Lending Club Loan Data Analysis

Springboard Capstone Project

J Slaga

Peer-to-peer lending sites such as Lending Club match lenders and borrowers through their online platforms. These businesses operate with a much lower overhead compared to traditional banks. Thus borrowers are able to get unsecured loans at lower interest rates and lenders are able to earn higher returns. The interest rates for the loans are determined by risk - borrowers deemed more risky will have a higher interest rate and those with less risk will have a lower interest rate.

Loans with higher interest rates (but more risk) provide a higher return on investment so tend to be more appealing to investors, however their risk of default is higher than the lower interest rate loans (with a lower return of investment).

Creating a machine learning model to predict which of the loans are more likely to default enables investors to gauge which of the high interest loans are more likely to be returned, and identification of defaulters will allow investors to better decide which loans to invest in or grant. The model will aid in finding the balance between risk and return on investment.

Client

Lending Club is the world's largest peer-to-peer lending (crowdlending) company, headquartered in San Francisco, California. It provides a link between investors who provide funds and borrowers seeking unsecured personal loans between $1000 and $40,000 at interest rates ranging from 5.6%-35.8%, depending on the loan term and borrower rating.The standard loan period is three years. The default rates vary from about 1.5% to 10%.

Investors can search and browse the loan listings on Lending Club platform and choose the loans that they want to invest in based on the borrower's information such as the amount of loan, annual income, number of open credit accounts, and loan purpose, and the Lending Club assigned loan grade. Investors make money from interest - higher interest (but higher risk of default) will have higher returns. Lending Club makes money by charging borrowers an origination fee and investors a service fee.

Lending Club Loan Data Analysis

Data

Data was obtained from https://www.kaggle.com/wendykan/lending-club-loan-data. It contains a data dictionary for all the feature columns and complete loan data for all loans issued through the 2007-2015. Current loan status (Current, Late, Fully Paid) in addition to features such as loan amount, amount funded, credit scores, number of finance inquiries, address including zip code and state, and collections are provided. In all, there are 60 numerical features and 18 categorical features used to assess default risk.

Feature Information

The feature information can be broken down into the following categories:

- Profile variables: These features describe the borrower's basic information such as address, marital status, and employment title and duration.
- Financial history variables: These features describe the borrower's credit history and include annual income, number of open credit accounts, total current balances, and number of bankruptcies.
- Loan variables: These features describe the loan including the loan amount, term, interest rate, and the Lending Club loan grade.

Data was imported and the dictionary file was used to name and determine which features to keep. Features with less that 65% available data were then dropped. This

brought the number of columns down from 145 to 63 columns for the 2,260,668 entries. Features containing a single unique value were also dropped.

