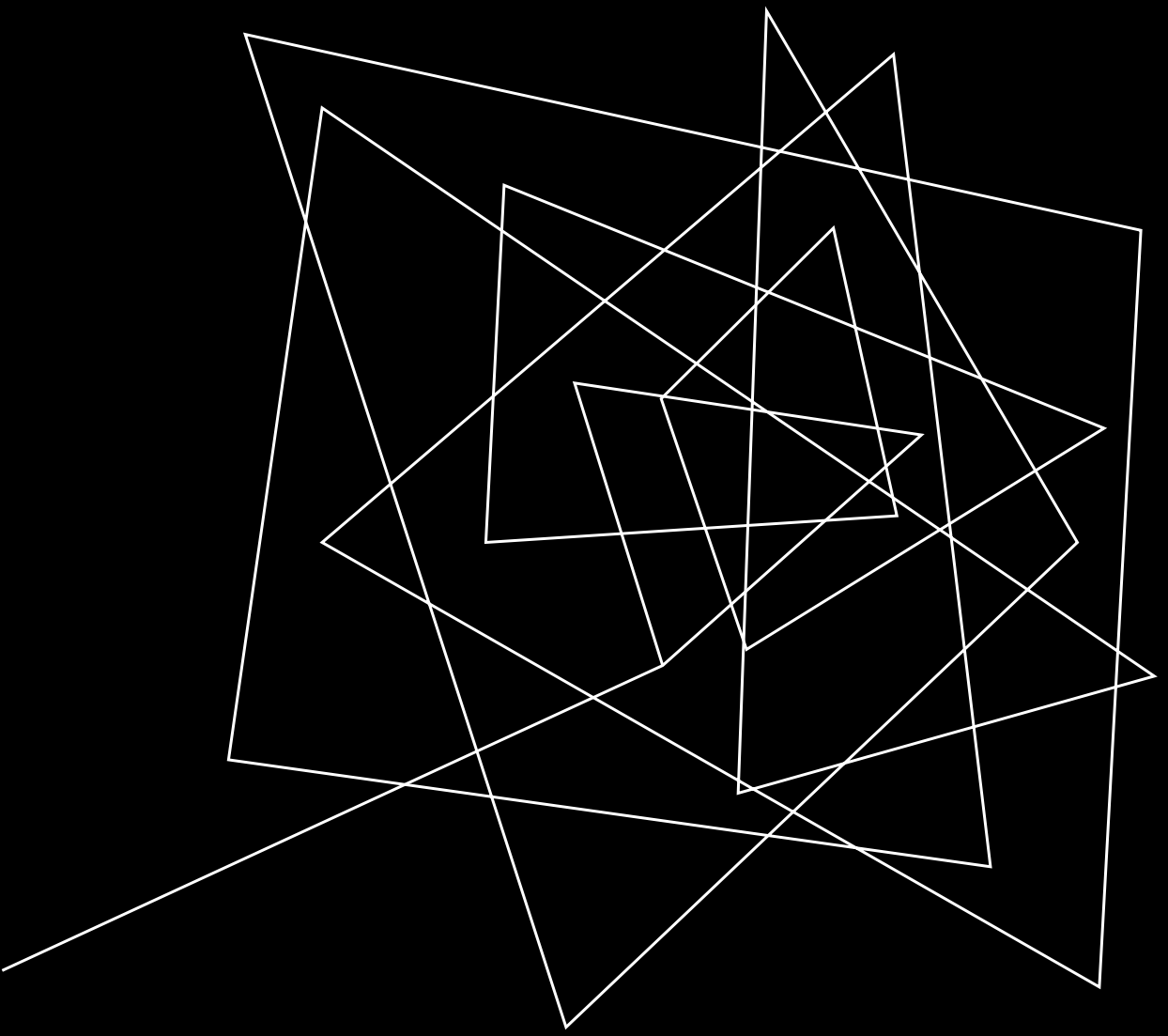
LOAN DEFAULT PREDICTION

CONTENTS

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

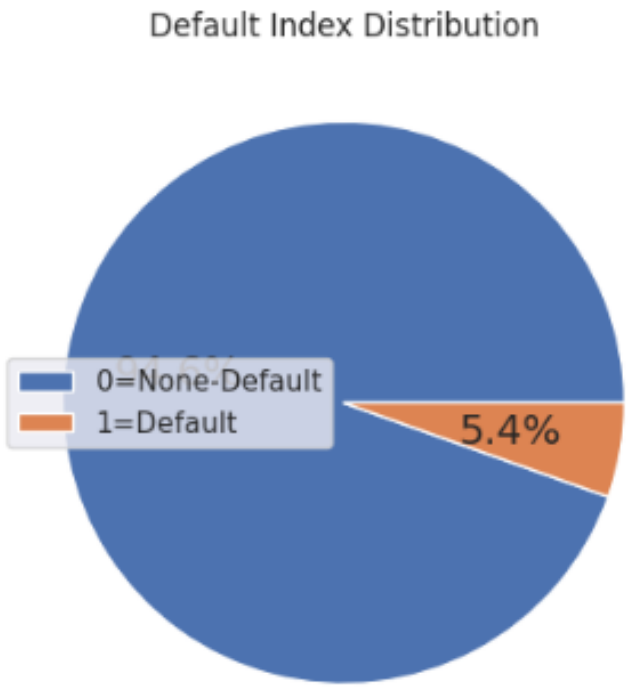
Original listing dataset had 855,969 rows and, 73 columns

