A Machine Learning Approach to Credit Card Default Prediction

BY ALISHBA TAHIR

1. Problem Statement

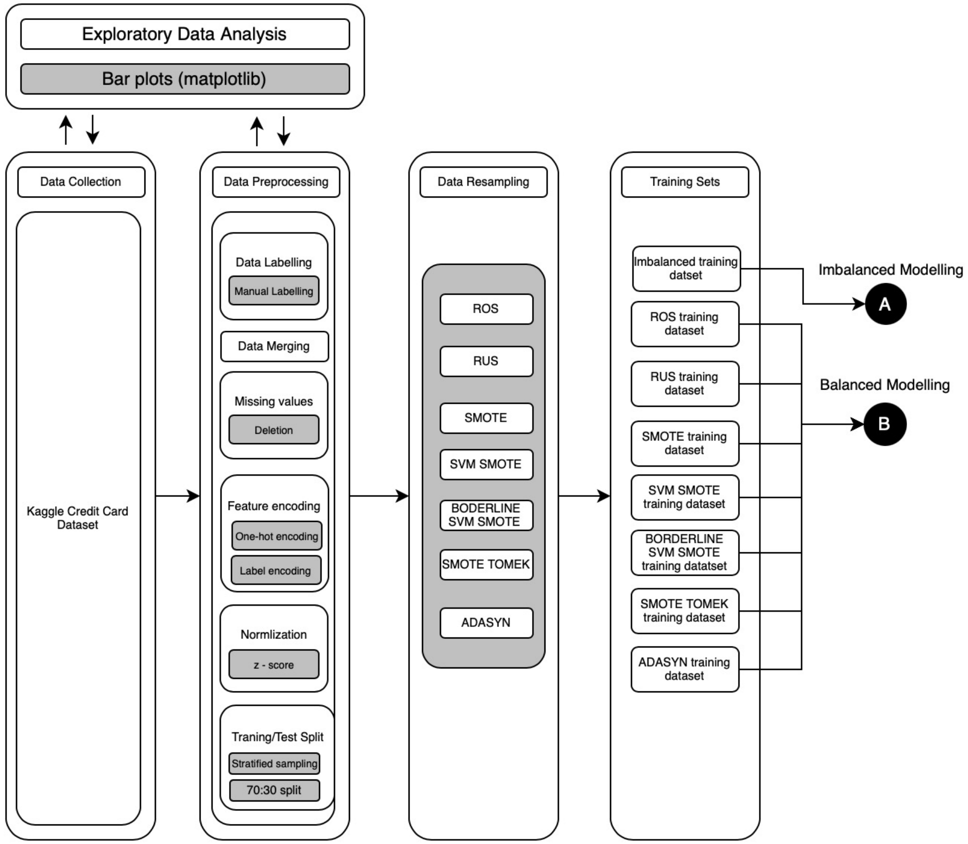
- The surge in credit card and bank loans has come with an increase in loan defaults, costing banks billions annually. To navigate this rising risk, accurate and efficient credit assessment has become more crucial than ever.

