Project 3 Summary:
Predicting Credit-Worthy Consumer Loans

Linda Ju

Domain

Motivation
Peer-to-peer lending ('P2P') is a relatively new form of financing in which individuals loan money directly to other individuals. Since its inception in 2005, the P2P industry has grown to an over $3 billion industry. The most popular P2P platforms include Prosper, Lending Club, and Upstart with Prosper being the first ever peer-to-peer lending marketplace in the U.S. I have an account with Prosper and I'd like to determine which loans are credit-worthy, as in, will pay in-full and on-time.

How does P2P work?
To take out a loan, a borrower typically fills out an online application describing how much they want to borrow (up to $40,000 on Prosper), for how long, (3 or 5 years), and for what purpose. Additionally, the lending platform will often perform a credit check. During the loan funding stage, investors can review the listing information such as a borrower's interest rate, amount, FICO range, number of past delinquencies, debt-to-income ratio, employment status, and income to decide which loans they want to fund (or invest in). Investments can be made in small increments, as low as $25, so investors can diversify their investments and put their money in many loan listings. A loan is considered funded once investors fund "enough" of the requested amount, as defined by the lending platform (70% for Prosper).

When a loan is fully funded, the borrower receives their funds and begin the loan repayment stage. The borrower makes fixed regular payments of principal and interest which is split proportionally among the investors and a small percentage in fees. If a borrower pays in full and on schedule, the investor will realize their expected return on investment. If, however, a borrower is late on payments or, worse, defaults on a payment, the investor will realize a lower, zero, or negative return.

Objective

The objective of my project is to identify credit-worthy loans to invest in with the goal of maximizing return on investment. I used classification techniques based on listing information such as interest rate, term, income, and FICO score.

Data

Descriptions
The dataset I used is from Prosper which contains information about 23,400 US loan listings from August 2009 to February 2014. The target variable is described below.

Target Description	**Value**	**LoanStatus Details**
"Good"/Credit-Worthy Loan	1	Completed

Target Description	Value	LoanStatus Details
"Bad" Loan	0	Defaulted, Chargedoff, Past Due (61-90 days), Past Due (91-120 days), and Past Due (>120 days)

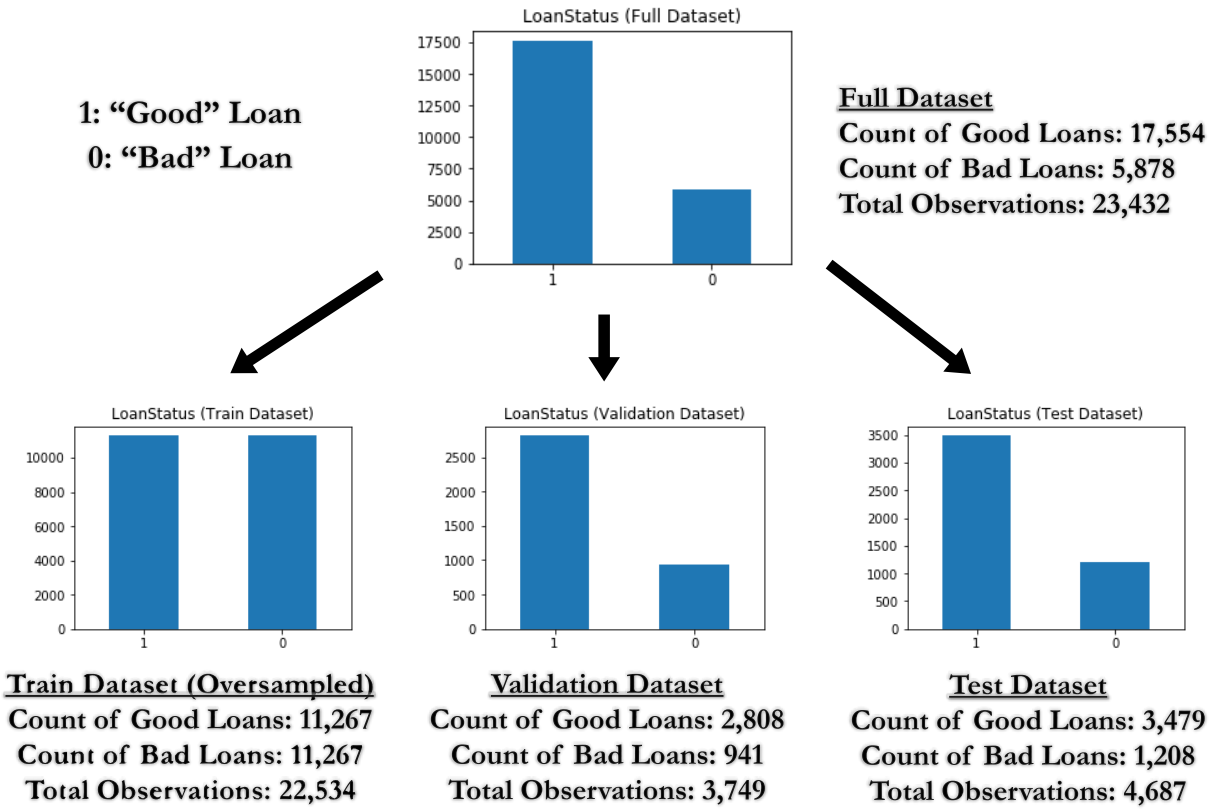
The 29 feature variables are described below.