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

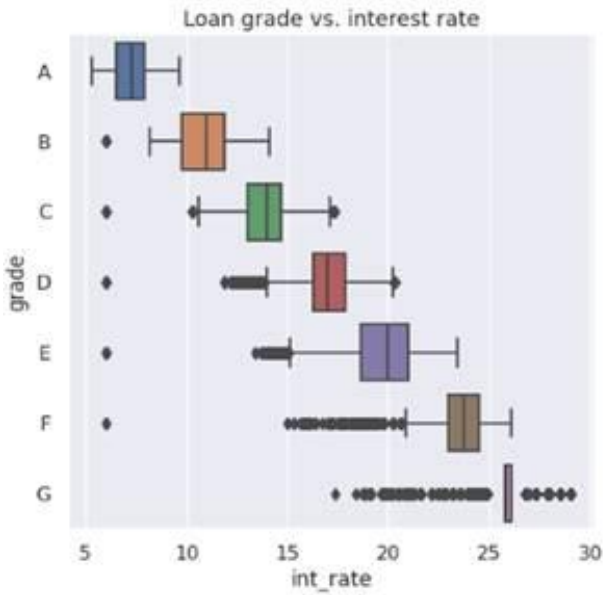
TARGET DEFAULT

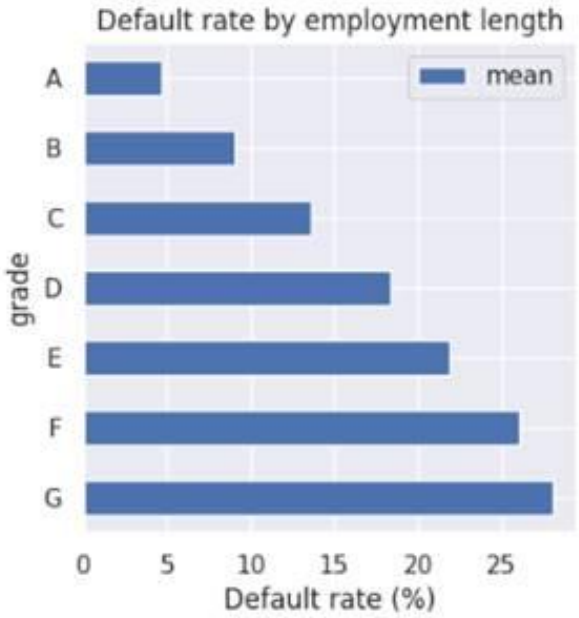


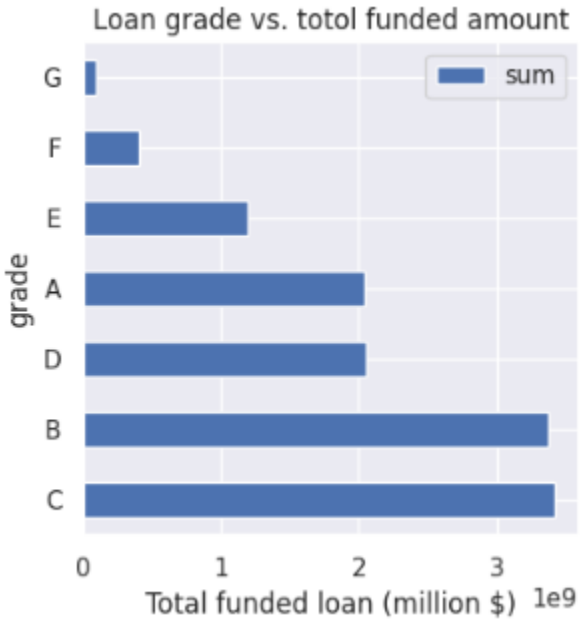
Why LendingClub default is higher than the average bank default of 3.9%?