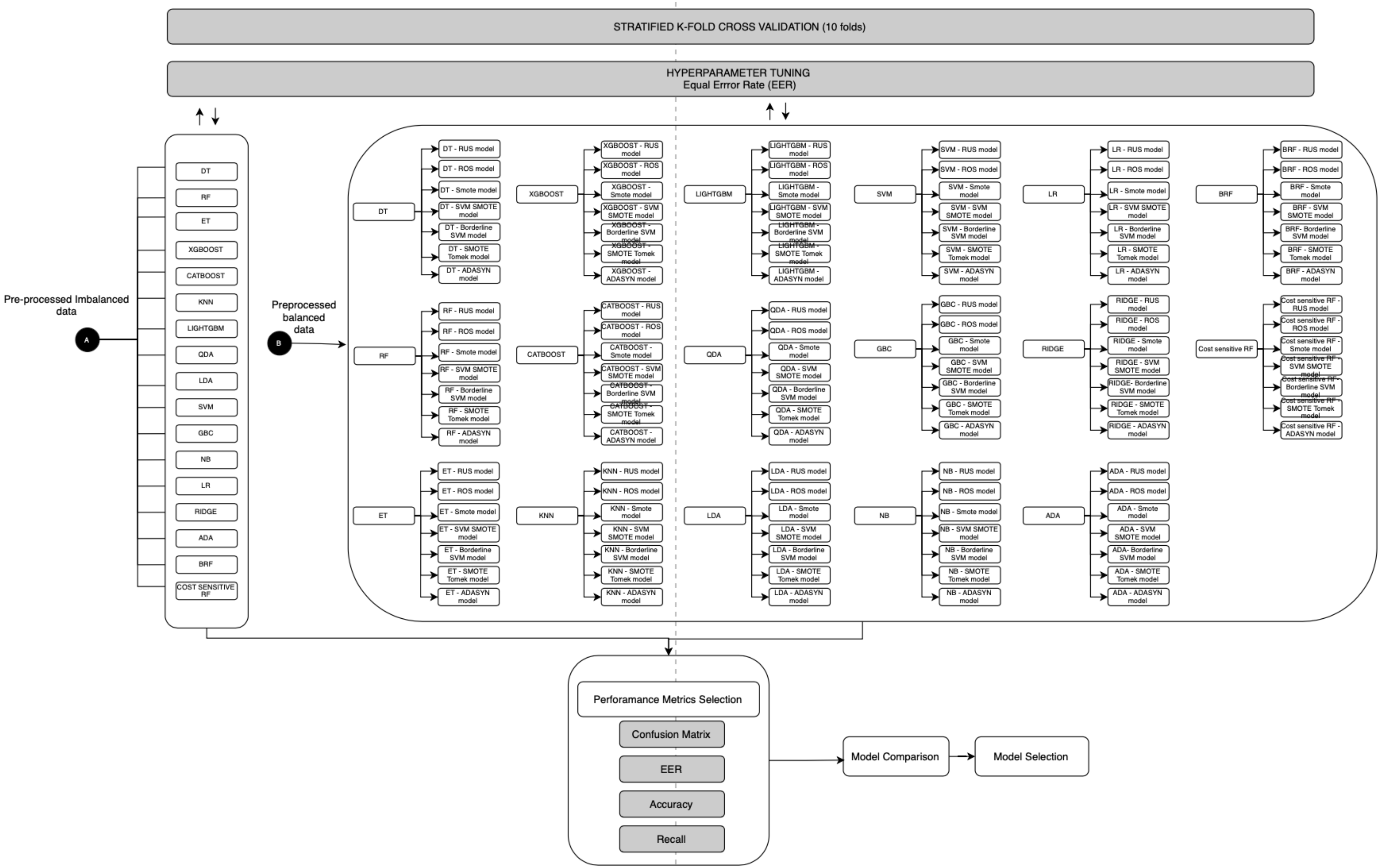
2. Project Aim

- This project aims to identify machine learning models that can effectively predict credit card and loan defaults.

