Janis Navarro

FINAL REPORT:
MINIMIZING RISK DURING LOAN PROCESS

Problem Statement

LendingClub is a peer-to-peer lending company, headquartered in San Franciso, California that offers loan trading on a secondary market. It is the world's largest peer-to-peer lending platform. Every time LendingClub receives a loan application, they must make a decision for loan approval based on the applicant's financial profile.

Using risk analytics in loan analysis to understand how data is used to minimize the risk of losing money while continuing to lend.

Can We Use Data to Minimize Risk and make changes to credit risk lending during the loan application process?

Data Wrangling

Dataset

The raw dataset came from LendingClub contains 5 rows with 27 columns and a total of 396030 entries. I started by looking at the dataset and familiarizing myself with the information for each row and column. I used the provided Words of the Wise LendingClub showing the description of each column.

With only 27 columns, immediately reducing dimensionality was not a critical step. As I worked through EDA, I considered each column and its data and the relation to other data. During data preprocessing, emp_title was removed as the column contained 173105 unique values. Emp_length and charge off rate were found to be very similar regardless of the number of years in the job title. That column was dropped as well.
During EDA purpose and title seemed to contain the same data. This was looked at further and determined they indeed had the same data. Therefore, the column named title was dropped.
In review of mort_acc, it showed that it contained 37795 missing values. It was filled in using fillna approach.

The dataset required categorical variables which changed term using unique from 36 and 60 months to 36, 60.

Based on results from exploring Grade and Sub_grade, I dropped Grade and kept Sub_grade. Sub_grade was kept because it contained the Grade information as well as the sub_grade breakdown which could be more useful when looking for data insights.

Feature Engineered zip code from the address by creating a column called zip_code that included the zip codes from the addresses. This could also help when targeting a certain zip code for data insights.

The dataset also included a column called issue_d which was defined as the month which the loan was funded. Dropped this feature from the dataset.

The final shape of my dataset was 393465 entries with 22 columns.

Exploratory Data Analysis

LendingClub's Words of the Wise description of each column served for definition and understanding of data in my analysis. The first reviewed current status of the loans. Current status provided how many were fully paid versus charged off loans.

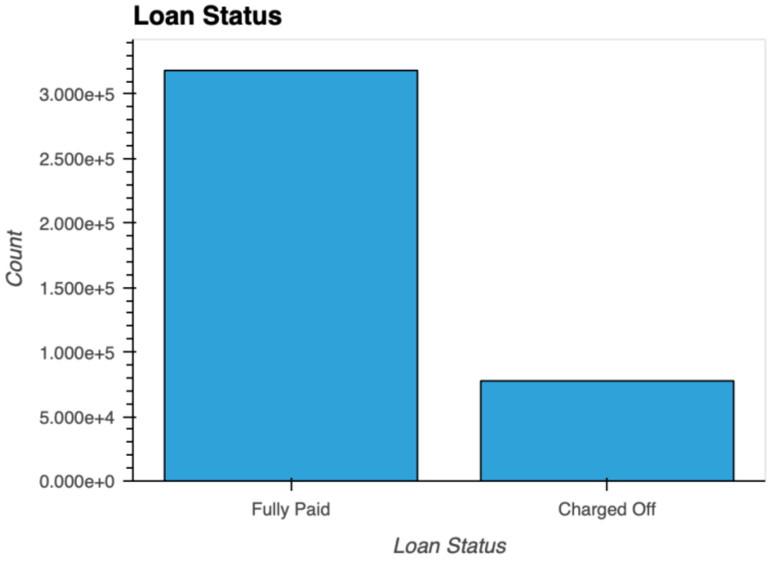
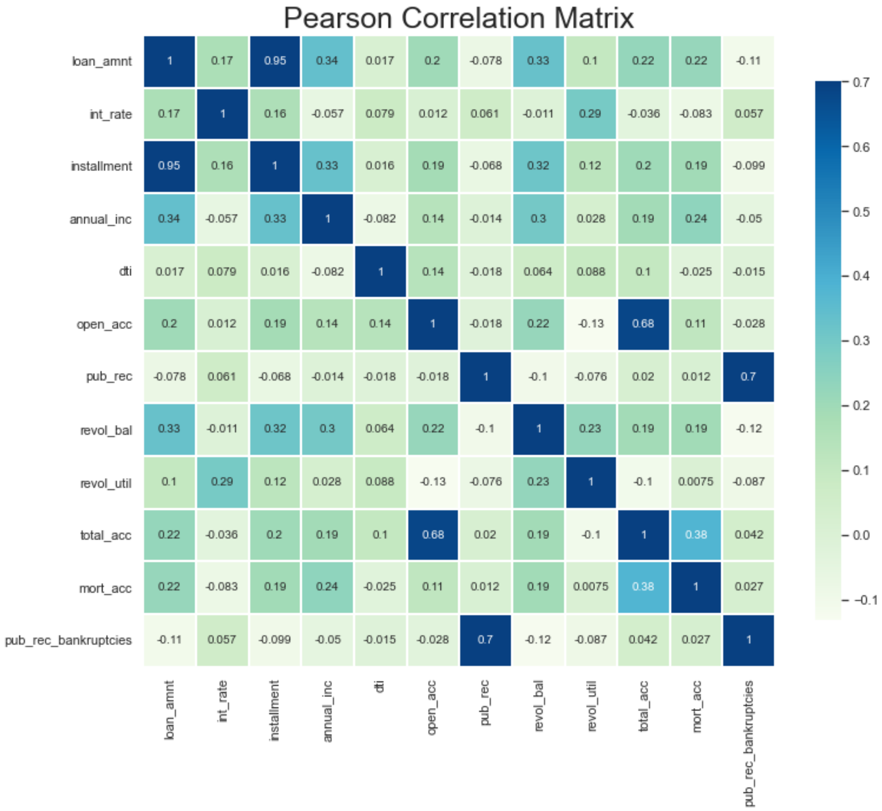


Figure 1: Bar Plot of Loan Status Fully Paid vs. Charged Off Loans

We can visualize there are less charged off than fully paid loans.

A heatmap was then completed to visualize the correlation between features.