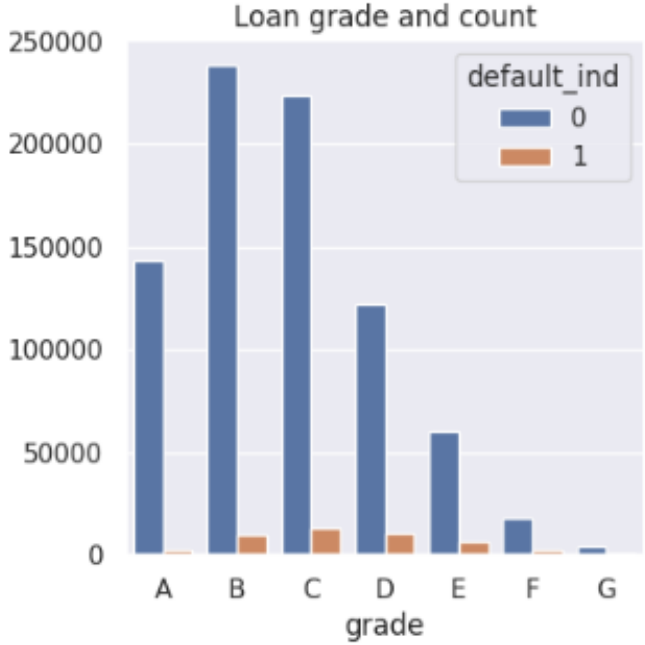
Are individual investors willing to tolerate extra risk for high returns with low-credit borrowers?

