Loan Default Prediction

Understanding the Problem statement

The bank's consumer credit department seeks to streamline the approval process for home equity lines of credit by adopting the guidelines of the Equal Credit Opportunity Act. They plan to create a statistically robust credit scoring model based on data from recent credit applicants who were approved. This model will use predictive modeling techniques but must be easily interpretable to explain any adverse decisions, such as credit application rejections.

Objective

The task involves building a classification model to identify clients who are at risk of defaulting on their loans. Additionally, recommendations on critical features to assess when approving loans should be provided to the bank.

Objective

Minimize False Positives (Minimize Loss)

OR

Maximize True Positives (Maximize Profit)

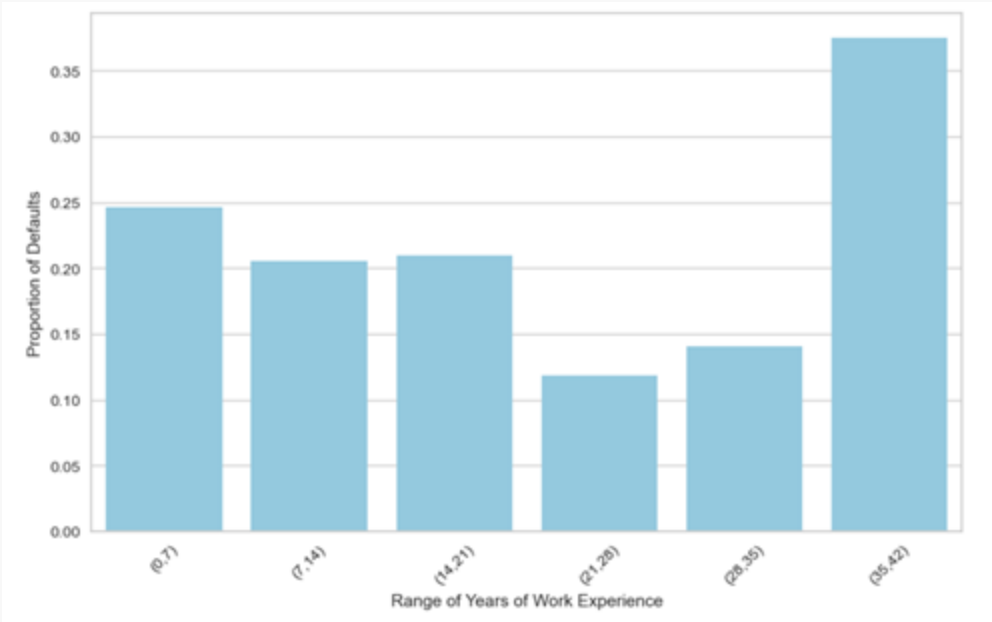
OR

A Balance of Both?

F1 score= 2 x Precision x Recall / (Precision + Recall)