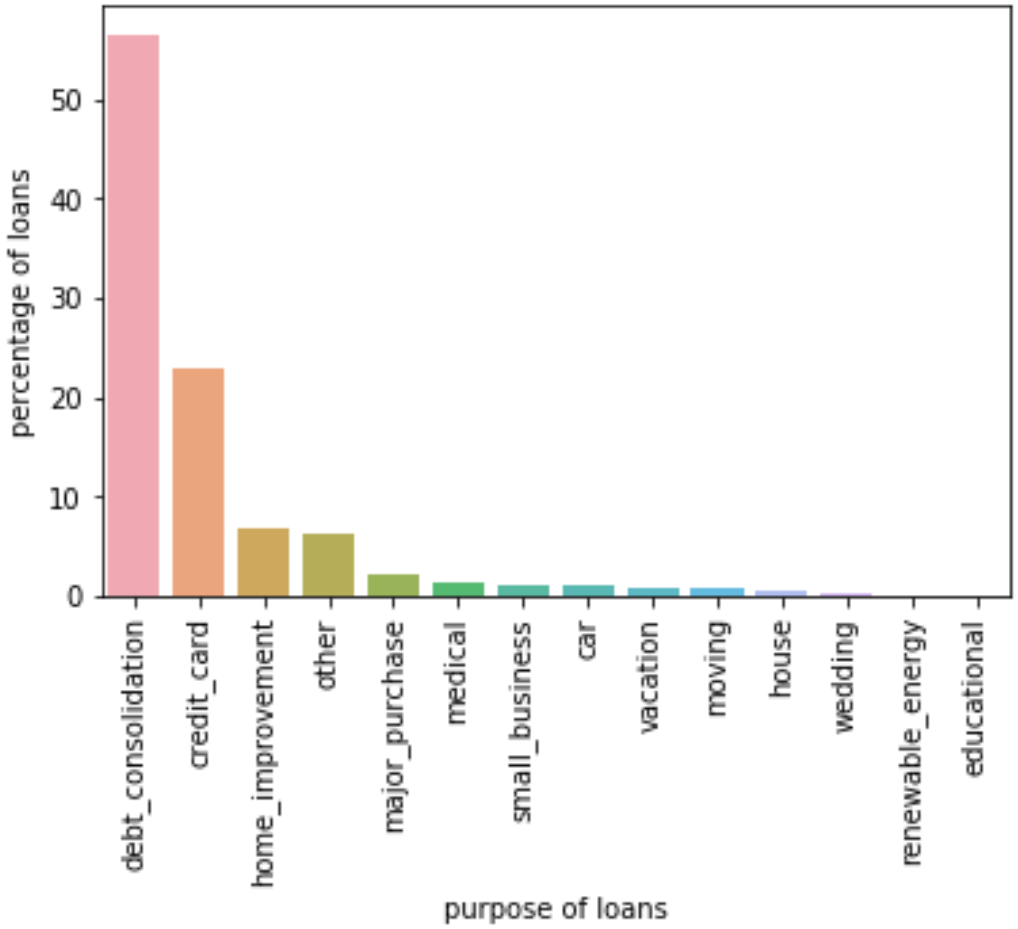
For our model, we're predicting default rate so we need only examine completed loans - those with the loan status of ‘charged off’ or ‘fully paid’. Loans that were current, late, or in grace periods were eliminated. Remaining loans were consolidated into either ‘charged off’ or ‘fully paid’ accordingly. The working data set was 1,306,356 entries and 61 features. Approximately 80% of the loans were fully paid and the remaining 20% charged off. The dataset was then split into train and test sets.

Data was separated into numerical and categorical parts to be further wrangled. Missing numerical data was filled with the feature's median values. Outliers were removed and data was scaled. I encoded categorical features using one-hot encoding and using a ‘Rare’ category for group frequencies of less than 1% in category. This brought the number of feature columns up to 109.

Lending Club Loan Data Analysis

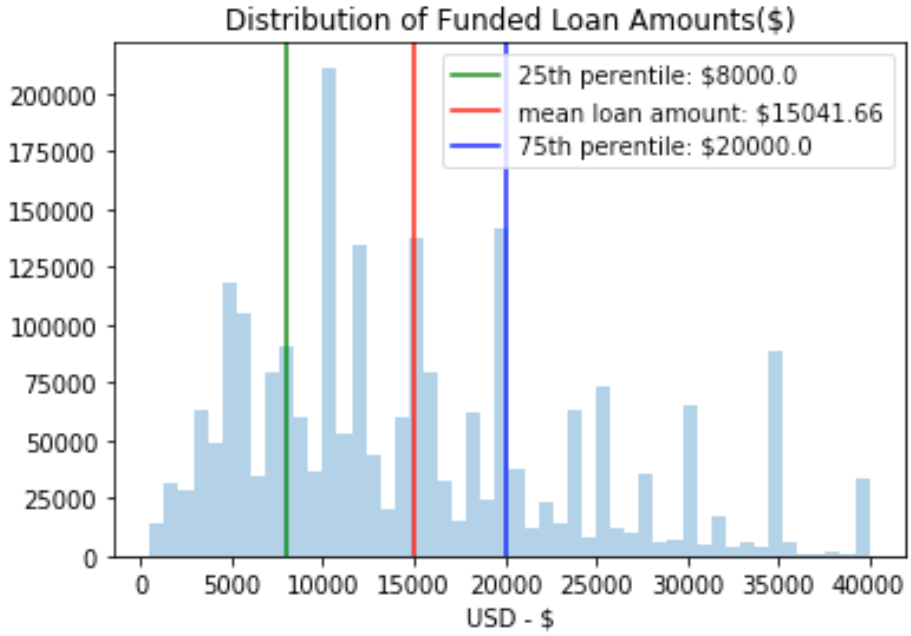
Exploratory Analysis:

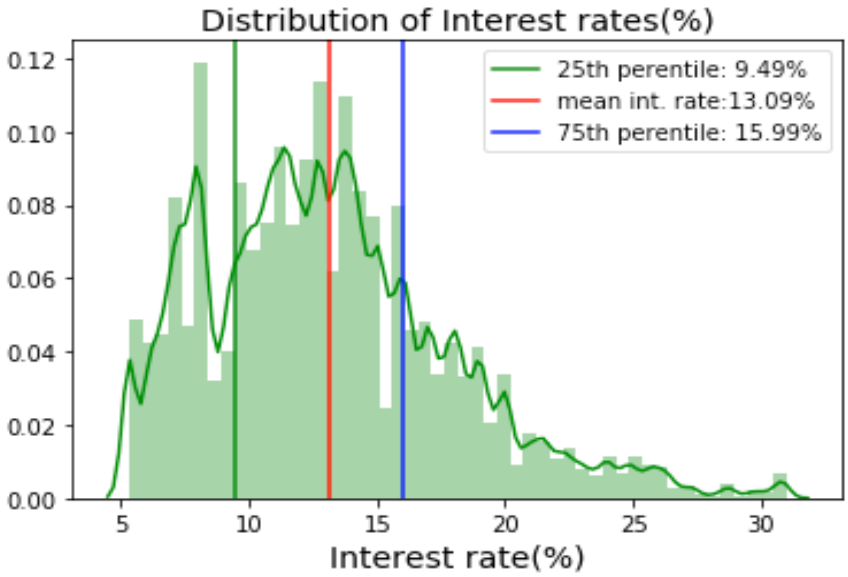
Lending club tracks what borrowers will use their loans for and debt consolidation (~60%) and paying off higher interest credit cards (~25%) were overwhelmingly the top reasons cited. Home improvement was the next most cited reason but by less than 10% of the borrowers.