Pearson Correlation Matrix

	loan_amnt	int_rate	installment	annual_inc	dti	open_acc	pub_rec	revol_bal	revol_util	total_acc	mort_acc	pub_rec_bankruptcies
loan_amnt	1	0.17	0.95	0.34	0.017	0.2	-0.078	0.33	0.1	0.22	0.22	-0.11
int_rate	0.17	1	0.16	-0.057	0.079	0.012	0.061	-0.011	0.29	-0.036	-0.083	0.057
installment	0.95	0.16	1	0.33	0.016	0.19	-0.068	0.32	0.12	0.2	0.19	-0.099
annual_inc	0.34	-0.057	0.33	1	-0.082	0.14	-0.014	0.3	0.028	0.19	0.24	-0.05
dti	0.017	0.079	0.016	-0.082	1	0.14	-0.018	0.064	0.088	0.1	-0.025	-0.015
open_acc	0.2	0.012	0.19	0.14	0.14	1	-0.018	0.22	-0.13	0.68	0.11	-0.028
pub_rec	-0.078	0.061	-0.068	-0.014	-0.018	-0.018	1	-0.1	-0.076	0.02	0.012	0.7
revol_bal	0.33	-0.011	0.32	0.3	0.064	0.22	-0.1	1	0.23	0.19	0.19	-0.12
revol_util	0.1	0.29	0.12	0.028	0.088	-0.13	-0.076	0.23	1	-0.1	0.0075	-0.087
total_acc	0.22	-0.036	0.2	0.19	0.1	0.68	0.02	0.19	-0.1	1	0.38	0.042
mort_acc	0.22	-0.083	0.19	0.24	-0.025	0.11	0.012	0.19	0.0075	0.38	1	0.027
pub_rec_bankruptcies	-0.11	0.057	-0.099	-0.05	-0.015	-0.028	0.7	-0.12	-0.087	0.042	0.027	1

Figure 2: Heatmap for Correlation of Features

The heatmap corroborated there is a great correlation between Loan Amount and Installment.

Loan amount is described as loan_amnt= the listed amount of the loan applied for by the borrower.

Installment is described as installment= the monthly payment owed by the borrower if the loan originates.

In-Depth Analysis

Now that I know the information my data holds and understand the relationships between the different variables, I needed to determine further the relationship between loan amount and installment.

I wanted to investigate the amount of the loan, the length of the loan for correlation and into how many are fully paid versus charged off loans. As well as start looking to get a sense of their distribution.

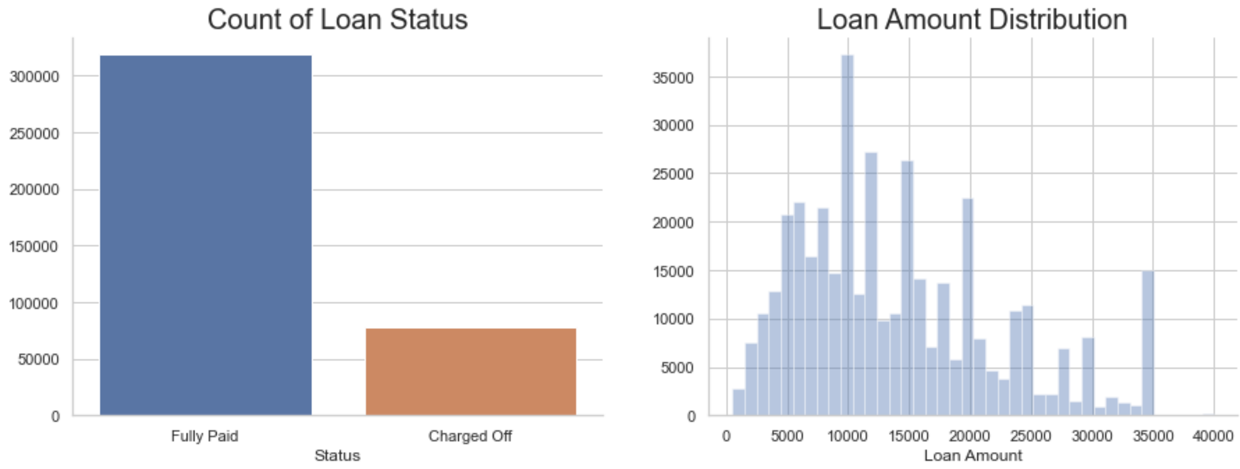


Figure 3: Bar Plot Count Loan Status and Loan Amount Distribution

Both of these bar plots demonstrate the distribution of each amongst fully paid versus charged off loans. It demonstrate that the charged off loans are greater between approximately for loans in the amount of $7,000.00 or less.

There may be additional features that correlate and help find a pattern with charged off loans within other variables that can help to lower the loan risk.

A scatterplot view between Loan Amount and Installment was hard to make comparison or extract deep insights from it. However, a box plot provided a better distribution view.

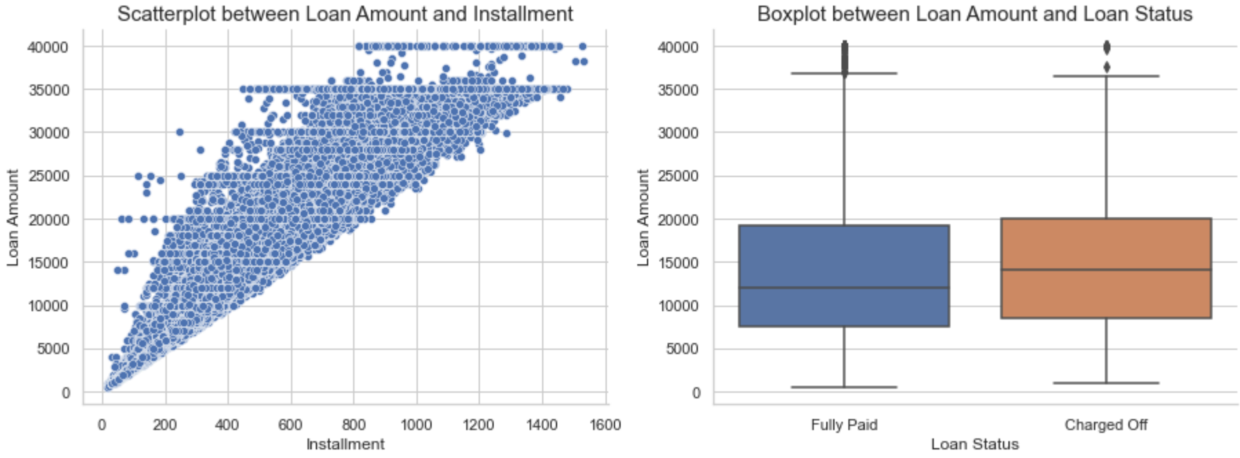
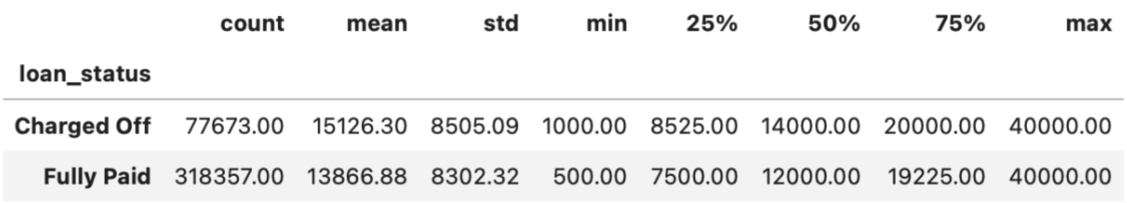


