Loan Default Capstone

Problem Statement

The finance sector focuses 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 buy and sell financial products. LendingClub, a financial services company headquartered in San Francisco, California, enables borrowers to create unsecured personal loans between $1,000 and $40,000 and investors to search and browse the loan listing on LendingClub website. This puts ordinary people in the same position as banks, allowing them to select loans they want to invest in based on the information supplied about the borrower, loan amount, loan grade, and loan purpose.

But individuals acting as banks then have the same problem as banks – how can they accurately assess the risk of giving a loan to maximize their return? 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 link: https://www.kaggle.com/datasets/adarshsng/lending-club-loan-data-csv

Data Wrangling

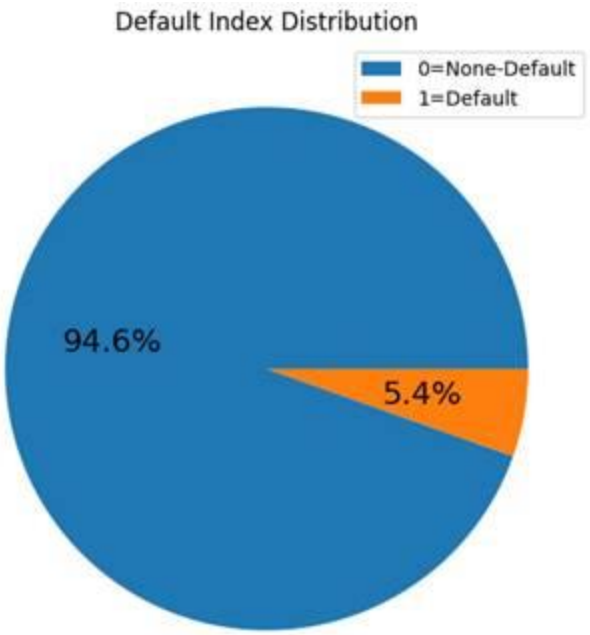
The dataset had 855,969 rows and 73 columns. After carefully reviewing the data, I removed 45 un-meaningful columns to keep the width of the dataset at a manageable size with good information for problem-solving. Next, I removed unwanted texts from the employment-length column and converted it from object to numeric to avoid creating more dummies for modeling. As I wanted to keep as much data as possible for data analysis, I will deal with missing data later. By the end of data wrangling, the total number of columns is 25.

Exploratory Data Analysis (EDA)

EDA is the most crucial task in data science. Without understanding the data and their correlation, there is no meaning in training machine learning as good data in good data out. We will uncover any patterns, trends, and correlations between the target variable with independent features and detect outliers or anomalous data to help answer the business questions.

Target variable default

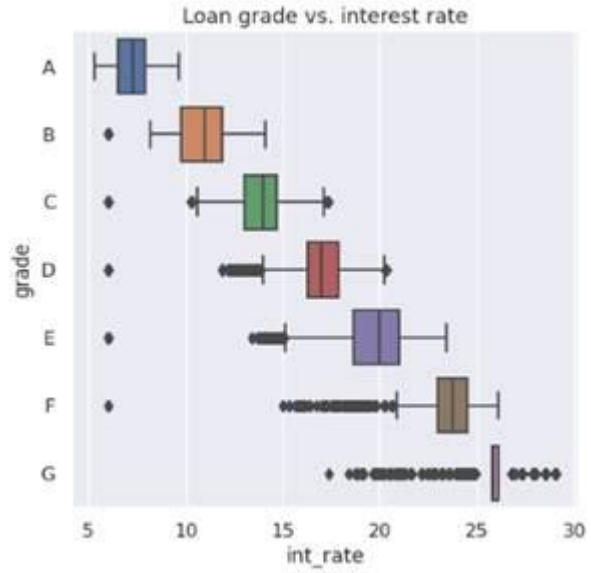
First, let's look at the default, a target variable of the problem-solving. Default happens when a borrower stops making required payments on a debt. According to TransUnion, in the first quarter of 2023, 3.91% of personal loan borrowers were late on their loan payments. The default rate of LendingClub is 5.4%, as shown in the pie chart below, a ratio of 20 to 1. Why LendingClub's default rate is so high? Is the platform more advantageous for borrowers with economic difficulties and low credit scores than traditional financial institutions, or are individual investors willing to tolerate the extra risk for higher returns?