Lending Club Loan Data Analysis

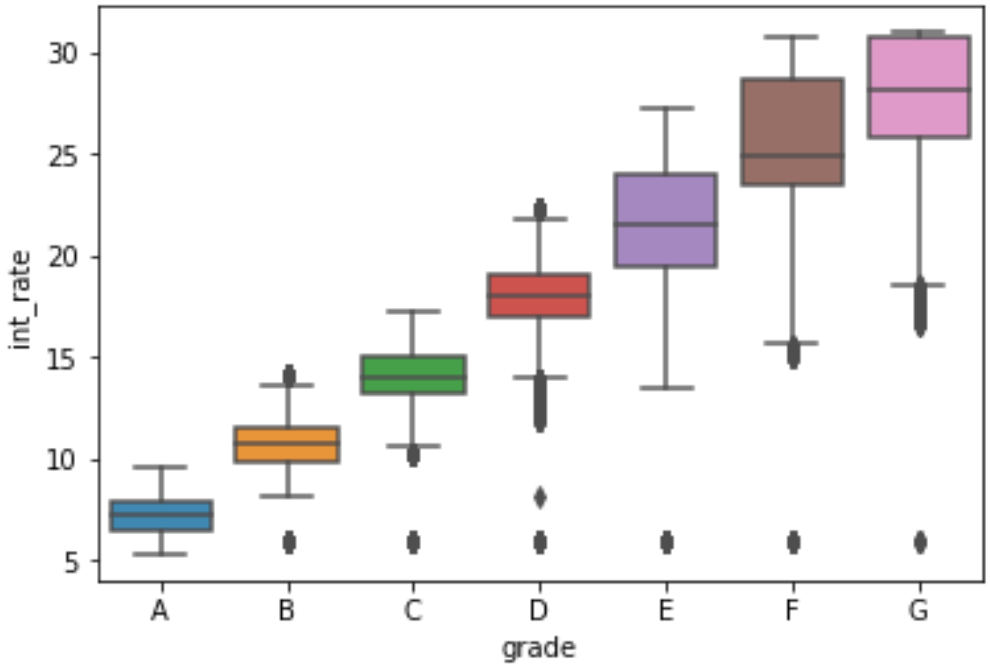
Next, I examined the amounts for the funded loans. These range from $1,000 to $40,000 skewing to the left with most loans being between $8,000 and $20,000. The median loan amount was $15,041.66. The distribution of interests rates has a similar shape skewing to the left with a range of 5.6%-35.8%. Most loans have an interest rate between 9.49% and 15.99% with a median of 13.09%.





Lending Club Loan Data Analysis

After assessing each loan, Lending Club assigns it a grade (A-G) to each loan according to its perceived credit risk. For example, if the borrower has a weaker credit profile they would get a lower grade, and if the borrower has a stronger credit history, they'd get a higher grade. As one would expect, the interest rates of the loan directly correspond to the Lending Club loan grade.



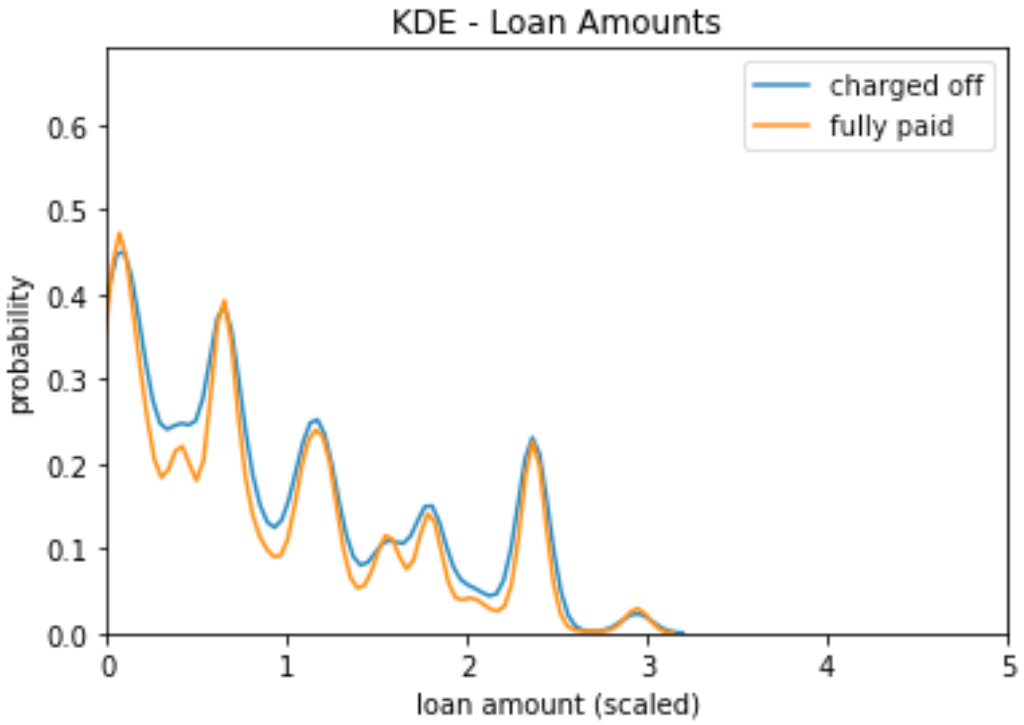
Lending Club Loan Data Analysis

Statistical Analysis:

Data was subset into two separate dataframes - ‘charged_off’ and ‘fully_paid’ - to analyse any statistical difference occurring between the groups for certain features.

