Back to the Future - Predicting P2P Loan Default, Loss Given Default, Prepayment, and Return

Mark Zhao
05/11/2020

Executive Summary

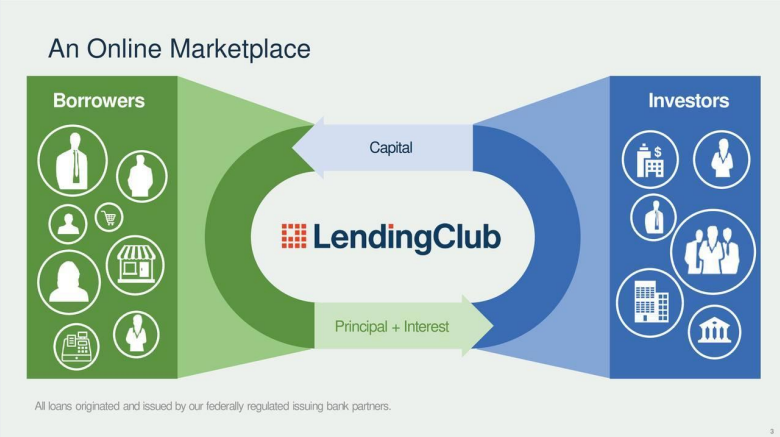
I built 4 machine learning models to predict the default probability, loss given default (LGD), prepayment speed, and return respectively for Lending Club P2P loans.

- Important features include: loan interest rate, term, amount, DTI, and # of installment accounts opened recently.
- XGBoost models perform the best with predictions of 1) **default** with an AUC of 0.67 and F1 score of 0.67; 2) **LGD** with 21.4% R^2 and 0.2 RMSE; 3) **prepayment** speed with 30% R^2 and 0.28 RMSE; 4) loan **returns** with 0.11 RMSE.
- An optimal portfolio was constructed with the highest ranked 1000 loans yielding an **excess realized return** of 5.65%

Business Applications

- Interest rates assigned by LC for the lower grade loans might not be high enough to account for the much higher default risk. The models can be used by LC to optimize its decision to **approve loans, and assign grade & interest rates.**
- By combining the predicted default rate and LGD, the model can be used to predict the **expected loss** of unsecured personal loans.
- Investors can use the model to construct an optimal portfolio that **maximizes returns given their risk tolerance**.
- For companies with different business models that securtize and sell the loans (ABS or MBS), with some adjustments, models can be built to predict the same metrics for **pricing and risk management purposes**.

Problem Statement



How does Lending Club P2P Lending Marketplace Work?

Lending Club (LC) is an online platform where people can either borrow money for multiple funding purposes or lend money as investors to the former group.

Based on the borrower's credit profile and term of the loan, approved loans are assigned with a grade and an interest rate to account for the risk.

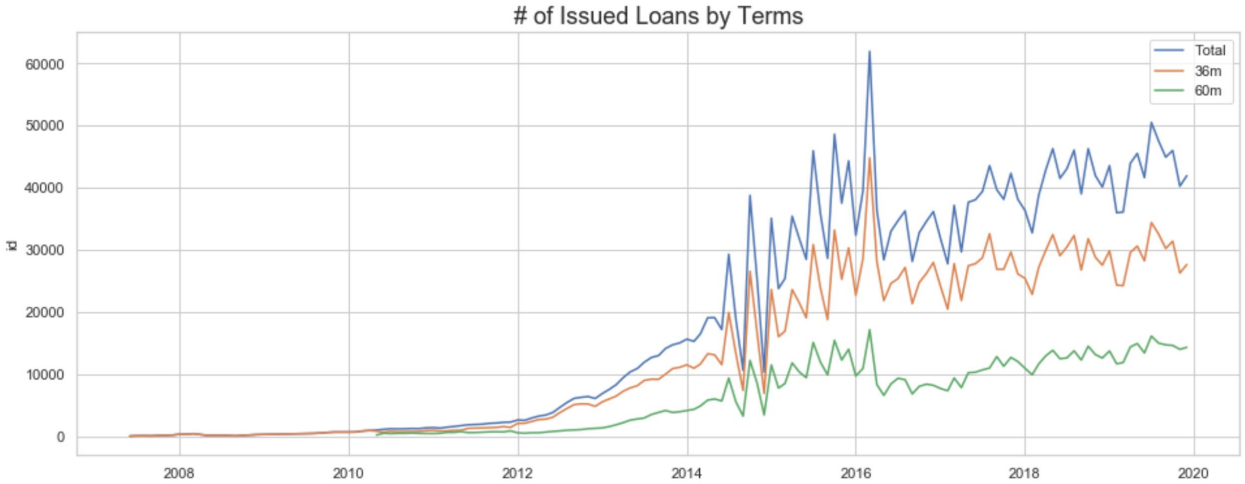
Goals

How do we know which loans are more likely to default? If they default how much can be recovered? How fast are loans prepaid? And what are the expected returns for the loans? These are exactly the goals of this study:

1) Identify important features known at the loan initiation date that can predict the default probability, loss given default, prepayment speed, and loan returns.
2) Predict the default probability and loss given default to get the expected loss.
3) Predict the prepayment speed.
4) Predict the returns of the loans and construct a portfolio that can beat the benchmark.

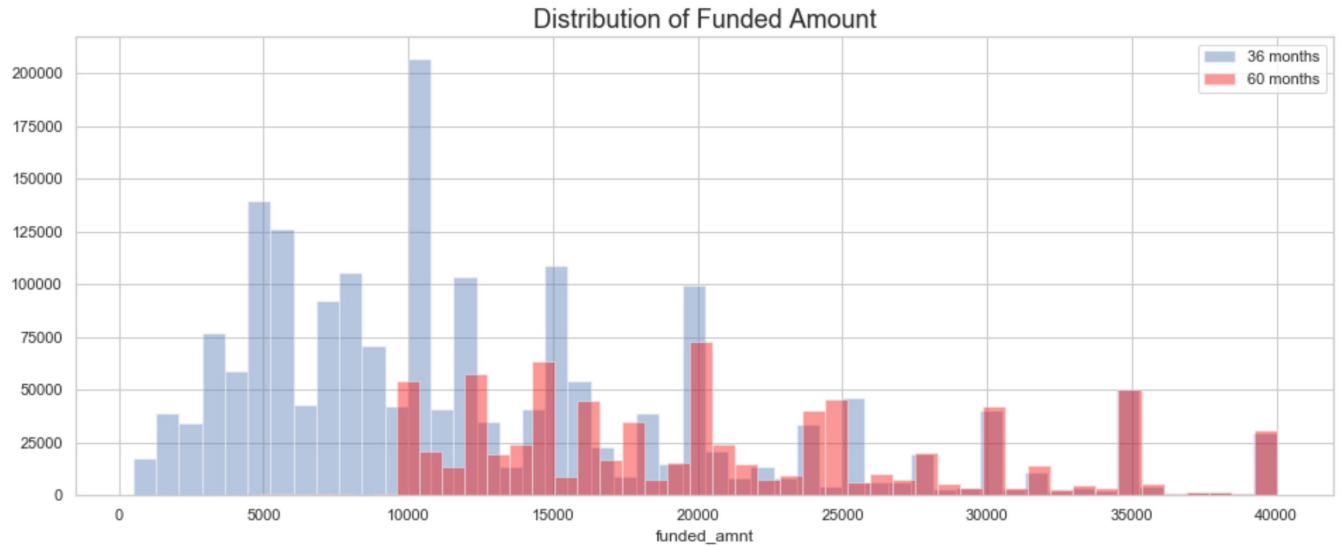
Data Wrangling

- The data were downloaded from Lending Club with 2.77 million loans and 150 features from 2007 to 2019.
- Loan initiations dropped severely in 2016 due to the infamous Lending Club scandal.



