



Business Process Improvement: SBA Loan Charge-Off Analysis

BA810: Supervised Machine Learning

Cohort B Team 13

Boyuan (Daniel) Zhang

Yutao (Peter) Luo

Yen-Chun (Albert) Chen

Xinyuan (Klaus) Xu

[Colab Notebook](#)

Background Information



Lending partner



Loan



Small business
(You)

SBA reduces risk and enables easier access to capital.

Project Goals



Our goals are to:

- **Predict** whether a SBA-backed loan to a US small business will be **paid in full or charged-off** using selected data,
- **Compare** prediction results across different models,
- **Make recommendations** to the SBA about updating their existing model on choosing which loans to guarantee on.

Potential Audience

- U.S. Small Business Administration (SBA)
- Commercial Banks (Fidelity, Capital One, etc.)
- Credit Unions
- Any government entities or lending organization.



Data Source



MIRBEK TOKTOGARAEV · UPDATED 3 YEARS AGO



127

New Notebook

Download (49 MB)



Should This Loan be Approved or Denied?

A large and rich dataset from the U.S. Small Business Administration (SBA)



U.S. Small
Administration

Raw Data:

The raw data is from U.S. Small Business Administration Open Data (public):

[U.S. Small Business Administration \(SBA\) | Open Data Statistics](#)

Modified Data:

The data we analyzed in this project was modified by Mirbek Toktogaraev, a Junior Data Analyst at Ubiquim Code Academy (Madrid, Spain.) The processed data is accessible on Kaggle.com.

[Should This Loan be Approved or Denied? -by Mirbek Toktogaraev](#)

Variables Selected



Numeric Variables (4):

“Term” : Term of the loan

“NoEmp”: Number of employee of the business

“RetainedJob”: Total jobs retained

“SBA_Appv”: Amount approved by SBA.

Target Variables (1):

“MIS” : 1 = Loan was Charged Off

0 = Loan was Paid in Full

Categorical Variables (8):

“UrbanRural”: Location of the business

“NewOrExist”: Type of the business

“Franchised”: If the business was franchised

“StateSame”: Borrow went to the bank of their state

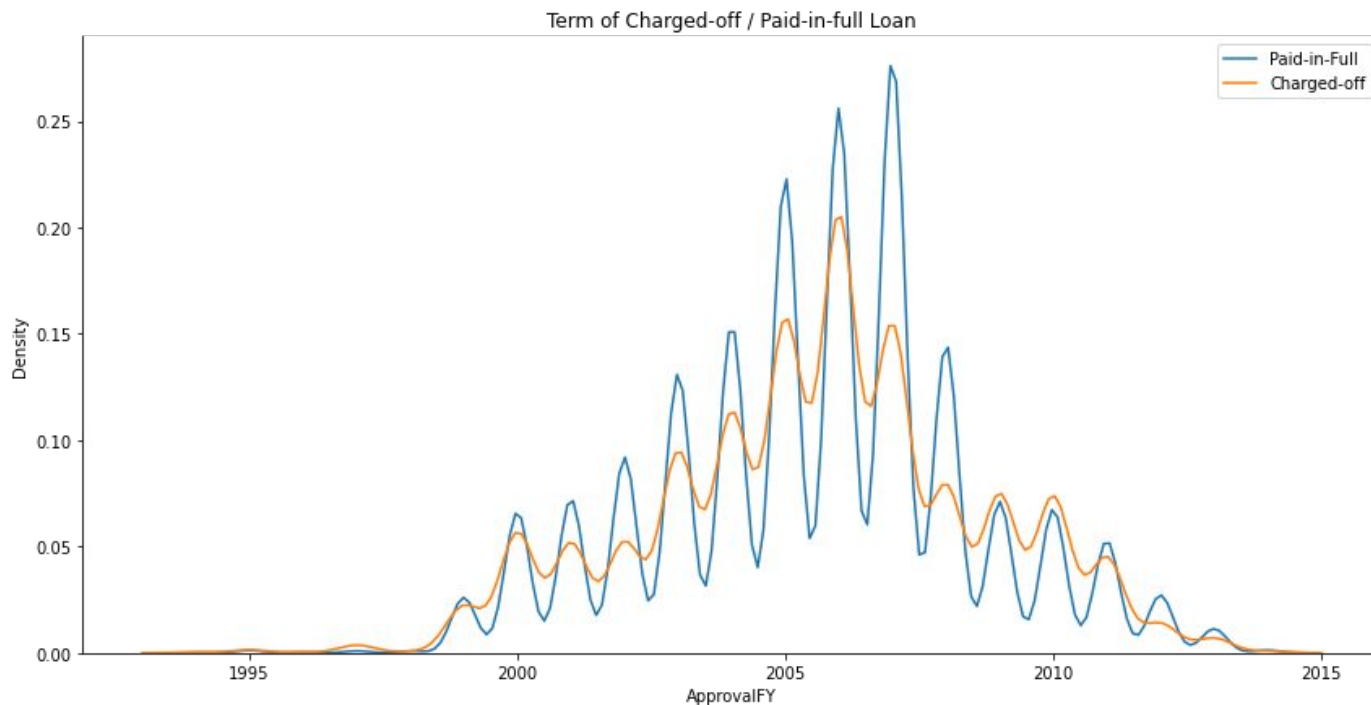
“FullDisbursed”: If the bank fully disbursed the amount that was approved by it.