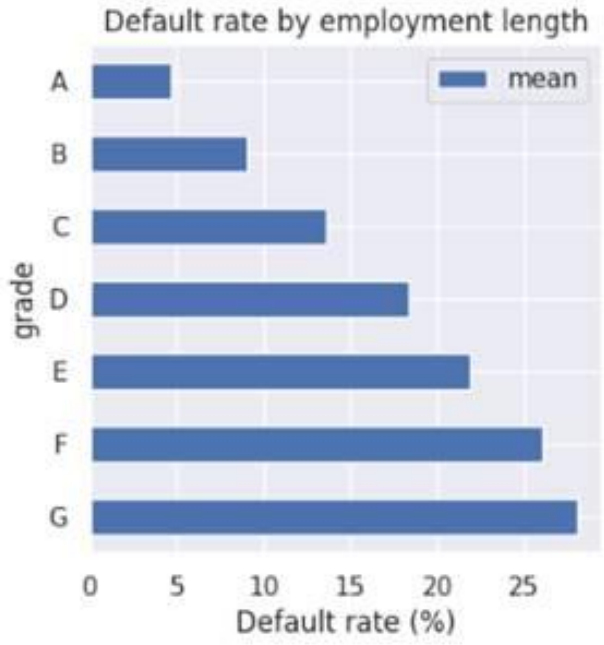


Independent features

Loan grade

Loan grade is a quality score to a loan based on a borrower's credit history, quality of the collateral, and the likelihood of repayment of the principal and interest. Loans grade A are the loans with the lowest expected risk of loss and, therefore, pay the lender the lowest interest. On the other hand, loans grade G pay the highest compensation to the lenders for the highest risk. The box plot below demonstrates the relationship between loan grade and interest. The average interest of grade A is 7% while grade G is 26%, almost a 19% difference between the lowest risk to the highest risk loans. A 30% APR is pretty high for personal loans, but it is what people pay for with their bad credit.

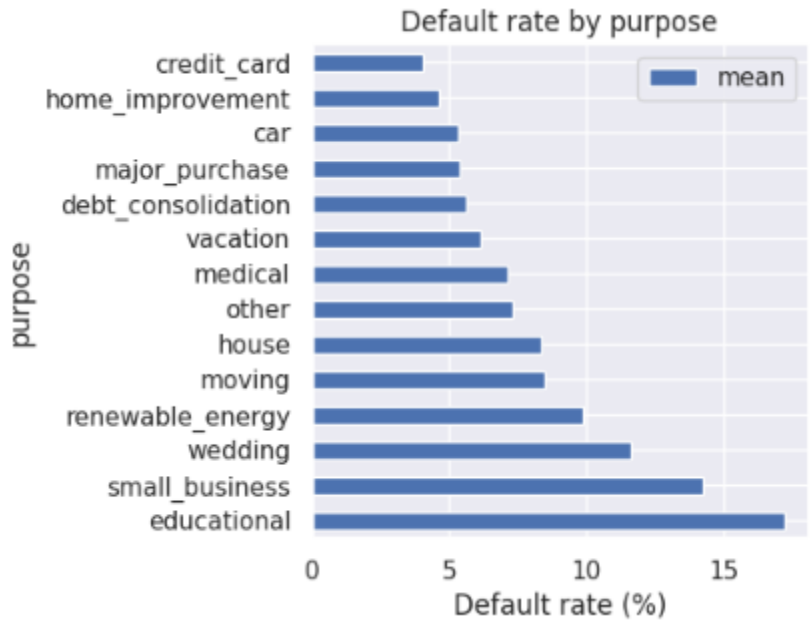




There is a risk associated with any loan. Even loans grade A have defaults, but the default rate is way less than lower loan grades. The best loan grade A has a 5% default rate, while loan grade G has a 28% default rate. Loan grades have a substantial direct correlation with defaults. Do high-interest rates compensate for the increased risks? No, it does not seem so.

