Home Loans Default Prediction

Xinyu Zhao

08/04/2022

Problem Definition

- **The context:**
- Non-Performing Loan (NPA) eats up a significant chunk of a bank's profits. We need to build a machine model that can automatically check a customer's creditworthiness with high efficiency and low bias.
- **The objectives:**
- We aim to simplify the decision-making process for home equity lines of credit. Our model predicts risky clients and recommends features to consider when approving a loan.
- **The problem formulation:**
- We will build an empirically derived classification model. The model will be based on the existing loan underwriting data, using predictive modeling techniques, and can justify any adverse behavior (rejections).
- **The key questions:**
- Should we approve clients based on the information they provided?
- What kinds of clients are likely to default on their loans?
- What are essential features to consider while approving a loan?

Variables

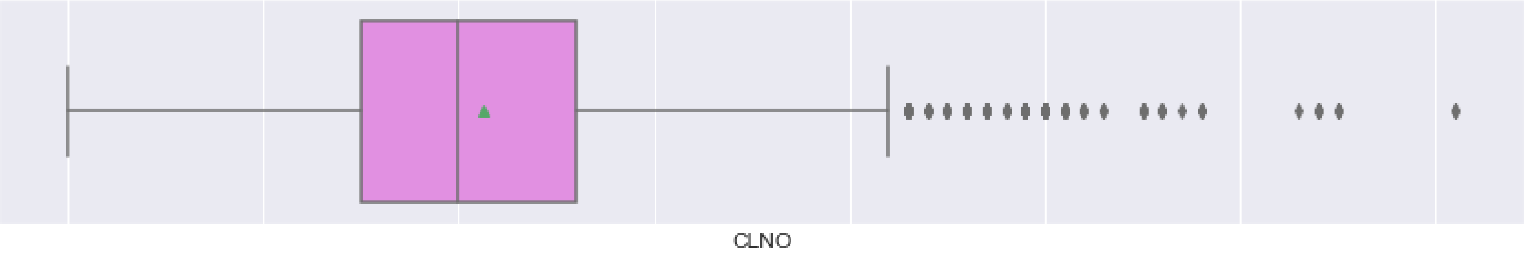
5960 Cases → 12 variables → 20% Default Cases

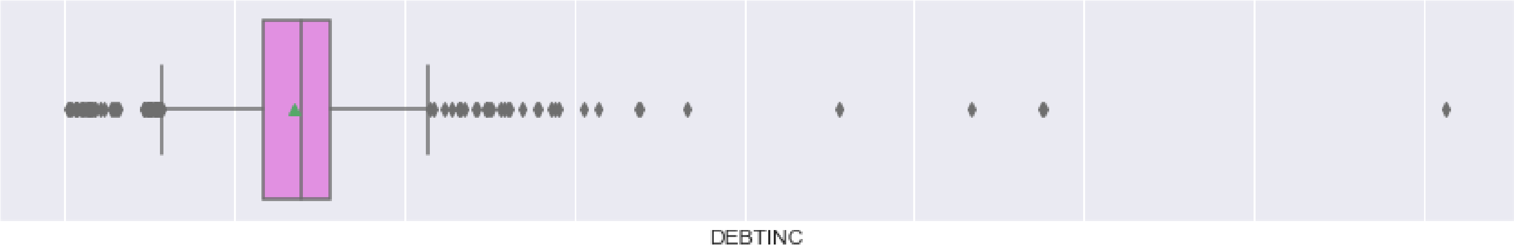
- **Target Variable – "BAD"- Binary:** 1 = Client defaulted on loan, 0 = loan repaid
- **LOAN:** Amount of loan approved.
- **MORTDUE:** Amount due on the existing mortgage.
- **VALUE:** Current value of the property.
- **REASON:** Reason for the loan request. (HomeImp = home improvement, DebtCon= debt consolidation)
- **JOB:** The type of job that loan applicant has
- **YOJ:** Years at present job.
- **DEROG:** Number of major derogatory reports (which indicates a serious delinquency).
- **DELINQ:** Number of delinquent credit lines (fail to make the minimum payments 30 to 60 days past the due date).
- **CLAGE:** Age of the oldest credit line in months.
- **NINQ:** Number of recent credit inquiries.
- **CLNO:** Number of existing credit lines.
- **DEBTINC:** Debt-to-income ratio (monthly debt payments divided by gross monthly income -- Measurement Ratio)

