Lending Club Project Report

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

Mark Zhao

05/11/2020

Executive Summary

In this research, I designed 4 machine learning models that predict the **default probability, loss given default (LGD), prepayment speed, and return** for the Lending Club P2P loans. By combining the predicted default rate and LGD, the model can be used in practice to predict the expected loss of unsecured personal loans for pricing and risk management purposes. Investors can also apply the model to construct an optimal portfolio to maximize returns given risks.

The key takeaways include:

- An XGBoost model was built to predict loan **default probability** with an AUC of 0.67 and F1 score of 0.67. Default probability is higher if the loan's interest rate is higher, if the loan amount is higher, if the term is longer (60m vs. 36m), if the borrower has a higher DTI and/or if they recently opened installment accounts.

- A **LGD** prediction model generates a performance of 21.4% R^2 and 0.2 RMSE. LGD is higher if the interest rate is higher, if the borrower recently opened installment accounts, if the loan amount is higher, if the borrower has a higher DTI and/or installment to income ratio.

- A **prepayment speed** prediction model generates 30% R^2 and 0.28 RMSE. Borrowers prepay faster if the interest rate is higher, if they recently opened installment accounts (suggesting they might have been refinancing), if the term is longer.

- A model is built to predict the **loan returns** with 0.11 RMSE. Constructing an optimal portfolio with the highest ranked 1000 loans yielding an excess realized return of 5.65% vs. the benchmark.

- **Interest rates assigned by LC for the lower grade loans might not be high enough to account for the much higher default risk.** This goes back to its business model. LC

is a platform and they charge an origination fee when the loan is approved. Later these loans are sold to retail or institutional investors so LC doesn't bear any credit risk (but charge a late fee if borrowers fail to pay on time). Therefore, LC has a motivation to approve many subprime loans even though they know the borrowers can't repay them. In fact, this is one of the issues in its 2016 2016 when the public learned many approved loans didn't meet criteria.

With insights learned from the study, the following practices can be applied into the LC business to act on investor's best interests. Investors can also use the model to optimize their investment strategies.

1) The models can be used by LC to optimize its decision to approve the loans, and assign grade & interest rates.

2) Investors can use the model to construct a diversified portfolio that maximizes returns given their risk tolerance.

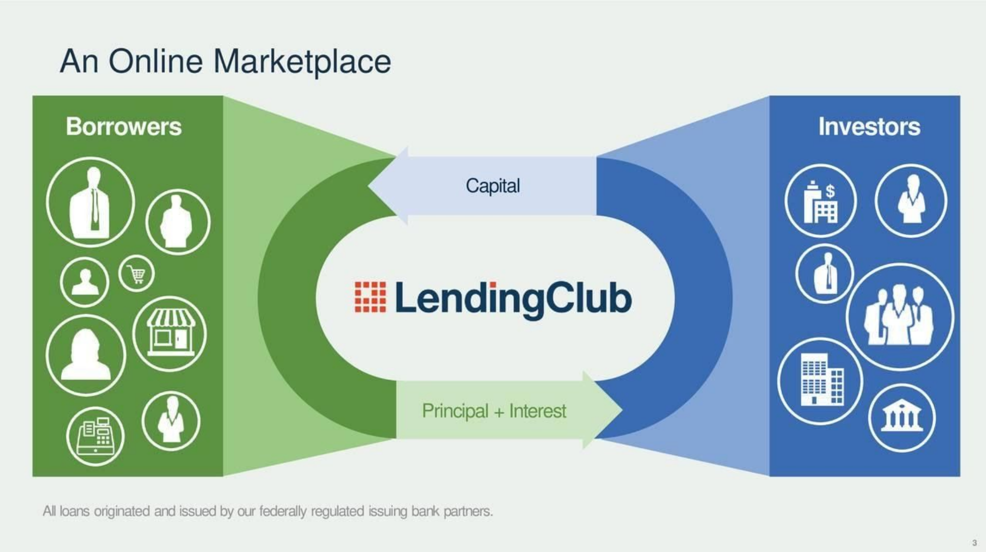
3) For companies with different business models that secturize and sell the loans (ABS or MBS), with some adjustments, models can be built to predict the same metrics for pricing and risk management purposes.