Figure 4: Scatterplot between Loan Amount and Installment & Boxplot between Loan Amount and Loan Status

I performed a groupby to obtain statistics on loan status and loan amount.



loan_status	count	mean	std	min	25%	50%	75%	max
Charged Off	77673.00	15126.30	8505.09	1000.00	8525.00	14000.00	20000.00	40000.00
Fully Paid	318357.00	13866.88	8302.32	500.00	7500.00	12000.00	19225.00	40000.00

Figure 5: Statistics on Loan Status and Loan Amount\

A deeper look at Grade and Subgrade and visualization of them against Fully Paid and Charged off Loans brought to light that Subgrade contains the same data as Grade. However, subgrade has more in-depth classification.

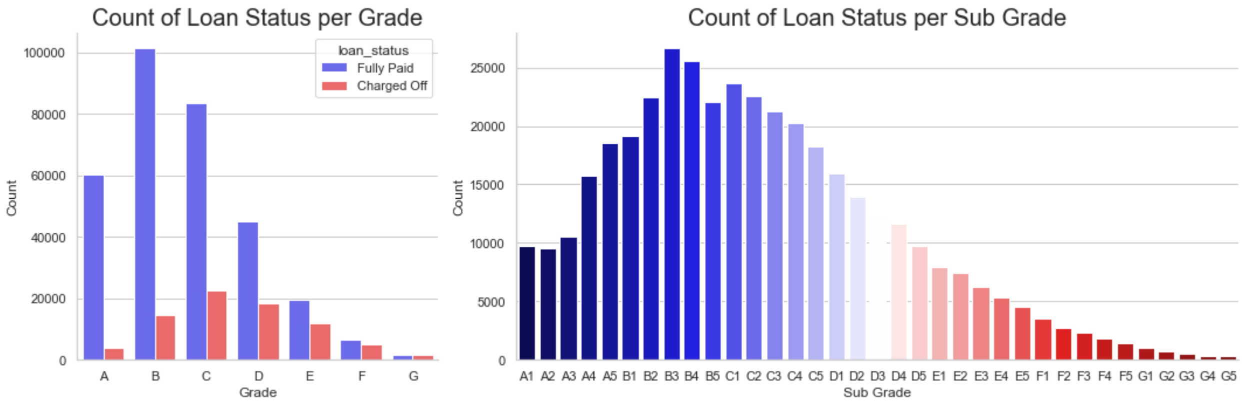


Figure 4: Bar Plot of Loan Status by Grade, Subgrade and Distribution

The bar plot provides a view of Loan Status, fully paid and charged off loans corresponding to a Grade. Grades and subgrades were provided by LendingClub.

Figure 4 allows us to see the distribution for Grade F & Grade G and all its corresponding subgrades do not get paid back as often as agreed. We should consider treating these grades as higher risk.

Looking into Grades F and G and its corresponding subgrades isolated from other Grades and subgrades.

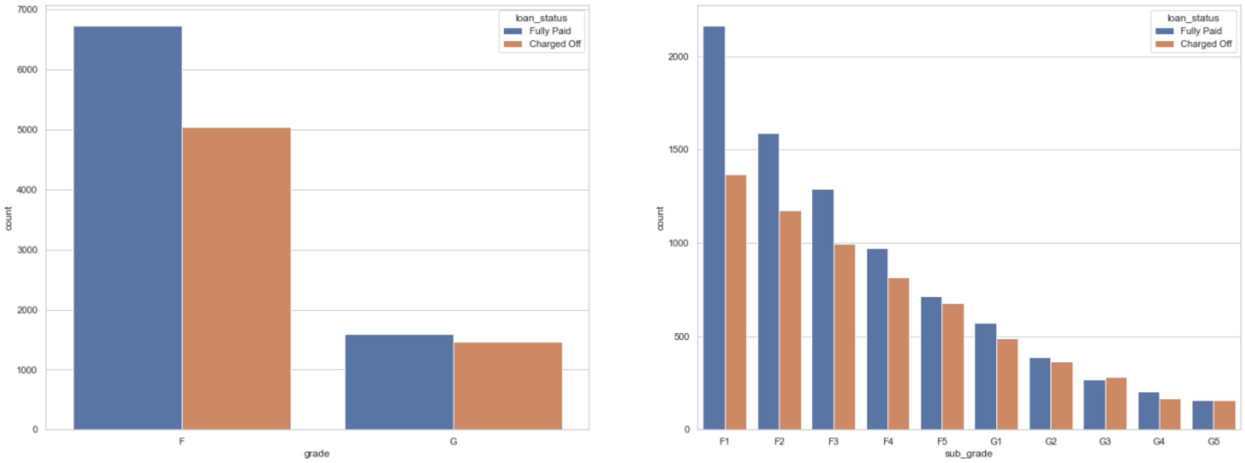


Figure 5: Grades F & G and its subgrades isolated other Grades and Subgrades

Figure 5 allows us to visualize the distribution in more depth for Grades F and G for its loan status of fully paid to charged off loans. Grade G demonstrates an almost even distribution of fully paid to charged off loans.

I then looked at Title, this is defined as the loan title provided by the borrower/applicant. I changed to all title to lowercase and pulled a value count.

I did the same for Purpose, which is defined as a category provided by the borrower for the loan request. Changed purpose to lowercase and pulled a value count.

This demonstrated that Title and Purpose included the same data. I used purpose going forward.

Following, I visualized loan status by home ownership to determine if homeowners are