Variable	Description
Term	The length of the loan expressed in months.
BorrowerAPR	The Borrower's Annual Percentage Rate (APR) for the loan.
BorrowerRate	The Borrower's interest rate for this loan.
ProsperRating (numeric)	The Prosper Rating assigned at the time the listing was created: 0 - N/A, 1 - HR, 2 - E, 3 - D, 4 - C, 5 - B, 6 - A, 7 - AA. Applicable for loans originated after July 2009.
ProsperScore	A custom risk score built using historical Prosper data. The score ranges from 1-10, with 10 being the best, or lowest risk score. Applicable for loans originated after July 2009.
IsBorrowerHomeowner	A Borrower will be classified as a homowner if they have a mortgage on their credit profile or provide documentation confirming they are a homeowner.
CreditScoreRangeLower	The lower value representing the range of the borrower's credit score as provided by a consumer credit rating agency.
CreditScoreRangeUpper	The upper value representing the range of the borrower's credit score as provided by a consumer credit rating agency.
CurrentCreditLines	Number of current credit lines at the time the credit profile was pulled.
OpenCreditLines	Number of open credit lines at the time the credit profile was pulled.
TotalCreditLinespast7years	Number of credit lines in the past seven years at the time the credit profile was pulled.
OpenRevolvingAccounts	Number of open revolving accounts at the time the credit profile was pulled.
OpenRevolvingMonthlyPayment	Monthly payment on revolving accounts at the time the credit profile was pulled.
InquiriesLast6Months	Number of inquiries in the past six months at the time the credit profile was pulled.
TotalInquiries	Total number of inquiries at the time the credit profile was pulled.
CurrentDelinquencies	Number of accounts delinquent at the time the credit profile was pulled.
AmountDelinquent	Dollars delinquent at the time the credit profile was pulled.
DelinquenciesLast7Years	Number of delinquencies in the past 7 years at the time the credit profile was pulled.
PublicRecordsLast10Years	Number of public records in the past 10 years at the time the credit profile was pulled.
PublicRecordsLast12Months	Number of public records in the past 12 months at the time the credit profile was pulled.
RevolvingCreditBalance	Dollars of revolving credit at the time the credit profile was pulled.
BankcardUtilization	The percentage of available revolving credit that is utilized at the time the credit profile was pulled.
AvailableBankcardCredit	The total available credit via bank card at the time the credit profile was pulled.
TotalTrades	Number of trade lines ever opened at the time the credit profile was pulled.