Data Preparation

- **Clean Up Missing Value** – Fill missing categorical value with mode and numerical value with median
- **Remove Outliers** – Outliers of features like “Derog” and “Delinq” are meaningful. Removing outliers from these features could lead to the loss of important information. We will keep the outliers.
- **No missing value and keep outliers -> Data is ready to be modeled**

0	BAD	5960	non-null	category
1	LOAN	5960	non-null	int64
2	MORTDUE	5442	non-null	float64
3	VALUE	5848	non-null	float64
4	REASON	5708	non-null	category
5	JOB	5681	non-null	category
6	YOJ	5445	non-null	float64
7	DEROG	5252	non-null	float64
8	DELINQ	5380	non-null	float64
9	CLAGE	5652	non-null	float64
10	NINQ	5450	non-null	float64
11	CLNO	5738	non-null	float64
12	DEBTINC	4693	non-null	float64





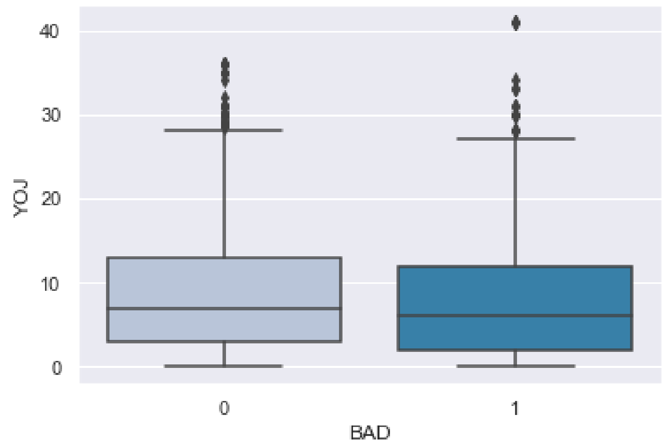
% of missing values in the each column

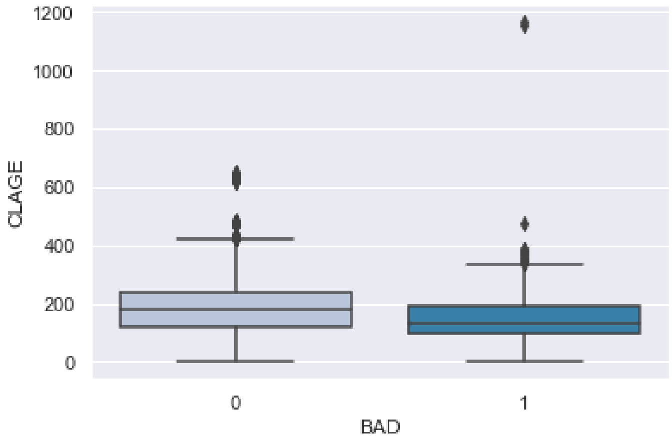
BAD	0.000
LOAN	0.000
MORTDUE	8.691
VALUE	1.879
REASON	4.228
JOB	4.681
YOJ	8.641
DEROG	11.879
DELINQ	9.732
CLAGE	5.168
NINQ	8.557
CLNO	3.725
DEBTINC	21.258

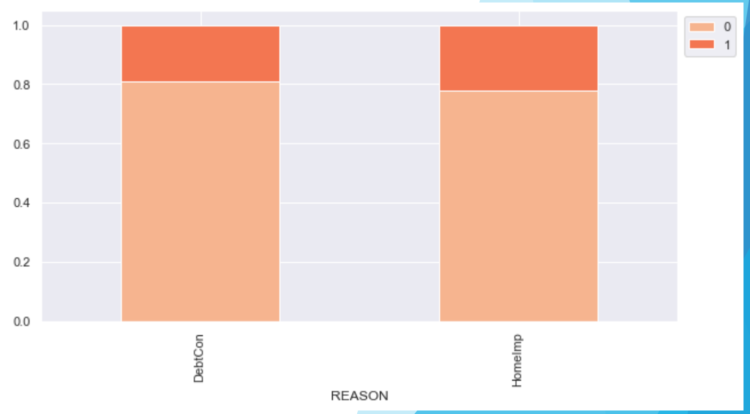
Data Exploration – Preliminary Findings

➢ **People are more likely to default if they have:**

- ↓ Working Experience
- ↓ Age of Credit History
- ↑ Debt to Income Ratio
- ↑ No. of Existing Credit Lines
- ↑Delinquency & Derogatory Reports
- For Home Improvement







Proposed Approach

Model

- logistic regression, Decision Tree, Random Forest

Design

- 30% Test and 70% Train
- Hyperparameter Tuning -> Find Best Model
- Features investigation and recommendation

Evaluate

- Confusion matrix, accuracy, precision, recall, and f1-score

- **Recall:** the proportion of actual occurrences captured by the model;
- **Precision:** the correctness of the model's prediction on a specific class
- **Accuracy:** the overall prediction correctness of all classes.