“RevLineCredit”: If the business had a revolving credit line.

“LowDocP: If the business was in the low document loan program.

New Added: “FinanCrisis”: If a loan was approved to the business during the great financial crisis.

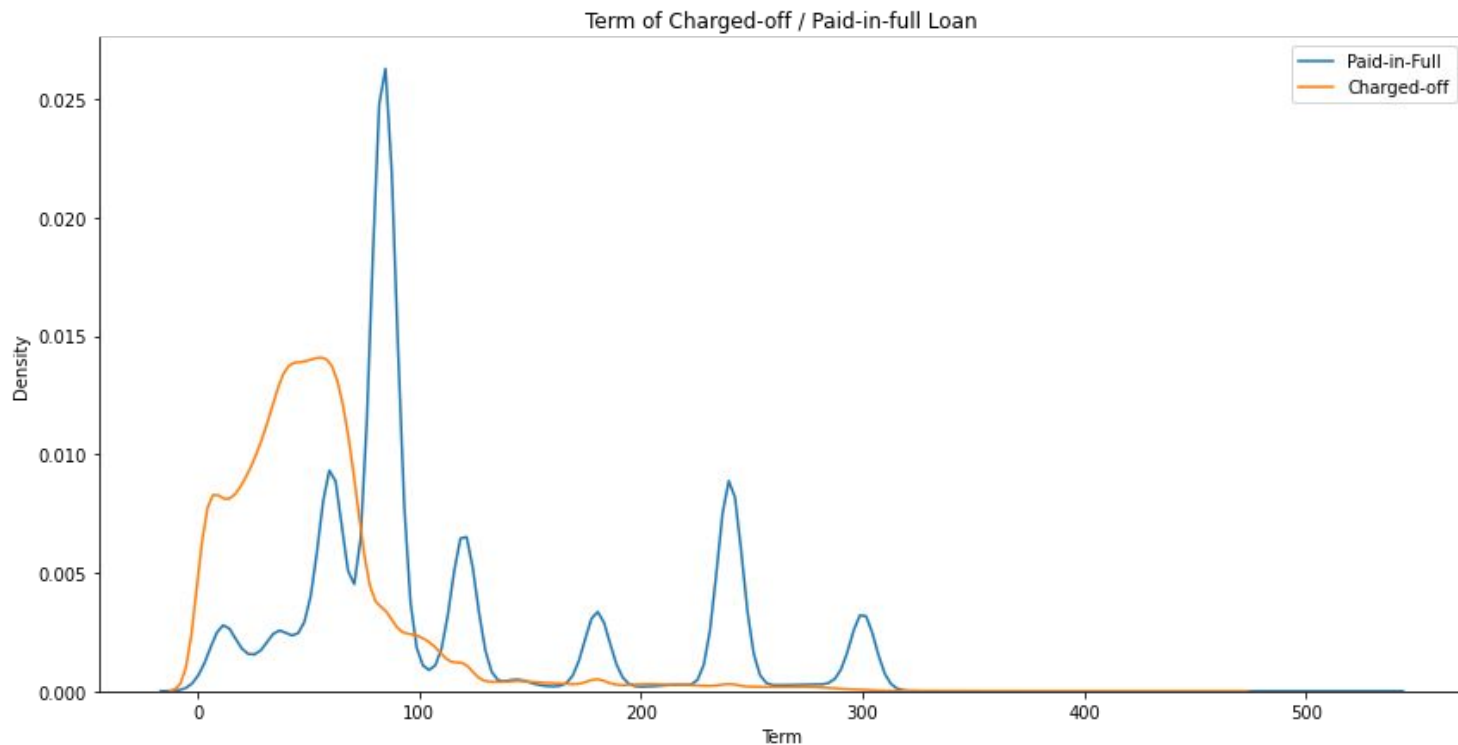
Selected Feature Exploring



1) Seasonality based on
Economic Cycles

2) 2008 Financial Crisis

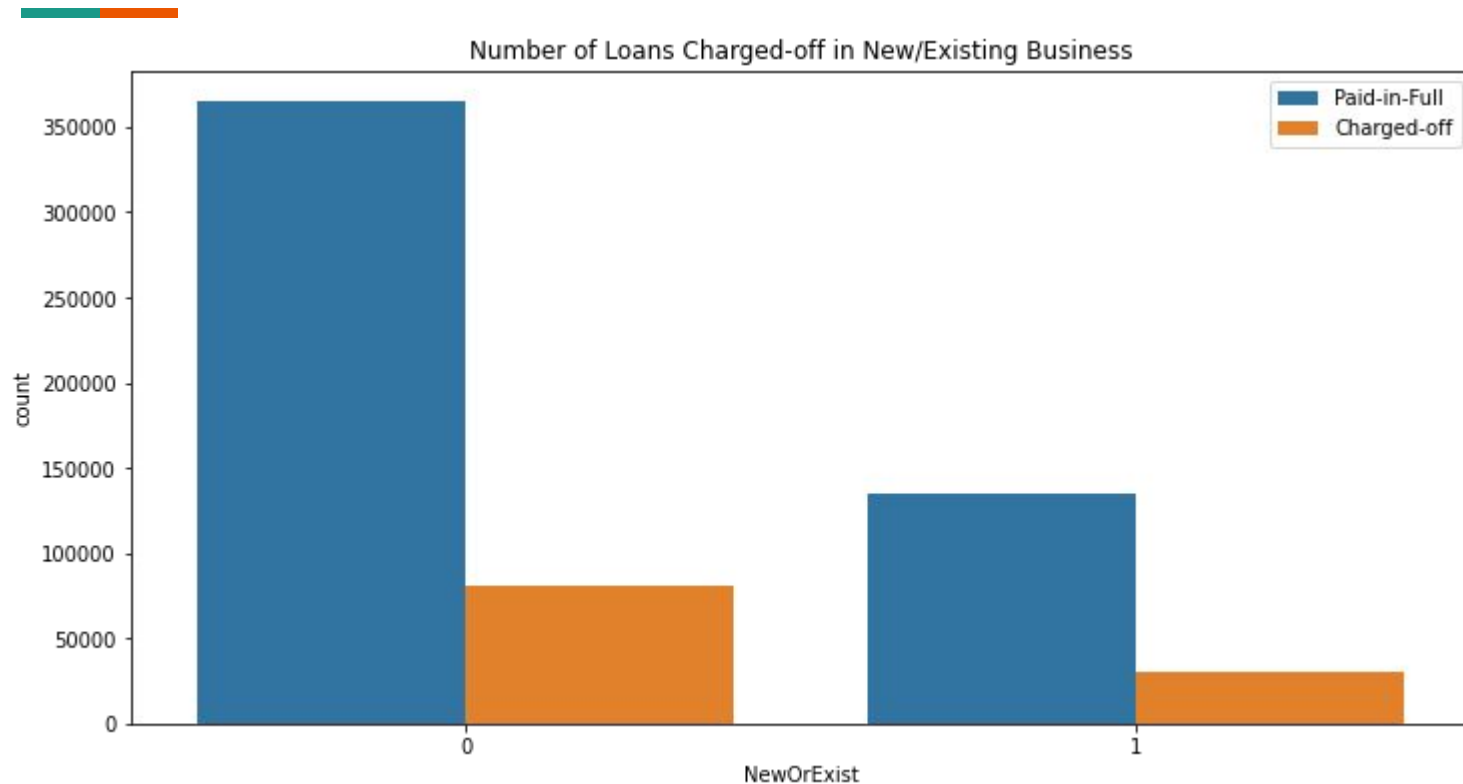
Selected Feature Exploring (Cont.)