Problem Statement

Lending Club (LC) is an online marketplace where people can either borrow money for multiple funding purposes (for example to consolidate credit card debt) or lend money as investors to the former group. Based on the borrower's credit profile and term of the loan, LC decides if the loan is approved or not. And if approved, the loan will be assigned with a grade from the highest A1 to the lowest G5.

The higher the grade, the lower the interest rate and also lower probability of default. Investors can choose to invest in loans given their risk tolerances. Interest rates are higher for lower grade loans but investors risk losing the interests and principals if the loan defaults. Therefore, realized returns are determined by the interest rate, default rate, loss given default, and prepayment speed.

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? And these are the goals of this study:



An Online Marketplace

Borrowers

Investors

Capital

LendingClub

Principal + Interest

All loans originated and issued by our federally regulated issuing bank partners.

3

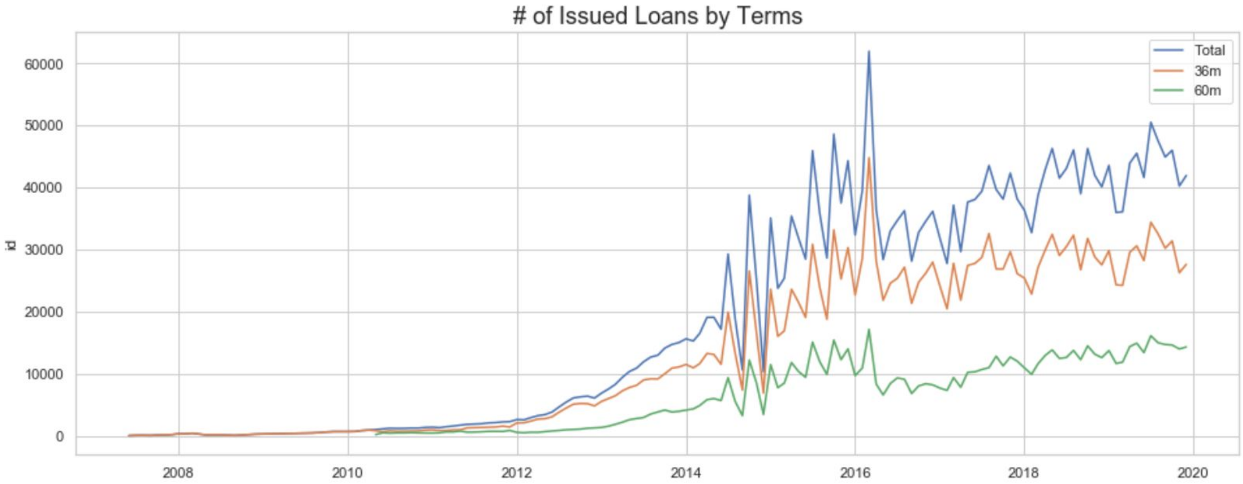
1) Identify important features known at the loan initiation date that can predict the default probability, loss given default, prepayment speed, and loan returns.
2) Predict the default probability and loss given default to get the expected loss.
3) Predict the prepayment speed.
4) Predict the returns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

Data Wrangling and Exploratory Data Analysis

Introduction

The data were downloaded from Lending Club and with 2.77 million loans and 150 features from 2007 to 2019. See the number of issued loans by terms in Figure 1.

Figure 1.



There are more loans with 36 months term than 60 months term (Figure 2) and interest rates tend to be higher for the longer term loans to reflect the higher risks (Figure 3). Note that loan initiations dropped severely in 2016 due to the infamous Lending Club scandal but slowly came back after it.

Figure 2.