3. Methodology

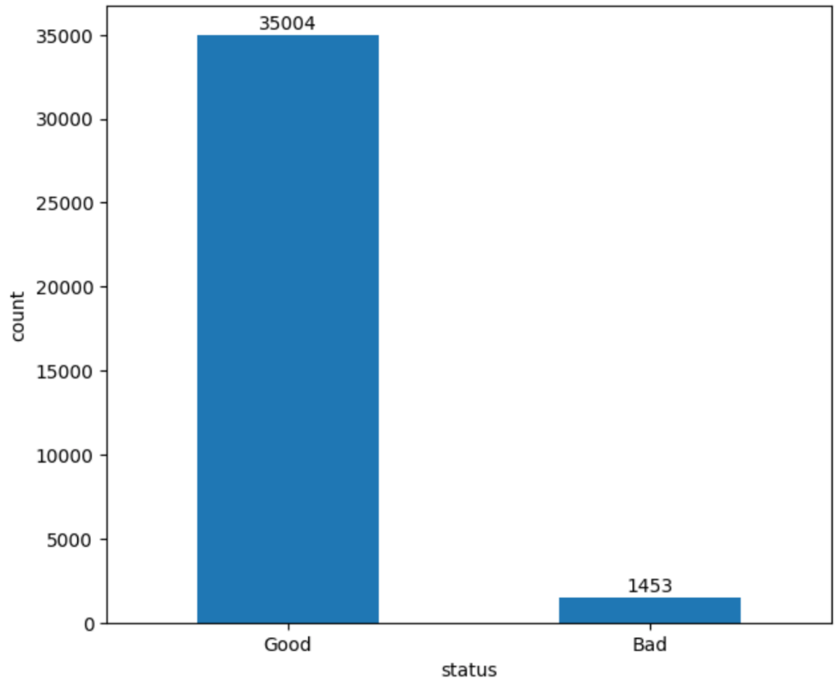
DATA PREPROCESSING



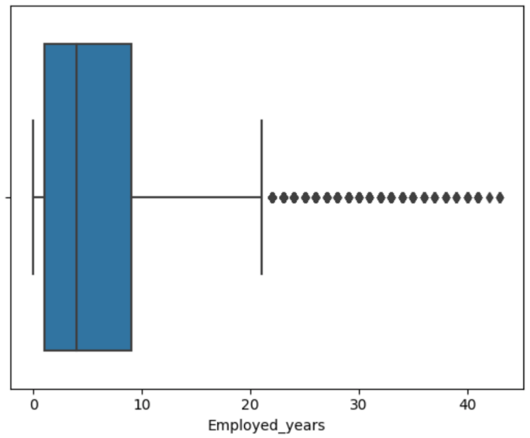

STRATIFIED K-FOLD CROSS VALIDATION (10 folds)
HYPERPARAMETER TUNING
Equal Errror Rate (EER)
Pre-processed Imbalanced data
A
DT
RF
ET
XGBOOST
CATBOOST
KNN
LIGHTGBM
QDA
LDA
SVM
GBC
NB
LR
RIDGE
ADA
BRF
COST SENSITIVE RF
Preprocessed balanced data
B
DT
DT - RUS model
DT - ROS model
DT - Smote model
DT - SVM SMOTE model
DT - Borderline SVM model
DT - SMOTE Tomek model
DT - ADASYN model
XGBOOST
XGBOOST - RUS model
XGBOOST - ROS model
XGBOOST - Smote model
XGBOOST - SVM SMOTE model
XGBOOST - Borderline SVM model
XGBOOST - SMOTE Tomek model
XGBOOST - ADASYN model
LIGHTGBM
LIGHTGBM - RUS model
LIGHTGBM - ROS model
LIGHTGBM - Smote model
LIGHTGBM - SVM SMOTE model
LIGHTGBM - Borderline SVM model
LIGHTGBM - SMOTE Tomek model
LIGHTGBM - ADASYN model
SVM
SVM - RUS model
SVM - ROS model
SVM - Smote model
SVM - SVM SMOTE model
SVM - Borderline SVM model
SVM - SMOTE Tomek model
SVM - ADASYN model
LR
LR - RUS model
LR - ROS model
LR - Smote model
LR - SVM SMOTE model
LR - Borderline SVM model
LR - SMOTE Tomek model
LR - ADASYN model
BRF
BRF - RUS model
BRF - ROS model
BRF - Smote model
BRF - SVM SMOTE model
BRF- Borderline SVM model
BRF - SMOTE Tomek model
BRF - ADASYN model
RF
RF - RUS model
RF - ROS model
RF - Smote model
RF - SVM SMOTE model
RF - Borderline SVM model
RF - SMOTE Tomek model
RF - ADASYN model
CATBOOST
CATBOOST - RUS model
CATBOOST - ROS model
CATBOOST - Smote model
CATBOOST - SVM SMOTE model
CATBOOST - Borderline SVM model
CATBOOST - SMOTE Tomek model
CATBOOST - ADASYN model
QDA
QDA - RUS model
QDA - ROS model
QDA - Smote model
QDA - SVM SMOTE model
QDA - Borderline SVM model
QDA - SMOTE Tomek model
QDA - ADASYN model
GBC
GBC - RUS model
GBC - ROS model
GBC - Smote model
GBC - SVM SMOTE model
GBC - Borderline SVM model
GBC - SMOTE Tomek model
GBC - ADASYN model
RIDGE
RIDGE - RUS model
RIDGE - ROS model
RIDGE - Smote model
RIDGE - SVM SMOTE model
RIDGE- Borderline SVM model
RIDGE - SMOTE Tomek model
RIDGE - ADASYN model
Cost sensitive RF
Cost sensitive RF - RUS model
Cost sensitive RF - ROS model
Cost sensitive RF - Smote model
Cost sensitive RF - SVM SMOTE model
Cost sensitive RF - Borderline SVM model
Cost sensitive RF - SMOTE Tomek model
Cost sensitive RF - ADASYN model
ET
ET - RUS model
ET - ROS model
ET - Smote model
ET - SVM SMOTE model
ET - Borderline SVM model
ET - SMOTE Tomek model
ET - ADASYN model
KNN
KNN - RUS model
KNN - ROS model
KNN - Smote model
KNN - SVM SMOTE model
KNN - Borderline SVM model
KNN - SMOTE Tomek model
KNN - ADASYN model
LDA
LDA - RUS model
LDA - ROS model
LDA - Smote model
LDA - SVM SMOTE model
LDA - Borderline SVM model
LDA - SMOTE Tomek model
LDA - ADASYN model
NB
NB - RUS model
NB - ROS model
NB - Smote model
NB - SVM SMOTE model
NB - Borderline SVM model
NB - SMOTE Tomek model
NB - ADASYN model
ADA
ADA - RUS model
ADA - ROS model
ADA - Smote model
ADA - SVM SMOTE model
ADA- Borderline SVM model
ADA - SMOTE Tomek model
ADA - ADASYN model
Perforamance Metrics Selection
Confusion Matrix
EER
Accuracy
Recall
Model Comparison
Model Selection