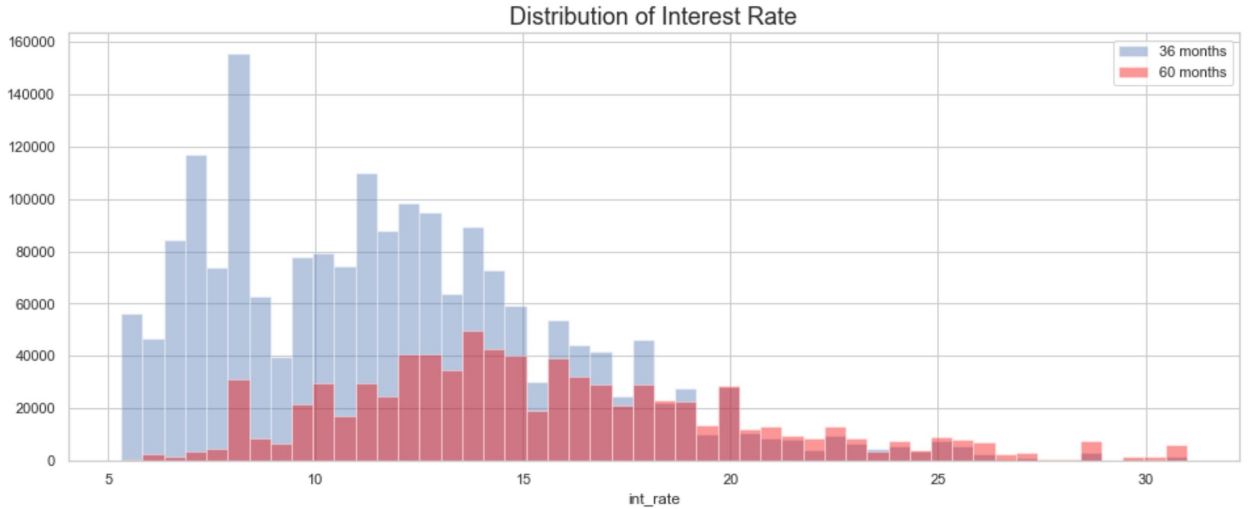
Loan Amount Distribution

- Loan amount ranges from $1,000 to $40,000 with the median as $13,000.



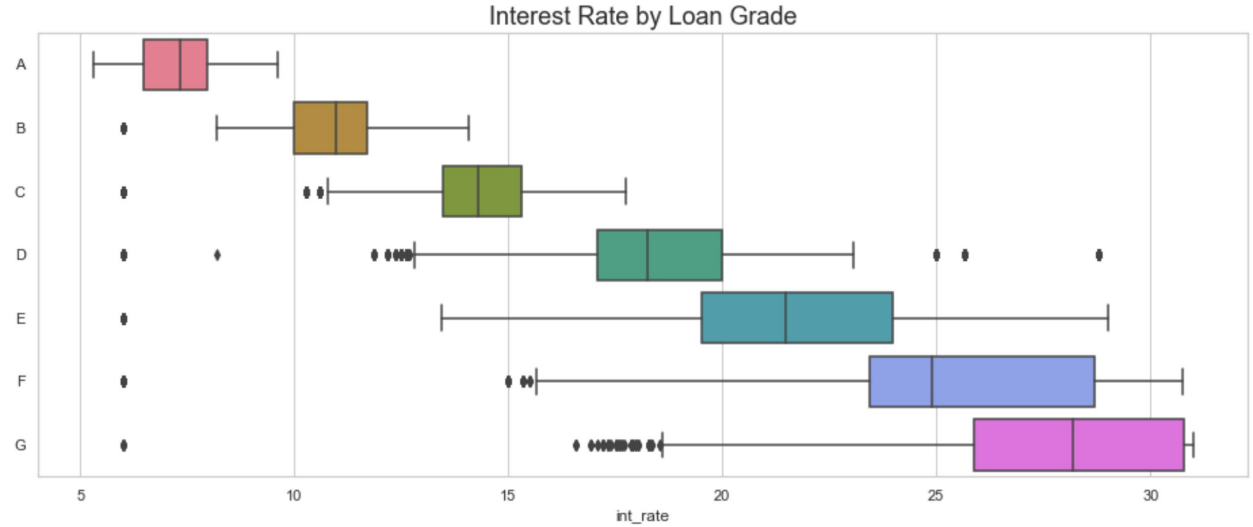
Loan Interest Rate Distribution

- Interest rates range from 5.31% to 30.99% with a median of 12.7%. Interest rates on the longer term loans (60m) are generally higher to account for the higher risk.



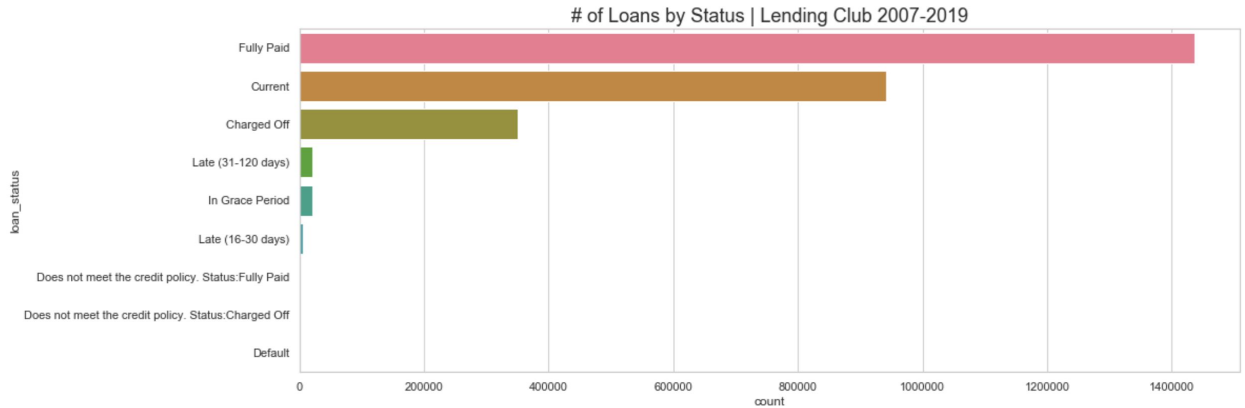
Interest Rate by Grade

- Interest rates are generally higher for lower grade loans.



Loan Status

- Only non-current loan types: “paid off”, “charged off” and “default” were used in the analysis and the others were removed. “Default” was recategorized as “charged off”.
- In the end there were around 1.7 million loans with 20% charged off and 80% paid off.



