LOAN DEFAULT PREDICTION

CONTENTS

Problem Statement

Data Wrangling

EDA

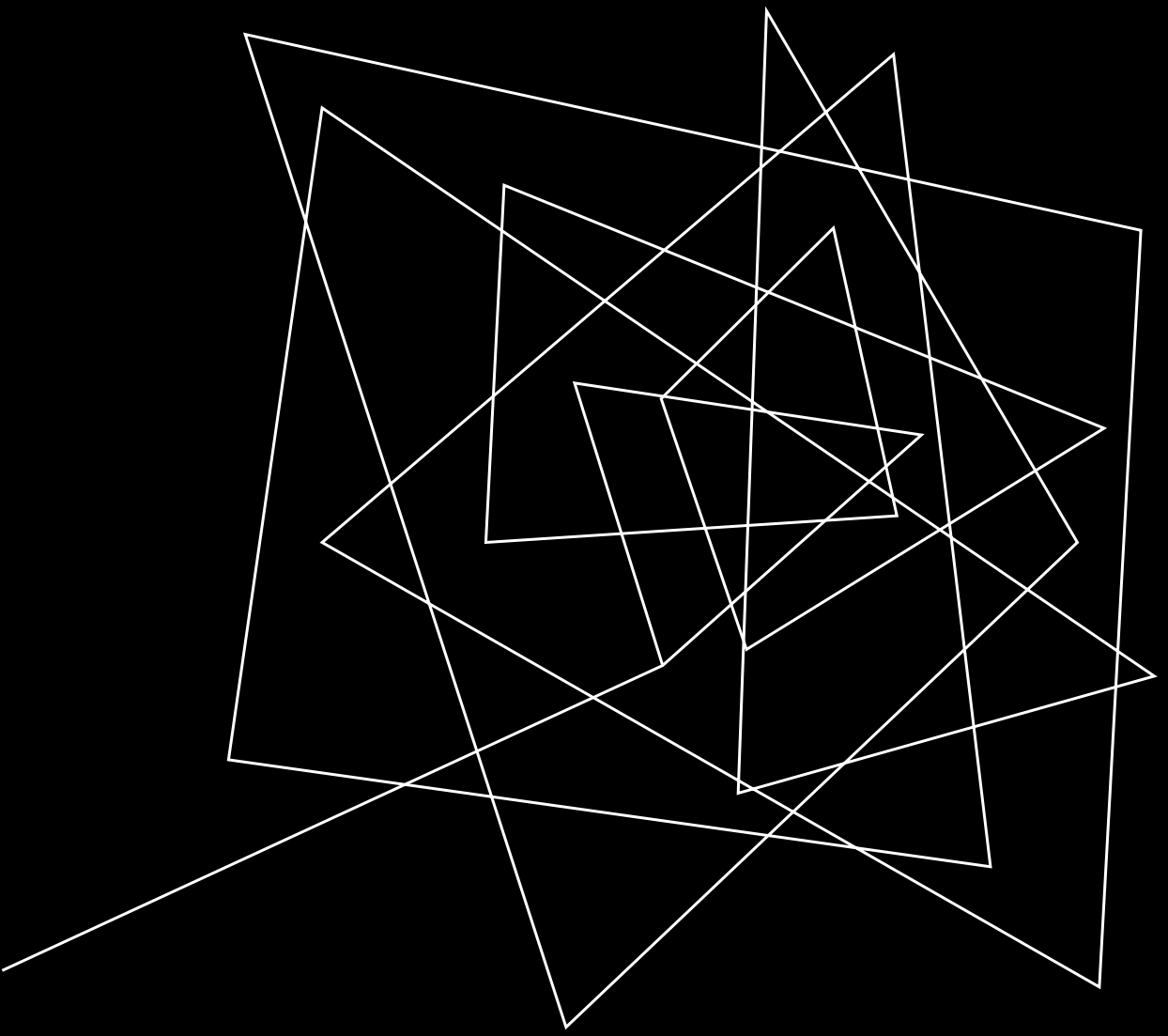
Modeling

Summary

PROBLEM STATEMENT

The finance sector is focused on one essential mathematical problem how can we assess and quantify risk? While this is usually calculated by large firms, in recent years more and more opportunities have arisen for individuals to not only buy but also sell financial products. LendingClub enables borrowers to create unsecured personal loans and investors to search and browse the loan listing on their website. This puts normal people in the same position as banks, allowing them to select loans that they want to invest in based on the information supplied about the borrower.