4. Challenges

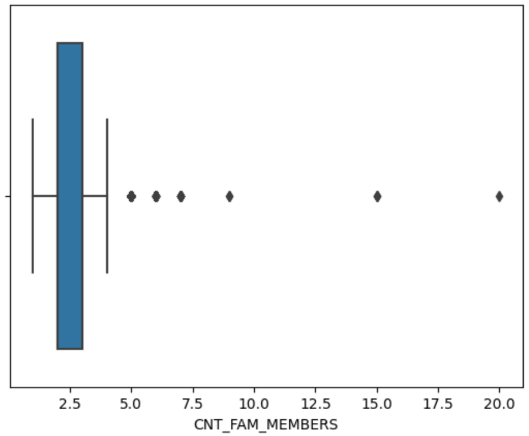
- 4.1. Imbalanced Data



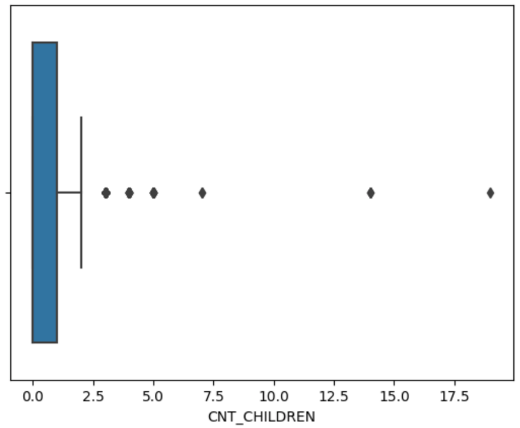
4.2. Outliers



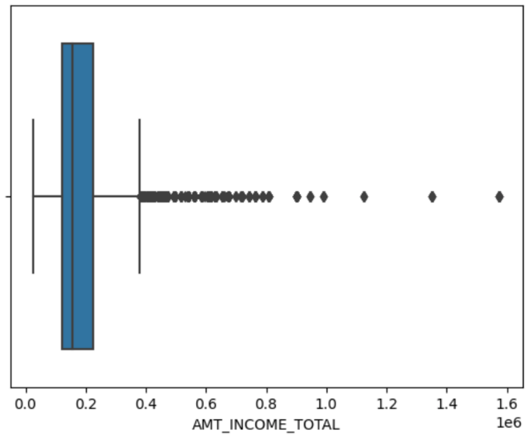
0
10
20
30
40
Employed_years



2.5
5.0
7.5
10.0
12.5
15.0
17.5
20.0
CNT_FAM_MEMBERS



0.0
2.5
5.0
7.5
10.0
12.5
15.0
17.5
CNT_CHILDREN



0.0
0.2
0.4
0.6
0.8
1.0
1.2
1.4
1.6
1e6
AMT_INCOME_TOTAL

4.3. Data Merging

application.csv

```
RangeIndex: 438557 entries, 0 to 438556
Data columns (total 18 columns):
 #   Column               Non-Null Count   Dtype
---  ------               --------------   -----
 0   ID                   438557 non-null  int64
 1   CODE_GENDER          438557 non-null  object
 2   FLAG_OWN_CAR         438557 non-null  object
 3   FLAG_OWN_REALTY      438557 non-null  object
 4   CNT_CHILDREN         438557 non-null  int64
 5   AMT_INCOME_TOTAL     438557 non-null  float64
 6   NAME_INCOME_TYPE     438557 non-null  object
 7   NAME_EDUCATION_TYPE  438557 non-null  object
 8   NAME_FAMILY_STATUS   438557 non-null  object
 9   NAME_HOUSING_TYPE    438557 non-null  object
 10  DAYS_BIRTH           438557 non-null  int64
 11  DAYS_EMPLOYED        438557 non-null  int64
 12  FLAG_MOBIL           438557 non-null  int64
 13  FLAG_WORK_PHONE      438557 non-null  int64
 14  FLAG_PHONE           438557 non-null  int64
 15  FLAG_EMAIL           438557 non-null  int64
 16  OCCUPATION_TYPE      304354 non-null  object
 17  CNT_FAM_MEMBERS      438557 non-null  float64
dtypes: float64(2), int64(8), object(8)
```