Variable	Description
TradesNeverDelinquent	Number of trades that have never been delinquent at the time the credit profile was pulled.
TradesOpenedLast6Months	Number of trades opened in the last 6 months at the time the credit profile was pulled.
DebtToIncomeRatio	The debt to income ratio of the borrower at the time the credit profile was pulled. This value is Null if the debt to income ratio is not available. This value is capped at 10.01 (any debt to income ratio larger than 1000% will be returned as 1001%).
StatedMonthlyIncome	The monthly income the borrower stated at the time the listing was created.
Recommendations	Number of recommendations the borrower had at the time the listing was created.

Exploratory Data Analysis
As part of my exploratory data analysis, I plotted the histograms for each variable.

The target is imbalanced as approximately 75% of the total observations are Good loans. To balance the data, I oversampled the training data with RandomOverSampler as demonstrated below.



From the plots of the features, it is apparent that there are several opportunities to clean the data. Though I analyzed these features "as is" for my analysis, I plan to explore these opportunities in the future. Several features, such as AmountDelinquent, CurrentDelinquencies, and DebttoIncomeRatio are right-skewed with extreme outliers. For example, over 98% of the borrowers in the dataset have 4 or fewer current delinquencies, however, there is one borrower in the dataset with 32 current

delinquencies. It may make sense to remove these outliers. Additionally, features such as InquiriesLast6Months, OpenCreditLines, and OpenRevolvingAccounts could be transformed to have a more "normal" shape, either with the Box-Cox or log transformation method.

Methodology

Simplifying Assumptions

To simplify my rate of return calculation, I made assumptions regarding the repayment behavior of Good and Bad loans.

- Good loans: full repayment of principal and interest on schedule and with no prepayments
- Bad loans: default immediately with zero return and no recovery

It is important to note that these simplifying assumptions are not completely realistic. Oftentimes, Good loans prepay, therefore, the full amount of interest is not realized over the life of the loan. Equally important, Bad loans oftentimes do not default immediately. There are payments at first with some of them possibly being late, and when the loan is defaulted, collection agencies may be able to recover some more of the principal.

Return on Investment ("ROI") Metric
In order to tune model parameters and evaluate model performance, I chose a standard metric used to evaluate investment performance, the return on investment ("ROI"). ROI is calculated as:

$$\left(\frac{\text{Interest from Good Loans} - \text{Amount Lost on Bad Loans}}{\text{Total Amount Invested in Good Loans and Bad Loans}}\right) \times 100\%$$

Modeling Summary

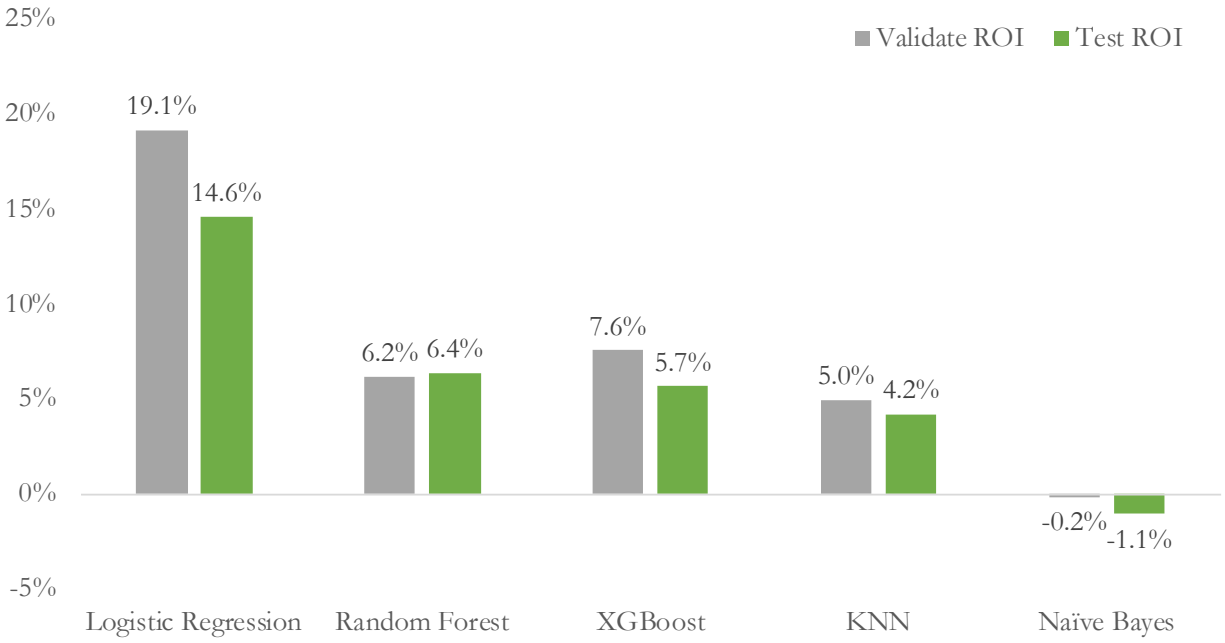
I trained, tuned, and validated five models: Random Forest, XGBoost, KNN, Logistic Regression, and Naïve Bayes.