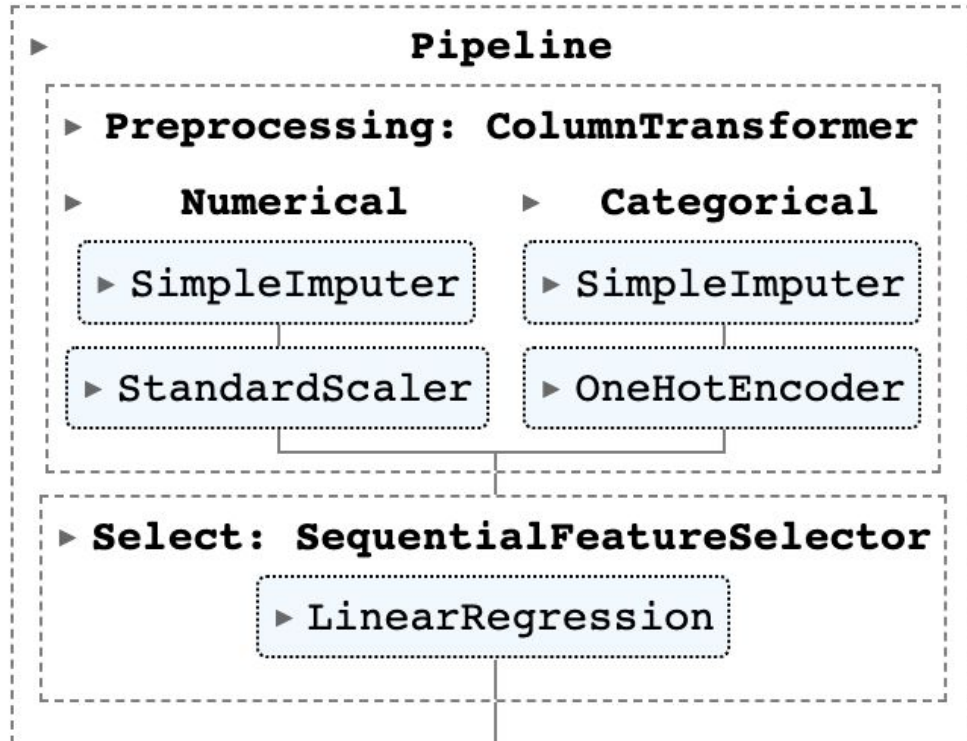
1) More shorter loans to be charged off.

2) More longer loans to be paid in full.

Selected Feature Exploring (Cont.)