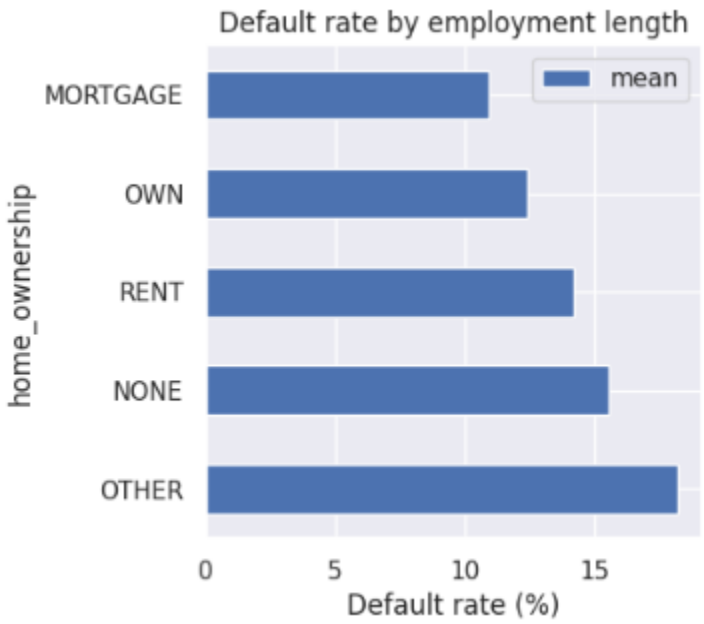
Loan purpose

Personal loans can be used for various significant purchases and expenses. Surprisingly, educational loans are at the top of the default. Still, the rate matches with the U.S. Department of Education reports that about 20 percent of borrowers default, and more than a million loans default yearly. The second purpose on the list is the small business loan, which is unsurprising as doing business comes with risks. Next is a wedding loan. Weddings are personal events tied to emotion; exceeding your budget is expected.



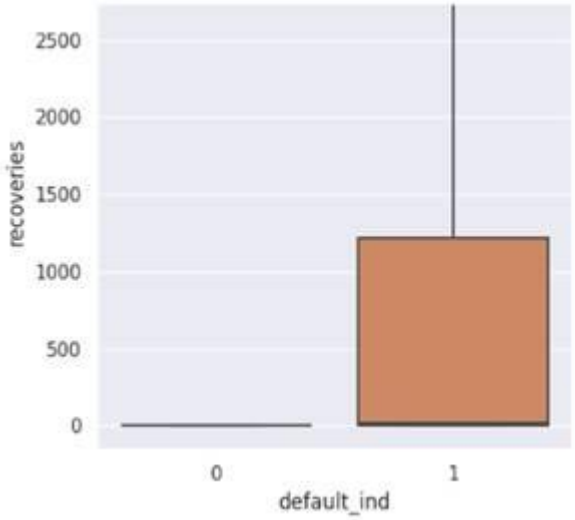
Home ownership

People who own a house or a property should think twice before defaulting on a loan. Besides damaging their credit, lenders or debt collectors can recover the money they hold by taking them to court, garnishing their wages, and putting a lien on their property. They may also lose their collateral, like the money in savings accounts. Therefore, the default rate is lower for these groups. On the other hand, people do not have property; they have less to lose, so the chance of default is higher.