Data Cleaning

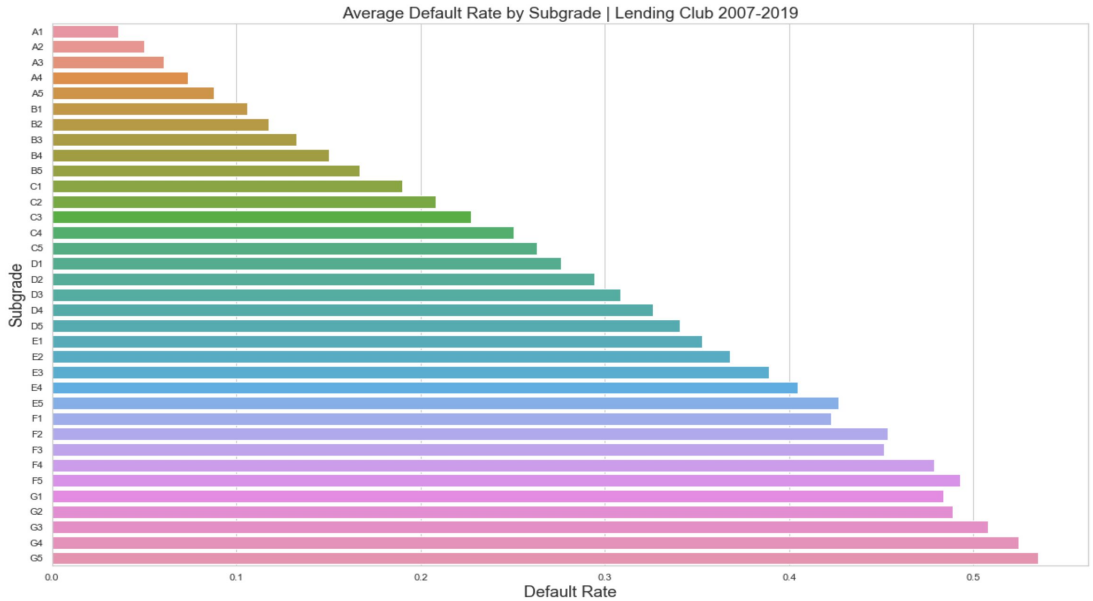
- Features unknown at the loan initiation date were removed to prevent data leakage.
- Joint application features were merged with single application features.
- Dummy variables were created for the following categorical features: term, sub_grade, home_ownership, verification_status, loan purpose, state, and application type etc.
- Employment length was converted to numerical (number of years).
- NAs for features like mths_since_recent_inq means there was never a delinquency and thus is replaced with 2 * maximum.
- For features like dti, employment length etc. the missing values were filled with the median.
- Annual income outliers below $8000 or above $300K were removed (<1% of the sample).
- Observations with negative DTI were also removed.

Feature Engineering

- A few new features were created: 1) length of the loan description provided by the applicants; 2) the history of credit line (# of months between the earliest credit line date and the loan initiation date); 3) average FICO score based on the FICO range low and high; 4) installment divided by monthly income which is the percentage of monthly income that is used to pay off the loan.
- New features that were added after 2015 by LC for example open_acc_6m were discarded after testing the post-2015 subsample where no significant power was found.
- Features were scaled for easier interpretation of coefficients or feature importances.

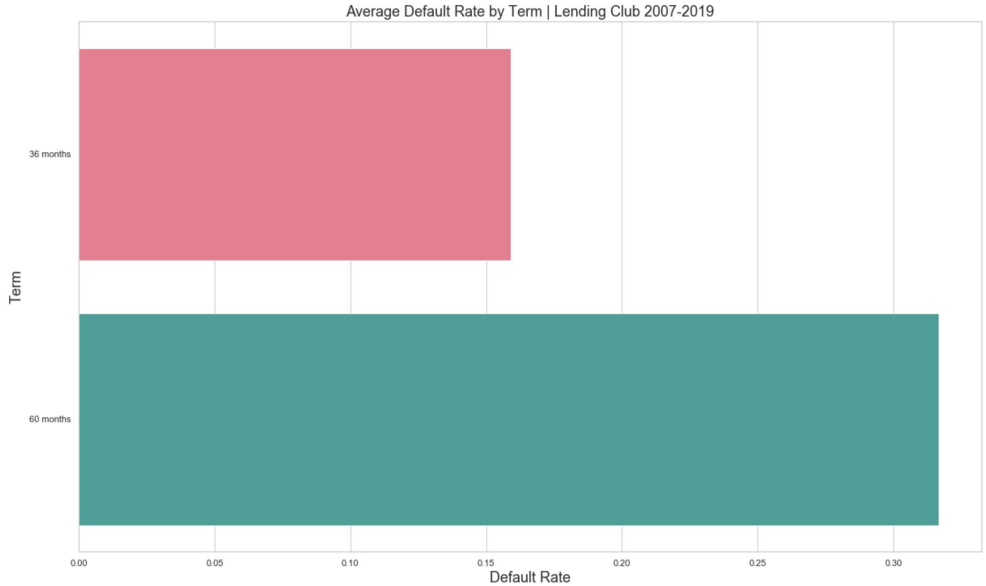
Default Rate by Subgrade

- Default rate increases from 4% for A1 to 55% for G5. Huge differences!



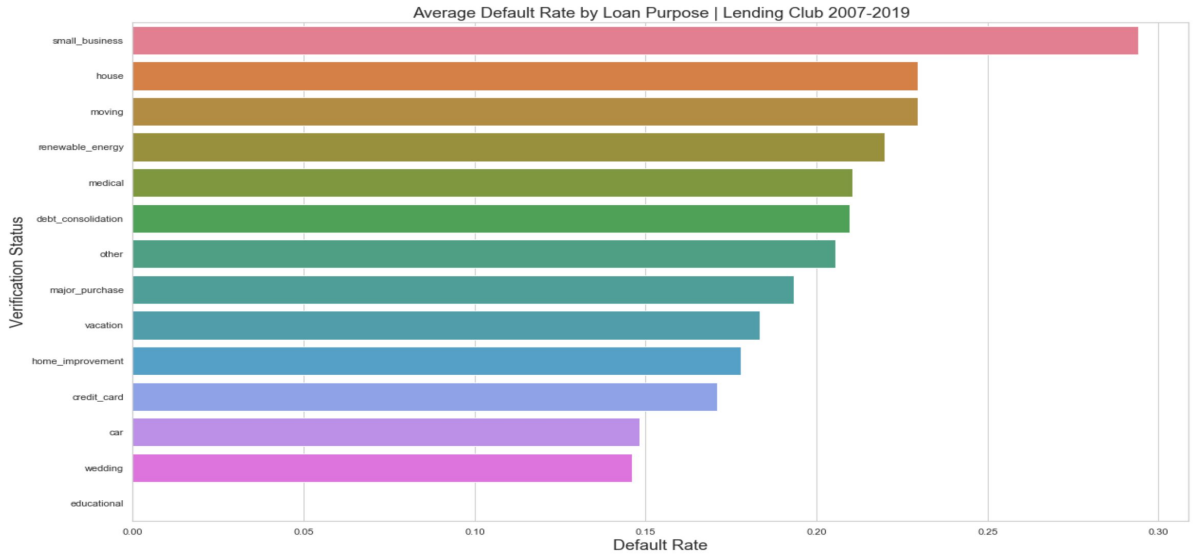
Default Rate by Term

- Default rate is higher for longer term loans reflecting its higher default risk.