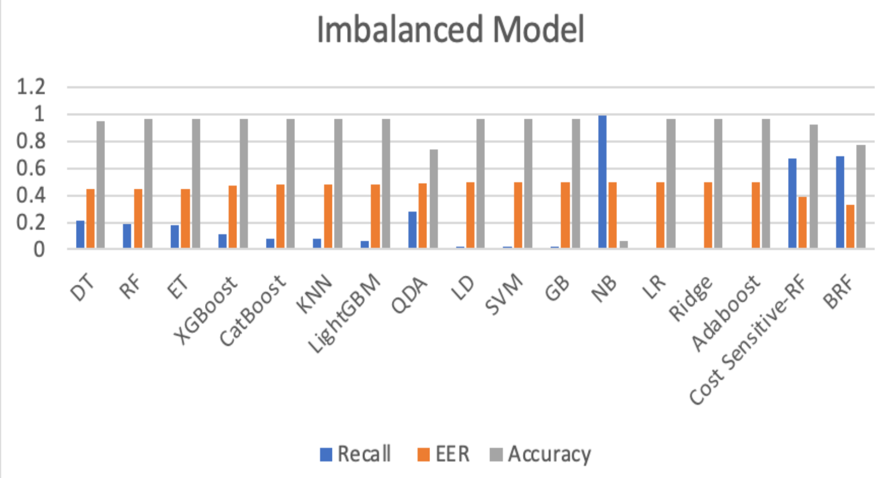
credit.csv

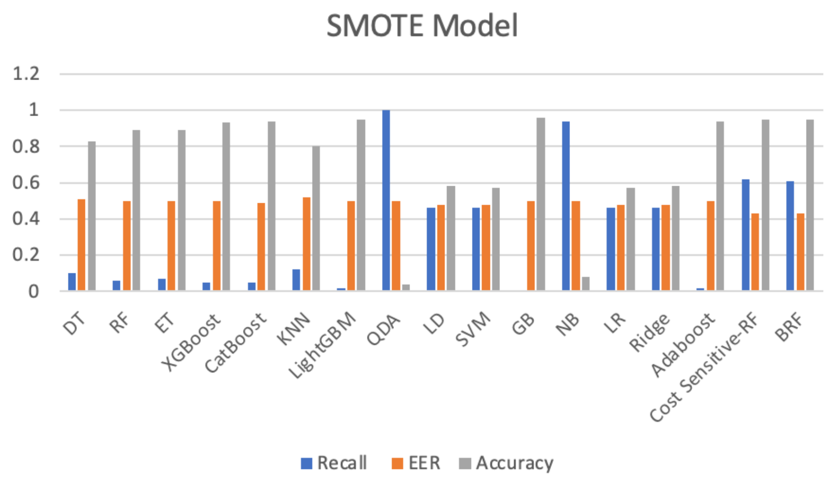
```
RangeIndex: 1048575 entries, 0 to 1048574
Data columns (total 3 columns):
 #   Column          Non-Null Count    Dtype
---  ------          --------------    -----
 0   ID              1048575 non-null  int64
 1   MONTHS_BALANCE  1048575 non-null  int64
 2   STATUS          1048575 non-null  object
```