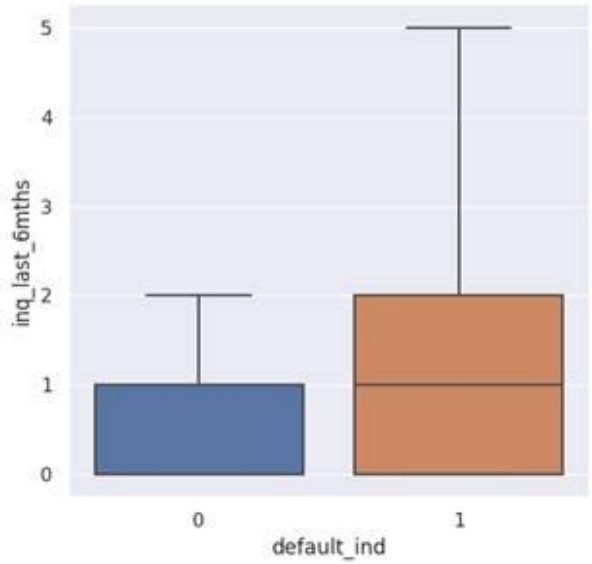
Recoveries

Recoveries are “post-charge-off gross recovery” when a company writes off debt as a loss it believes it can no longer collect as the borrower has become delinquent on payments. Charge-offs remain on the credit report for seven years. The borrowers are still responsible for the charge-off. As we can see, no examples of recovery occur in non-default cases. It seems anyone with recovery will default.

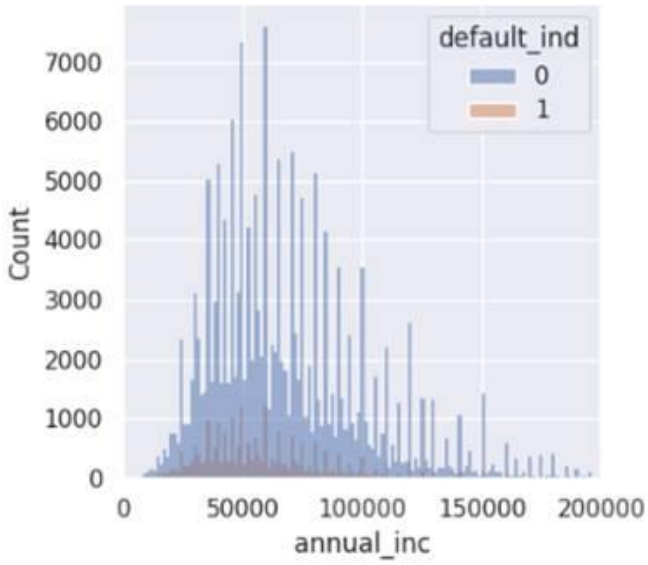


Inquiry

Inquiry is an essential aspect of the entire loan approval process by which lenders evaluate borrowers' credit history and credit score. There are two types of inquiries. The hard inquiry is made for serious purposes only as they leave a record on the borrower's credit report. In contrast, a soft inquiry is made by a borrower who wants to know their credit score or by a lender to offer pre-approved loans that do not affect your credit score. Frequent hard inquiries decrease credit scores as they can be seen as a symptom of financial insecurity of demand for extra credit. There is no way to know if inquiries shown in the data set are hard or soft, but my guess was that they are hard inquiries and showed a strong correlation to the default. The average is one inquiry, but some bad loans had up to 5 inquiries. Thus, it seems there is some relationship between the number of inquiries and the likelihood of defaults.



It seems like default loans have slightly lower income than the non-default loans.



Feature engineering

With a good understanding of all the features and their correlation, additional unwanted features like recoveries, collection fee, and total payment were dropped before the data was split into train and test data sets with 25% test size. Categorical features were transformed into numerical ones, which can be used in machine learning models. Remember during data wrangling, many missing values have been left untouched as we wanted to view the distribution from all loans. Those missing values were imputed with median values then standard scaler was applied to X_train and X_test.

Before moving on to modeling, both y_train and y_test values were inverted, good loans have the value of 1, and bad loans have the value of 0. The reason behind this is we will use precision and recall for model evaluation. Precision measures the accuracy of positive predictions and recall measure the completeness of positive predictions. More about precision and recall will be detailed in the section below.