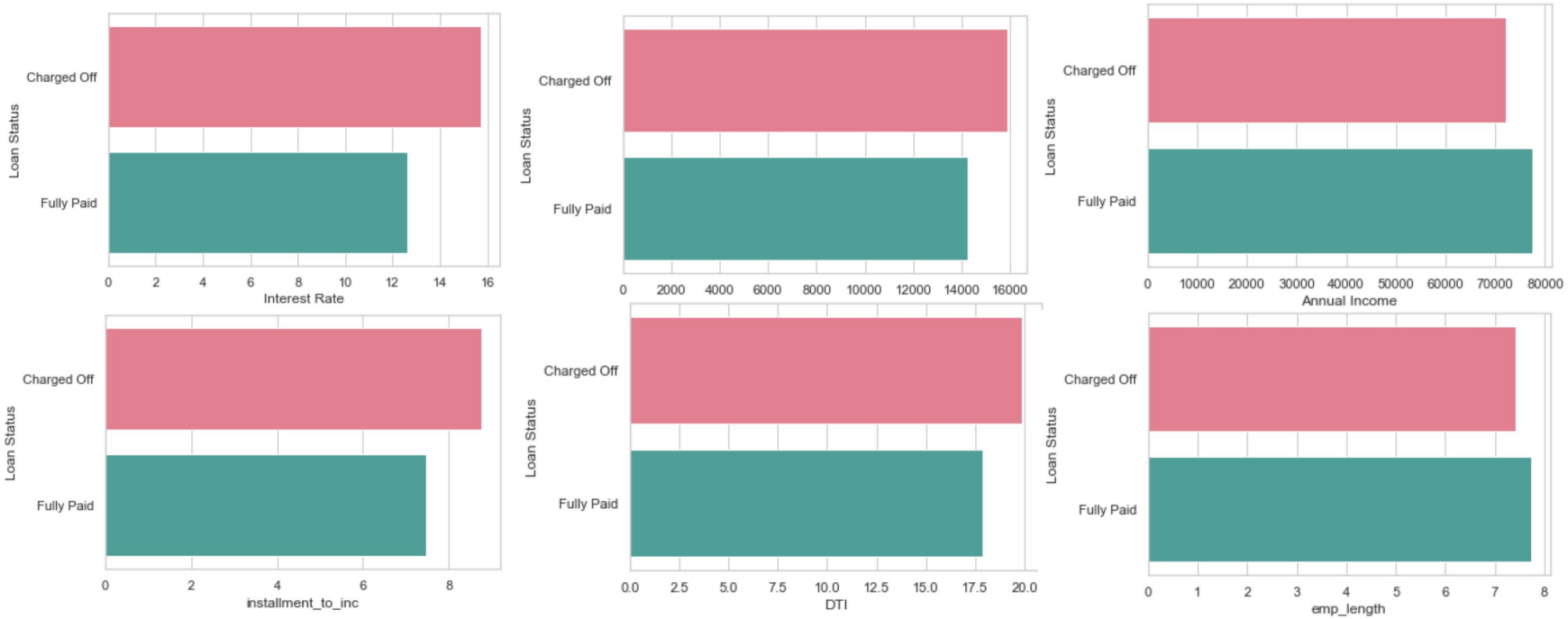


Default Rate by Loan Purpose

- Small Business is the riskiest! Wedding and Education have the lowest default rates.



Average Feature Values for Loans Charged Off vs. Fully Paid



Default Prediction Machine Learning

- Target variable: loan status with 1 being charged off and 0 being paid off.
- Features 95% or more correlated with other features were removed.
- With 20% of the loans charged off, this is a typical imbalanced data and resampling was done to reduce the imbalance. Under-sampling (randomly selecting paid off loans so the number of paid off loans is the same as charged off loans) was used in the end.
- A 70%/30% train and test split with stratification approach was used for model training and testing. A different approach using data from 2012 to 2017 for training and testing on 2018-2019 data generated similar results.
- Before running logistic regressions, one of the categories need to be dropped for each category to avoid perfect correlation of features.
- 676,734 loans with 152 features in the end.

Default Prediction - Logistic Regression

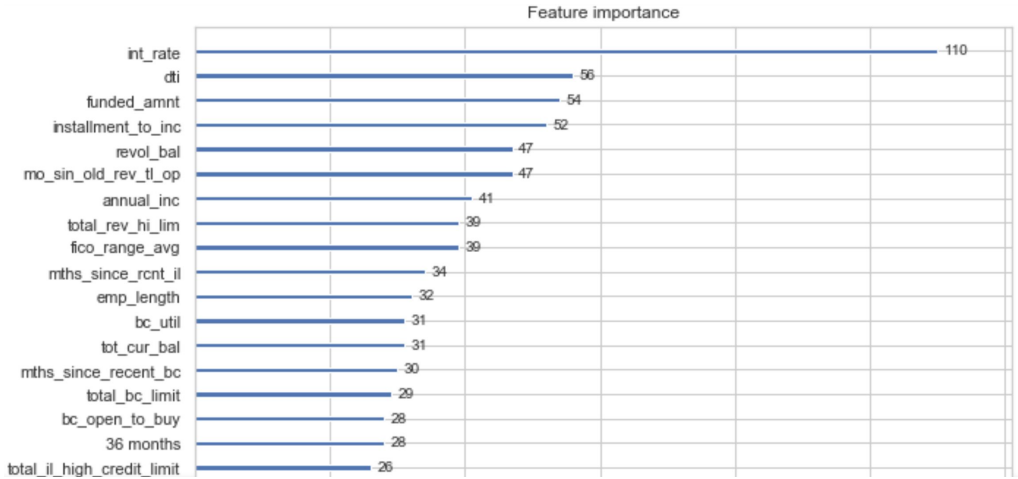
- Recursive feature elimination (RFE) was used for regularization. Top 30 ranked features were used in the first step and then features with p-values higher than 0.05 were removed next.

- Default probability is lower for a shorter term loan (36 month vs. 60 month), higher rated loans (A1 through B3), lower DTI, and less accounts opened recently.

- This model leads to AUC of 0.64.

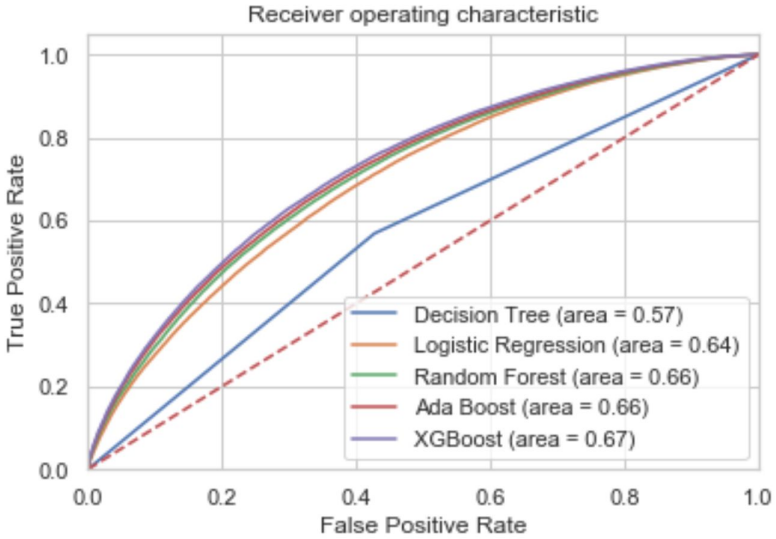
Default Prediction - XGBoost

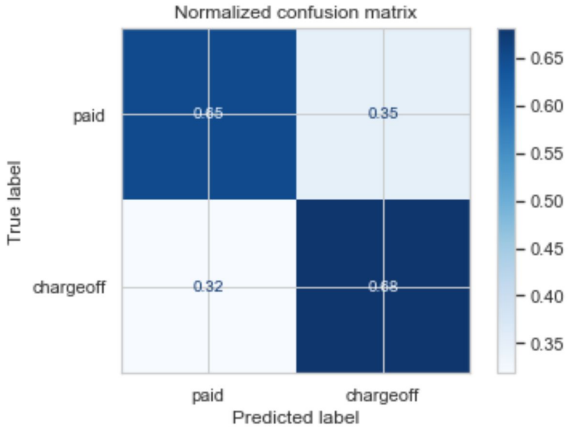
- Decision Tree, Random Forest, AdaBoost, and XGBoost were trained and tested.
- The most important features in the XGBoost model include interest rates, DTI, funded amount, installment to income, revolving balance etc.