I first examined borrowers’ loan amounts (the total amount funded for each borrower). The scaled means for the charged off and fully paid accounts are 0.1302069983214234 and -0.03186664773276771, respectively. The scaled variances are 1.0239344377642705 and 0.9900112342208185. Using the stats.ttest from the scipy library, I obtained a t-value of 73.5915 and p = 0. The high t-value indicates that there is a difference between the charged_off and fully_paid accounts. This is reinforced by p=0 (i.e., not by chance).

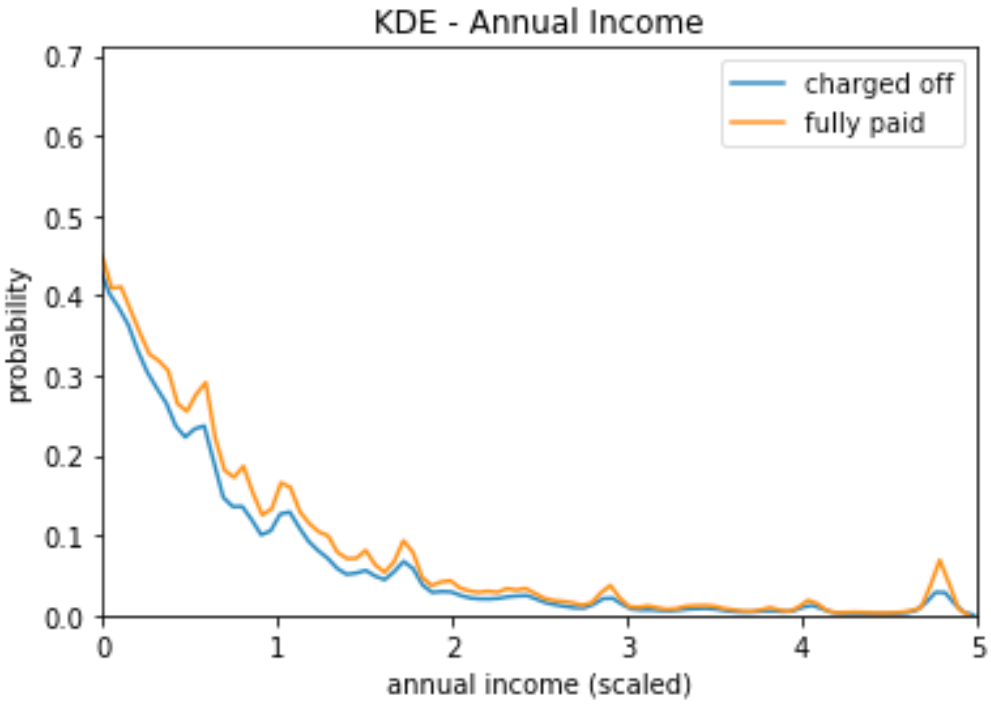
Bootstrap hypothesis testing using the mean as the test statistic produced a p-value of zero furthering this finding. The 95% confidence interval for the difference between replicant charged off and fully paid accounts is: [0.15773027, 0.1664066]. We reject the null hypothesis. There is a statistical difference in loan amount sought between charged_off and fully_paid loans.



Lending Club Loan Data Analysis

The second feature I examined was the annual income of the borrowers. The scaled means for the charged off and fully paid accounts are -0.12468682384120139 and 0.03187923073861866, respectively, and the corresponding scaled variances are 0.8436454101206241 and 1.0376205908584824. A t-value of -76.3158 and p = 0 were calculated using scipy.stats.ttest indicating that there is no difference between the charged_off and fully_paid datasets.