Modeling

The data is labeled which means we will be solving a supervised classification problem. After determining the type of machine learning, I chose Logistic regression (LR), Random forest (RF), and Gradient boosting (GB) for modeling. I used GridSearchCV to find best estimators before training and predicting. Scores will be used to evaluate the model performance are:

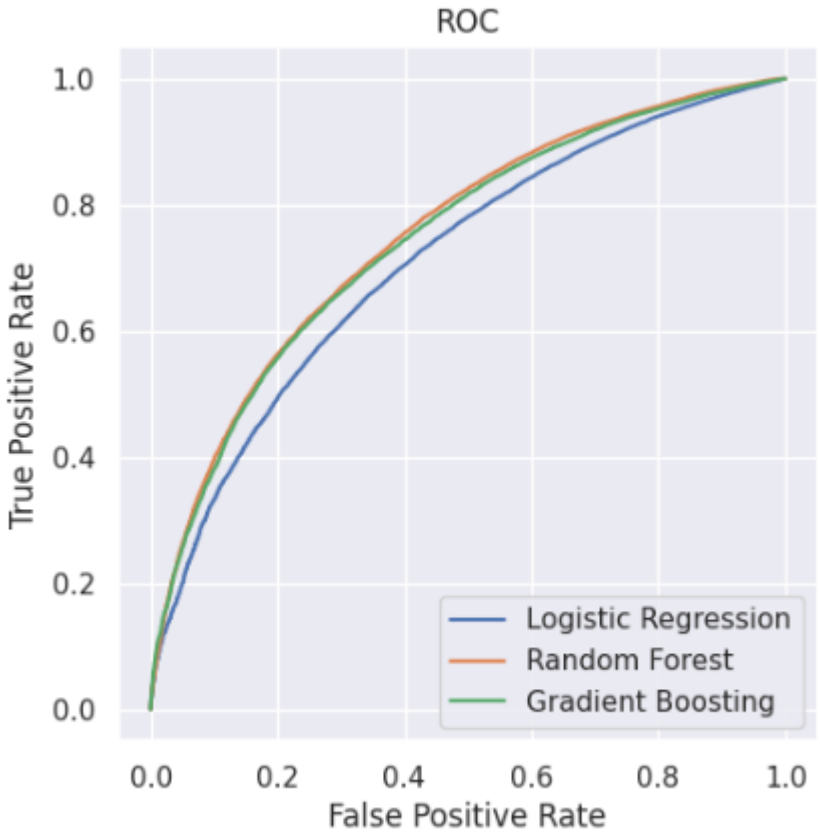
- Precision = TP/(TP+FP). It is a measurement of good loans that are correctly identified as good loans.
- Recall = TP/(TP+ FN). In simple terms, recall tells us how many good loans are classified as bad loans.
- ROC is receiver operating characteristic curve used to summarize the model's performance tradeoff with all classification thresholds between true positive rate (TPR) and false positive rate (FPR). TPR=TP/(TP+FN) and FPR = FP/(FP+TN)
- AUC is area under the ROC curve which used to measures the entire two-dimensional area underneath the entire ROC curve across all possible classification thresholds

Terminology from confusion matrix to help us gain insight into how correct our predictions were compared to the actual values**.**

- TP= True Positive (good loan and classified as good loan)
- TN= True Negative (bad loan and classified as bad loan)
- FP= False Positive (bad loan and classified as good loan)
- FN= False Negative (good loan and classified as bad loan)

ROC/AUC

ROC curve typically feature TPR on Y axis and FPR on X axis to show the trade-off between them. Models that give curves closer to the top left corner is the ideal, larger under the curve (AUC). Notice that random forest and gradient boosting have a AUC of 0.75 and is better than logistic regression.