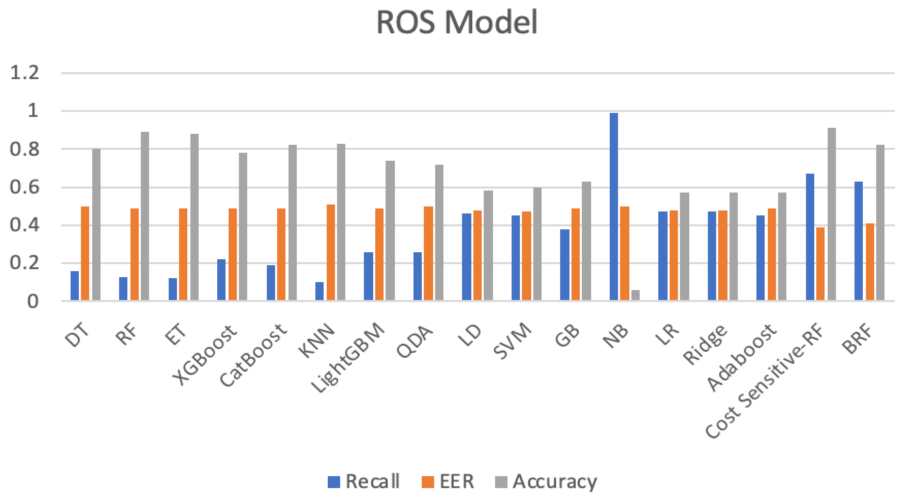
4.4. Hyperparameter Tuning

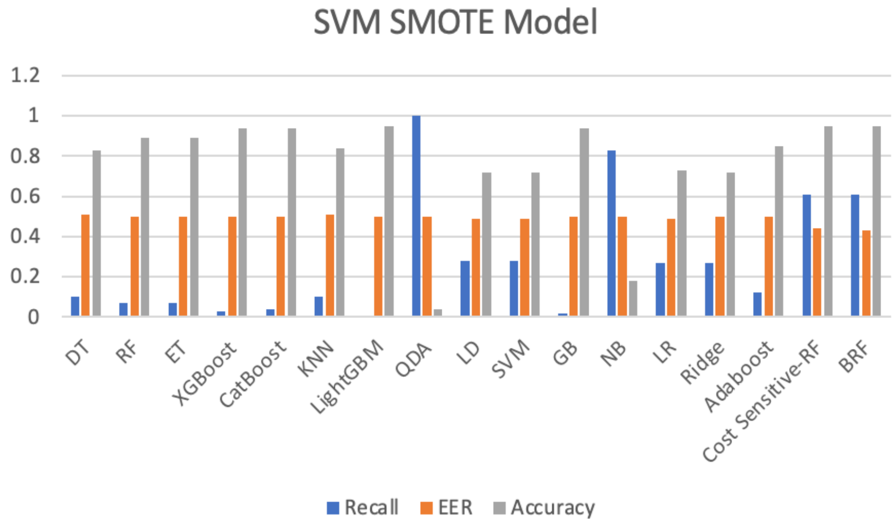
- Navigating the vast parameter space and identifying the most optimal set of parameters to maximize model performance and avoid overfitting was a complex task.
- Fine-tuning ensemble models such as random forest using GridCVSearch proved to be highly resource-intensive, requiring a substantial 5 hours to optimize a single model within the constraints of a limited parameter set.