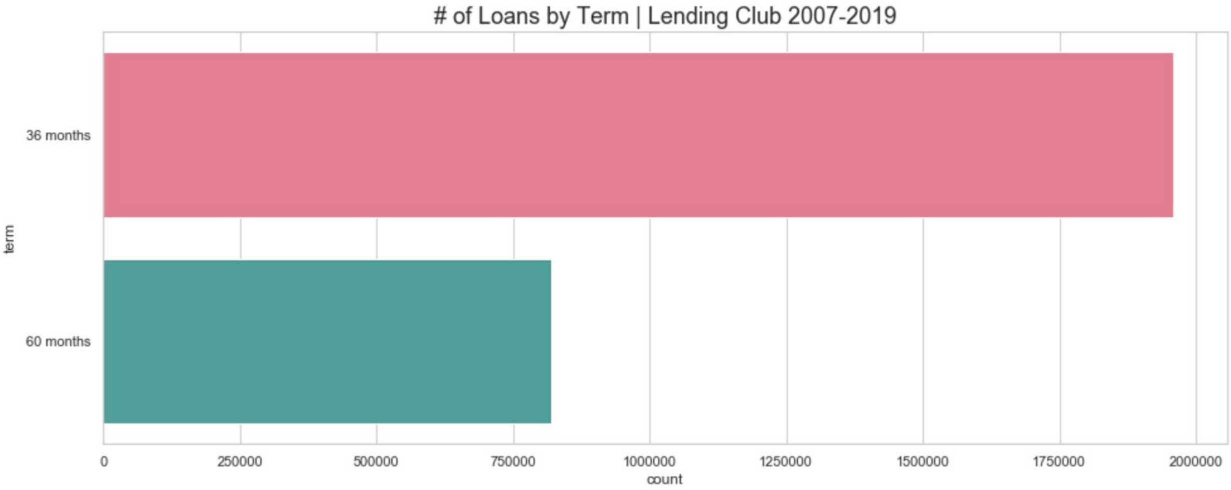
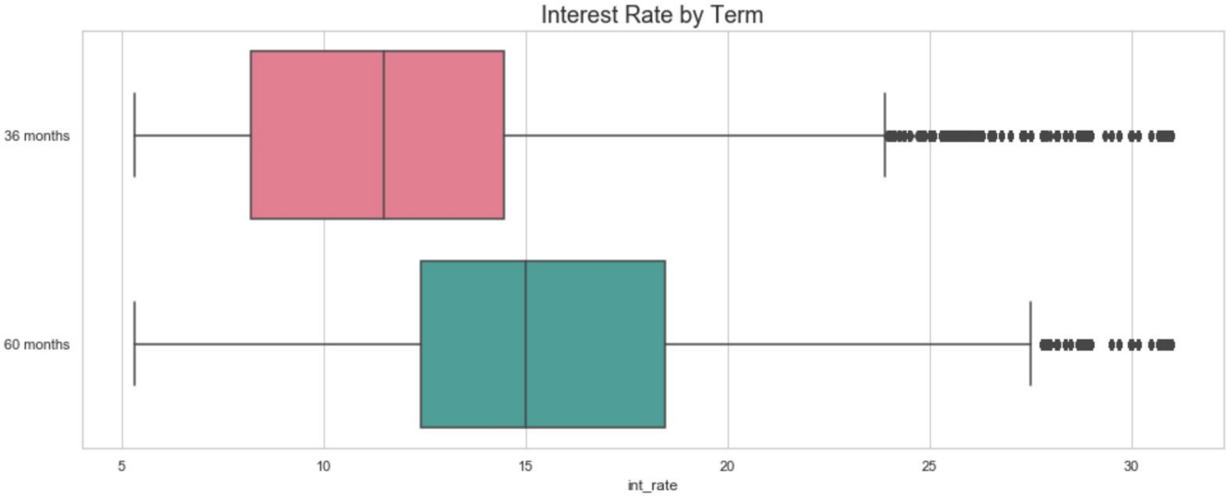


Figure 3.



There are 7 main tranches of loans with grades from A to G. The higher the grade, the lower the interest rates, expected return, expected loan losses and volatility. On the other hand, loans with lower grades have higher interest rates, expected returns, expected loan losses and volatility.

Figure 4 shows the distribution of funded loan amounts. It ranges from $1,000 to $40,000 with the median as $13,000. It's interesting to notice the peaks at rounded numbers like $10,000, $20,000 etc.

Figure 4.