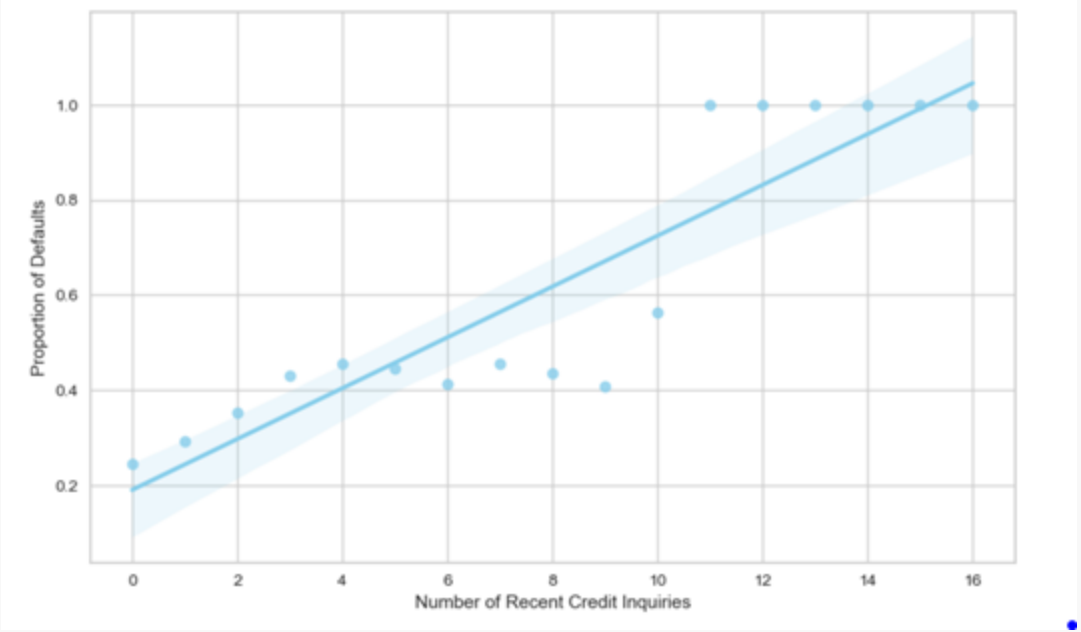
Business Insights from EDA

Do Applicants with more work experience tend to default loans less?



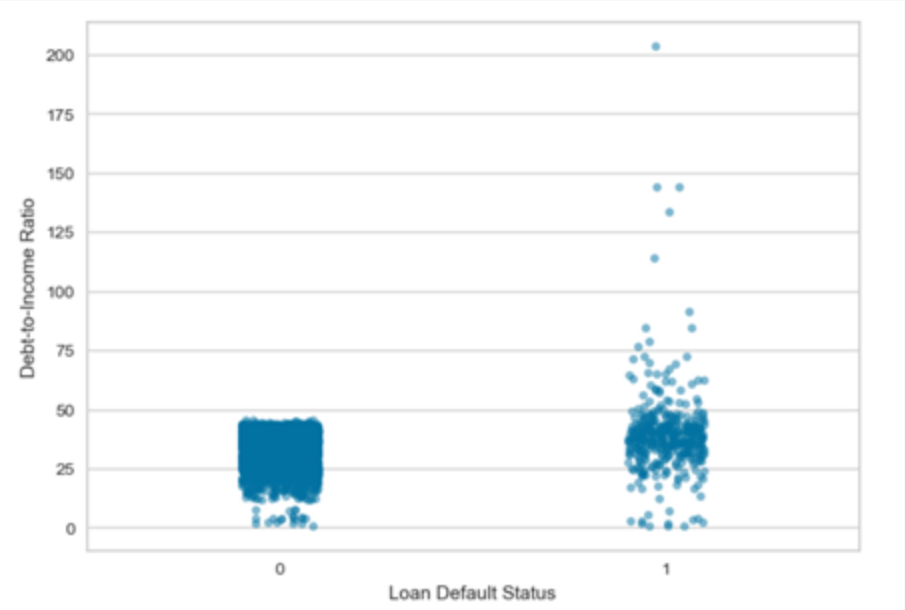
Business Insights from EDA

Does someone who inquires more often for a loan tend to default loans more often?