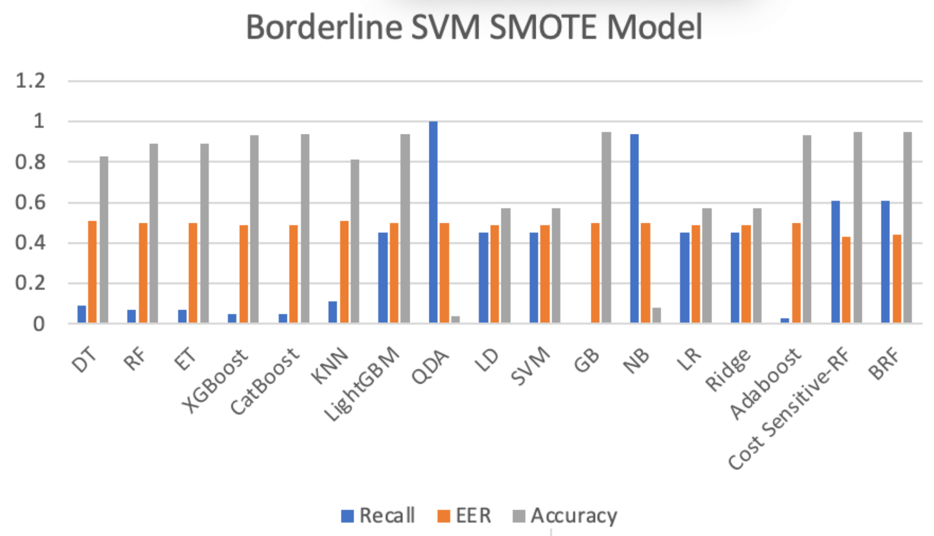
Model Results

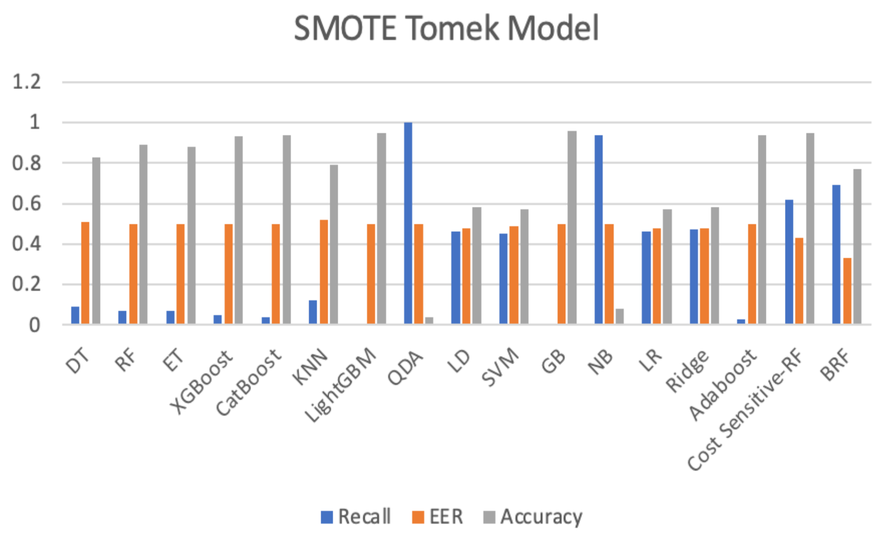


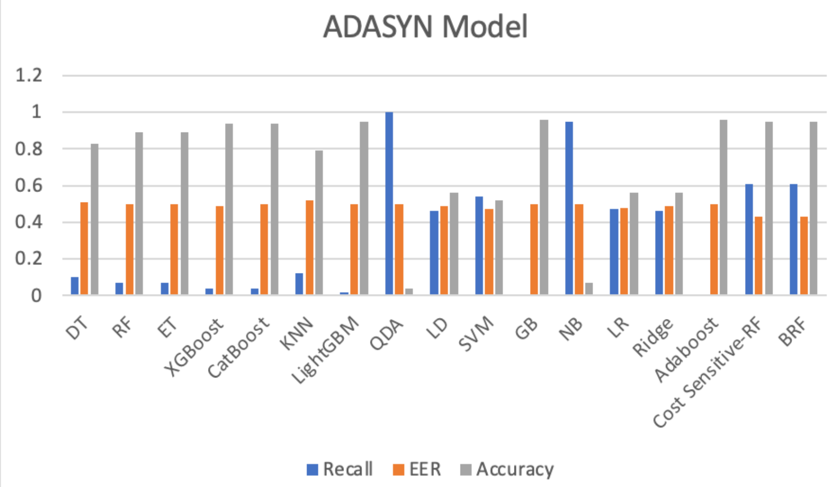






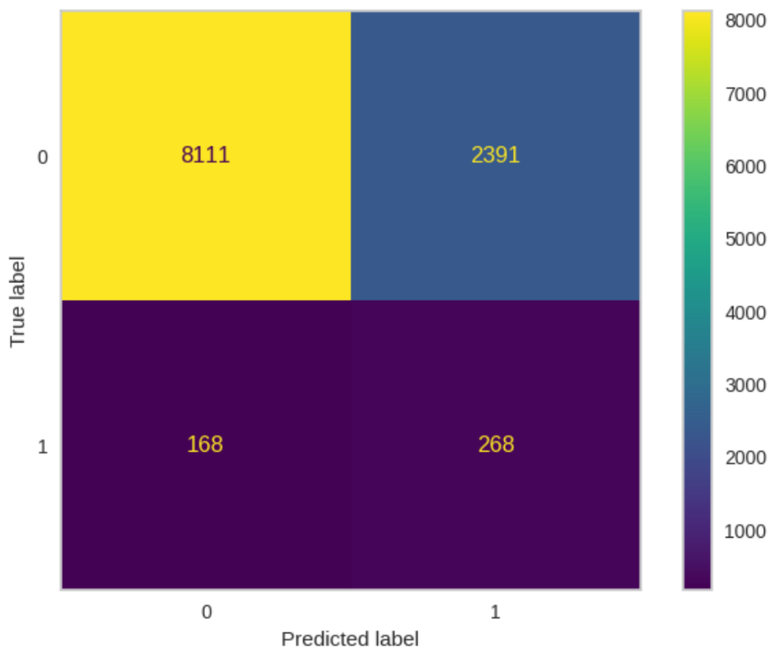

Borderline SVM SMOTE Model
1.2
1
0.8
0.6
0.4
0.2
0
DT
RF
ET
XGBoost
CatBoost
KNN
LightGBM
QDA
LD
SVM
GB
NB
LR
Ridge
Adaboost
Cost Sensitive-RF
BRF
Recall
EER
Accuracy