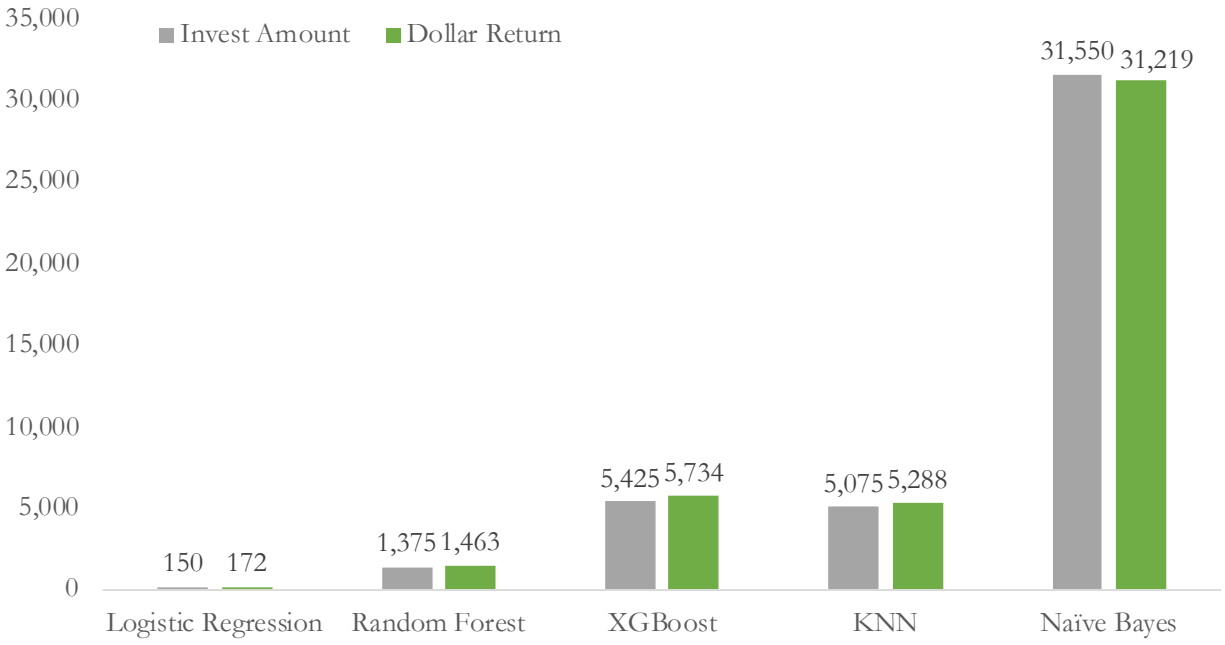
Model	**Tuned Parameters for Max ROI**
Random Forest	• n_estimators = 100 • max_depth = 6 • Probability threshold = 95%
XGBoost	• max_depth = 5 • n_estimators = 100 • Probability threshold = 97%
KNN	• n_estimators = 54 • Probability threshold = 93%
Logistic Regression	• C = 15 • Probability threshold = 99%
Naïve Bayes	• Probability threshold = 99%