LOAN GRADE



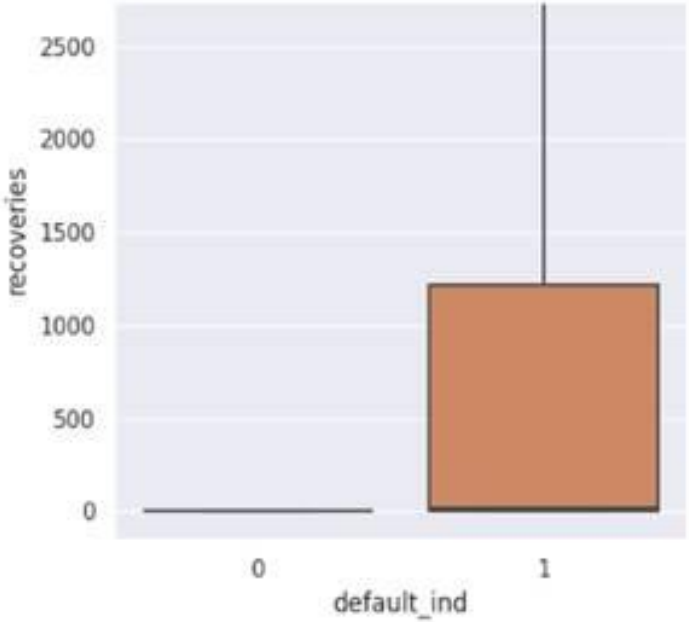


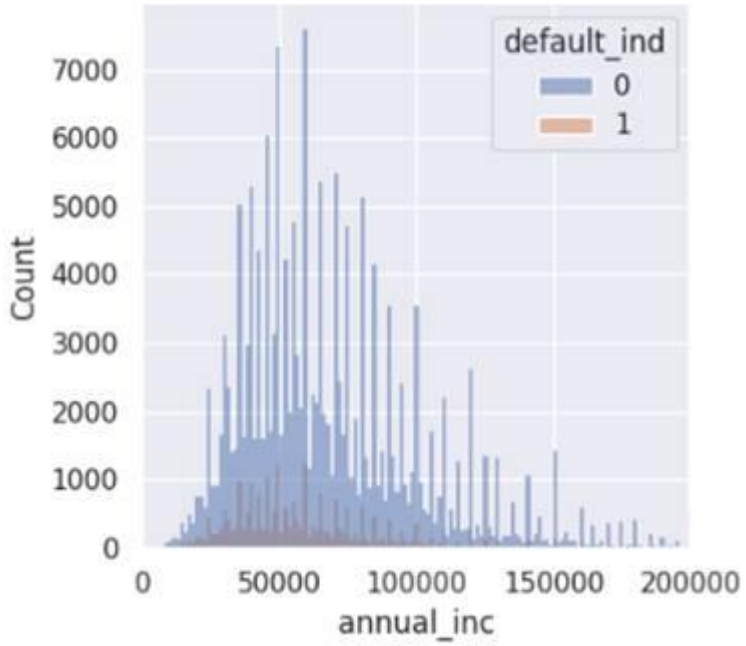


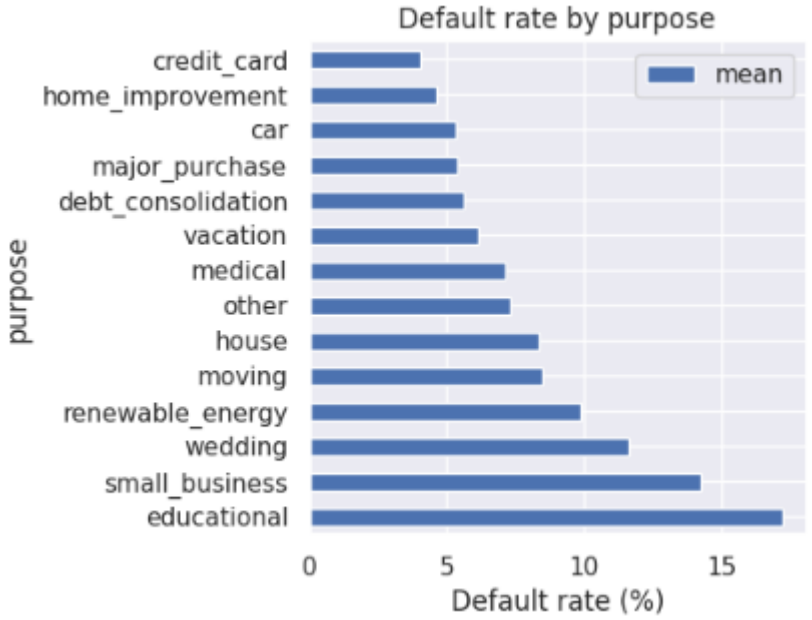


Lower grade higher interest and higher default

RECOVERIES AND PURPOSE







It seems anyone with recovery will default. Loan purpose does show a correlation to default rate.