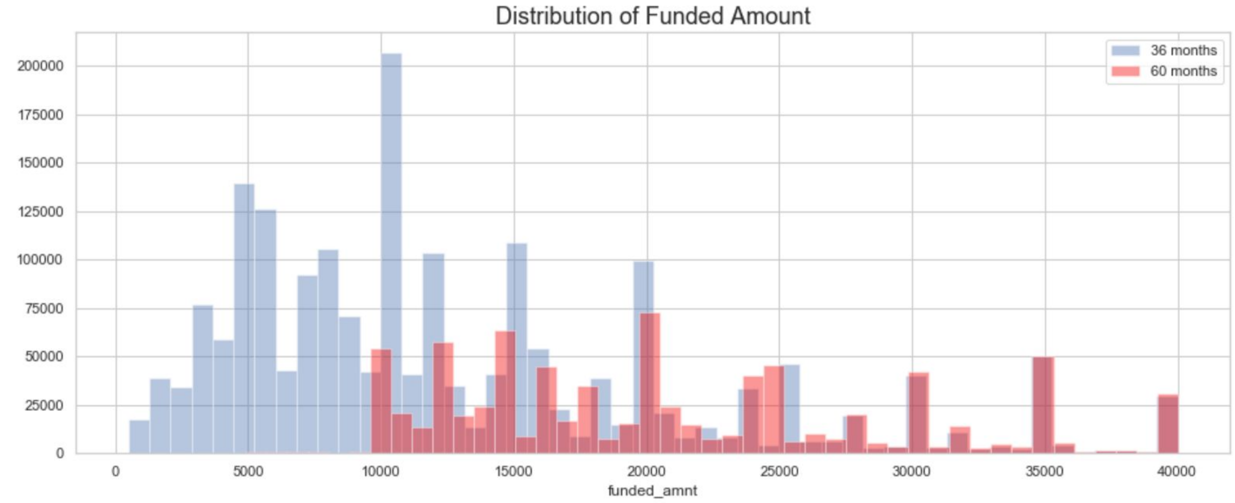
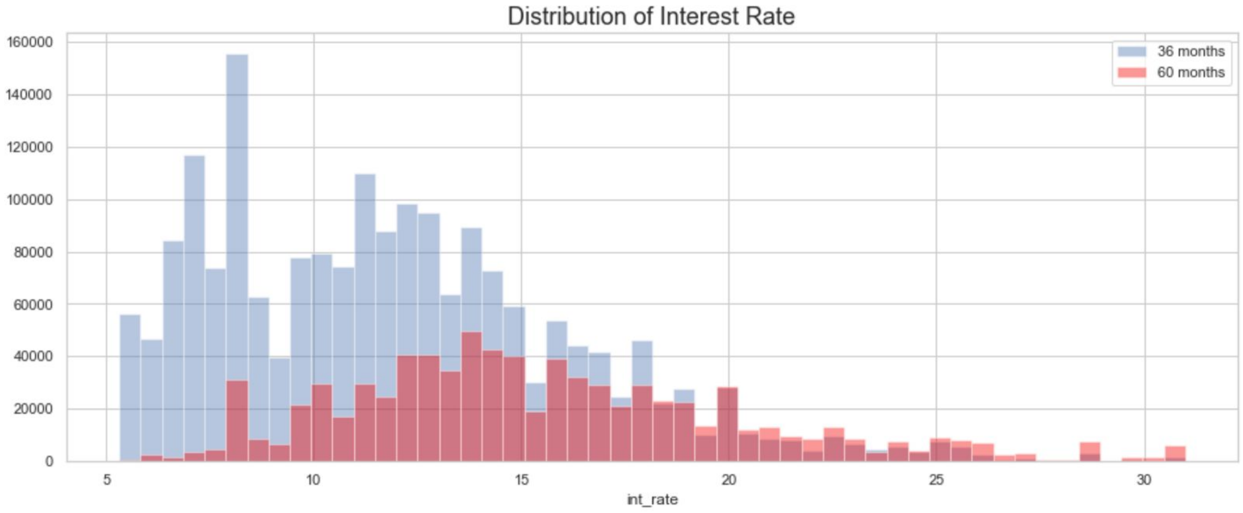


Figure 5 shows the distribution of interest rates ranging from 5.31% to 30.99% with a median of 12.7%. Interest rates on the longer term loans (60m) are generally higher than the shorter term loans to account for the higher risks.

Figure 5.