Default Prediction - XGBoost

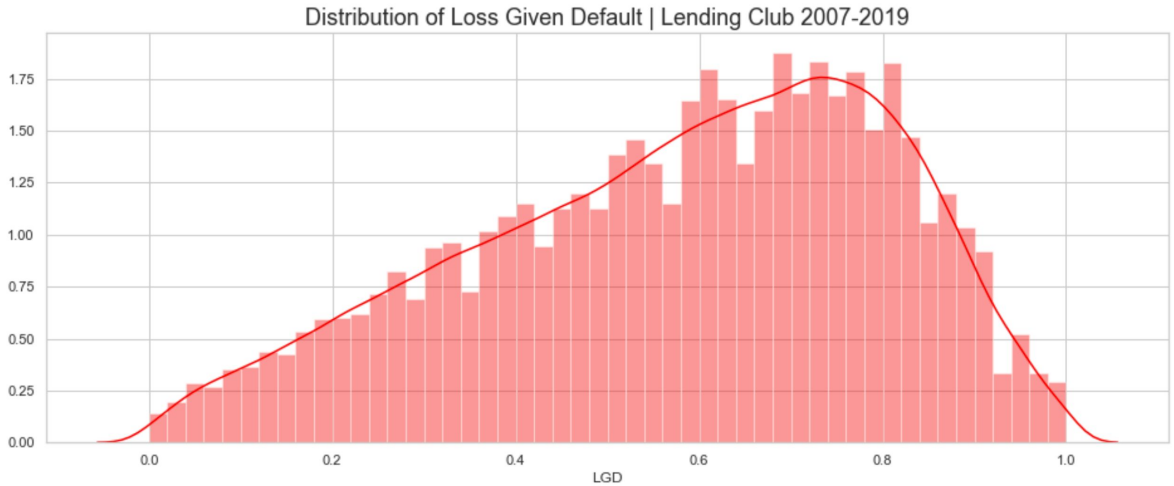
- AUC is 0.67 for XGBoost with n_estimator of 200 and max_depth of 3. This is a winner!
- The XGBoost model yields a precision 0.66, recall 0.68, and F1 score of 0.67.





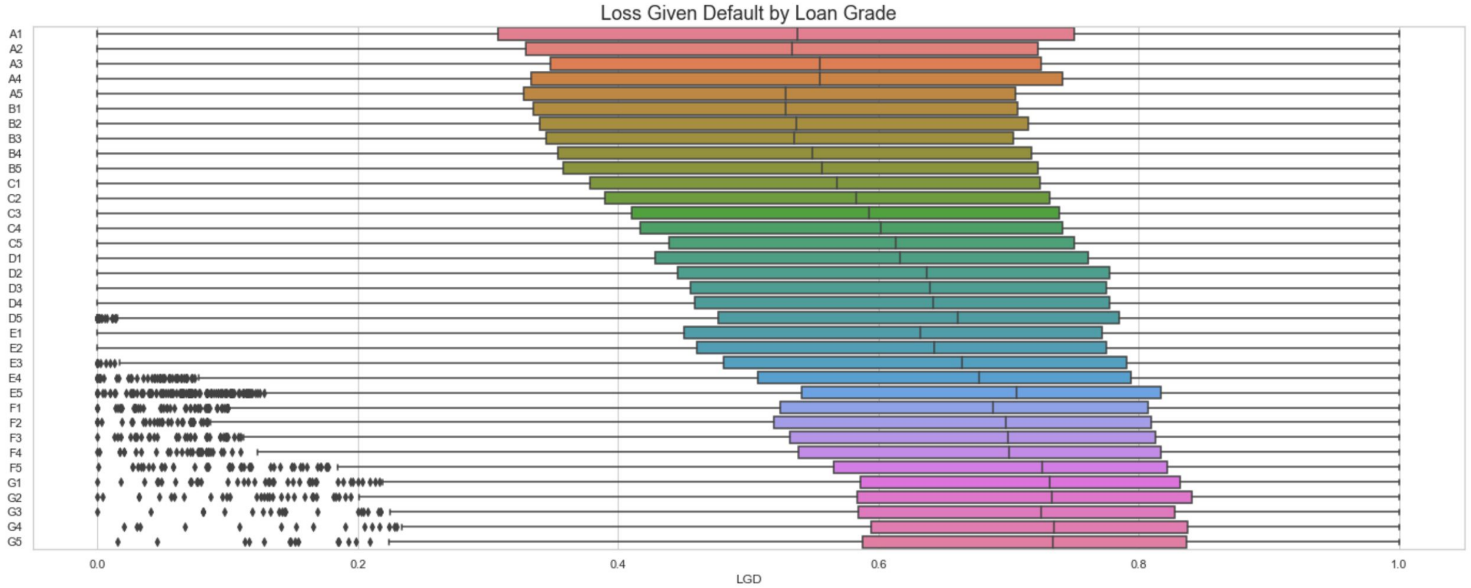
Loss Given Default (LGD) Prediction

- LGD captures the magnitude of the loss given the loan is in default. This was rarely studied.
- LGD is defined as the **total payments made by borrowers (including any recoveries) divided by the total amount of principal and interests.**



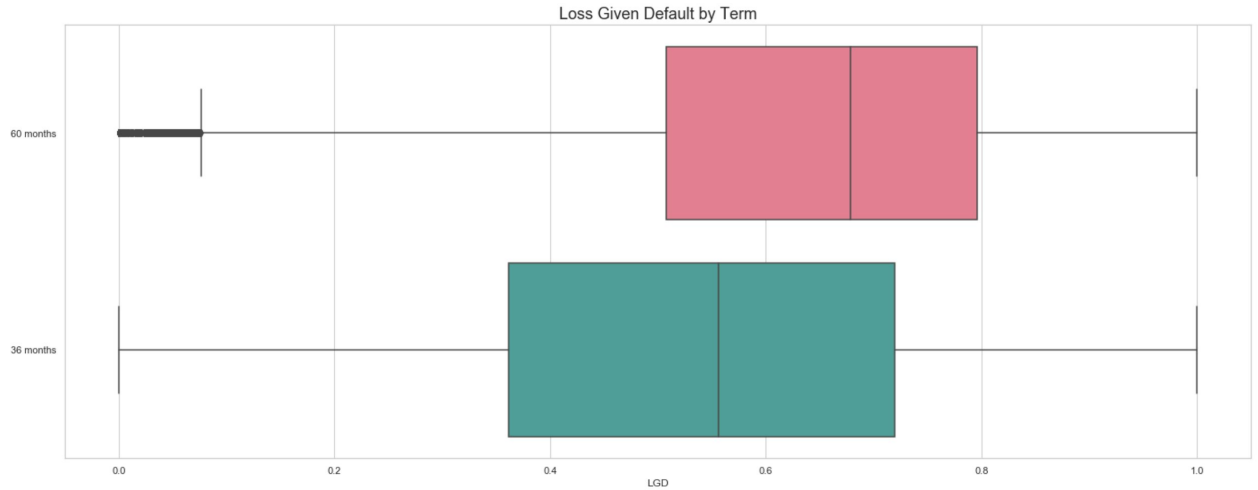
LGD by Subgrade

- LGD is generally higher for lower grade loans.



LGD by Subgrade

- LGD is generally higher for longer term loans.



LGD Prediction - Linear Regression

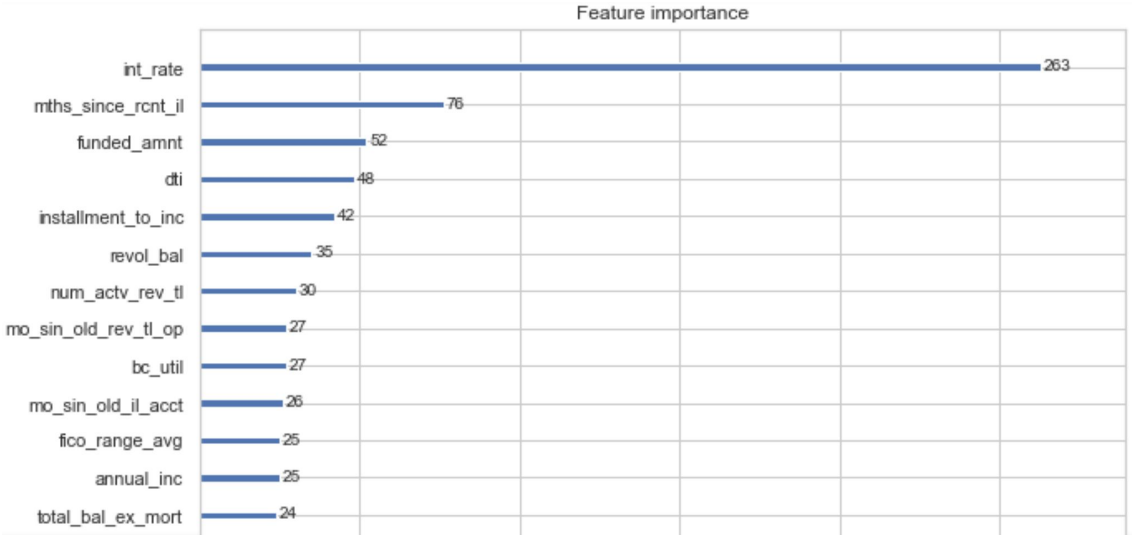
- Recursive feature elimination (RFE) was used for regularization. Top 30 ranked features were used in the first step and then features with p-values higher than 0.05 were removed next.