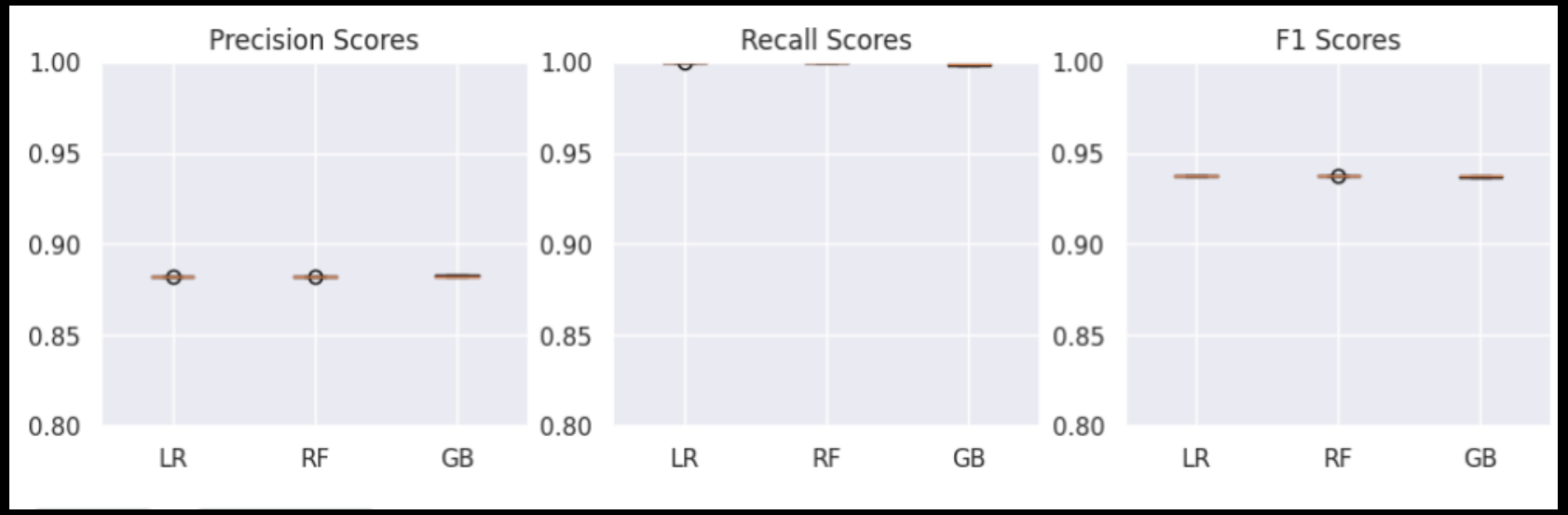
MODELING

Logistic Regression

Random Forest

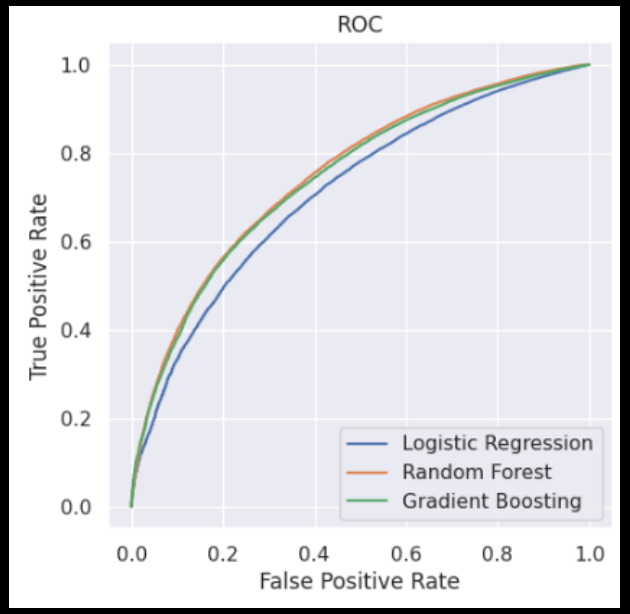
Gradient Boosting

PRECISION/RECALL



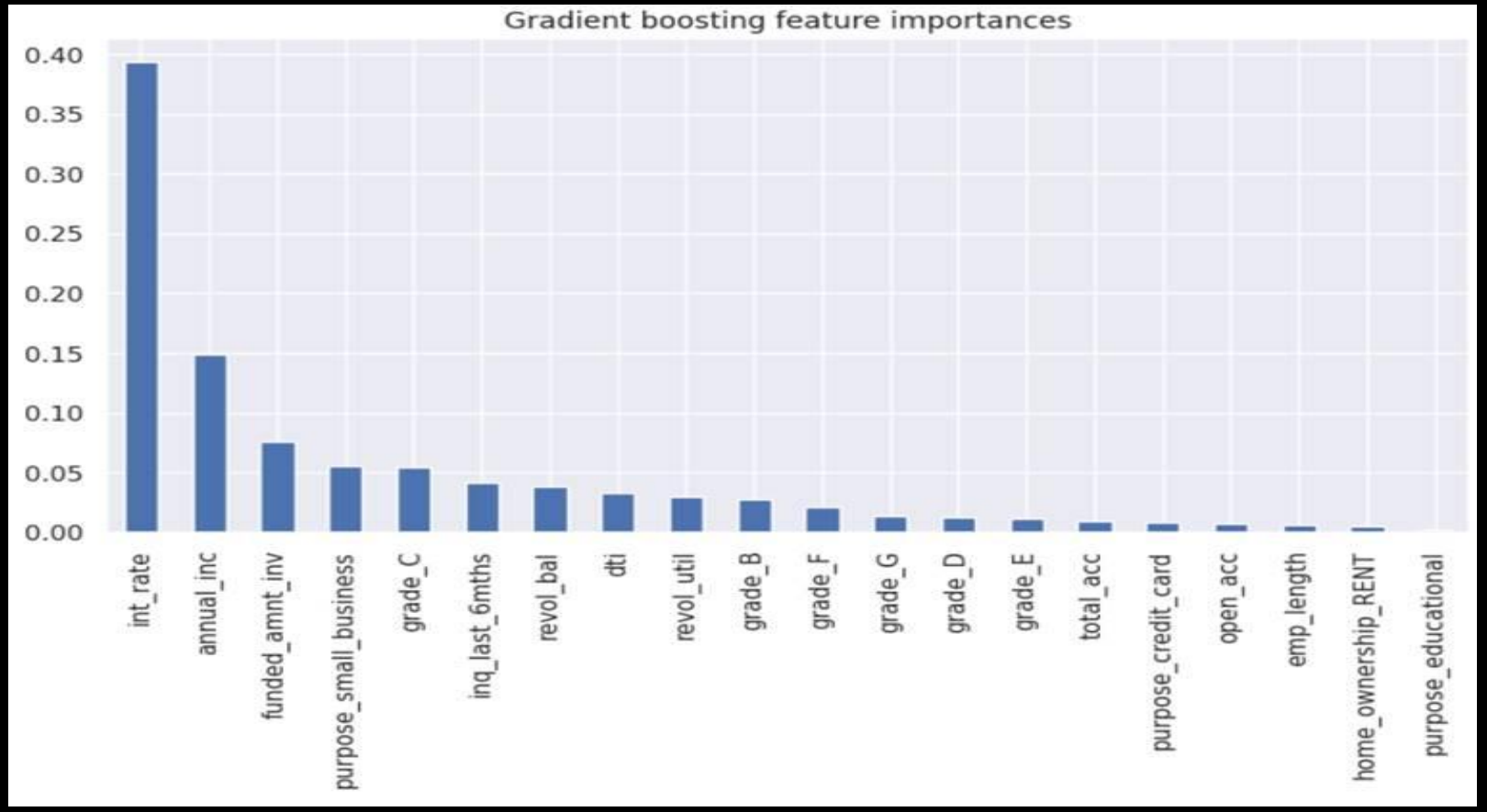
LR= Logistic Regression, RF= Random Forest, GB= Gradient Boosting