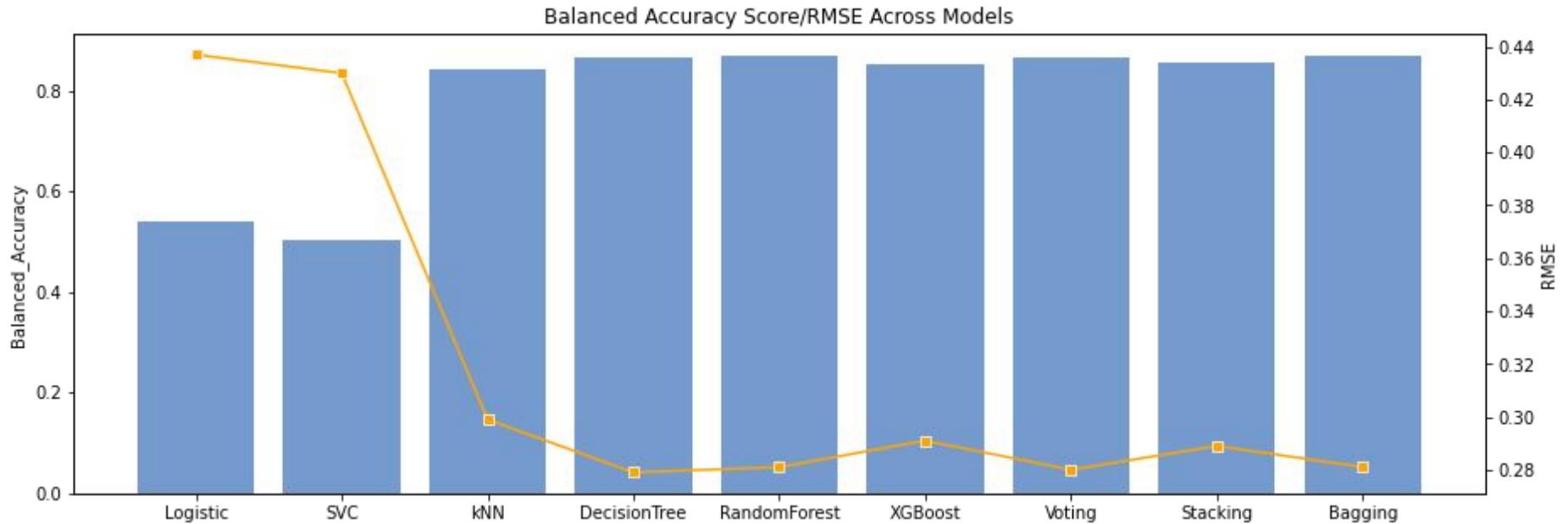


Pre-Processing Pipeline

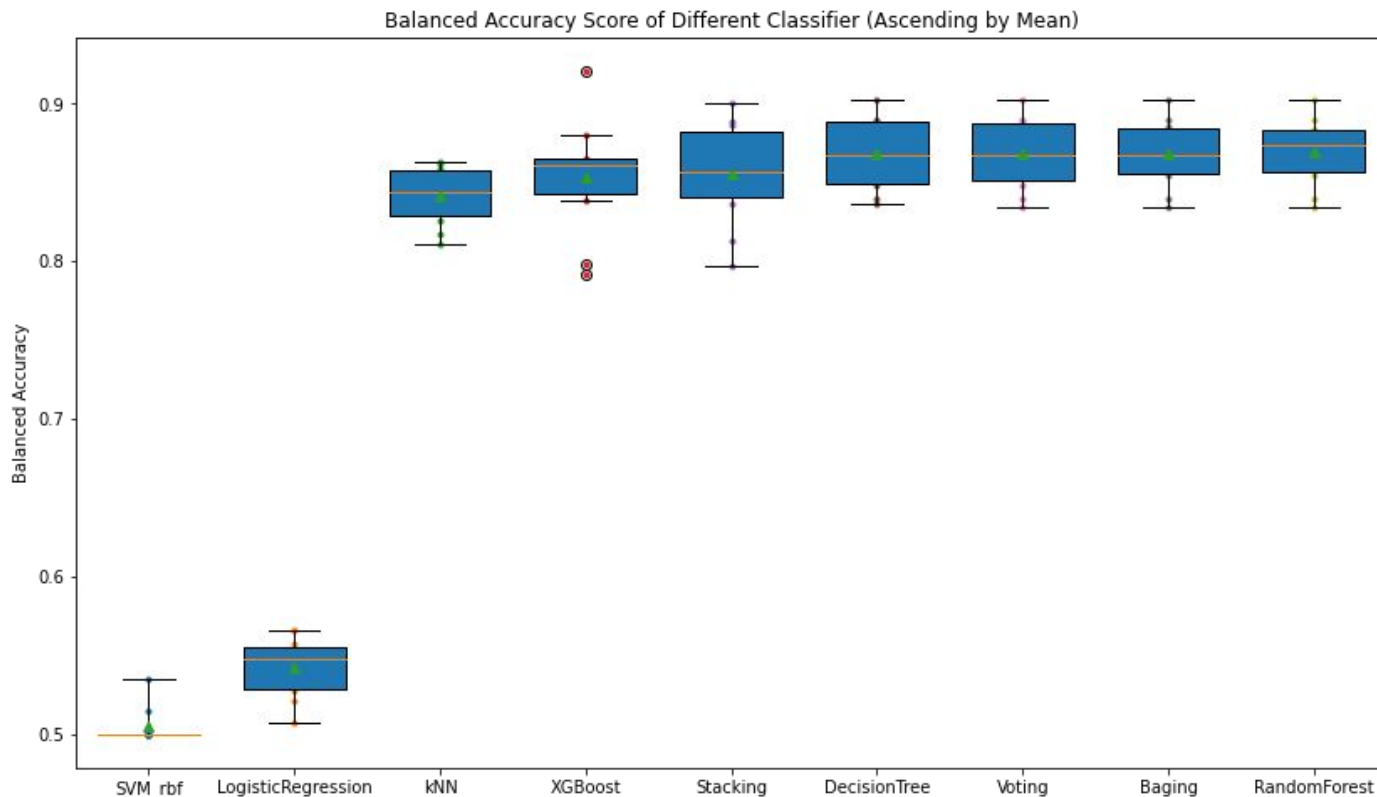


- No full pipeline until hyper-parameter tuning (for model comparison purposes.)