Using the mean as the test statistic for bootstrap hypothesis testing and a 95% confidence interval of [-0.16054925 -0.15259419] for the difference between replicant charged off and fully paid accounts, bootstrap hypothesis testing produced a p-value of 1.0. For α = 0.05, we keep the null hypothesis: There is no statistical difference in the annual incomes between charged_off and fully_paid loans.



Lending Club Loan Data Analysis

Creating Our Models:

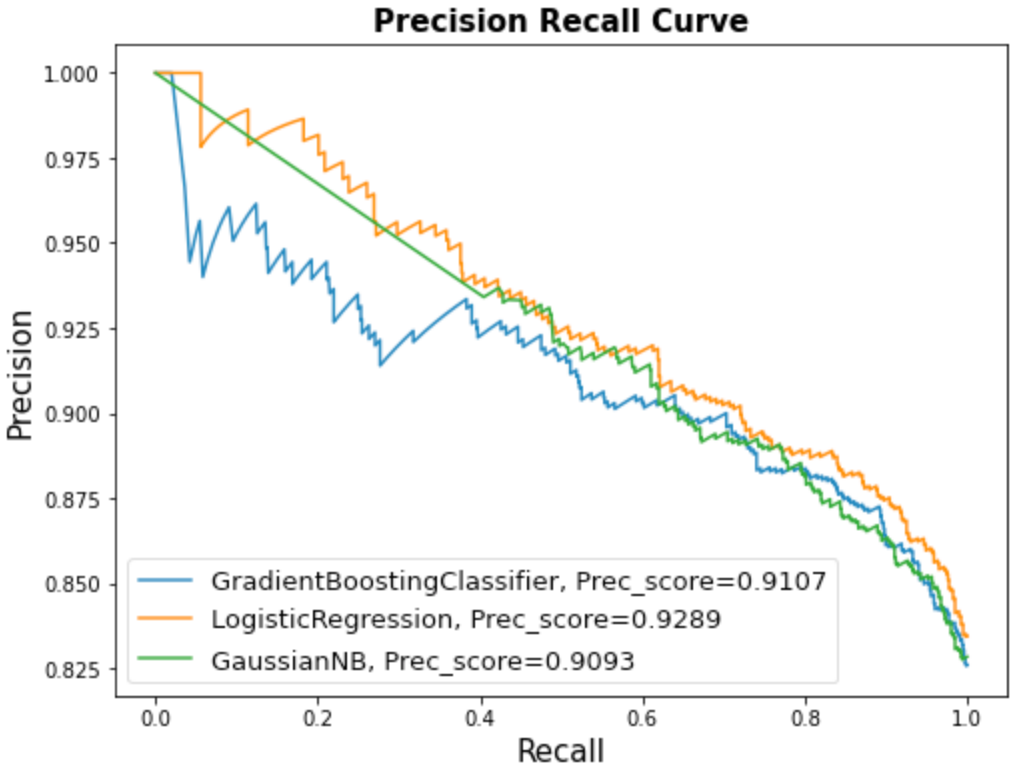
Three different classification models were applied to the data: Gradient Boosting Classifier (GBC), Logistic Regression (LR), and Guassian NB (GNB). For each, false positive rate (fpr), true positive rate (tpr), area under curve (auc), precision, recall, and precision score were calculated.