Outliers and Missing Value Flag?

Outliers

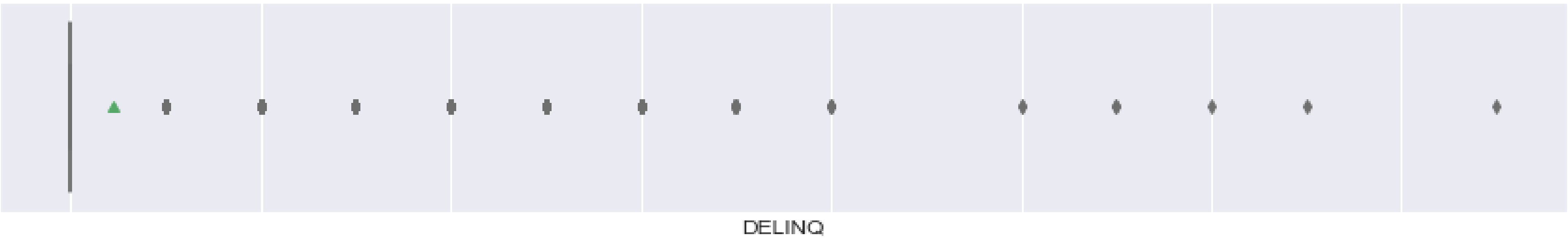
- Outliers of “Derog” and “Delinq” are meaningful
- Treat outliers -> Lose crucial information
- Will not treat outliers

Missing Value Flag

- Fill missing value with mode or mean
- Flag missing value could brings additional insights
- Train models on both flagged and unflagged data

Evaluate

- banks want to capture as many risky loans as possible
- which dataset is the best?
  Recall>Accuracy/Precision



Logistic Regression

Labels	Class 1 -> default, Class 0 -> not default Model's threshold is 0.5, prediction >0.5 label as 1
Simple Model with no regularization	Default model only predicts 3% of actual defaulted loans
Lasso Regression (L1 regularized)	Standardized the dataset Lasso regression predicts 33% unflagged vs 62% flagged
Top Features (Highest Coefficient)	Flagged: 'DEROG', 'CLNO_flag', 'JOB_flag', 'DEROG_flag', 'DEBTINC_flag', 'DEBTINC', 'DELINQ' Unflagged: 'NINQ', 'CLAGE', 'DELINQ', 'DEROG', 'DEBTINC'

Flag – Simple Logistic Regression

	precision	recall	f1-score	support
0	0.79	1.00	0.88	1416
1	0.50	0.01	[illegible]	[illegible]
accuracy			0.79	1788
macro avg	0.65	0.50	0.45	1788
weighted avg	0.73	0.79	0.70	1788

Much Higher Recall Score

Flag – L1 Regression

	precision	recall	f1-score	support
0	0.91	0.96	0.93	1416
[illegible]	[illegible]	0.62	0.70	372
accuracy			0.89	1788
macro avg	0.86	0.79	0.82	1788
weighted avg	0.89	0.89	0.89	1788

Decision Tree

Utilize quantitative selection criteria like Entropy to form selection criteria that can classify data into separated groups

Simple Tree	Overfitting with perfect score on training dataset
Class Weighted Tree 0.8 for 1, 0.2 for 0	Overfitting with perfect score on training dataset
Tuned Tree	(0.2,0.8), max_depth ->12, min_samples_leaf-> 24, random_state=5 Not overfitting
Top Features	DebtInc_Flag, DebtInc, Delinq, Clage, Value, YOJ