- LGD is lower for a shorter term loan, for borrowers who didn't open installment accounts recently, and for a lower interest rate.

- The R^2 is 15% and RMSE is 0.2

LGD Prediction - XGBoost

- XGBoost is still the winner here! It achieved the highest R^2 of 21% and 0.2 RMSE.
- The most important features: interest rates, if recently opened installment, funded amount, and dti.



Default and LGD Prediction - Caveats

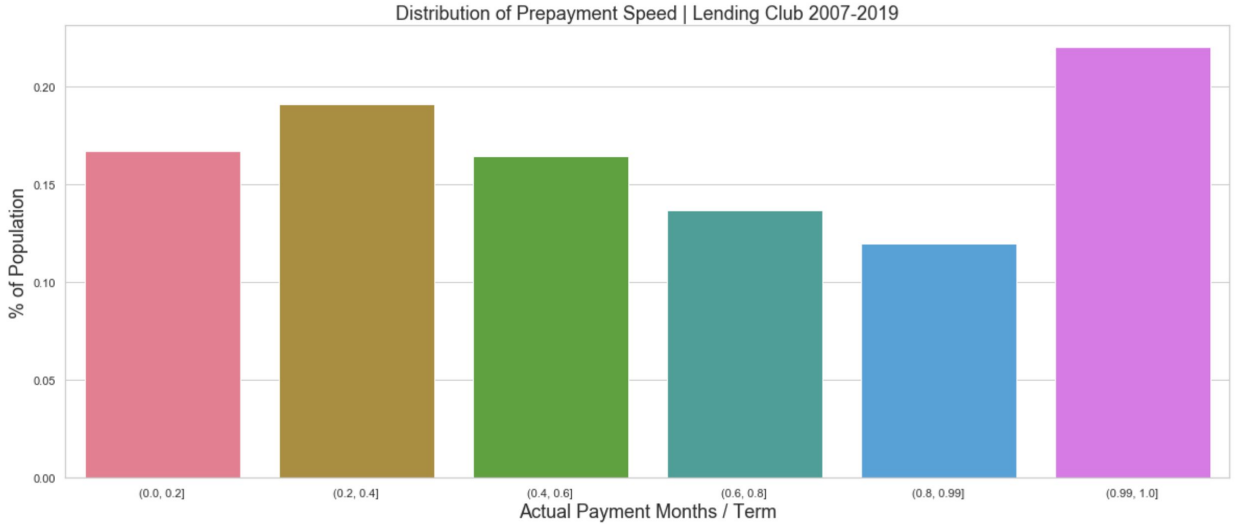
- There are apparently important missing features for example the macroeconomic conditions: the yield curve, inflation, unemployment and GDP that could cause the **systematic default risk** rise or fall across all applicants.
- Similarly, the loan applicant's own credit condition might change from time to time and cause defaults.
- Another shortfall is when calculating LGD, cash flows were not discounted to the present value due to the lack of the paths of cash flows and the corresponding dates
- Since most of the data span from 2012 to 2019 when the economy and market performed relatively well it's less of an issue for this data but could definitely change if a major economic shock like the COVID-19 period is included in the analysis in future studies.

Prepayment Speed Prediction

- There is another risk that's also important to investors which is the prepayment risk. This was also rarely studied for this data.
- If the prevailing interest rates fall or the borrowers' credit improves, then they tend to prepay the loan and borrow at a lower rate.
- Investors have to reinvest proceeds usually in a lower interest rate and thus lower returns.
- Since this data lack the information of the paths of prepayment and only the last payment date and total payment were known, prepayment speed was defined as **actual # of months pay off / loan term**. So if a 36 month loan was paid off in 18 months then its prepayment speed is 50%. So **the lower the metric, the faster the prepayment**.
- However this measure overlooks the amount paid in each period and is not able to distinguish cases when someone paid off more in the beginning vs. towards the end due to lack of the paths of prepayment data.

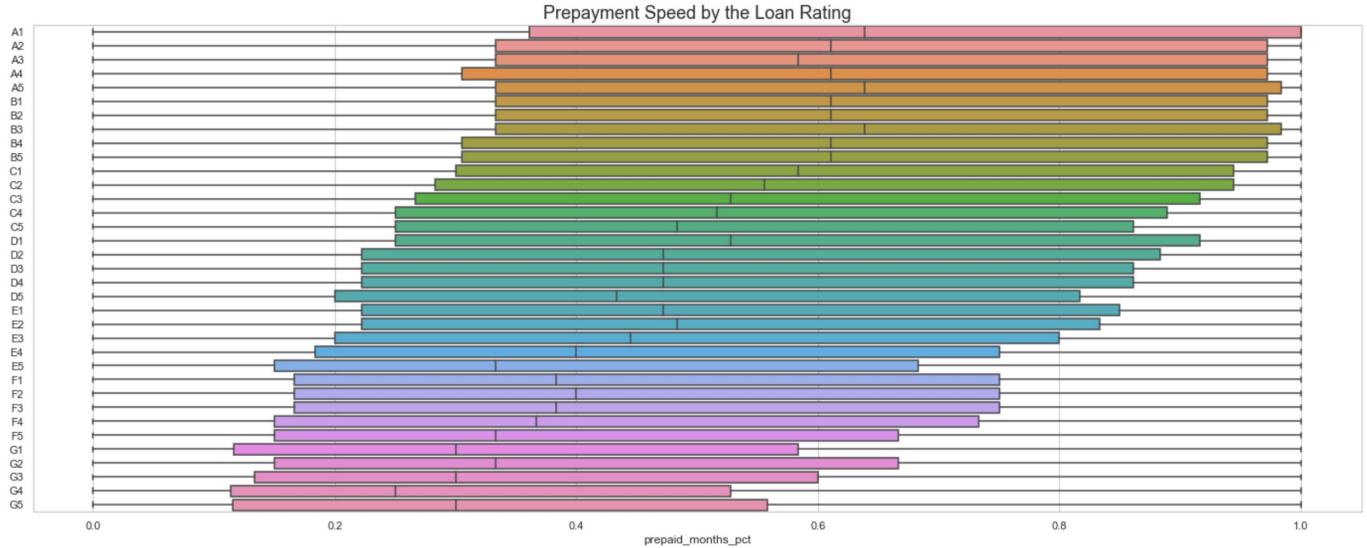
Prepayment Speed Distribution

- Only around 25% of total loans were not prepaid! And 75% of the loans were prepaid at different speeds probably because refinancing on LC is fairly easy.