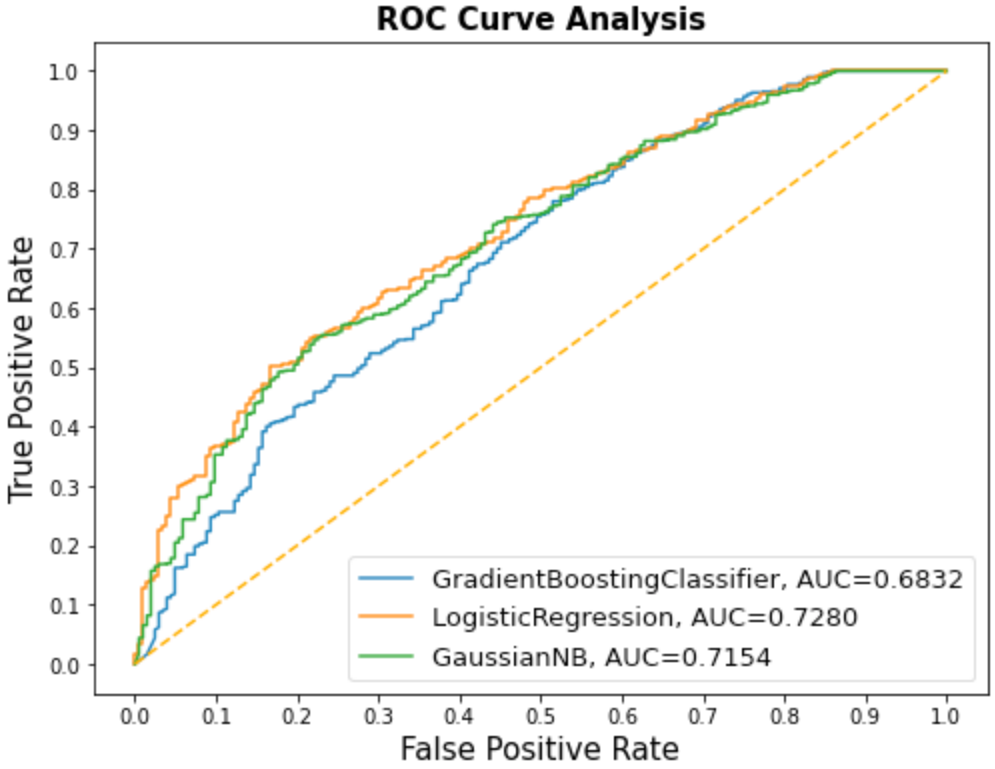
classifiers	fpr	tpr	auc	precision	recall	precision score
GradientBoostingClassifier	[0.0, 0.014705882352941176, 0.0245098039215686...	[0.0, 0.016331658291457288, 0.0439698492462311...	0.683238	[0.819773429454171, 0.8195876288659794, 0.8204...	[1.0, 0.9987437185929648, 0.9987437185929648, ...	0.874181
LogisticRegression	[0.0, 0.0, 0.0, 0.004901960784313725, 0.004901...	[0.0, 0.001256281407035176, 0.0163316582914572...	0.727978	[0.819773429454171, 0.8195876288659794, 0.8204...	[1.0, 0.9987437185929648, 0.9987437185929648, ...	0.907232
GaussianNB	[0.0, 0.004901960784313725, 0.0049019607843137...	[0.0, 0.017587939698492462, 0.0439698492462311...	0.715360	[0.8189300411522634, 0.8187435633367662, 0.818...	[1.0, 0.9987437185929648, 0.9974874371859297, ...	0.898822

All three models have high precision scores ranging between 0.874181(GBC) and 0.907232 (LR). The AUC for the three classifiers are in acceptable range ~0.7. Again, the Logistic Regression model scores the best at 0.714913. For further analysis of the performance of the models, the precision-recall and ROC curves were computed.

For all three models, there is a high area under the precision-recall curve representing both high recall and high precision (where high precision relates to a low false positive rate, and high recall relates to a low false negative rate). For the ROC curve representing a relation between sensitivity (recall) and specificity (not precision), all three of these models fall within an acceptable range.

Lending Club Loan Data Analysis





Lending Club Loan Data Analysis

Conclusions

The best performing classifier is the Logistic Regression. However, exploration of this dataset is preliminary. With the large number of features available for evaluation, future work should focus on evaluating which features provide the most insight and have the largest effect on outcome. Next steps would be to incorporate feature importance and feature interdependence in refining the models.