With Machine Learning, I aim to help answer this question by building a model that can evaluate and learn from previous loans to help recommend the best loans for individuals to invest in.



DATA WRANGLING

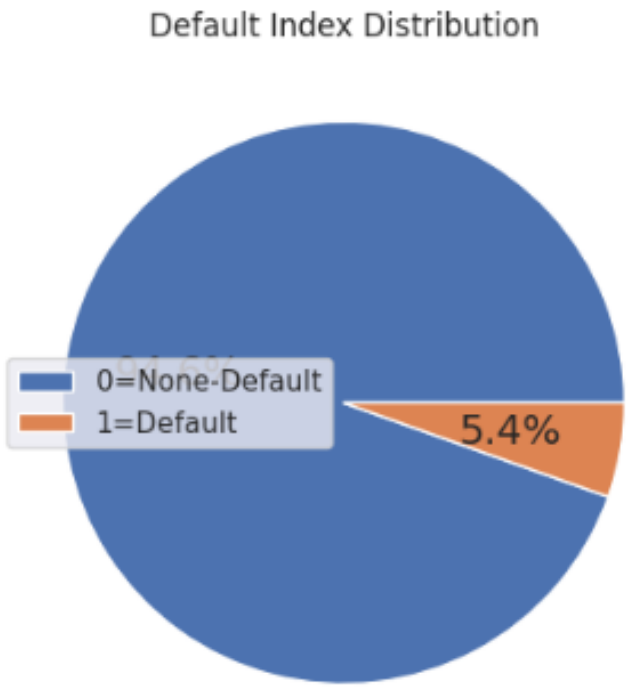
Original listing dataset had 855,969 rows and, 73 columns

Removed un-useful columns

Converted categorical column to numeric to avoid creating more dummies for modeling