ROC/AUC

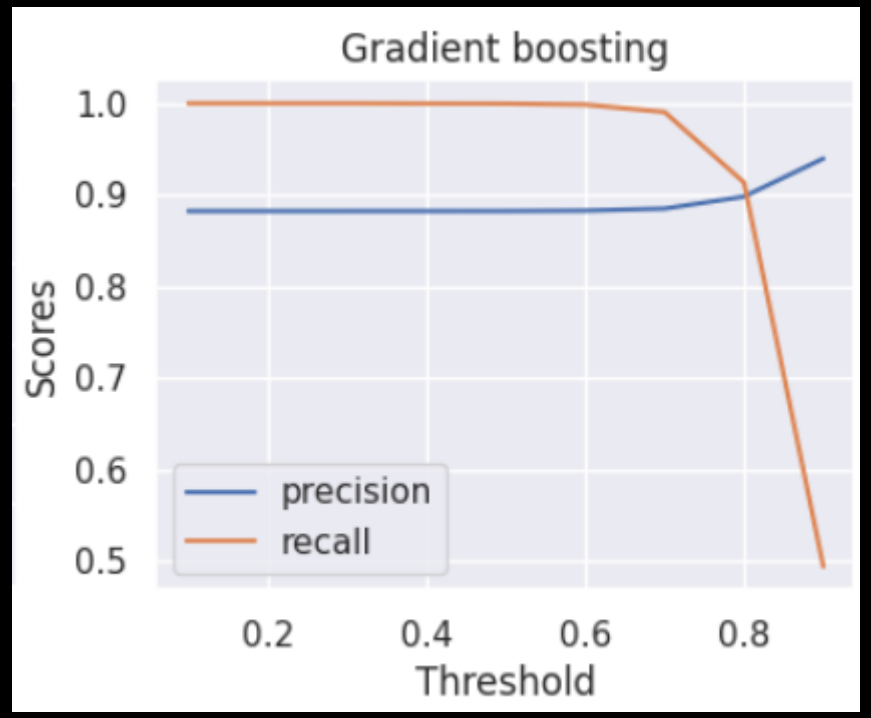


	Logistic Regression (LR)	Random Forest (RF)	Gradient Boosting (GB)
AUC	0.71	0.75	0.75
Best Estimators	'C': 0.001 'max_iter': 50	max_depth=5	Learning-rate=1

