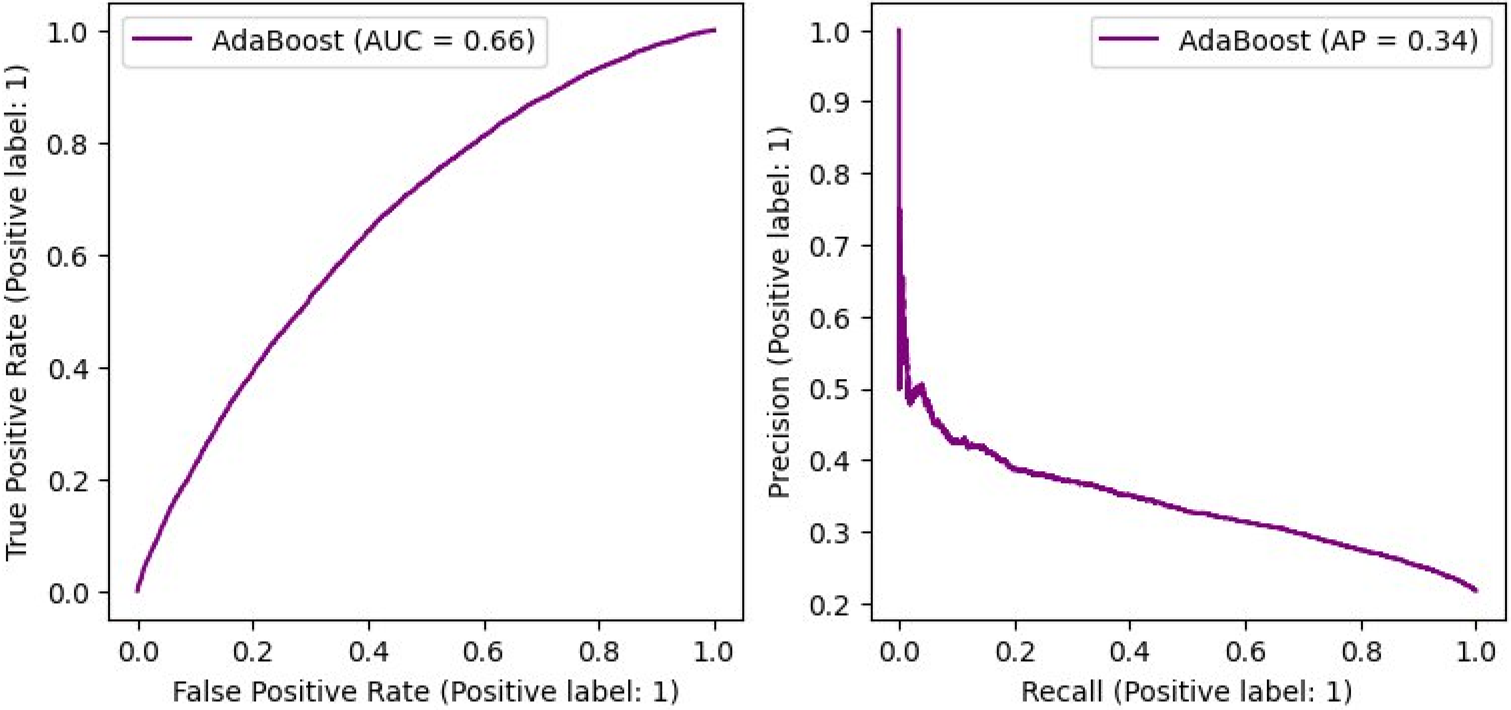
Optimal Hyperparameters

Hyperparameter	**Value**
AdaBoost__learning_rate	0.8026
AdaBoost__minority_class_weight	5.0
AdaBoost__n_estimators	1700
AdaBoost__estimator__criterion	gini
mutualinfoclassif__k	42