The final dataset is 855,969 rows and, 25 columns

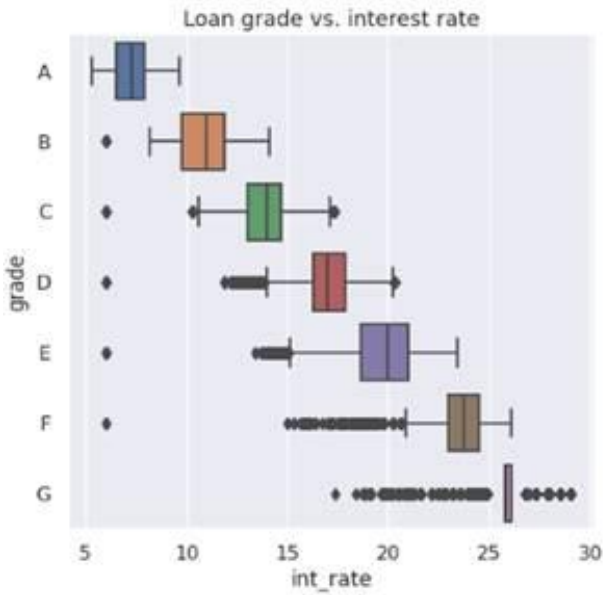
TARGET DEFAULT

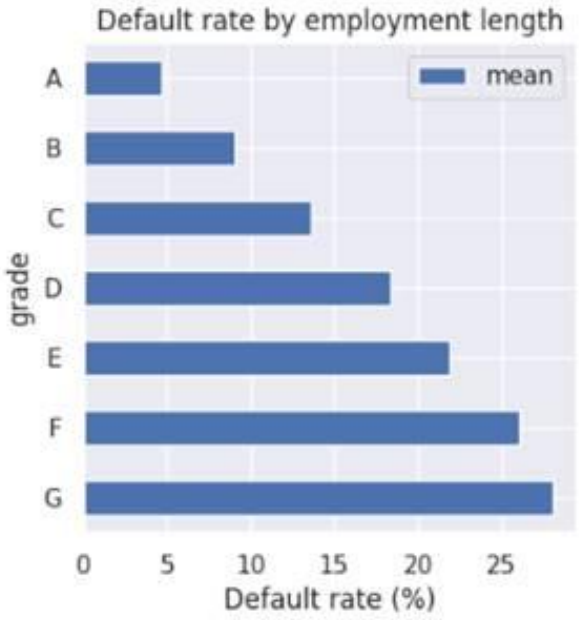


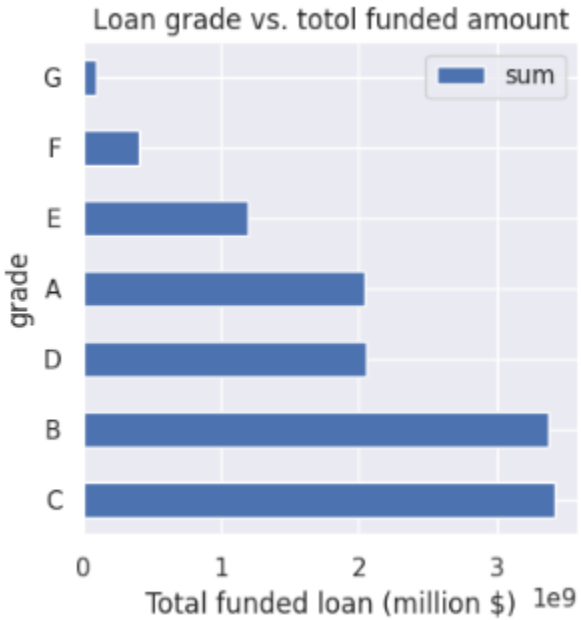
Why LendingClub default is higher than the average bank default of 3.9%?

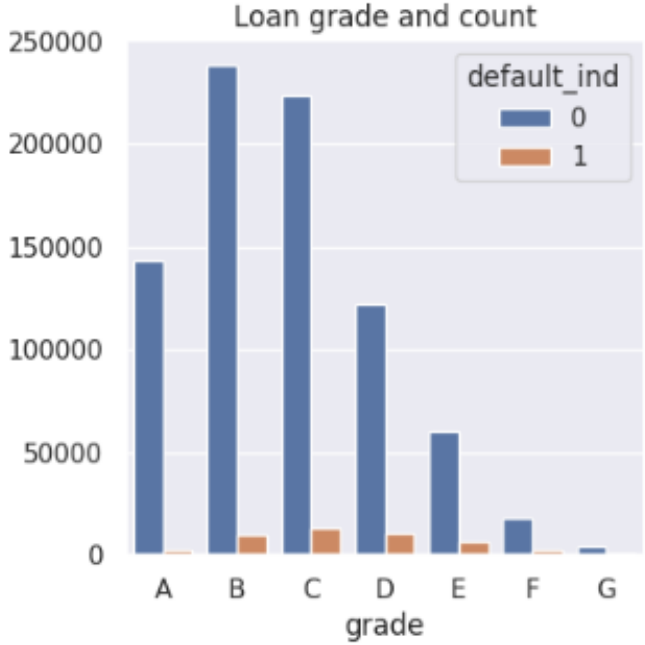
Are individual investors willing to tolerate extra risk for high returns with low-credit borrowers?