Results

Model Comparisons
Assuming investments of $25 on each predicted Good loan, the model that generated the highest ROI for the test set was Logistic Regression with a 14.6% ROI. Random Forest, XGBoost, and KNN performed similarly at 6.4%, 5.7%, and 4.2% ROI, respectively. Lastly, Naïve Bayes performed the worst with a -1.1% return. For context, if I naïvely invested in all loans, I would have realized a loss of 10.67%. Therefore, these models are useful.

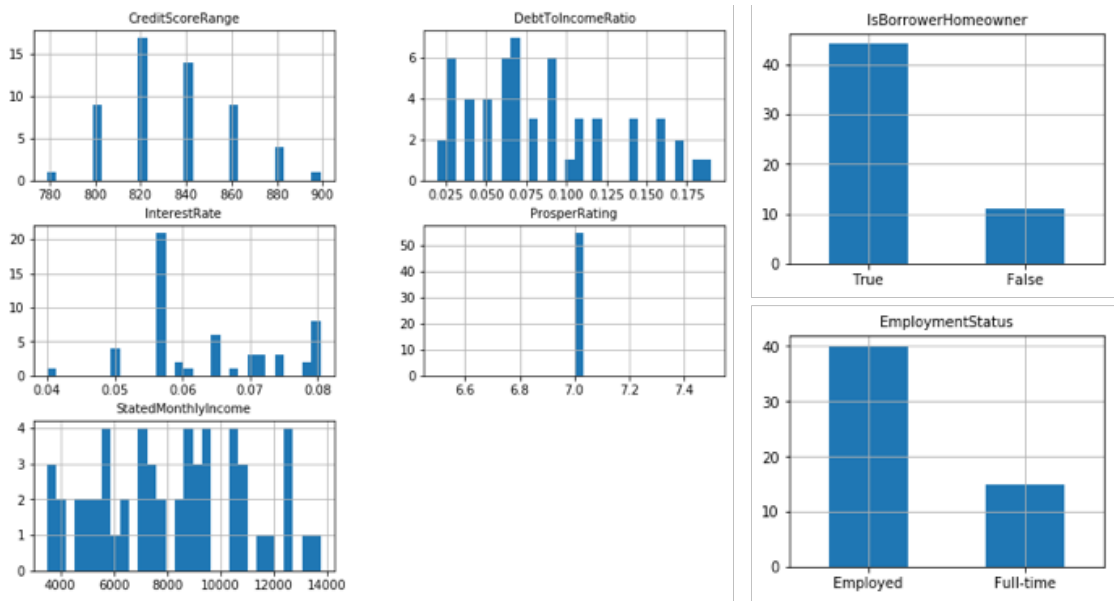
Taking investment amount into account, Random Forest, XGBoost, and KNN perform reasonably well because they require investments of about $1,400, $5,400, and $5,000, respectively. Logistic Regression, on the other hand, was extremely picky and chose only 6 loans to invest in for a total investment of $150. On the opposite end of the spectrum, Naïve Bayes required an investment of $31,550, only to generate a negative return.