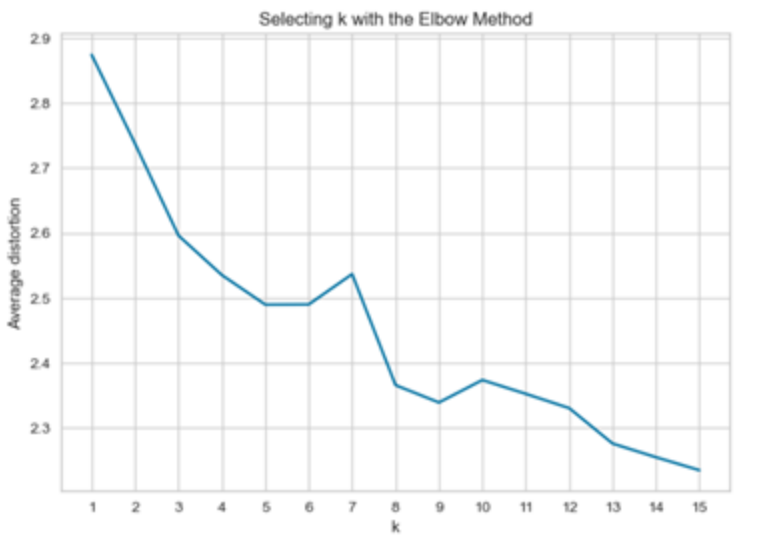
Business Insights from EDA

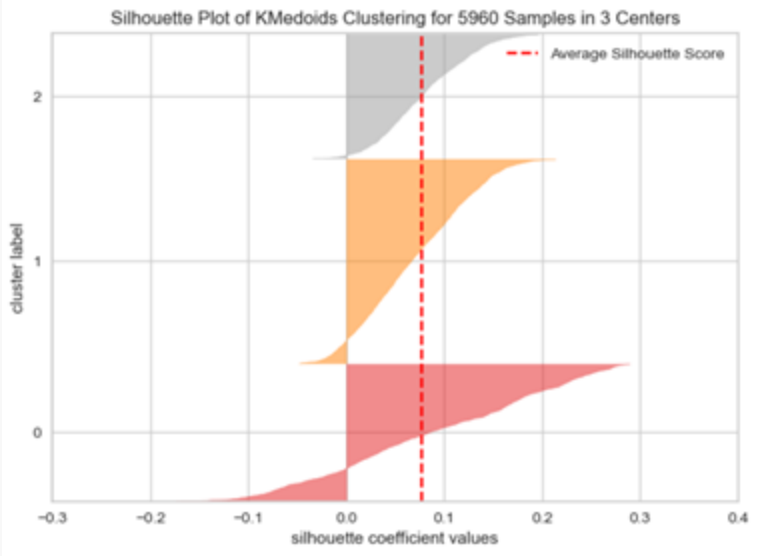
Do applicants with higher debt-to-income ratio tend to default loans more often?



Clustering - KMedoid

Why not KMeans?





Clustering - KMedoid

What do these clusters signify

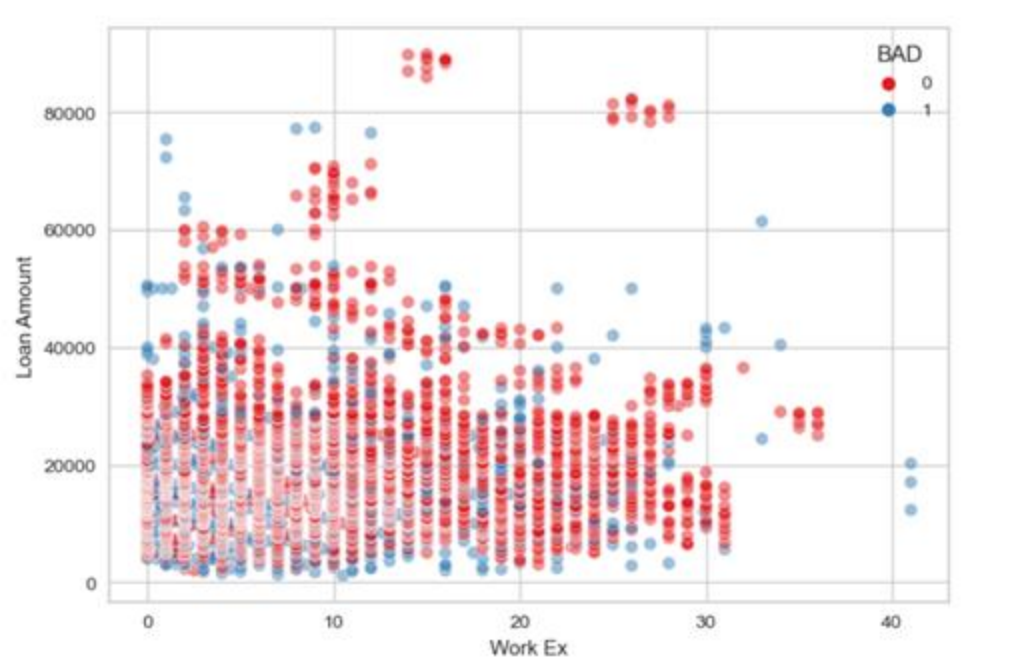
Cluster	BAD	LOAN	MORTDUE	VALUE	YOJ	DEROG	DELINQ	CLAGE	NINQ	CLNO	DEBTINC
0	0.132915	21872.104963	116556.987476	158909.805046	10.460463	0.099060	0.268701	202.225454	0.969589	27.764773	35.864496
1	0.291444	13834.186402	54043.724073	73922.497489	7.143027	0.496947	0.831555	139.494199	0.818100	17.821844	31.419094
2	0.121460	22872.120831	53230.195877	83260.309930	10.059860	0.079188	0.113900	215.657604	1.938580	19.463138	34.668936