LOAN GRADE

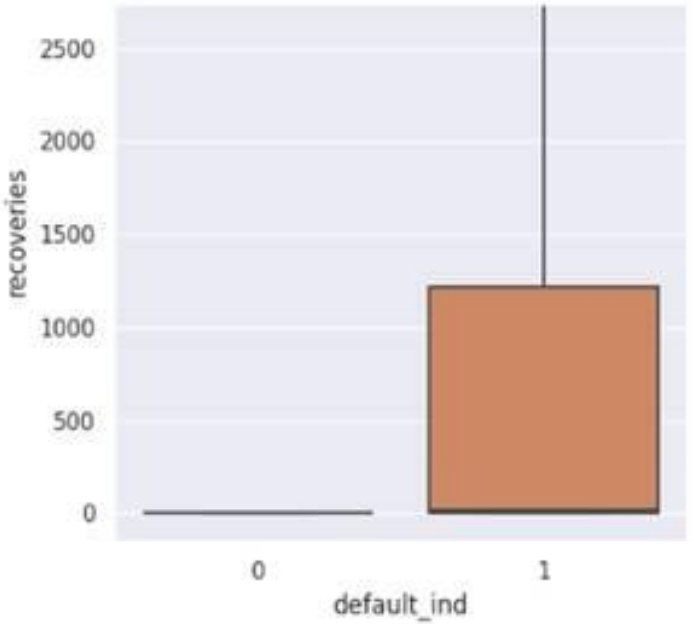

Loan grade vs. interest rate
grade
A
B
C
D
E
F
G
5
10
15
20
25
30
int_rate

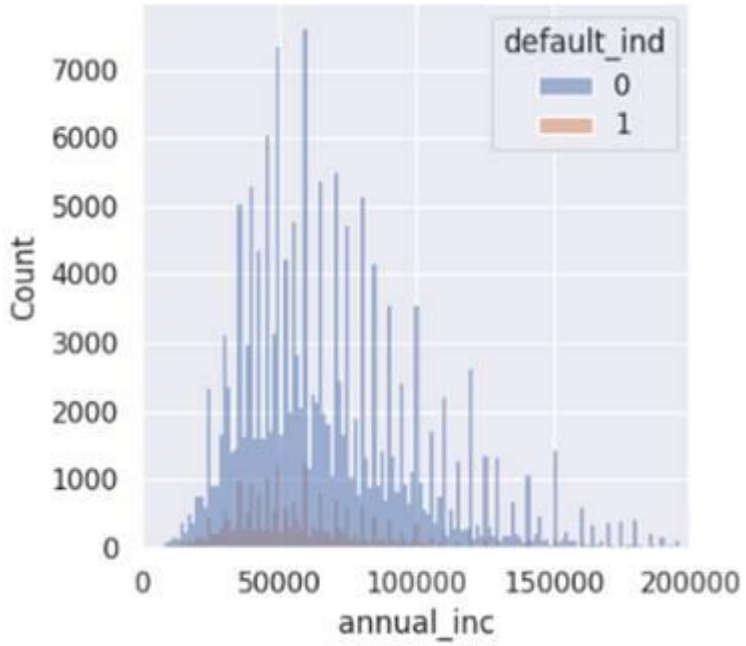

Default rate by employment length
mean
grade
A
B
C
D
E
F
G
0
5
10
15
20
25
Default rate (%)