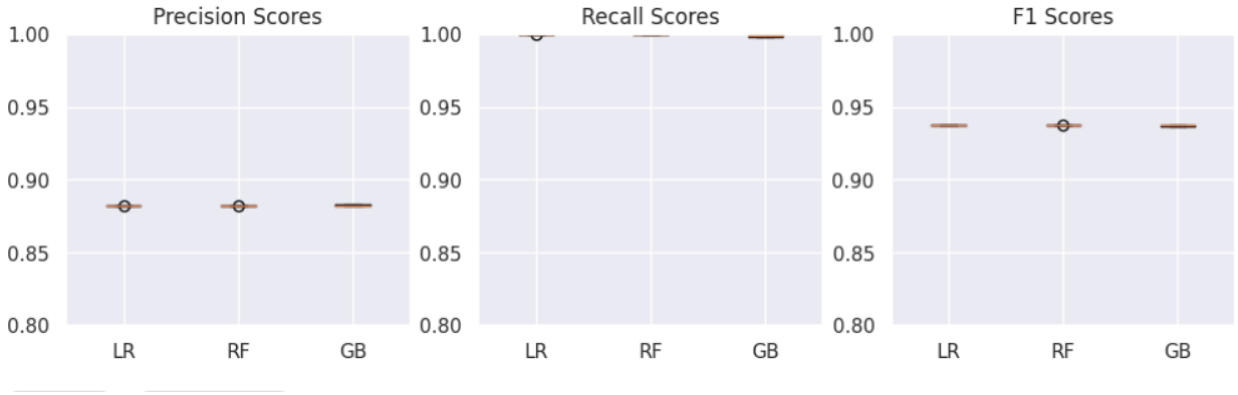


	Logistic Regression (LR)	Random Forest (RF)	Gradient Boosting (GB)
AUC	0.71	0.75	0.75
Best Estimators	'C': 0.001 'max_iter': 50	max_depth=5	Learning-rate=1

Precision and Recall

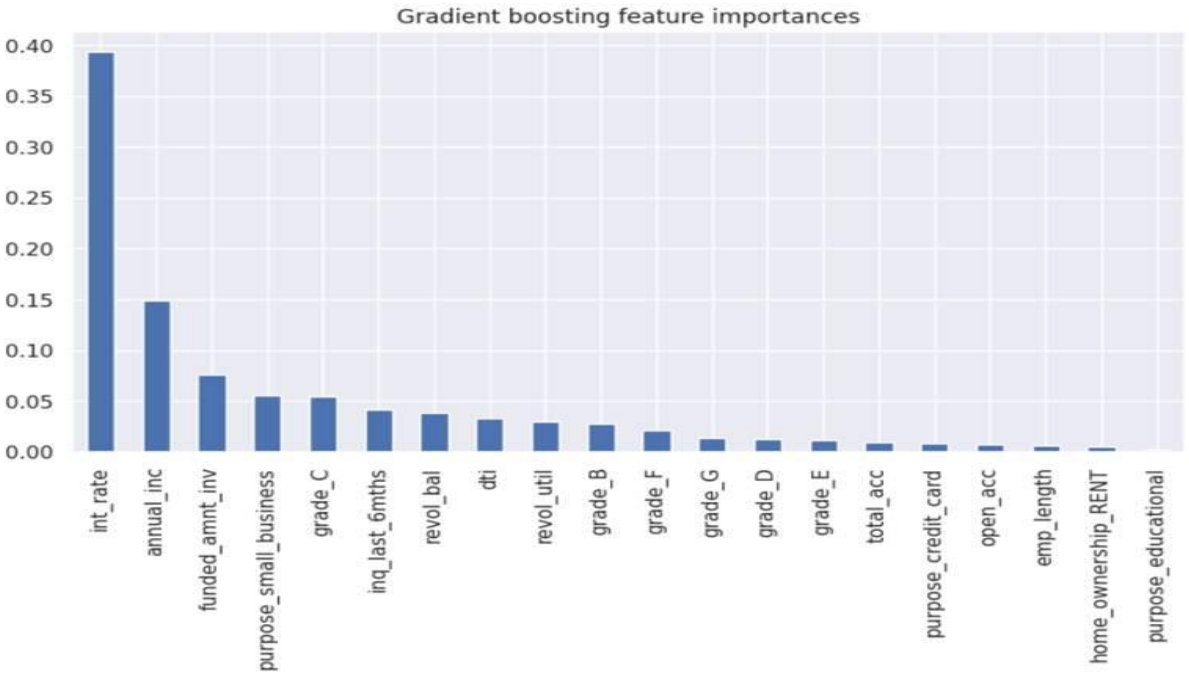
Since, the data set is imbalanced, more good loans than bad loans, Precision and recall are essential metrics to use in our case. As shown in the plots below, all 3 models have similar performance. Precision score of 0.87 means 13% of bad loans were incorrectly classified as good loans which is not good. Investor will lose their money for these loans. A recall score of 1means there was no good loans classified as bad loans. Unfortunately, recall of 1 is not enough to minimize the default risks. F1 is the