- Cluster 0 : High loan amount, high mortgages remaining, high number of credit lines, high work ex
- Cluster 1 : Low loan amount, high number of delinquent credit cards and derogatory reports
- Cluster 2 : High loan amount, less mortgage remaining, low number of credit cards and derogatory reports, high work ex

Revisiting Business Insights from EDA

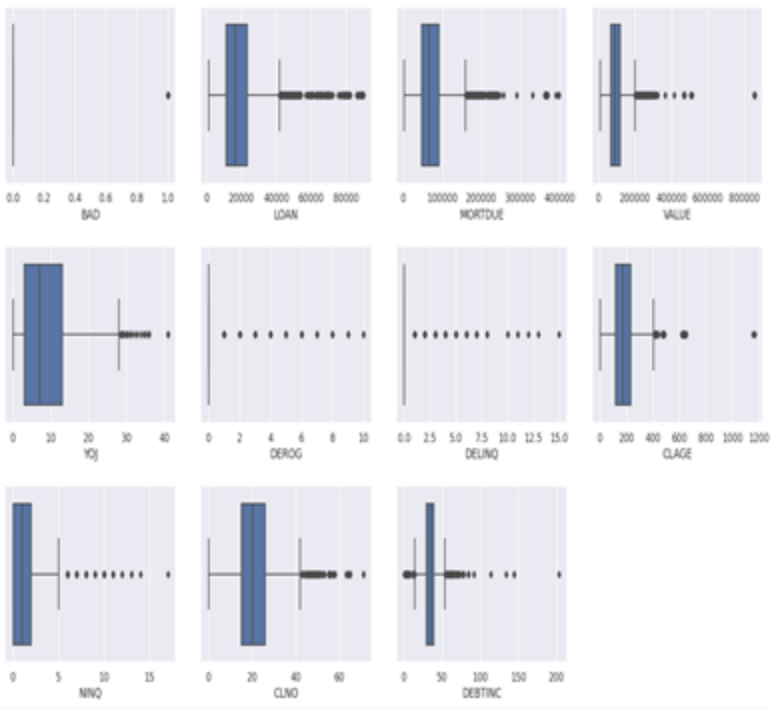
When someone takes a lower amount of loan for a highly valued property, is that person more likely to default?