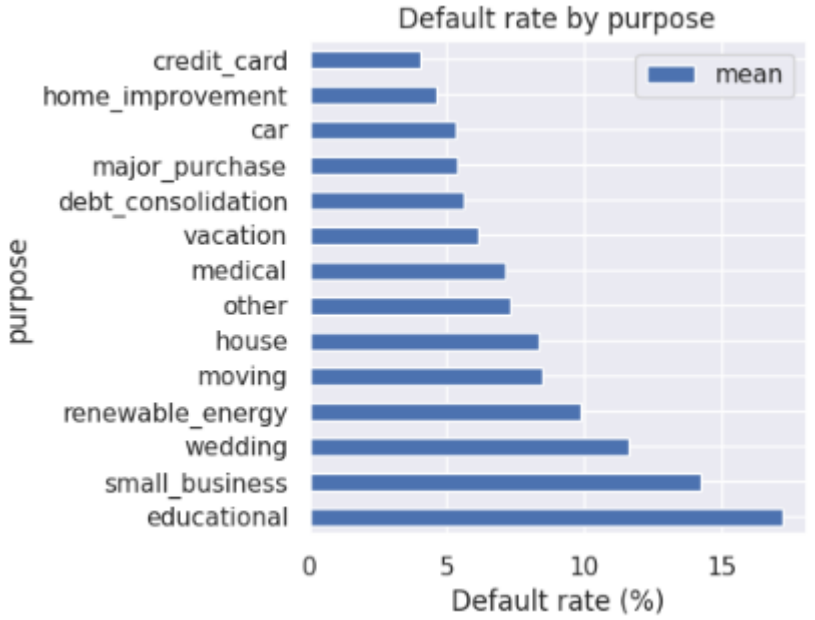

Loan grade vs. totol funded amount
sum
grade
G
F
E
A
D
B
C
0
1
2
3
Total funded loan (million $)
1e9


Loan grade and count
default_ind
0
1
250000
200000
150000
100000
50000
0
A
B
C
D
E
F
G
grade