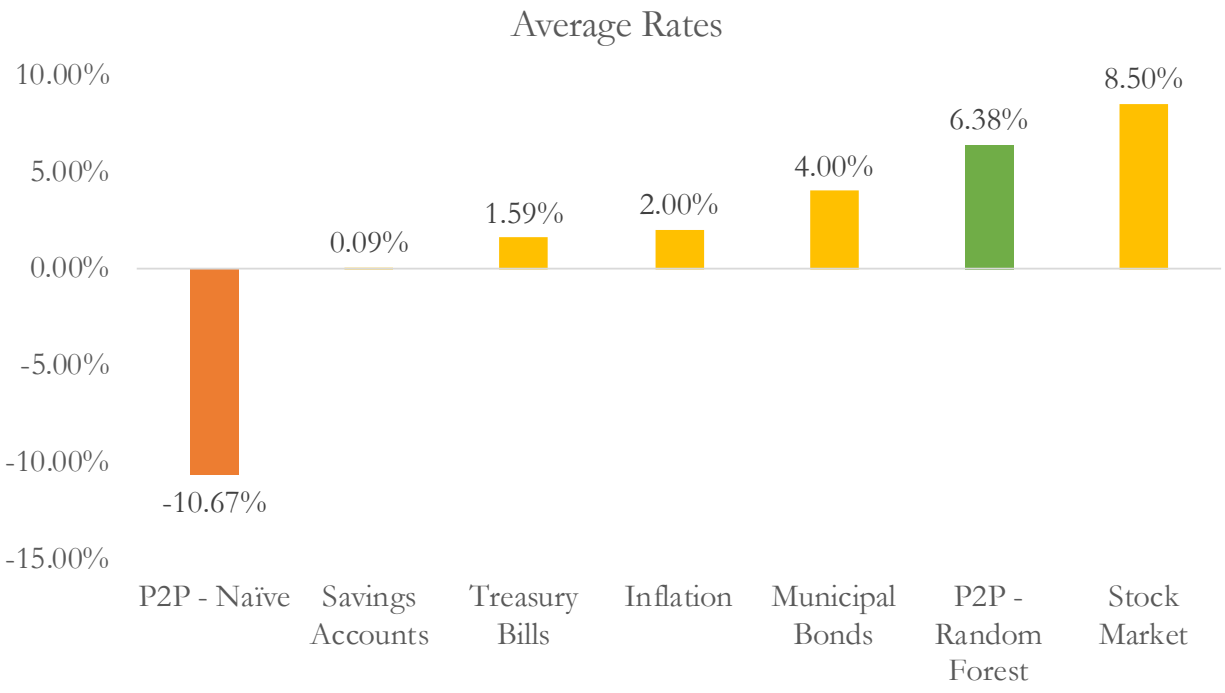


Further Insights

I was interested in exploring the types of loans the Random Forest model predicted to be Good loans. The results make sense. Credit scores were all above 780, debt-to-income ratios were below 18%, stated monthly incomes were above 4,000, and all borrowers were employed.



In addition, compared to other popular investment options, P2P is a decent option, with higher returns than bonds, treasuries, and savings accounts on average.



Final Comments

Future Enhancements

I have many more improvements to make before I make any investment decisions based on my modeling efforts. As mentioned already, I could try further cleaning the data by removing outliers and

testing out transformations. I would also like to explore using other oversampling methods, such as SMOTE and ADASYN.

I must refine my assumptions for the ROI calculation. Good loans may not fully realize their yield even if the entire principal is repaid as completed loans may have had late payments or prepayments; bad loans may not be a complete loss as it may take a few payment cycles for the default to occur. Furthermore, I will constrain the models to find enough loans for an investment between $1000-$2000. Perhaps Logistic Regression will find enough loans to beat the 6.4% from the Random Forest model.