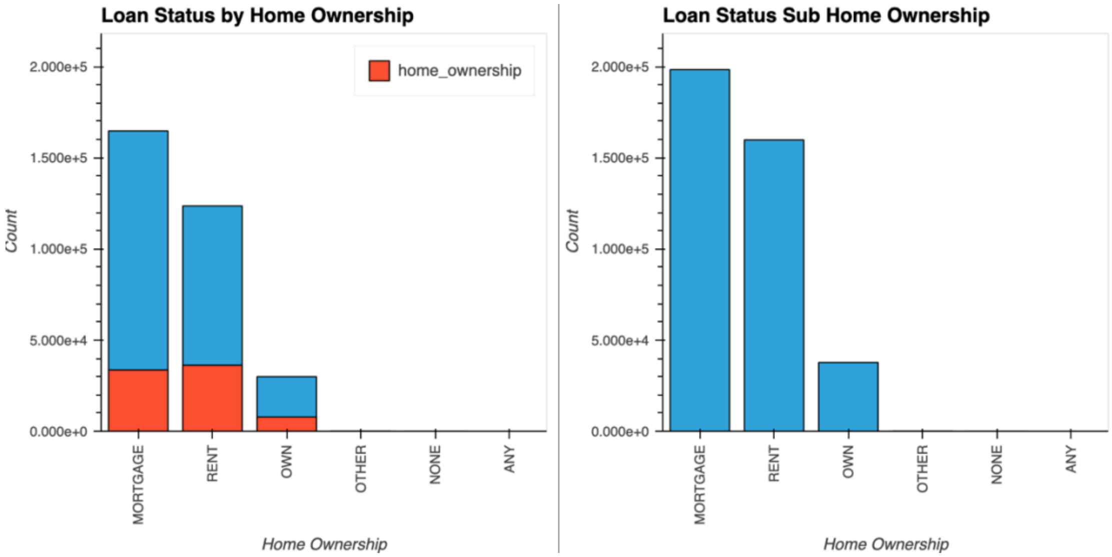


Figure 6: Loan Status by Home Ownership and Sub Home Ownership

The visualization shows us there are a lower number of homeowners to renters. Most loans are from applicants that do not own a home.

I wanted to see the relationship of loan status to term. Term is defined by LendingClub as the number of payments on the loan. Values are in months and can be either 36 or 60.

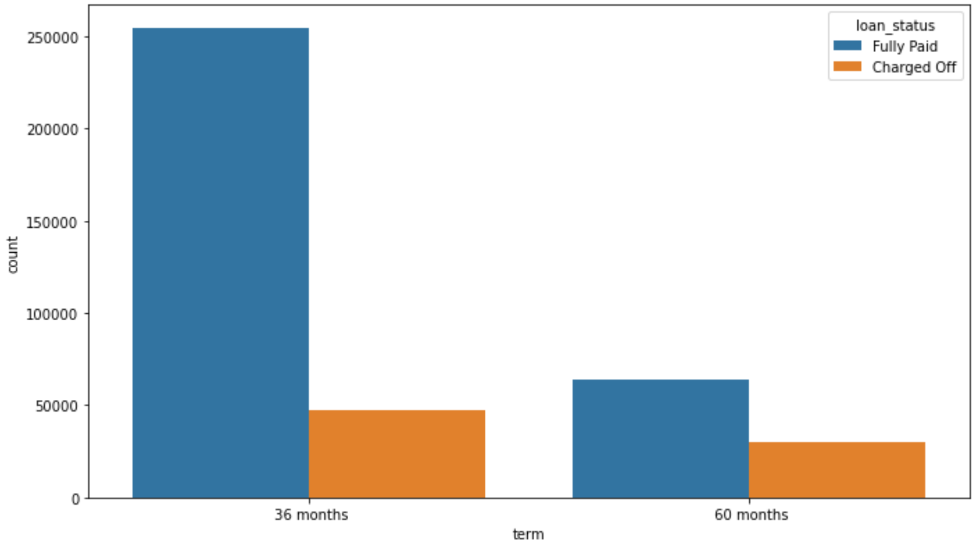


Figure 7: Loan Status to Term 36 or 60 months

Figure 7 provides us the insight that 36 months term loans have a higher distribution to be fully paid when compared to 60 month term loans.

We can corroborate that 60 term loans and smaller dollar approximately $5,000.00 or less are higher risk of being charged off.

Verification Status as defined by LendingClub indicates if income was verified by LendingClub, not verified, or if the income source was verified.

I wanted to visualize if there was a higher risk within verification status of the loans.

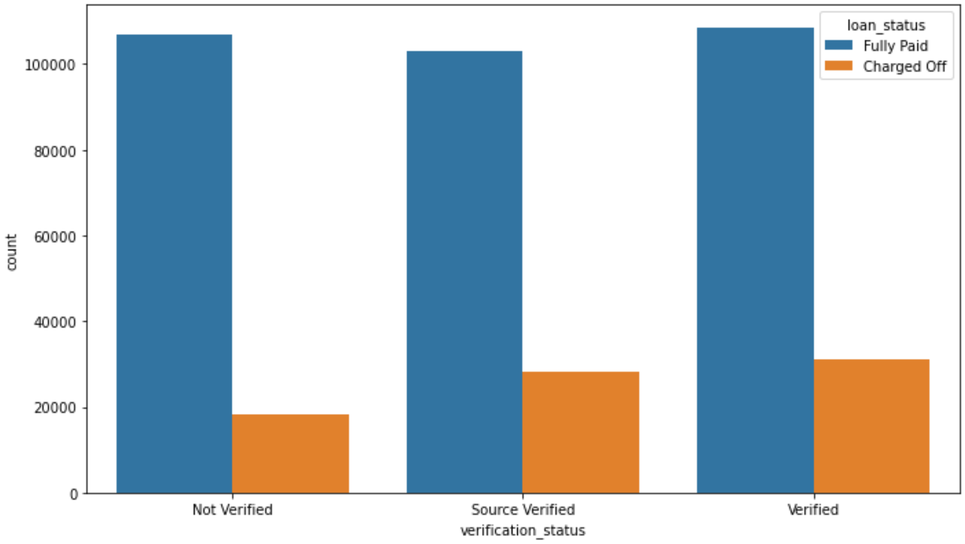


Figure 8: Verification Status to Loan Status

We see that majority of the loans are either source verified or LendingClub verified. There is a good amount of loans that are not verified and in charged off status.

Home Ownership included a category of other. I explored that further by performing a value count of loan status and deemed majority are fully paid.

Out of purpose for the loan, I explored educational and renewable energy and amounts returned were dismal.

I next explored Loan Status by interest rate and annual income.