RECOVERIES AND PURPOSE







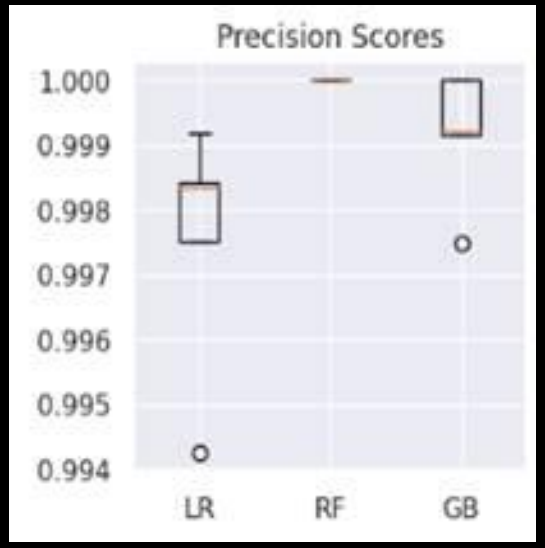
MODELING

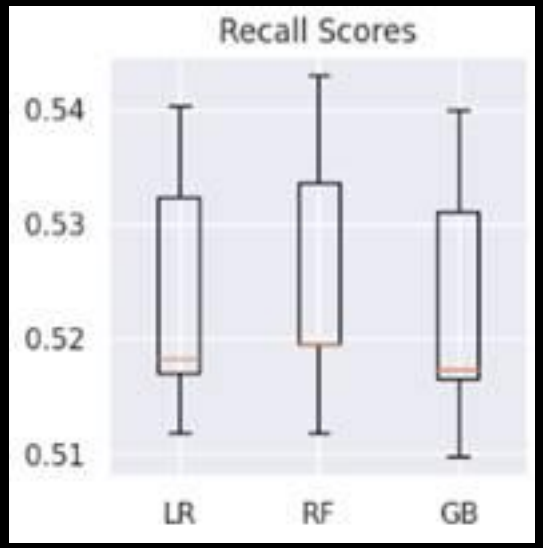
Logistic Regression

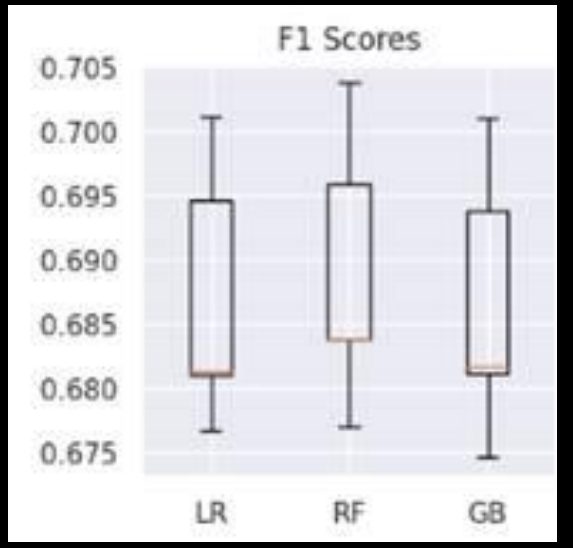
Random Forest

Gradient Boosting

PRECISION/RECALL

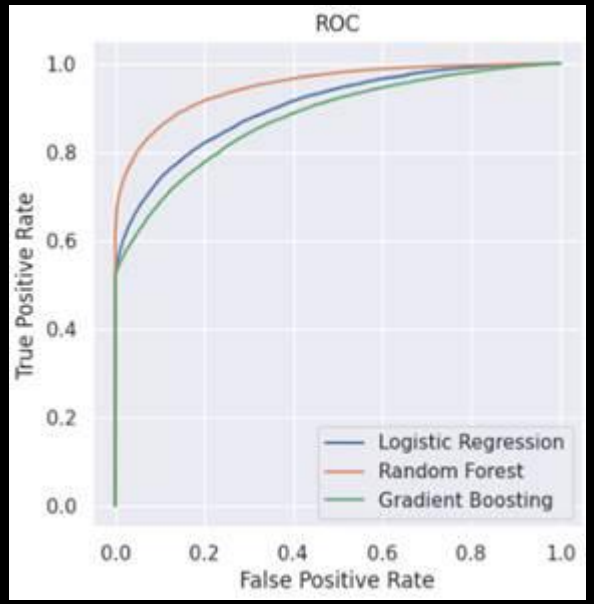






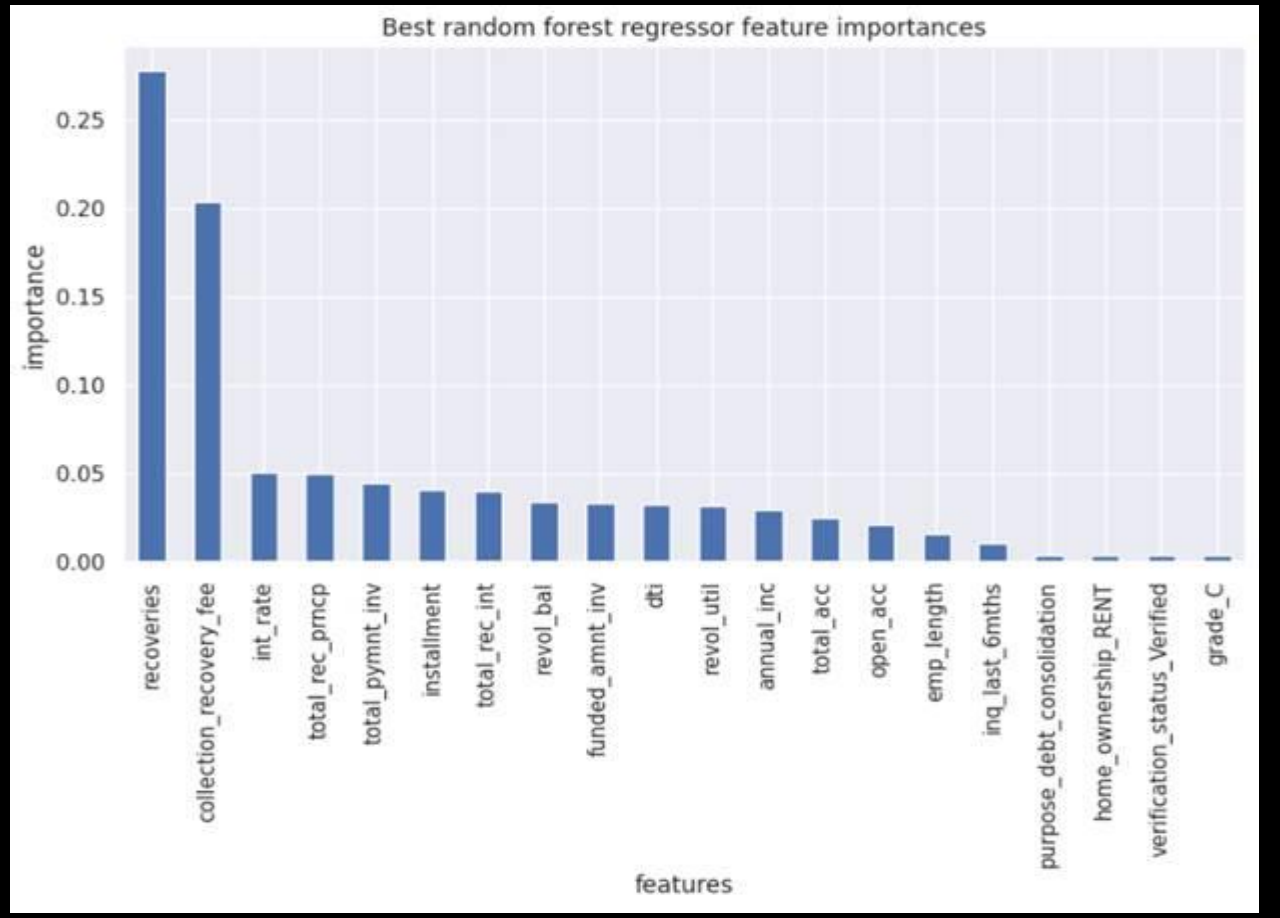
LR= Logistic Regression, RF= Random Forest, GB= Gradient Boosting