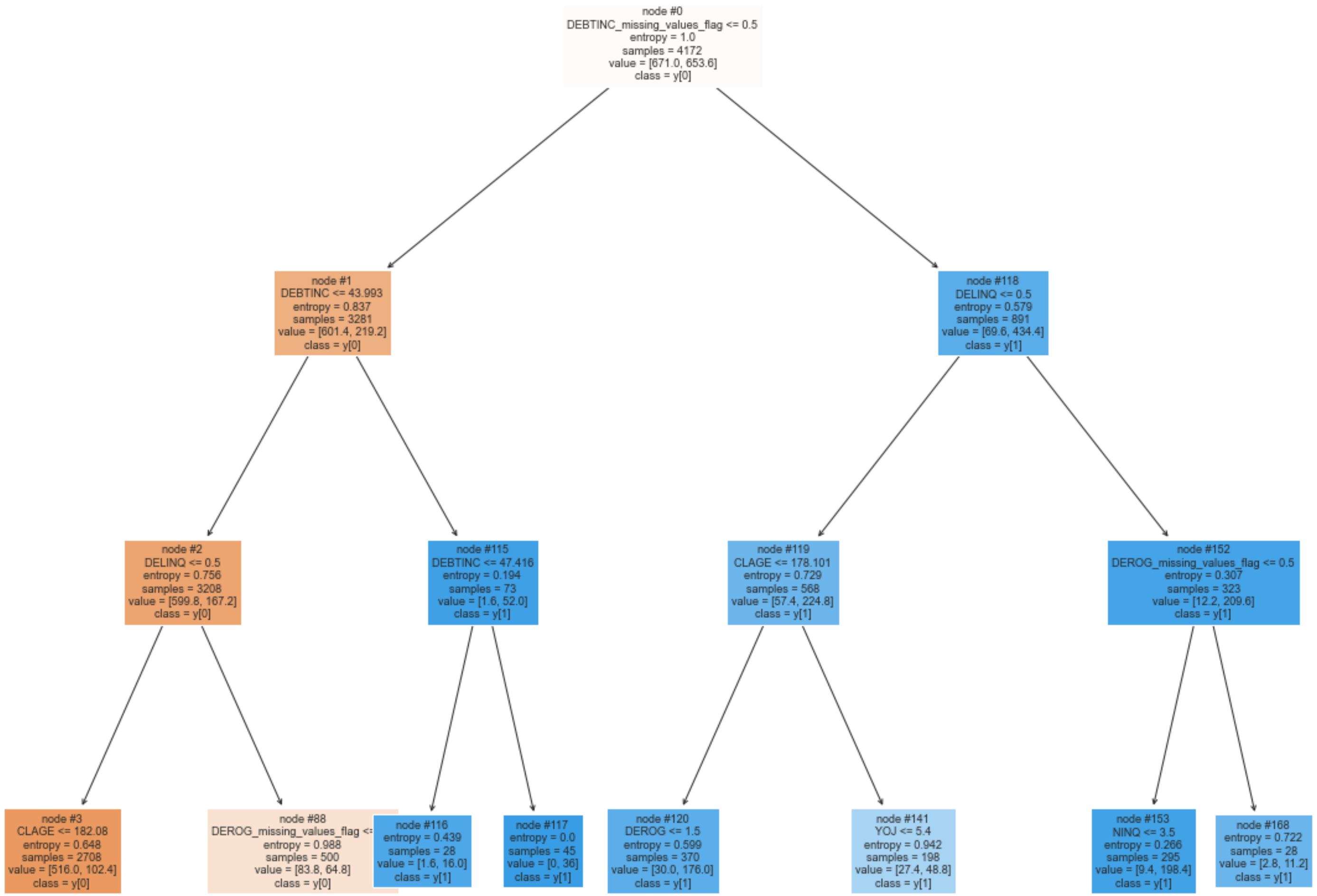
	precision	recall	f1-score	support
0	1.00	1.00	1.00	[illegible]
1	1.00	1.00	1.00	[illegible]
accuracy			1.00	4172
macro avg	1.00	1.00	1.00	4172
weighted avg	1.00	1.00	1.00	4172