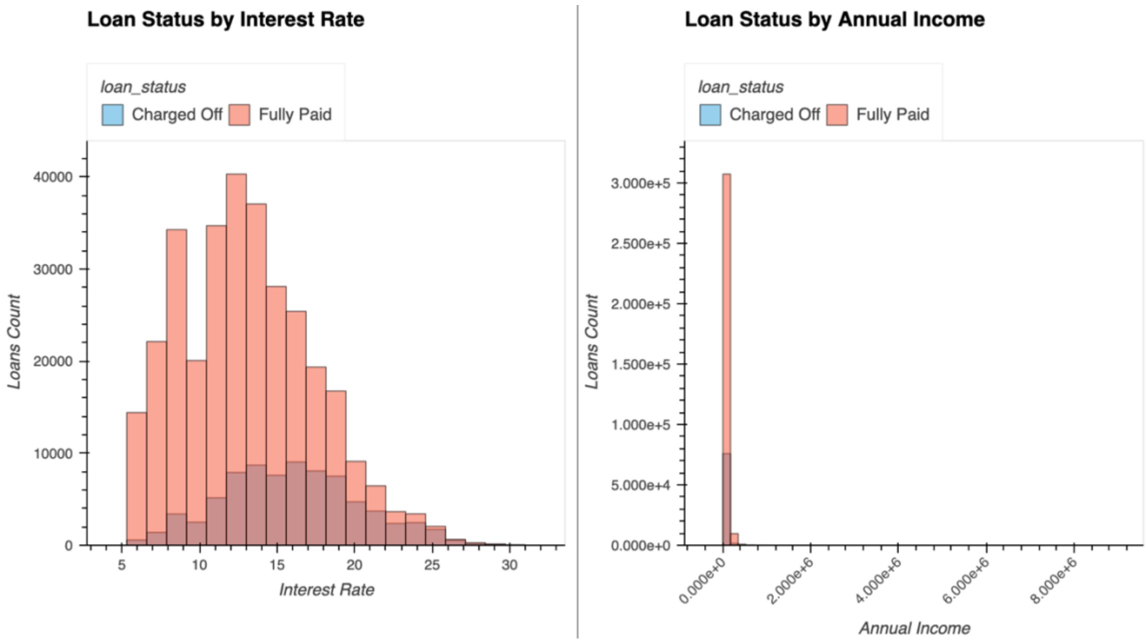


Figure 9: Loan Status by Interest Rate and Annual Income

Based on this visualization we can gather that the higher the interest rate, the more likely the loan will not be paid and will be charged off.

Exploring annual income further to see if there is a pattern within a set amount of income.

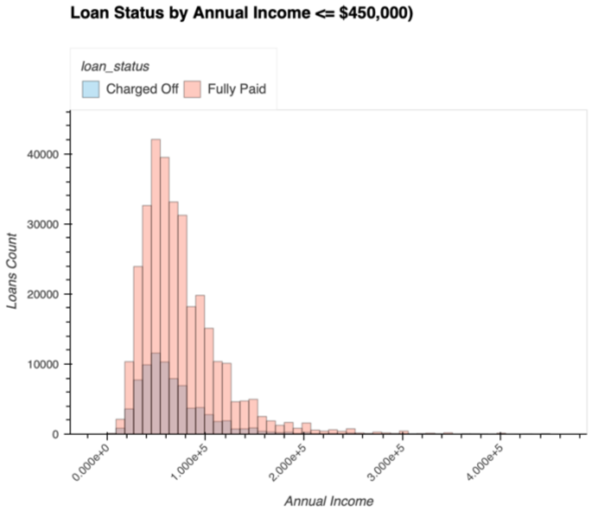


Figure 10: Loan Status by Annual Income <= $450,000

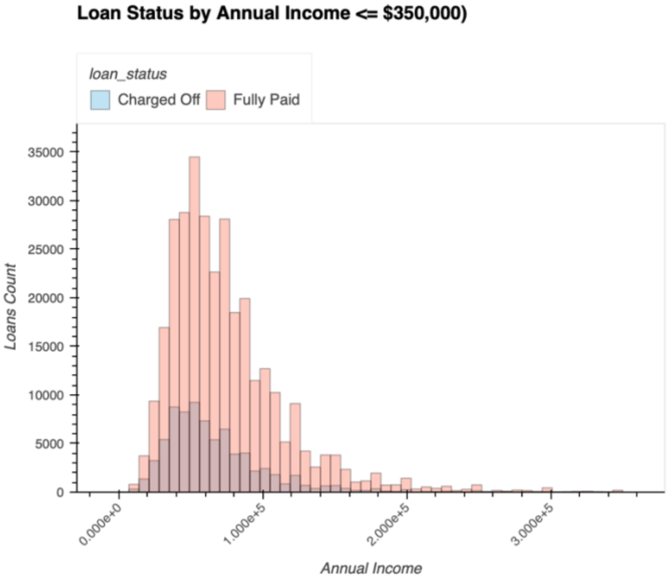


Figure 10: Loan Status by Annual Income <= $450,000

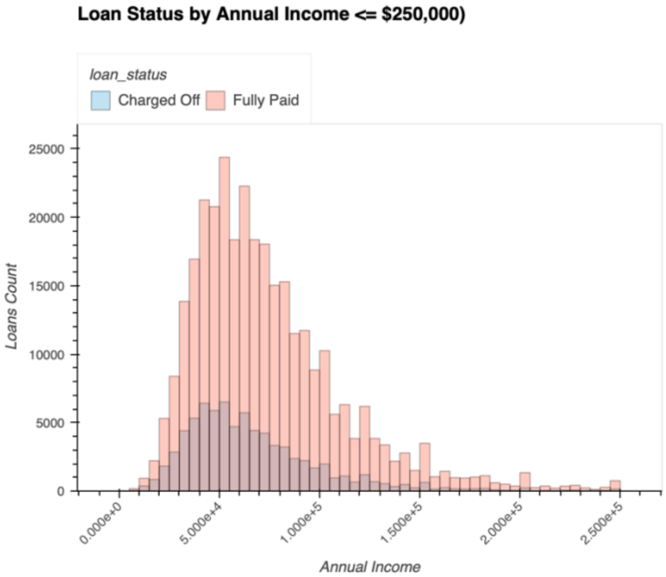


Figure 11: Loan Status by Annual Income <= $250,000

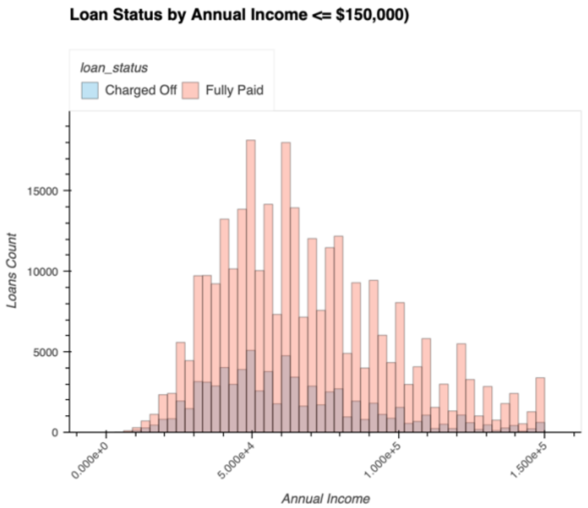


Figure 12: Loan Status by Annual Income <= $150,000

In visualizing income we can see that at the Annual income of:

Greater than or equal to 1 Million, there are only 75 borrowers
Greater than or equal to $250,000, there are 4077 borrowers

I also explored employment title and employment length to look if there are titles or length of employee that are a lower risk.