ROC/AUC

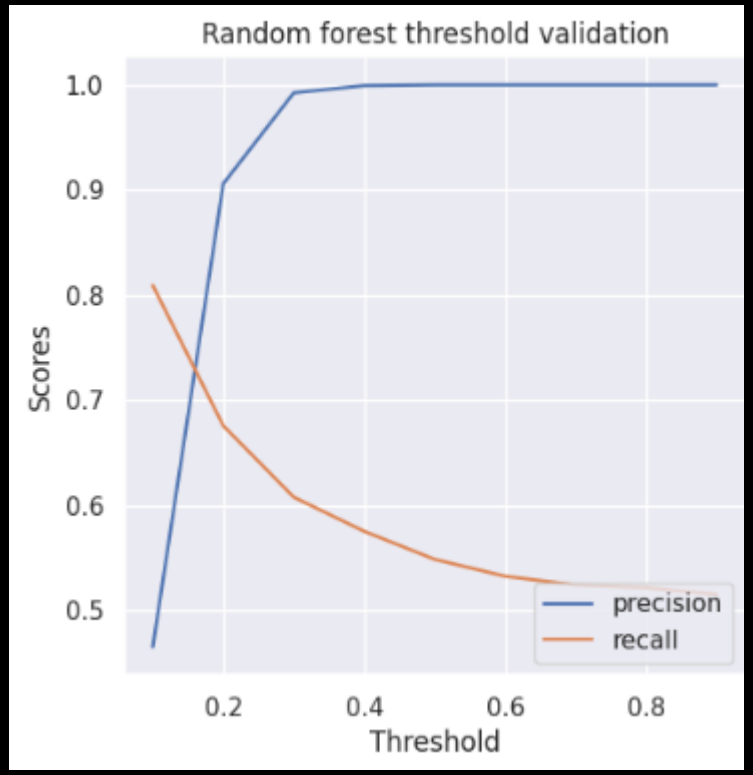


	Logistic Regression (LR)	Random Forest (RF)	Gradient Boosting (GB)
AUC	0.9	0.95	0.88
Best Estimators	Penalty=l2 Standardscaler=None	N_estimators=300 Standardscaler=None	Learning-rate=1 Standardscaler=standardscaller()