Overfitting

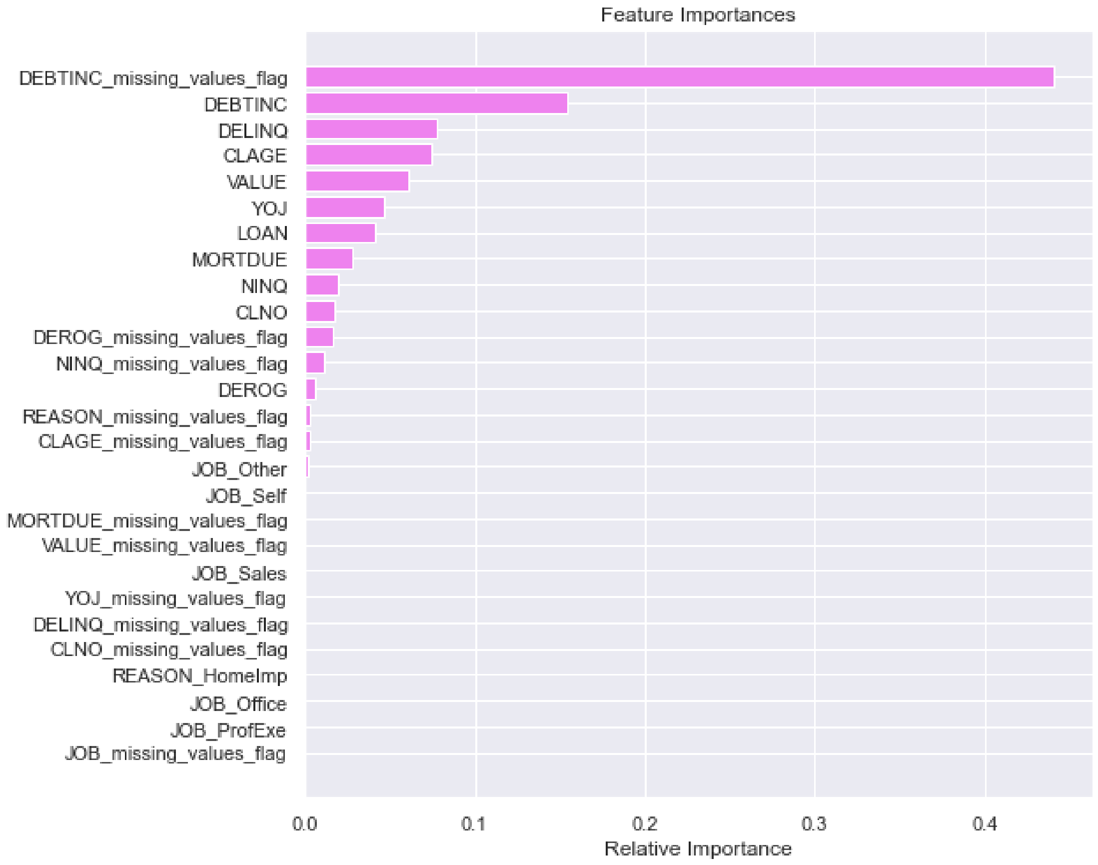
	precision	recall	f1-score	support
0	0.90	0.92	0.91	1416
1	0.68	0.61	0.64	372
accuracy			0.86	1788
macro avg	0.79	0.77	0.78	1788
weighted avg	0.85	0.86	0.86	1788