- Included **feature selection** for overfitting problem encountered.

Model Comparison (Classifier Comparison)



Classifier Comparison (Cont.)

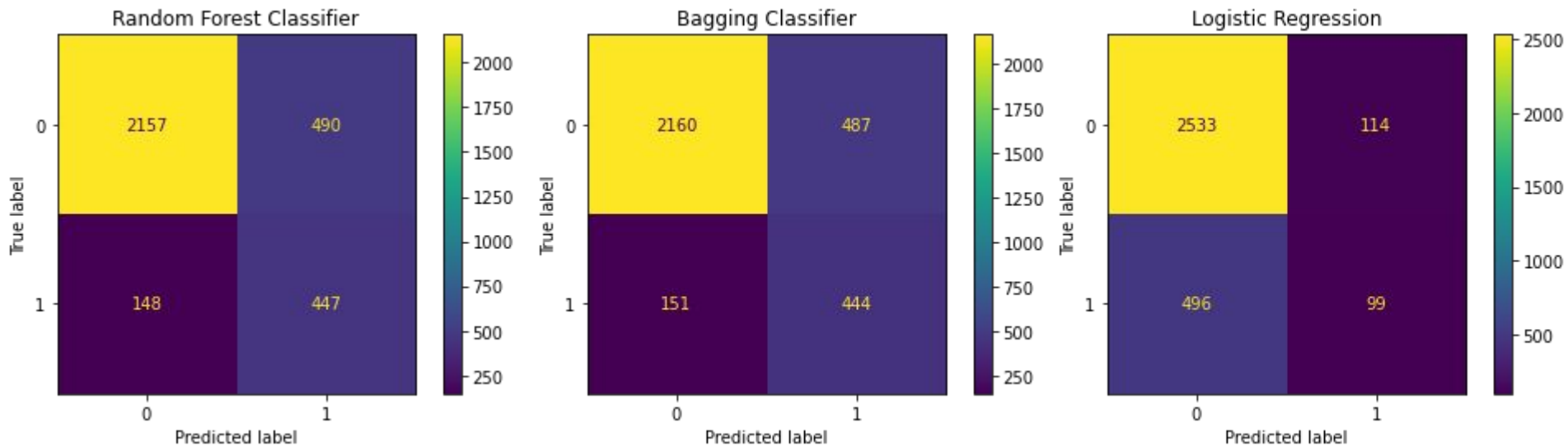


Classifier Comparison (Cont.)