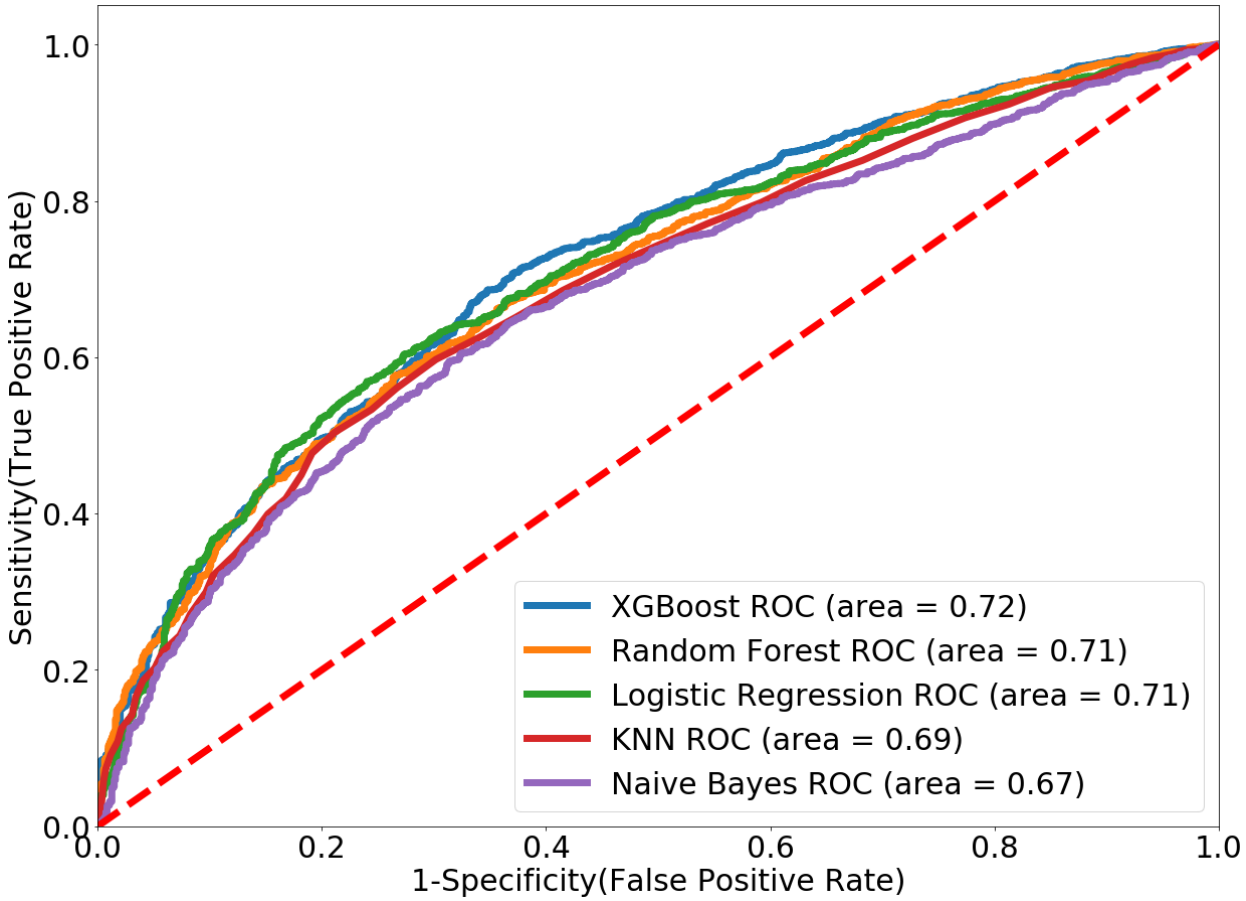
Lastly, I would like to train more models such as SVM and try more modeling techniques such as ensembling.

Further Considerations

P2P is a relatively new investment product and thus the industry has not been through many economic cycles. Financial products may perform wildly different in declining versus growing economic conditions. Therefore, it is important to gather more observations at different points throughout the cycle to better predict loan performance in the future.

Exhibits

ROC Curves



Confusion Matrices

Random Forest

Actual \ Predicted	Bad	Good
Good	1208	0
Bad	3424	55

XGBoost

Actual \ Predicted	Bad	Good
Good	1203	5
Bad	3267	212

KNN

Actual \ Predicted	Bad	Good
Good	1201	7
Bad	3283	196

Logistic Regression

Actual \ Predicted	Bad	Good
Good	1208	0
Bad	3473	6

Naïve Bayes

Actual \ Predicted	Bad	Good
Good	1062	146
Bad	2363	1116