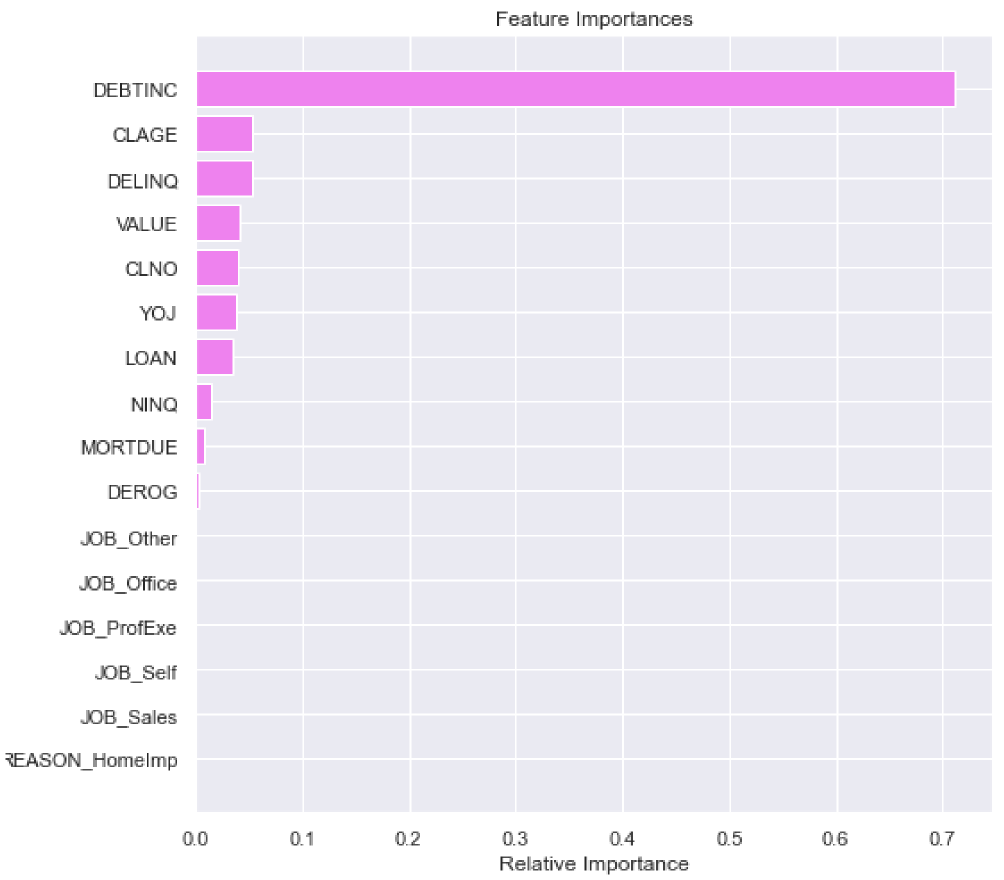
Decision Tree – Tree at Step 3 - Flagged

1. If clients **do not disclose their Debt-to-Income ratio**, they are more likely to default
2. If clients have **a Debt-to-income ratio < 44**, no delinquent account, and have account age > 182 months, they are less likely to default
3. If clients have >**47 Debt-to-income ratio**, they are very likely to default



Decision Tree – Feature Importance



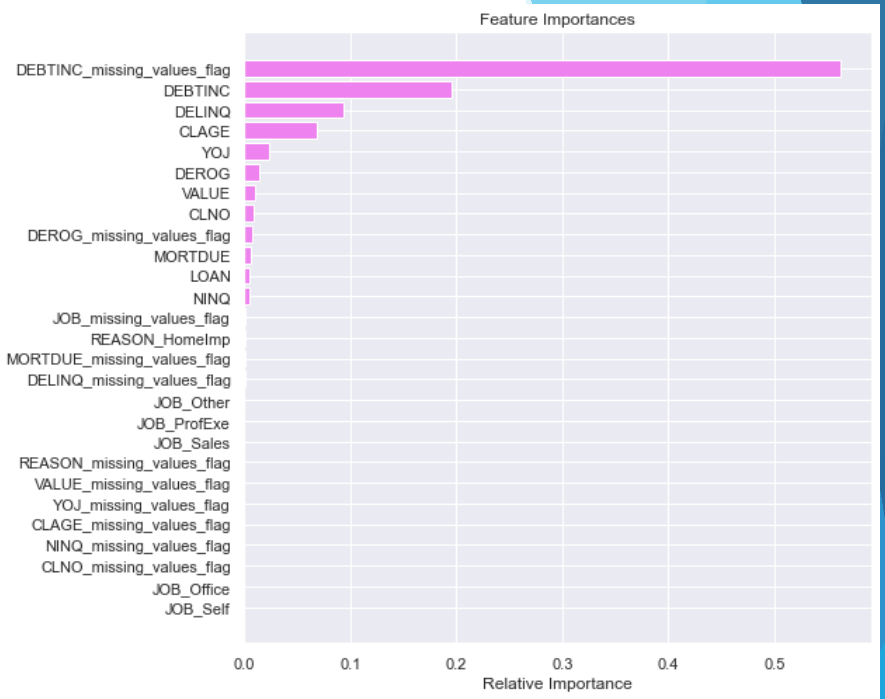


Random Forest

Constructs many decision trees using resampling techniques like bootstrapping and random selection of features

Less interpretable than decision tree

Default Forest	Overfitting with perfect score on training dataset
Class Weighted Forest 0.8 for 1, 0.2 for 0	Overfitting with perfect score on training dataset
Tuned Tree	class_weight={0: 0.2, 1: 0.8}, criterion='entropy’, max_depth=4, max_features=0.7, min_samples_leaf=30, n_estimators=90, random_state=7
Top Features	DebtInc_Flag, DebtInc, Delinq, Clage, YOJ



	precision	recall	f1-score	support
0	0.94	0.89	0.92	
1	0.64	0.79	0.70	
accuracy			0.87	4172
macro avg	0.79	0.84	0.81	4172
weighted avg	0.88	0.87	0.88	4172