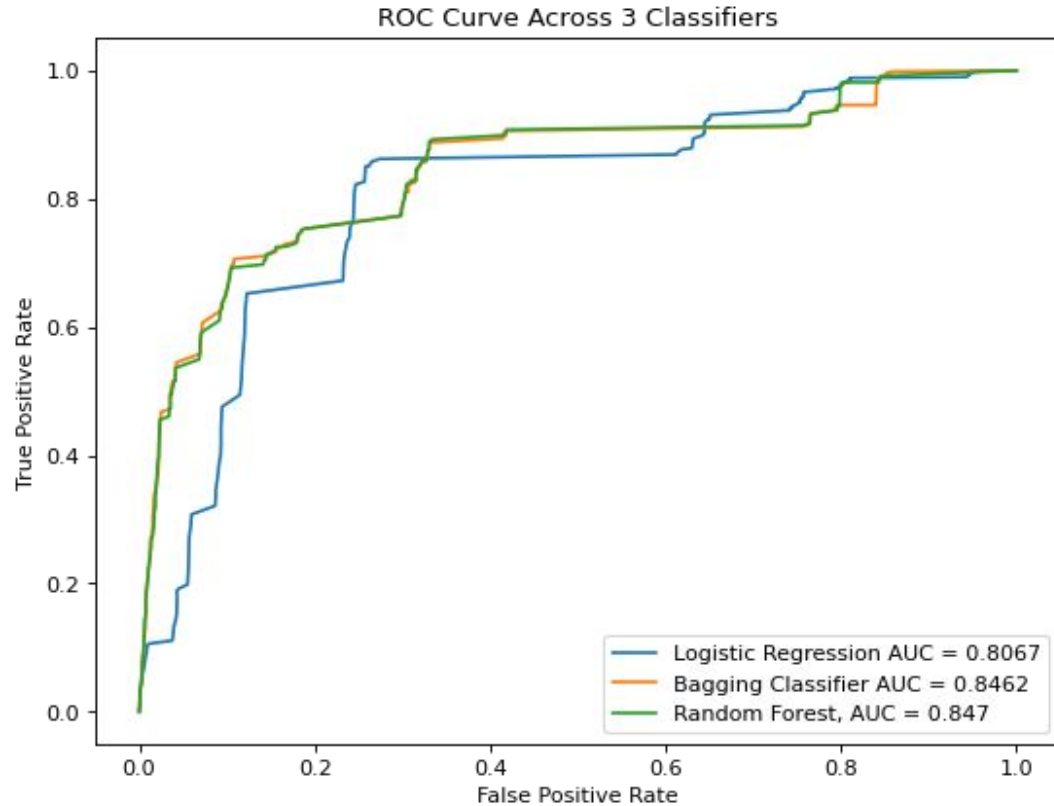


Classifier	Balanced_Accuracy_Score
Logistic Regression	0.542
SVC	0.505
kNN	0.842
Decision Tree	0.868
Random Forest	0.870
XGBoost	0.854
Voting	0.868
Stacking	0.856
Bagging	0.869

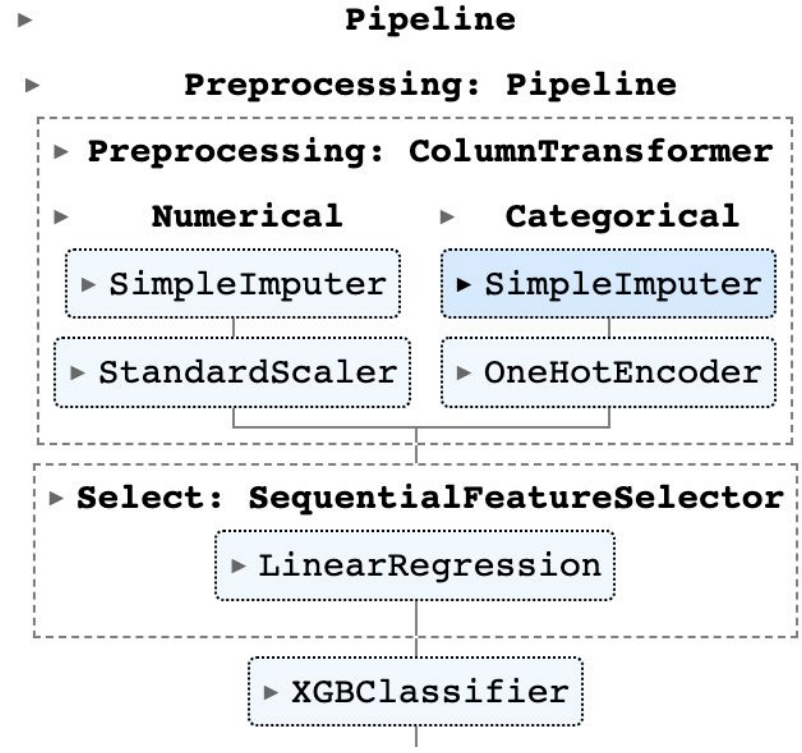
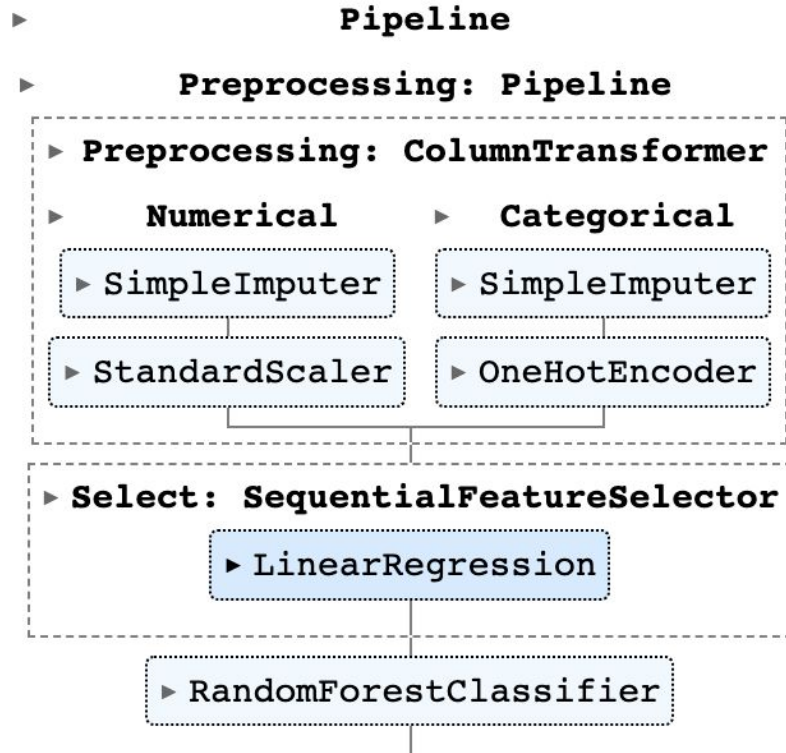
Model Evaluation