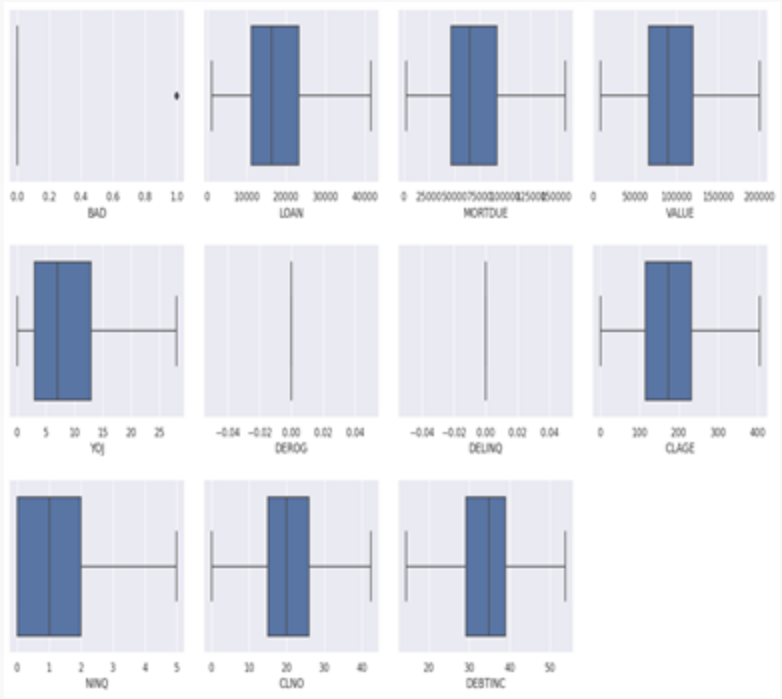
Revisiting Business Insights from EDA



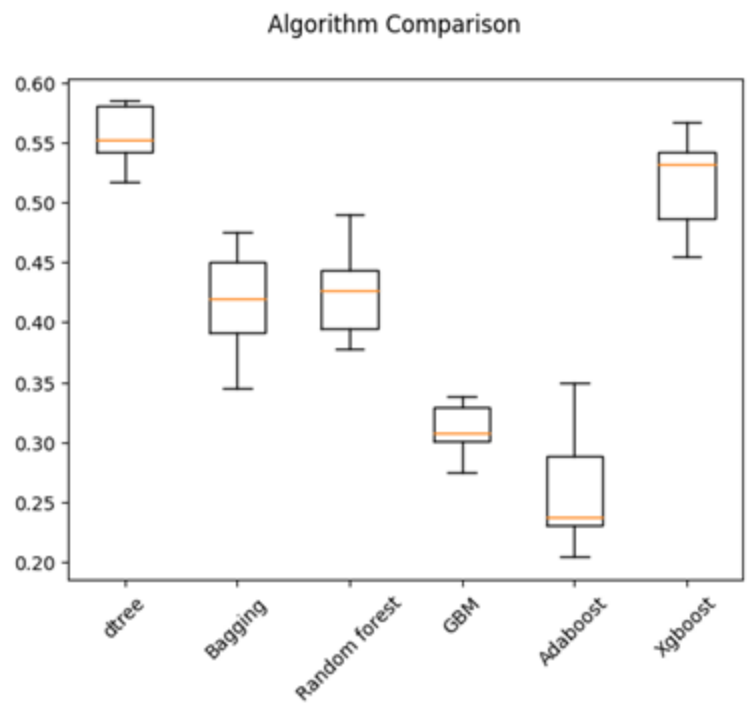
Model Building and Selection

Treating outliers with IQR





Model Building with original data



- The cross validation training performance scores are similar to the validation perfromance score. This indicates that the default algorithms on original dataset are able to generalize well.

- We can see that the dtree (~ 55%) is giving the highest cross-validated recall followed by XGBoost (~ 51%) then Random Forest (~ 42%)

- Bagging have one outlier as can be observed from the boxplot

- We will tune the best three models i.e. `dtree, XGBoost and Random Forest` and see if the performance improves