Not Overfitting

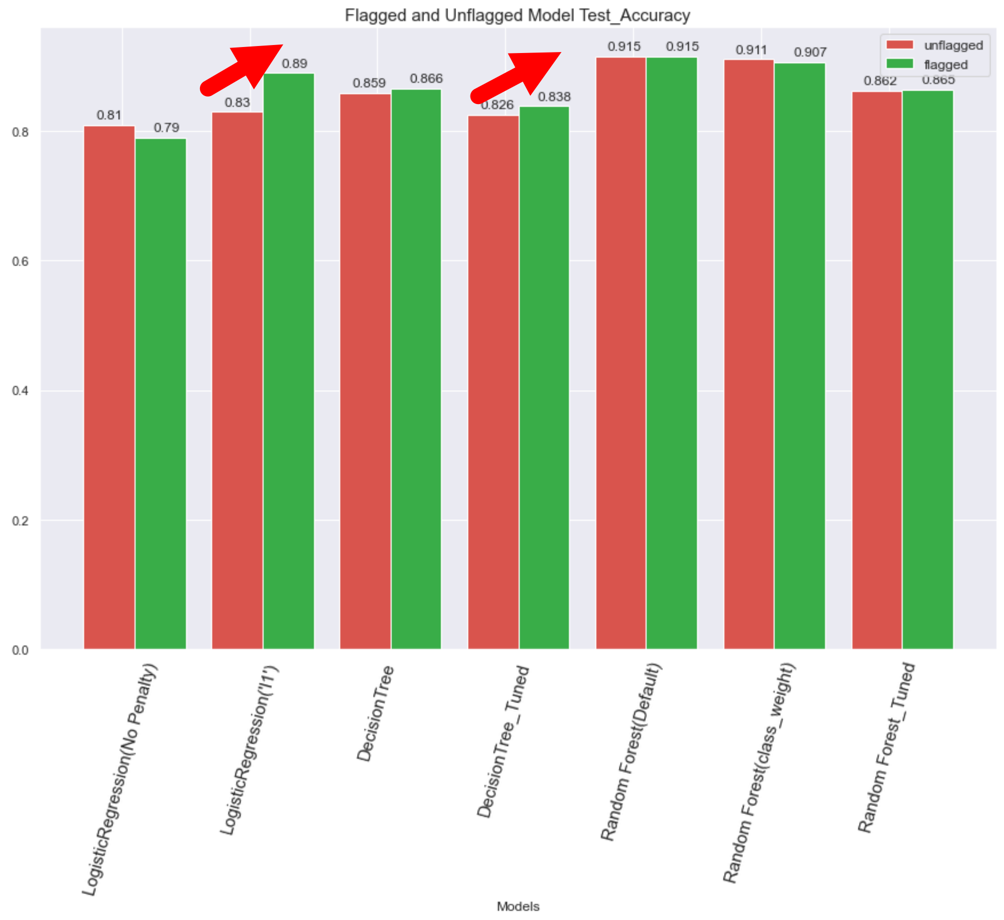
	precision	recall	f1-score	support
0	0.93	0.90	0.91	1416
1	0.65	0.74	0.70	372
accuracy			0.86	1788
macro avg	0.79	0.82	0.80	1788
weighted avg	0.87	0.86	0.87	1788