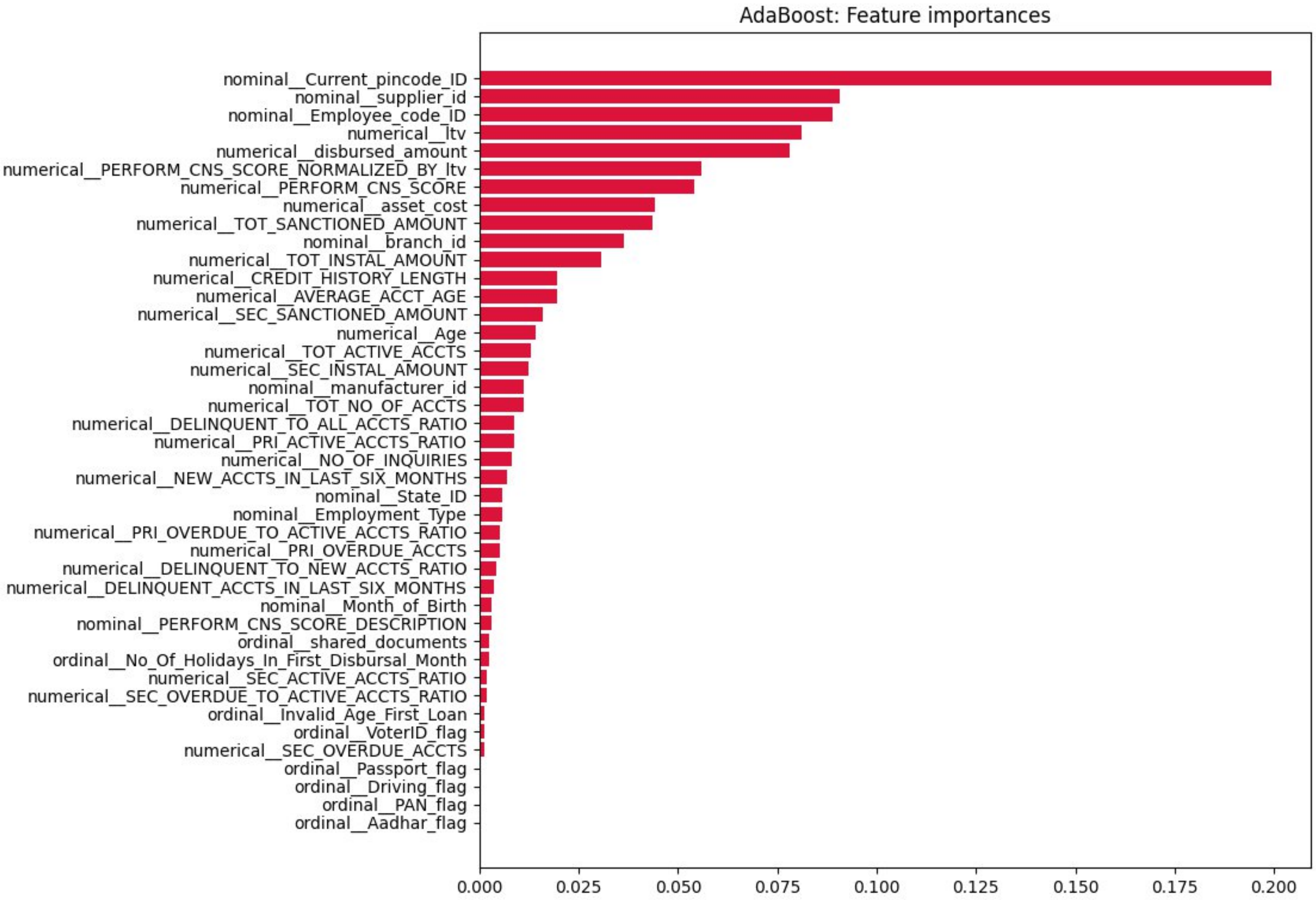
Evaluation

Dataset	**AUC**
CV	65.888%
Train Set	67.144%
Test Set	**66.302%**
Entire Set	67.069%

ROC & PR curve


AdaBoost (AUC = 0.66)
1.0
0.8
0.6
0.4
0.2
0.0
True Positive Rate (Positive label: 1)
0.0
0.2
0.4
0.6
0.8
1.0
False Positive Rate (Positive label: 1)
AdaBoost (AP = 0.34)
1.0
0.9
0.8
0.7
0.6
0.5
0.4
0.3
0.2
Precision (Positive label: 1)
0.0
0.2
0.4
0.6
0.8
1.0
Recall (Positive label: 1)

Feature Importance


AdaBoost: Feature importances
nominal__Current_pincode_ID
nominal__supplier_id
nominal__Employee_code_ID
numerical__ltv
numerical__disbursed_amount
numerical__PERFORM_CNS_SCORE_NORMALIZED_BY_ltv
numerical__PERFORM_CNS_SCORE
numerical__asset_cost
numerical__TOT_SANCTIONED_AMOUNT
nominal__branch_id
numerical__TOT_INSTAL_AMOUNT
numerical__CREDIT_HISTORY_LENGTH
numerical__AVERAGE_ACCT_AGE
numerical__SEC_SANCTIONED_AMOUNT
numerical__Age
numerical__TOT_ACTIVE_ACCTS
numerical__SEC_INSTAL_AMOUNT
nominal__manufacturer_id
numerical__TOT_NO_OF_ACCTS
numerical__DELINQUENT_TO_ALL_ACCTS_RATIO
numerical__PRI_ACTIVE_ACCTS_RATIO
numerical__NO_OF_INQUIRIES
numerical__NEW_ACCTS_IN_LAST_SIX_MONTHS
nominal__State_ID
nominal__Employment_Type
numerical__PRI_OVERDUE_TO_ACTIVE_ACCTS_RATIO
numerical__PRI_OVERDUE_ACCTS
numerical__DELINQUENT_TO_NEW_ACCTS_RATIO
numerical__DELINQUENT_ACCTS_IN_LAST_SIX_MONTHS
nominal__Month_of_Birth
nominal__PERFORM_CNS_SCORE_DESCRIPTION
ordinal__shared_documents
ordinal__No_Of_Holidays_In_First_Disbursal_Month
numerical__SEC_ACTIVE_ACCTS_RATIO
numerical__SEC_OVERDUE_TO_ACTIVE_ACCTS_RATIO
ordinal__Invalid_Age_First_Loan
ordinal__VoterID_flag
numerical__SEC_OVERDUE_ACCTS
ordinal__Passport_flag
ordinal__Driving_flag
ordinal__PAN_flag
ordinal__Aadhar_flag