Model Evaluation (Cont.)



Full Pipelines




Hyper-Parameter Tuning



	Best Parameters	Score
Random Forest with Halving Grid Search	<code>{'Random_forest__max_depth': 10, 'Random_forest__min_samples_leaf': 8}</code>	0.866
Random Forest with Random Search	<code>{'Random_forest__max_depth': 47, 'Random_forest__min_samples_leaf': 6}</code>	0.869
XGBoost with Random Search	<code>{'XGBoost__max_depth': 21, 'XGBoost__max_leaf_nodes': 15}</code>	0.866

Cost Matrix Application & Threshold Tuning



	Score	CM-Score	Threshold	Cost Before	Cost After
Random Forest with Halving Grid Search	0.866	0.868	0.240	862	650
Random Forest with Random Search	0.869	0.869	0.169	859	644
XGBoost with Random Search	0.866	0.866	0.199	913	648

Conclusion



We would make the following suggestions to SBA:

1. After model comparison, we have located that **Decision Tree, Voting, Stacking, and Random Forest** have been the best classifiers that achieve roughly the same balanced accuracy score in charge-off prediction.
2. If we need to pick one for any lending entity, we would choose **Random Forest Classifier** for producing the highest score (**0.87**).
3. When locating the best hyper-parameters for random forest, the result is `{'max_depth': 47, 'min_samples_leaf': 6}`, although it did not improve the accuracy score too much (**0.869**).
4. When applying the cost matrix, the accuracy score did not change either, we would recommend banks to focus on how to balance false positives and false negatives for a better result.
5. In threshold tuning, it is noticeable that the thresholds are around **0.20**, which might suggest that banks should be less tolerant of any “red flags” when it comes to borrowers defaulting on their loans.

Challenges



Overfitting:

↳	precision	recall	f1-score	support
Chargedoff	0.972	0.967	0.969	24683
PaidInFull	0.991	0.992	0.992	89554
accuracy			0.987	114237
macro avg	0.981	0.979	0.980	114237
weighted avg	0.987	0.987	0.987	114237

Solutions:

1. Applied feature selection (linear regression) during preprocessing_pipelines
2. Dropped columns that we actually did not need.

(Originally, we had calculated a lot of columns (ChargeOffDays, DisbursedDays) without taking into the consideration that some of theses columns have a correlation of 1 with our y-variable.



THANK YOU

[Colab Notebook \(Same with First Page\)](#)