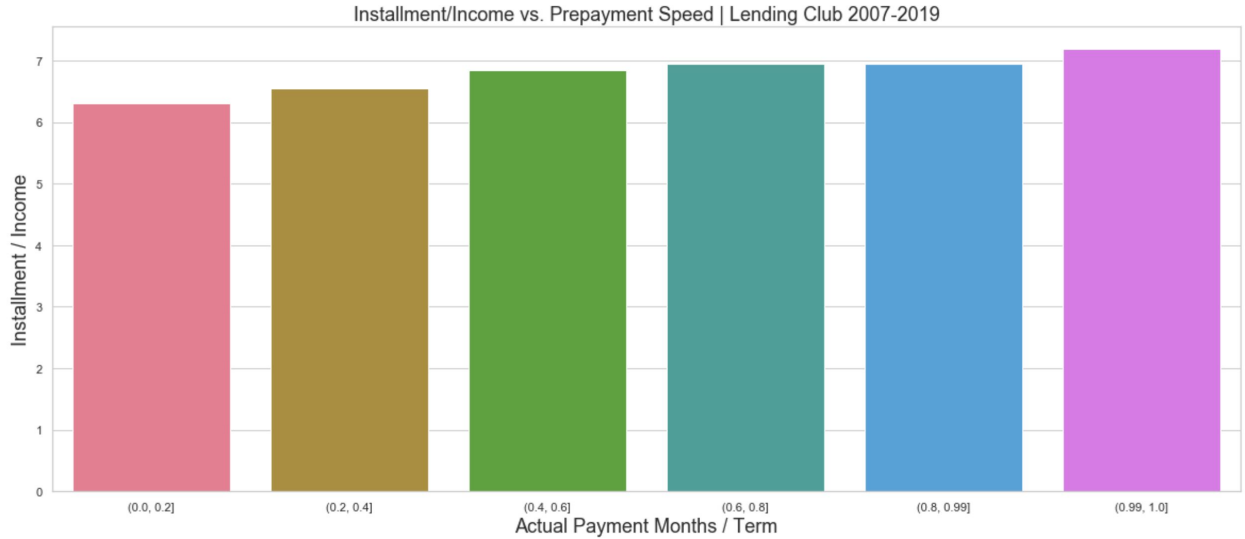
Prepayment Speed by Subgrade

- The higher interest rates /lower grade loan borrowers have more motivation to pay off (if they can) and refinance at a lower rate.



Prepayment Speed by DTI

- Having the motivation to prepay is one story, having the ability to prepay is another story. The lower the debt to income ratio, the faster the prepayment happens on average.



Prepayment Speed Prediction - Linear Regression

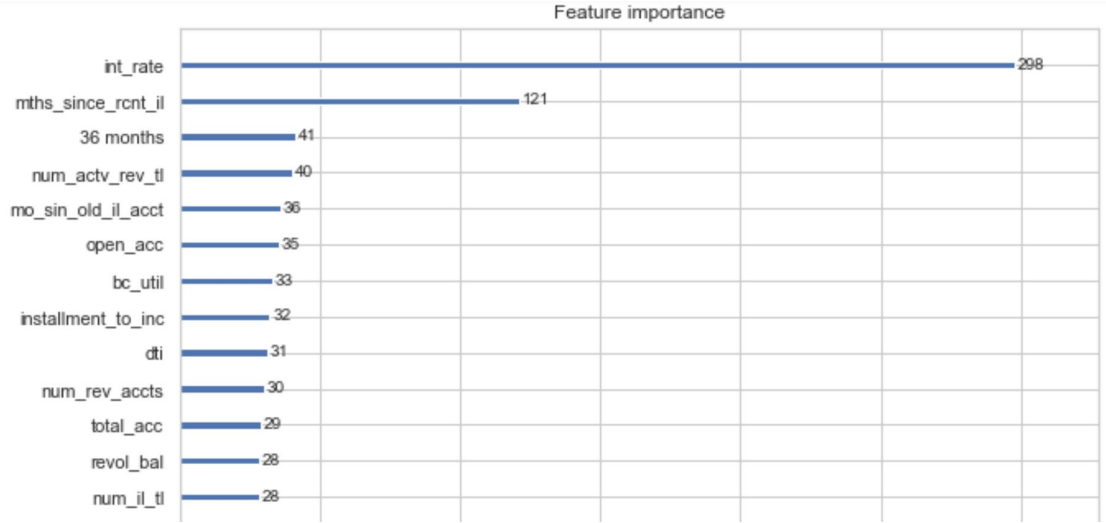
- Recursive feature elimination (RFE) was used for regularization. Top 30 ranked features were used in the first step and then features with p-values higher than 0.05 were removed next.

- Prepayment is faster if loans have a longer term, if the interest rate is higher, if more accounts were opened recently (suggesting applicants might open other accounts to pay off the existing higher interest loans).

- The R^2 is 21.5% and RMSE is 0.28

Prepayment Speed Prediction - XGBoost

- Again, XGBoost beat all the other models including random forest and delivered a performance of 30% R^2 and 0.28 RMSE.



Return Prediction

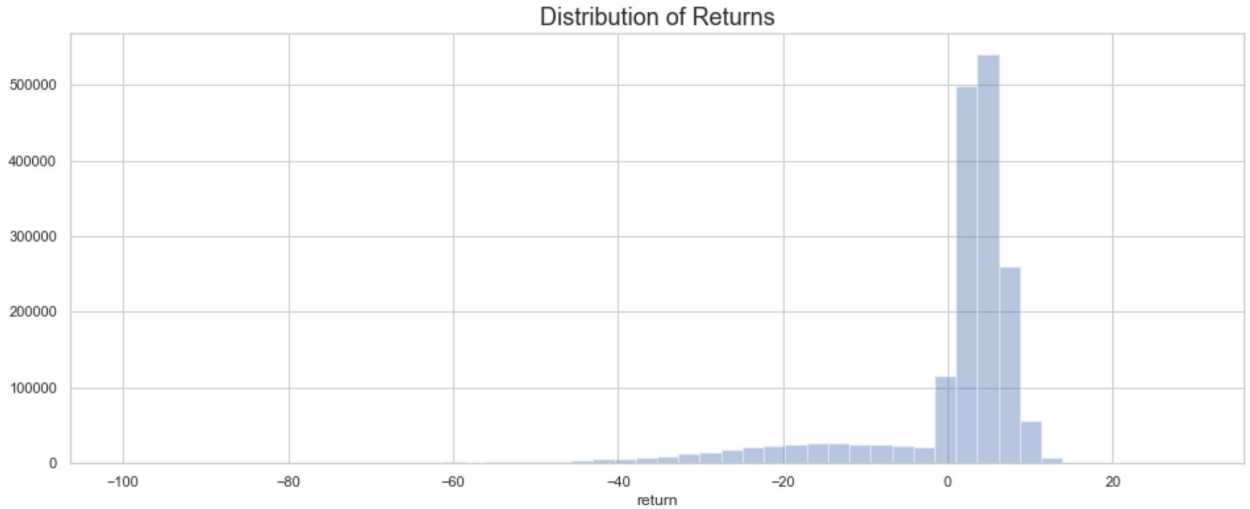
- Return prediction is the “Holy Grail” for investors and it depends on all the components from previous studies: default probability, LGD, and prepayment.

- Loan return is defined as the **total payments / funded loan amount (the principal of the loan) - 1** and then annualized using the # of months in the term.

- This is assuming no reinvestments once the loan defaulted or got prepaid due to the lack of data when and what reinvestment happened. Even with this constraint the model performs very well.