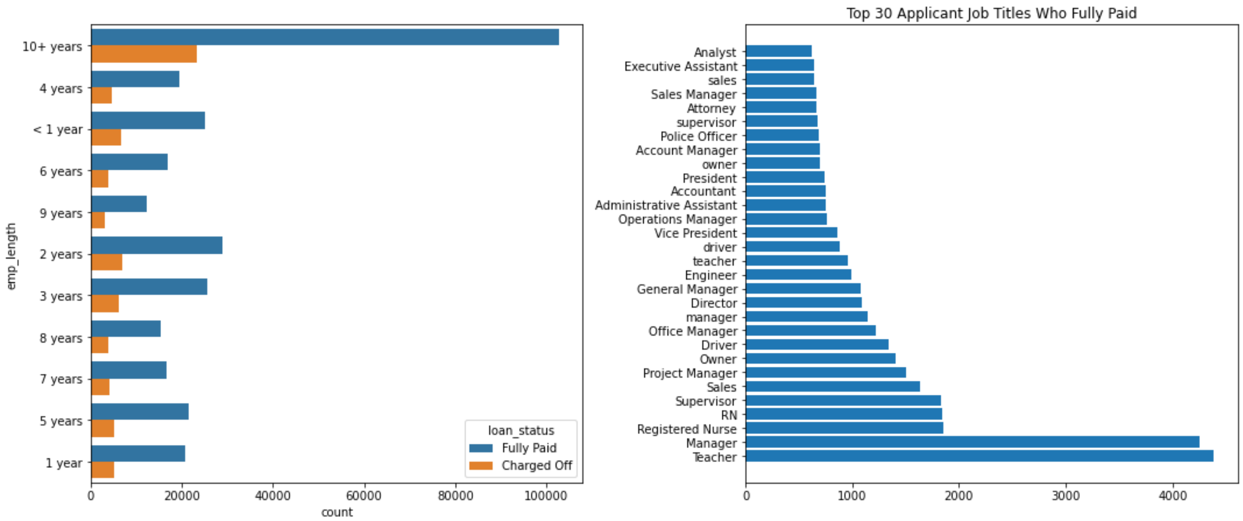


Figure 13: Loan Status by Employment Length and Top 30 Applicant Job Titles Fully Paid

We can visualize that the longer the employment length the lower the risk. I also compiled a list of the top 30 applicant job titles that can be considered a lower risk. Top 30 applicant job titles show as the most fully paid were Teacher and Manager.

I next wanted to explore and visualize issued date and earliest credit line and its relation to Loan status for fully paid and charged off loans.

LendingClub defined issued date as the month the loan was funded.
Earliest credit line is defined as the month the borrower's earliest reported credit line was reported.
I first used to datetime method to convert string Dates time into date time objects.

I used a bar hist to visualize the distribution and impact of fully paid versus charged off based on issue date and earliest credit line.

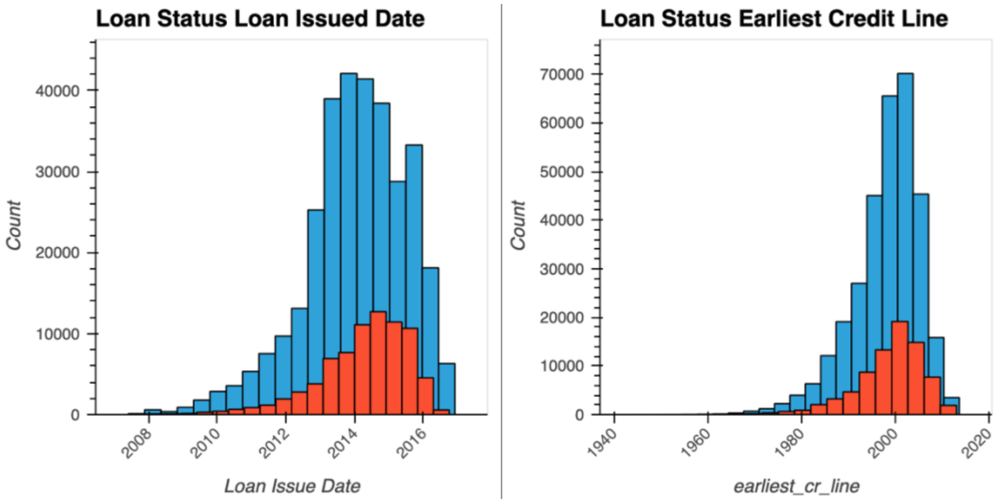


Figure 14: Loan Status Loan Issued Date and Earliest Credit Line

From this visualization we can gather there was a spike in charged off accounts for loans with issue date of 2012 to 2015. I would recommend LendingClub review their credit policy at the time to see if there were changes made and if lending was more lenient.

I then reviewed debt to income ratio, the number of open credit lines, total credit revolving balance, revolving line utilization rate, and the total number of credit lines.