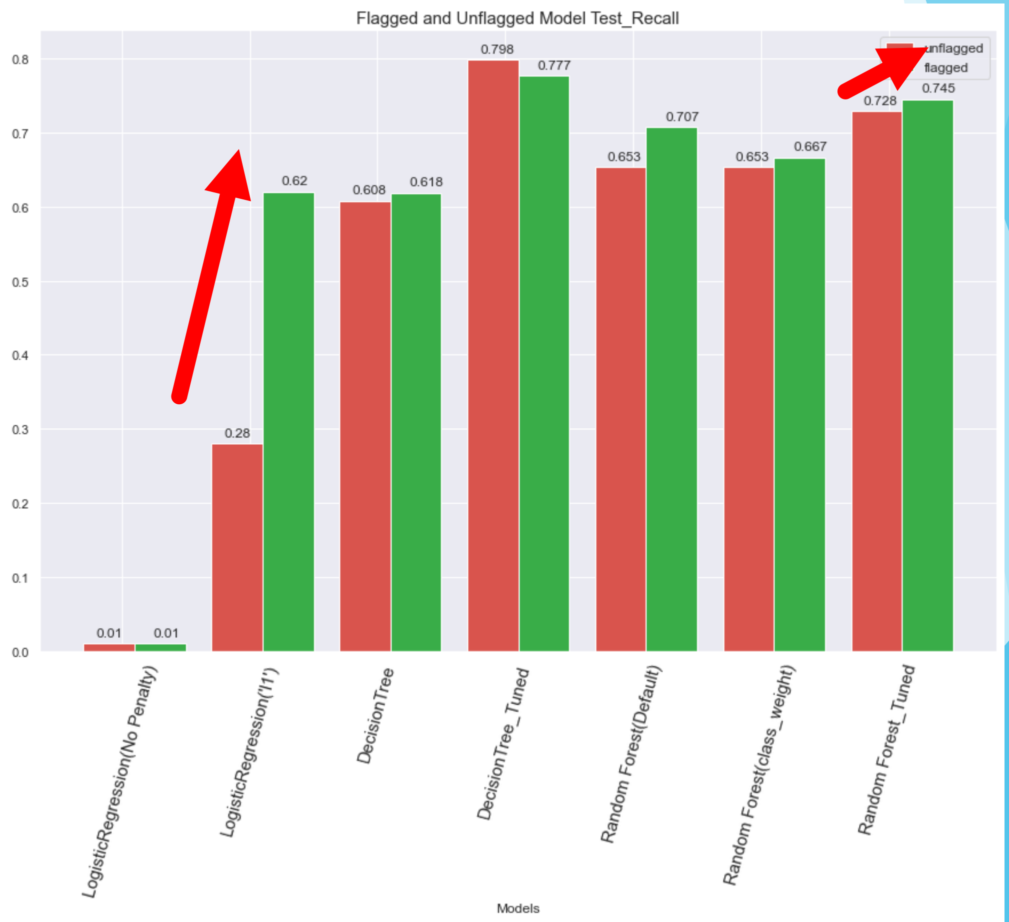
Which Dataset to Choose?

Use flagged data as the result indicate a better fitted model with more important insights

Logistic regression works much better with the flagged data

Overall flagged data perform better than unflagged for many models





Which Model to Choose?

High Accuracy?
Logistic Regression
90% accuracy

Balanced Approach

- 3% lower recall but 7% higher precision than DecisionTree
- 15% lower precision but 12% higher recall than logistic regression

Comprehensiveness?
Decision Tree
78% recall

	Model	Train_Accuracy	Test_Accuracy	Train_Recall	Test_Recall	Train_Precision	Test_Precision
0	LogisticRegression(No Penalty)	0.810000	0.790000	0.040000	0.010000	0.740000	0.500000
1	LogisticRegression('l1')	0.890000	0.890000	0.630000	0.620000	0.780000	0.820000
2	DecisionTree	1.000000	0.865772	1.000000	0.618280	1.000000	0.701220
3	DecisionTree_Tuned	0.863135	0.838367	0.884945	0.776882	0.602500	0.583838
4	Random Forest(Default)	1.000000	0.914989	1.000000	0.706989	1.000000	0.859477
5	Random Forest(class_weight)	1.000000	0.907159	1.000000	0.666667	1.000000	0.855172
6	Random Forest_Tuned	0.870566	0.864653	0.787026	0.744624	0.637265	0.653302

Conclusion

- **Most Important Features:**
- Debt-to-income Ratio, Number of Delinquent Accounts, Age of the oldest credit line in months, number of delinquent credit lines and derogatory reports, years of job experience are the most important factors. Loan value does not have a substantial effect on default loan
- **Best Model:**
- Depends on the needs for accuracy or comprehensiveness. A balanced model of random forest is recommended. It can capture 75% of actual defaulted loan while giving an overall prediction accuracy of 86%
- **Future Actions:**
- Use other models like SVM and KNN to find the best model
- The lack of interpretability limits the random forest model's use. Feature importance plot itself is not sufficient to identify the best features
- Logistic regression can be improved if we consider imbalance sample threshold and other regularization measures