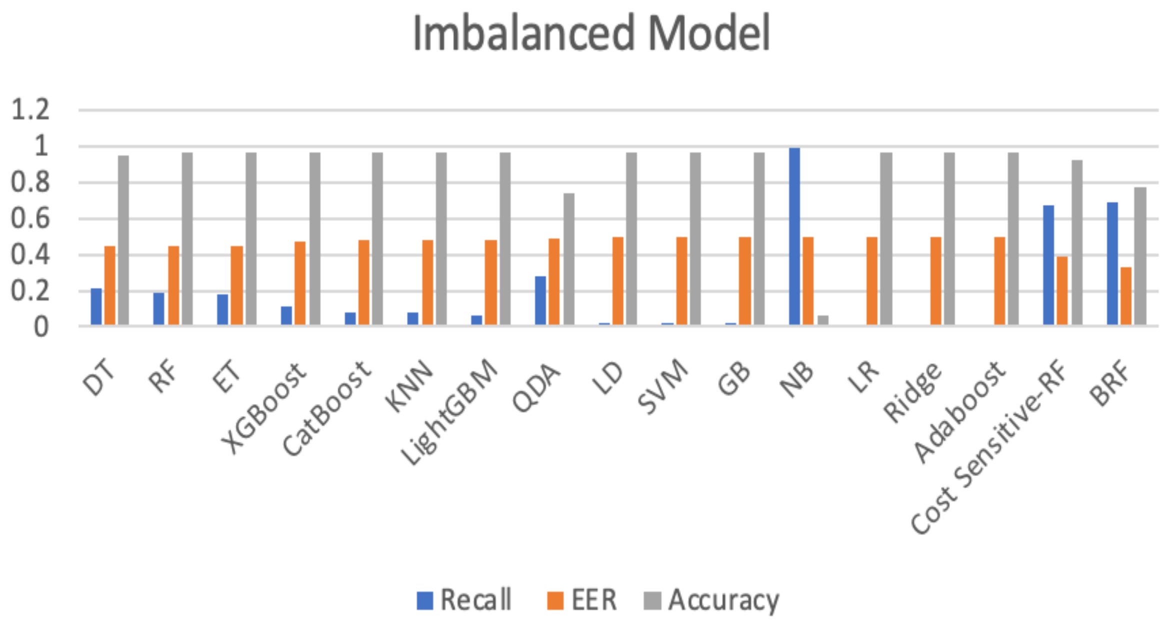

SMOTE Tomek Model
1.2
1
0.8
0.6
0.4
0.2
0
DT
RF
ET
XGBoost
CatBoost
KNN
LightGBM
QDA
LD
SVM
GB
NB
LR
Ridge
Adaboost
Cost Sensitive-RF
BRF
Recall
EER
Accuracy


ADASYN Model
1.2
1
0.8
0.6
0.4
0.2
0
DT
RF
ET
XGBoost
CatBoost
KNN
LightGBM
QDA
LD
SVM
GB
NB
LR
Ridge
Adaboost
Cost Sensitive-RF
BRF
Recall
EER
Accuracy

Model Selection

- Balanced Random Forest





Conclusion

- To find the optimal model to predict loan defaulters we built seventeen machine learning classification models Decision Trees, Random Forest, Light Gradient Boosting Machine, Linear Discriminant Analysis, Ridge Classifier, Logistic Regression, CatBoost, AdaBoost, Extreme Gradient Boosting, K-Nearest Neighbors, SVM with a Linear Kernel, Quadratic Discriminant Analysis, Extra Trees, Gradient Boosting, Naïve Bayes, Balanced Random Forest and Cost Sensitive Learning (Random Forest).
- To tackle the imbalance data problem we used different variations of datasets using seven different data sampling techniques namely SMOTE, SVM SMOTE, BORDERLINE SVM SMOTE, SMOTE TOMEK, ADASYN, Random Undersampling, and Random Oversampling.
- Balanced Random Forest with imbalanced data and default parameters stands out with its well-rounded performance with an accuracy of 77% and lowest EER score of 0.33.

Limitations

- We did not investigate any feature selection methods.
- We could not conduct a more extensive search for optimal hyperparameters for the selected models.
- We did not incorporate deep learning algorithms.

THANK YOU!