FEATURE IMPORTANCE



PRECISION/RECALL VS. THRESHOLD



SUMMARY

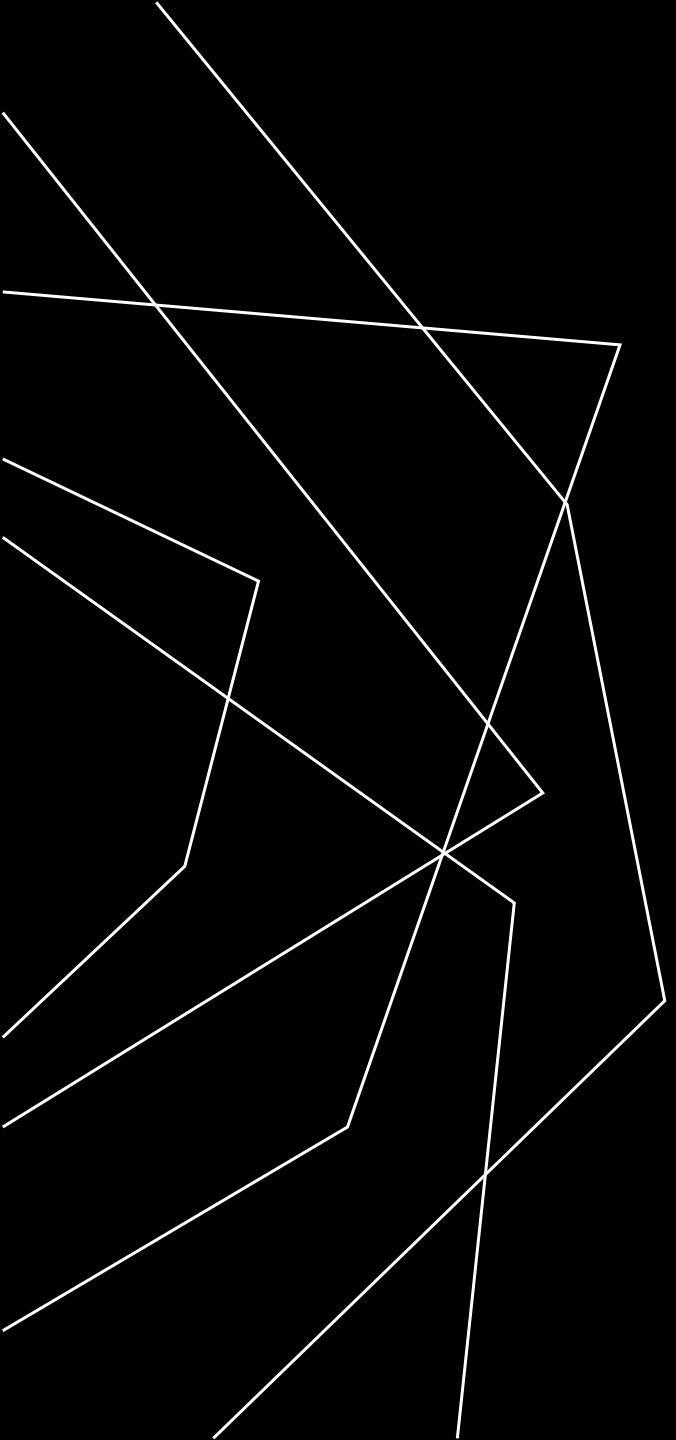
All three models have very similar performance. The important features of all models seem to agree with what we saw during EDA section that interest rate and loan grades have the strongest correlation to defaults.

For loan business, precision score is more important than recall as higher precision higher prediction rate for good loans. At the threshold of 0.95, gradient boosting just give us the best precision and recall, 0.986 and 0.353 respectively. You can argue recall is not so good, but recall is not critical because when loans are incorrectly classified as default it will only impact loan acceptance rate which leads to lower returns. Investors would be happier with slightly lower return than losing money for defaults.

FUTURE IMPROVEMENT

What loans are safe for investors? Stay away from loan grades E, F, and G as they have the highest default rates, with the least funded amount. Also watch out for education and small business loan purposes. They are the top two in the default list.

We are in the middle of a challenging macroeconomic environment characterized by the highest inflation and the highest interest rate in 22 years. A weak economy increases default risks. Now is a perfect time to be more conservative.



THANK YOU