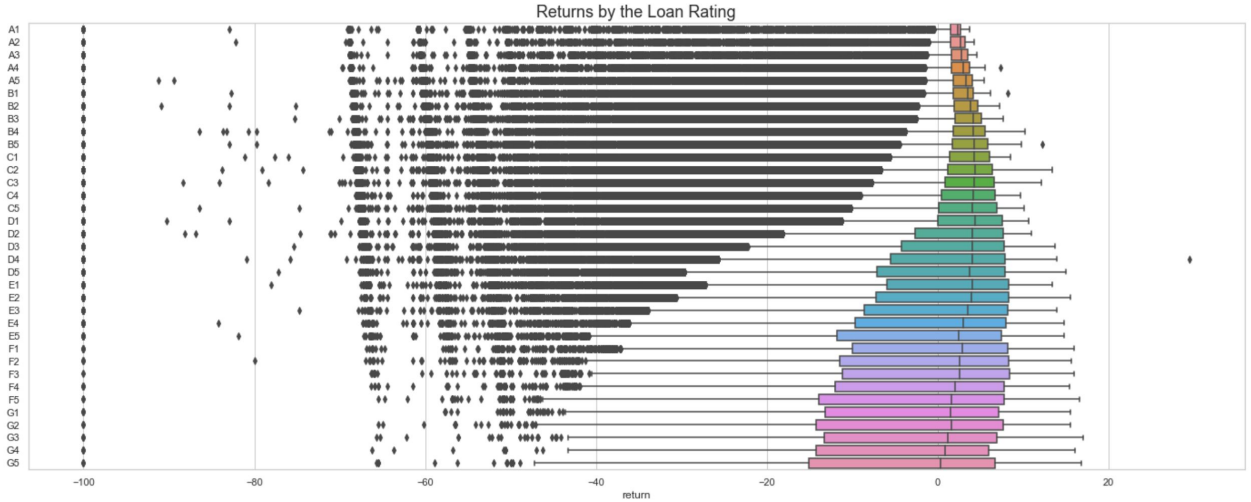
Return Distribution

- Returns range from -100% (when not a single dime was paid after loan initiation) to around 29% (likely to be a high interest loan that was paid off till the end of the term).



Returns by Subgrade

- The higher the grade, the less spread out the returns are and less risks for investors. As grade gets lower, returns are more spread out and investors start bearing more risks (but not necessarily getting more returns!)



LC Marketed Returns

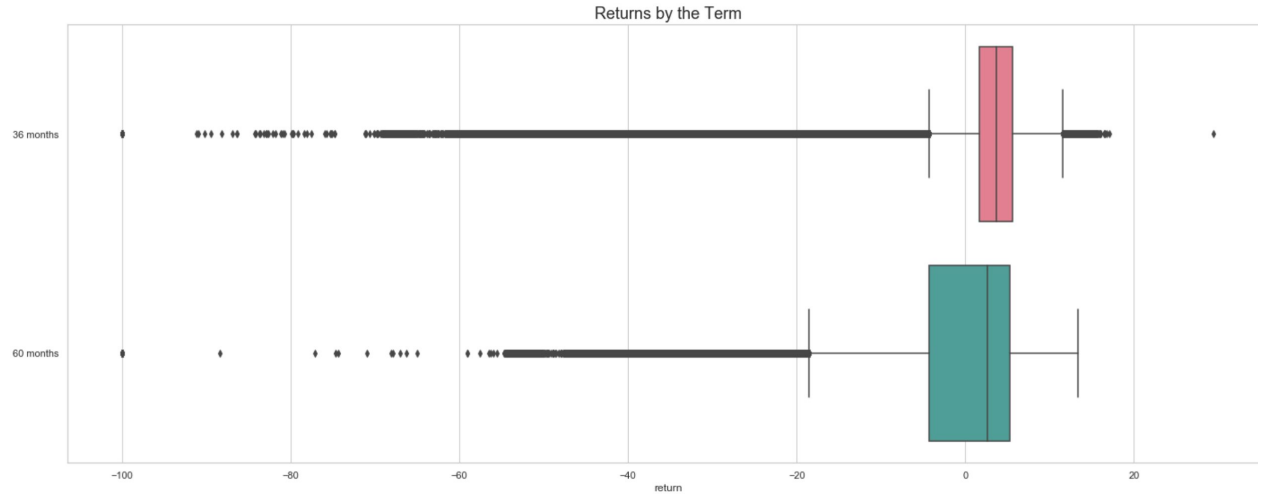
How does the realized returns compare with the marketed returns on LC's website (see left)?

It claims the expected return is 26% for loans with a grade of G, if they don't default (or get prepaid). That's a big "if"!

In fact, the median return for grade G was around 1%! LC probably didn't charge high enough interest rates to compensate investors in Grade D and beyond.

Returns by Term

- Returns are not necessarily higher for longer term loans in this case, probably due to higher default probability, LGD, and prepayment when term is longer as shown previously.



Return Prediction - Linear Regression

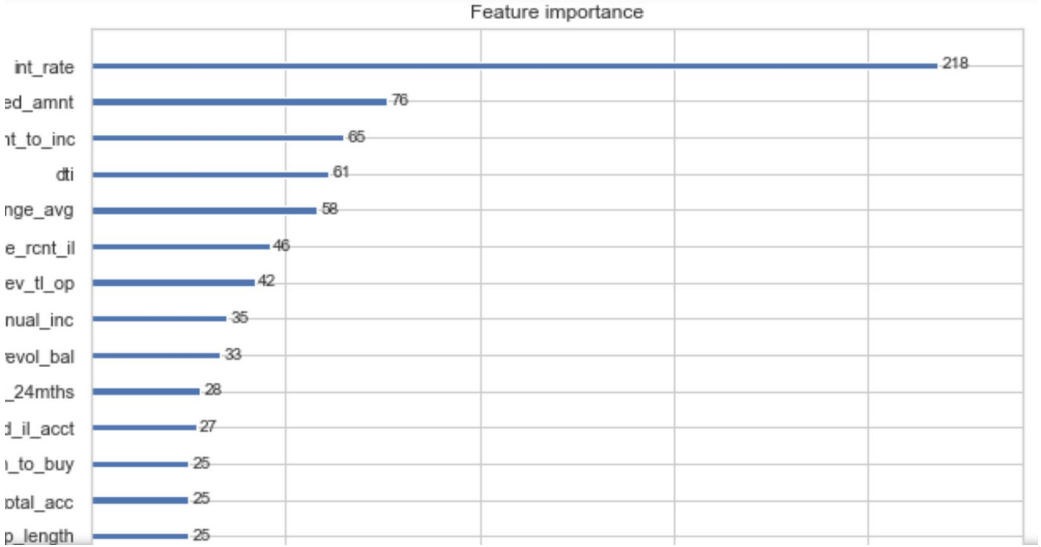
- Recursive feature elimination (RFE) was used for regularization. Top 30 ranked features were used in the first step and then features with p-values higher than 0.05 were removed next.

- The expected returns are higher for shorter term loans, lower DTI, less accounts opened in the last 2 years and less recent installment accounts opened (not hungry for credit). The relationship between returns and the interest rates is not linear.

- The R^2 is 4.4% and RMSE is 0.11

Return Prediction - XGBoost

- Once again, XGBoost was the best performing model, delivering a R^2 of 7.2% (vs. 5.9% using Random Forest) and RMSE of 0.11.



Optimal Portfolio Construction

- To see if our model can generate extra returns, we first randomly selected 1000 loans from the test sample of 511,371 loans and ran this experiment for 10000 times. The average return (assuming equal-weighting) was around 0%.

- What if we select 1000 loans with the highest predicted returns using the XGBoost model? The realized return of this portfolio was 5.65% which is also the Alpha!.

- We could make it even fancier - by predicting returns for charged off loans and paid off loans separately and then combine them with the default prediction model to potentially achieve a better result.

Conclusions

- Using the classic Lending Club P2P loan data, I built machine learning models that cover different aspects and can be valuable to the business and investors.

- This study corrects several mistakes made by earlier default prediction works and explores subjects that were rarely studied before: LGD, prepayment risk, and returns.

- With predicted default probability and LGD, we are able to predict the expected loss of a loan. And by combining default risk and prepayment risk, returns can be predicted to construct an optimal portfolio for investors.

Future Improvements and Applications

- Measure and predict the risk of the loans and design a mean variance optimization to construct a portfolio that maximizes expected returns given investor's risk tolerances.

- Include periodically updated borrowers and economic features and run a cohort study to predict the default risk and prepayment risk in a more precise and timely manner.

- Analyze and predict the default risk, prepayment risk, and returns for different tranches of loans by incorporating the joint default correlation of loans in each tranche into the model.