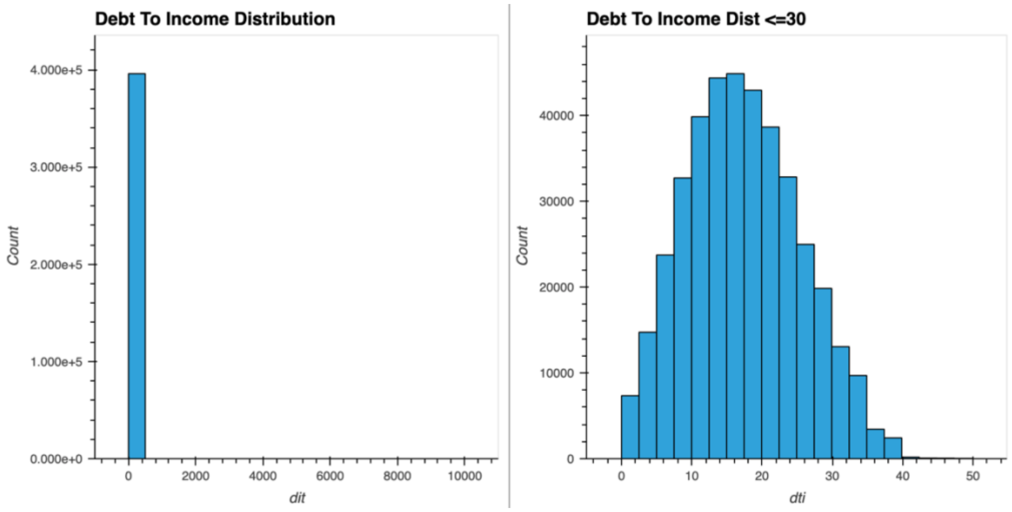
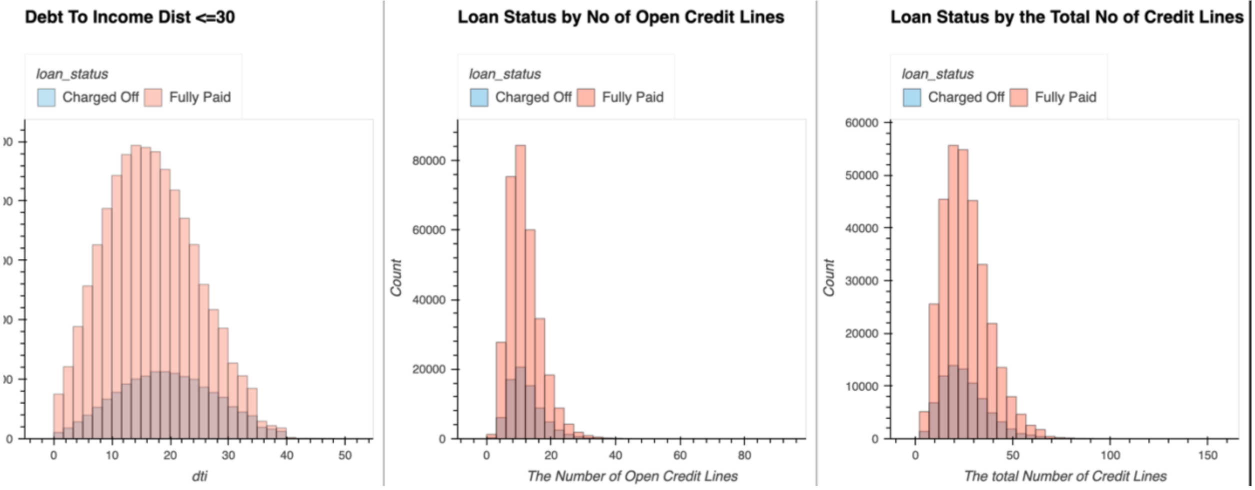


Figure 15: Debt to Income Distribution and Debt to Income Distribution less than or equal to 30

The information yielded was very general, therefore I completed a bar plot with distribution to look deeper into Debt to Income Distribution less than or equal to 30 with fully paid and charged off accounts. I did the same for loan status by number of open credit lines and by the total number of open credit lines.


Debt To Income Dist <=30

Loan Status by No of Open Credit Lines

Loan Status by the Total No of Credit Lines

Figure 16: DTI <=30, Loan Status by Number of Open Credit Lines and Loan Status by Total Number of Credit Lines

From this figure we can see that the greater the Debt to Income Ratio the more likely the borrower will not pay the loan.

I looked further into total open accounts at 40, total accounts at 80 and revolving utilization rate at 120 and learned the following:

- There are 217 borrowers that have more than 40 open credit line accounts
- There are 266 borrowers that have more than 80 open credit line accounts

As we learned earlier that the greater the debt to income, the greater the probability the borrower would not pay the loan, I wanted to review the revolving utilization rate a bit deeper.

Visualization at revolving utilization rate at over 150 yielded very small results, so I looked at under 120. This provided us a better visualized distribution and the charged off accounts increased as the number of revolving lines increased. The alarming increasing to probability of charge off began at approximately 30 and continued through approximately 90.

I then proceeded to review the total credit revolving accounts in terms of balance. This yielded a right skewed distribution when total credit revolving balance is under $250,000 as seen in Figure 17 below.

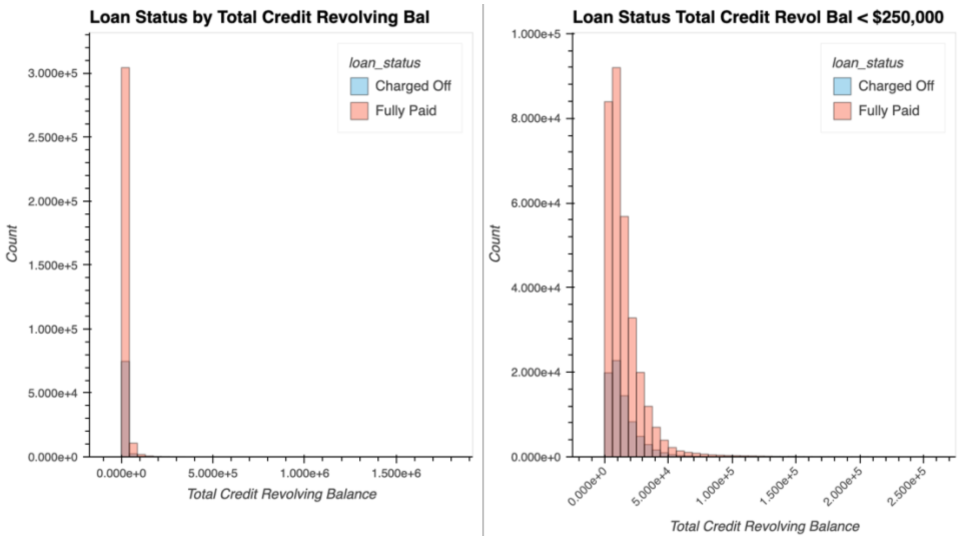


Figure 17: Loan Status by Total Credit Revolving Balance

Looking at the number of derogatory public records visualized as one can expect the greater the number of derogatories, the more likely they are to not fully pay their loan.
I wanted to look into this further and see how many were fully paid and charged off who had derogatory public records at the time of application. This provided information that both fully paid and charged off may have derogatory public records as seen in Figure 18.

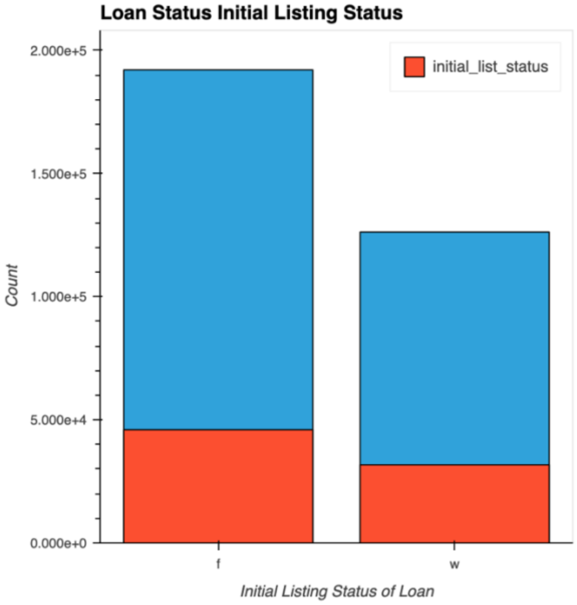


Figure 18: Loan Status Initial Listing Status of Derogatory Public Records

I looked at applicant type to possible look further into individual, joint or direct pay. However, all data yielded individual loan applications.