Model Selection

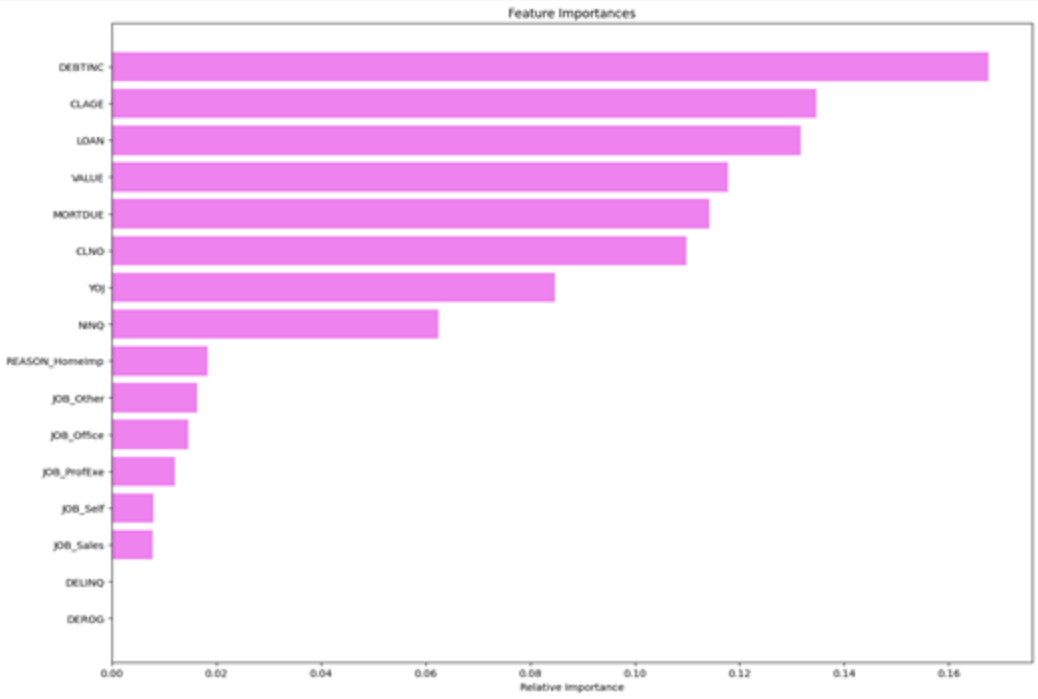
- Models built on original dataset have given generalized performance on cross validation training and validation sets unlike models built on oversampled or undersampled sets.

- We will tune the dtree, XGBoost, and Random Forest models and see if the performance improves

Comparison Report

Models	Accuracy	Recall	Precision	Pros	Cons
DTree Tuned	0.58	0.71	0.28	-Interpretable - Good Recall	-Low Precision
XGBoost Tuned	0.80	0.79	0.5	-Better than DTree - Similar Recall to DTree	-Low Precision -Not Interpretable
Random Forest	0.87	0.38	0.91	-Good Precision	-Low Recall -Not Interpretable
SVM	0.92	0.64	0.96	-Highest Precision	-Low Recall -Not Interpretable

Factors affecting the loan default



This graph shows the factors affecting the loan default and their relative importance.

1. The type of job an applicant has seems to have less impact than other features
2. Debt-to-income ratio seems to have the most impact on our model

Summary

71% of the time when a borrower comes to request a home loan, a decision tree algorithm can predict who will default on the loan.

This model lends itself easily to interpretation.

The most crucial components for accurately predicting are DELINQ, CLAGE, and DEBINC.

The most significant factor is the debt-to-income ratio, which is also the one with the greatest amount of missing data (21.3%), comparable to the percentage of defaulting customers (20%).

Investigating the possibilities of developing a different business procedure to handle and make decisions for clients who do not have access to a debt-to-income ratio is advised.

Recommendations

Examine the viability of developing a different business procedure to handle and make decisions for clients who don't have access to a debt-to-income ratio.

Investigate other machine learning strategies, including neural networks, engineering features, and dropping columns.

Risk and Challenges

The main risk involved in this project is underperforming in comparison to the manual, present procedure.

Changing the bank's internal culture to accommodate the new model will be a significant problem.

To surpass the present incomes with the new model is another issue.

THANK YOU