measurement of accuracy which combined the precision and recall scores that is why F1 is right in the middle of precision and recall cores. I am not happy with the result but there is one more parameter that we can evaluate later on to find the good trade-off between precision and recall. Remember our goal is high precision as close to 1 as possible.



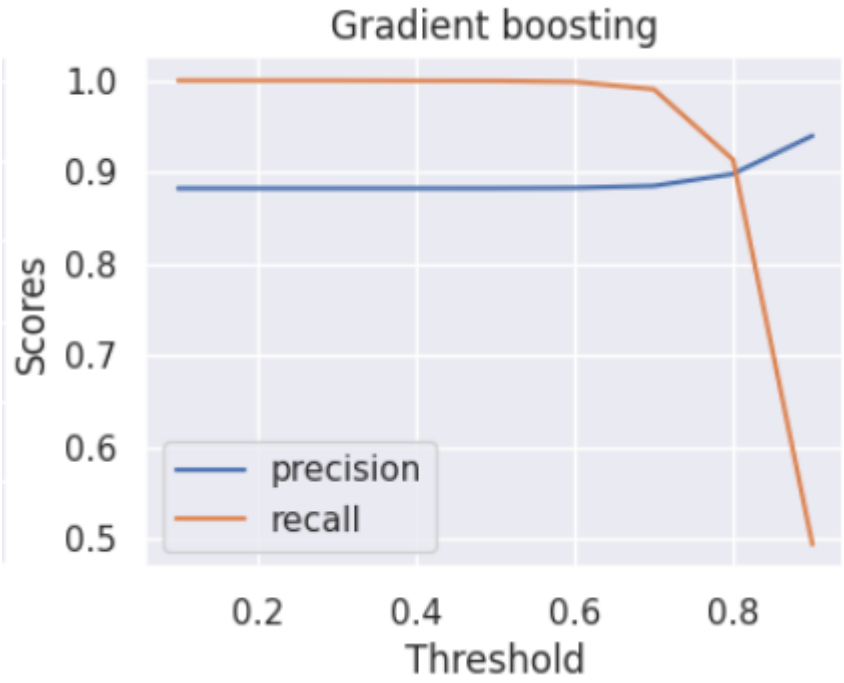
Feature importance

The important features of all models seem to agree with what we saw during EDA section which showed the interest rate is the most important feature that has a strong positive correlation to defaults, the lower the loan grade the higher the interest rate and higher default rate. I seem to prefer Random forest and gradient boosting more as they can pick up the annual income, purpose loans, and inquiry among top features that shows they handle highly correlated features well, whereas the logistic regression would not.



Model conclusion

All three models have very similar performance on precision, recall, and AUC scores through cross validation. 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. To further nail down a best model, I trained the models with thresholds vary from 0.1 to 1 to find a good trade-off between precision and recall. 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? Loan grades E, F, and G have the highest default rates, with the least funded amount. I calculated profits for each of them and could not find the reward for higher risks. Loan grades A, B, C, and D are safer. Also, I would recommend investors stay away from education and small business loan purposes as they are the top two in the default list.

As show during EDA, recoveries feature showed a solid correlation to defaults. Recoveries are writes off debt when the borrower has become delinquent on payments. 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 revisit the most important feature recoveries, refresh strategies, and establish more efficient fraud detection.