When comparing number of public records to Loan Status, the visualization provided that the higher the number of public records, the more likely to be charged off.

A look at the number of mortgage accounts by Loan Status, visually provided that the higher the number of mortgage accounts, the more likely the applicant will be fully paid.

A dive into number of public record bankruptcies also visually provided that the greater the number of public record bankruptcies an applicant had, the greater the likelihood the applicant to be charged off.

To understand the relationship all of these features and loan status I wanted a more in depth view. I completed a Correlation to grasp this further.

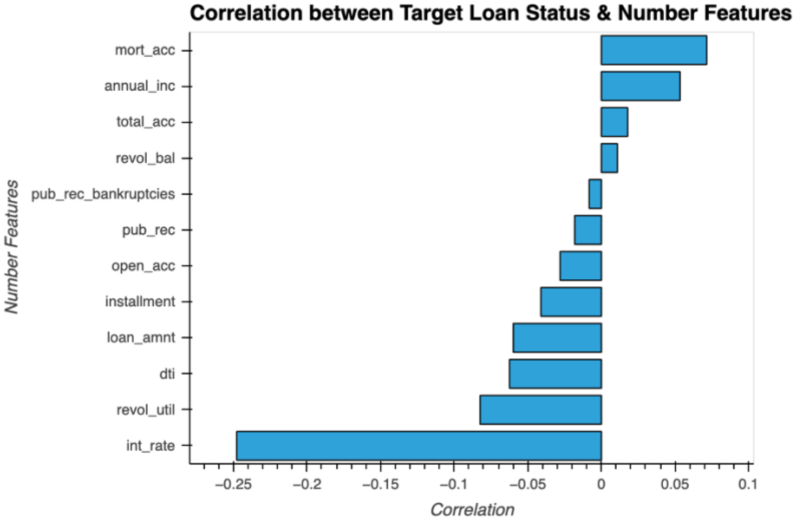


Figure 19: Correlation between Loan Status and Number Features

Figure 19 confirmed there are two categories for features.
Category 1- Features Correlated to Loan Characteristics.
As an example:

- Loan Amount
- Interest Rate
- Revolving Utilization Rate
- Debt to Income Ratio
- Installment
- Open accounts
- Public records

Category 2- Features Correlated to the Applicant.
As an example:

- Number of Mortgage Accounts
- Income
- Occupation
- Employment Title

Data Pre-Processing

I proceeded to data preprocessing by looking for missing values within my features. I looked into employment length and found that employment length and charge off rate is very similar regardless of the number of years in the Job Title. Therefore, I dropped employment length.
During EDA, purpose and title seemed to have similar data. I looked at these two features further and found they indeed had the same data. I dropped Title.

This was the same for Grade and Subgrade. I dropped Grade.

I pulled further details on mortage account and found it had 37795 missing values. I filled it in using the fillna approach.

I also looked into purpose, revolving utilization and public record bankruptcies for missing values and found missing values. I calculated percent of number of missing values, and as amounts were low, I dropped those with missing values.

I proceed to look at my categorical variables. Changed the term from 36 months, 60 months to integers of 36 and 60 respectively.

I also converted the year in date from earliest credit line and converted it to a numeric feature.

Model Selection

I tested 3 different machine learning classification models: Logistic Regression, XG Boost, and Random Forest Classifier. The metric I focused on when building my model was on precision. I wanted my model to predict areas with high charge off.

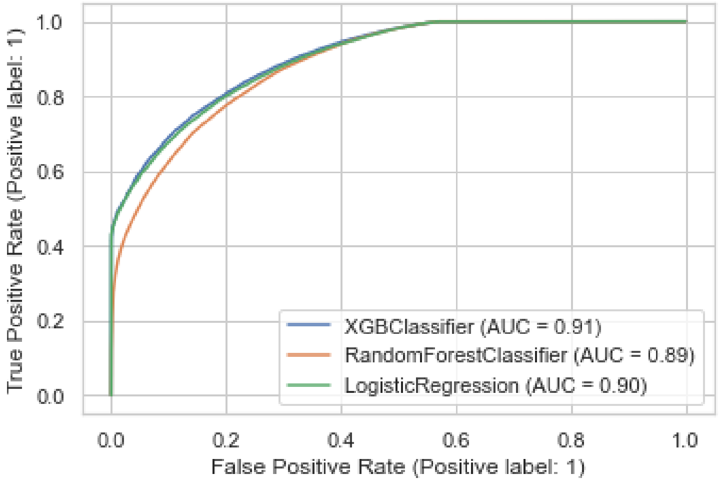
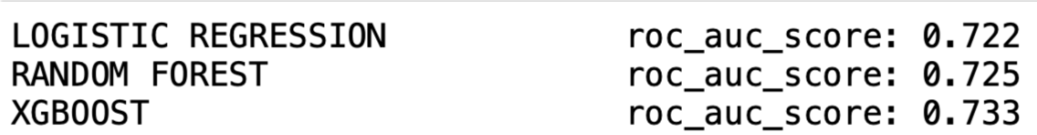


Figure 20: Precision Curve for Models

When it came to the models, the performance was very similar for all three. I completed a Performance Model Comparison of all three to and yielded:



```
LOGISTIC REGRESSION               roc_auc_score: 0.722
RANDOM FOREST                     roc_auc_score: 0.725
XGBOOST                           roc_auc_score: 0.733
```

Figure 20: Comparing Model Performance

XGBoost had the best performance while Logistic Regression had the worst.

Takeaways

XGBoost is the best model. It gave the highest precision which is highly important. False positives can cost LendingClub losses in approving loans to higher risk individual applicants.

The features with the highest correlation to fully paid should be the target for LendingClub to use as their set parameters within their lending policy. And although a lot of them are already in place as known features, we can use the target parameters that the data provides to fine tune the existing policy.

I would like to have expand or create separate models for data with other type of applicants. This dataset was limited to individual applicants. It would be interesting to obtain a joint applicant dataset and compare if the models to be used are the same or different, which yields the highest accuracy and of course what features are the most in correlation to loan status of fully paid versus charged off loans. This dataset also looked at data as far back as 2007. It would also be interesting to see if going back farther historically how the larger data impacts the models.

Future Research

This project gave me a better understanding and lots to think of in terms of the complexity of the lending loan process. While the models can be useful, it is difficult to gauge how useful they can be left at this point within the process. I would like to expand these models to include deep machine learning. I was only able to take this project to this level of modeling as this is applying the lessons the course has taken me now. In the future, given the opportunity I would like to apply AI, machine learning, and deep neural networks which I believe will help towards improving credit decisioning.