FEATURE IMPORTANCE



RANDOM FOREST VS. THRESHOLD

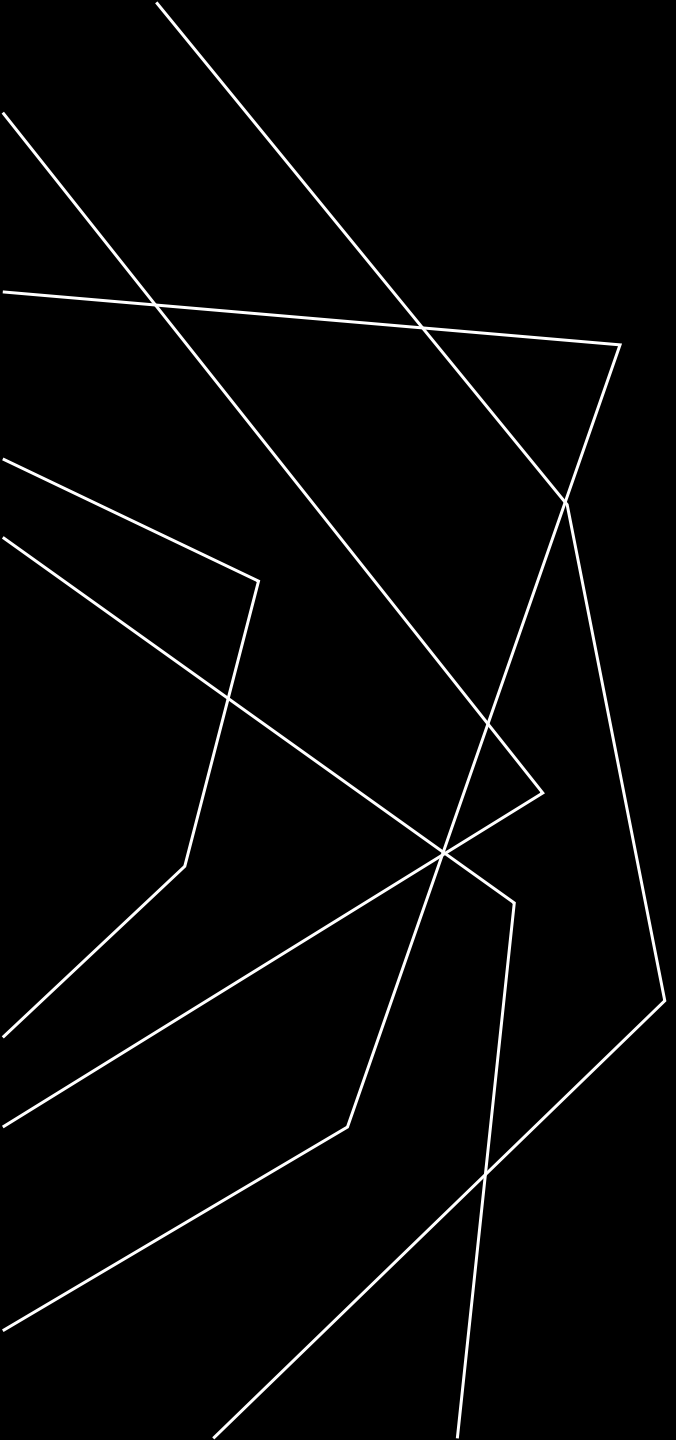


SUMMARY

Focusing on the need to maximize returns and minimize risks for investors, the random forest classifier seems like the best model as it has the best performance scores. Would recommend loan grades A and B to investors

FUTURE IMPROVEMENT

We are in the middle of a challenging macroeconomic environment with the highest inflation and the highest interest rate. As a result, default risk increases. Now is the time to be conservative. I would only recommend loan grades A, B, and C to investors as they have low default rates and are in the top three funded loans which means they have investors' trust.



THANK YOU