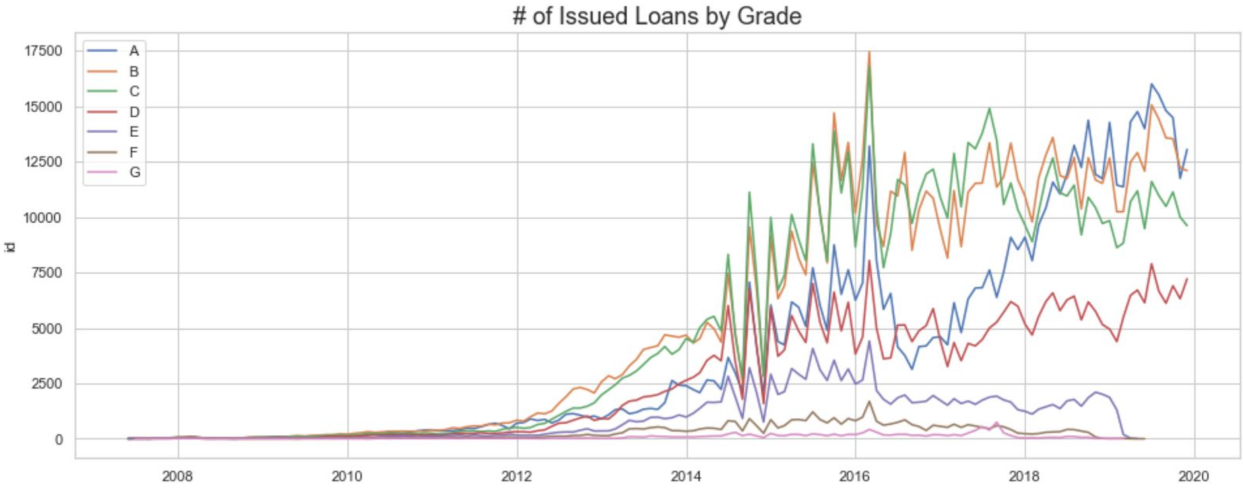


As shown in Figure 6, the most issued loans are of grade B and C in most of the time except recently grace A caught up after the 2016 scandal. Figure 7 shows the range of interest rates by grades and it's in line with our expectations.

Figure 8 shows the # of loans by loan purposes with **debt consolidation, credit card, and home improvement** as the top 3.

The top 5 states where loans initiated were: CA, TX, NY, FL, and IL as shown in Figure 9.

Figure 6.



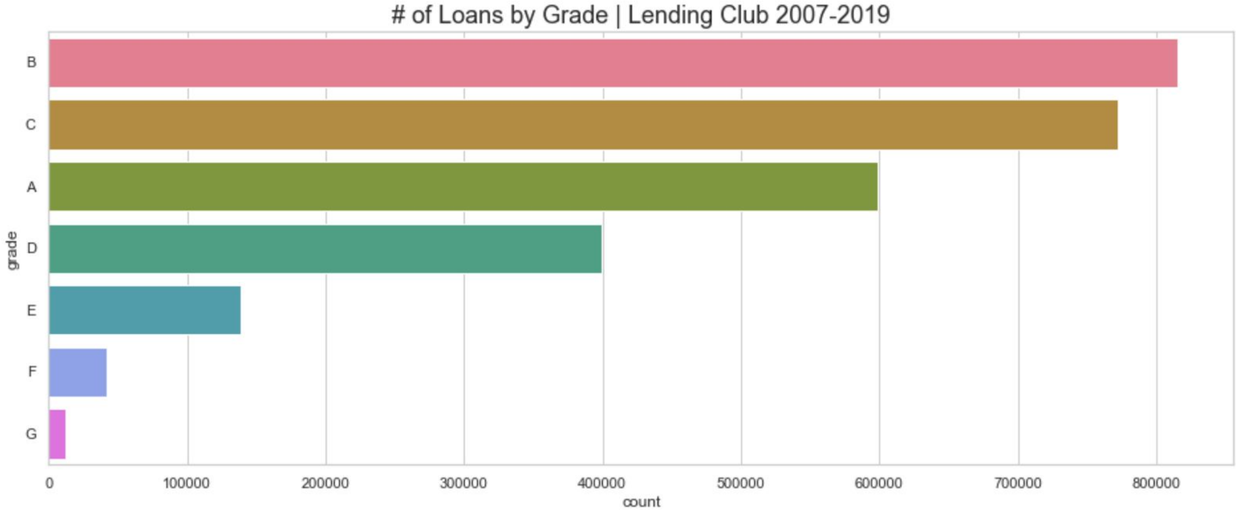


Figure 7.

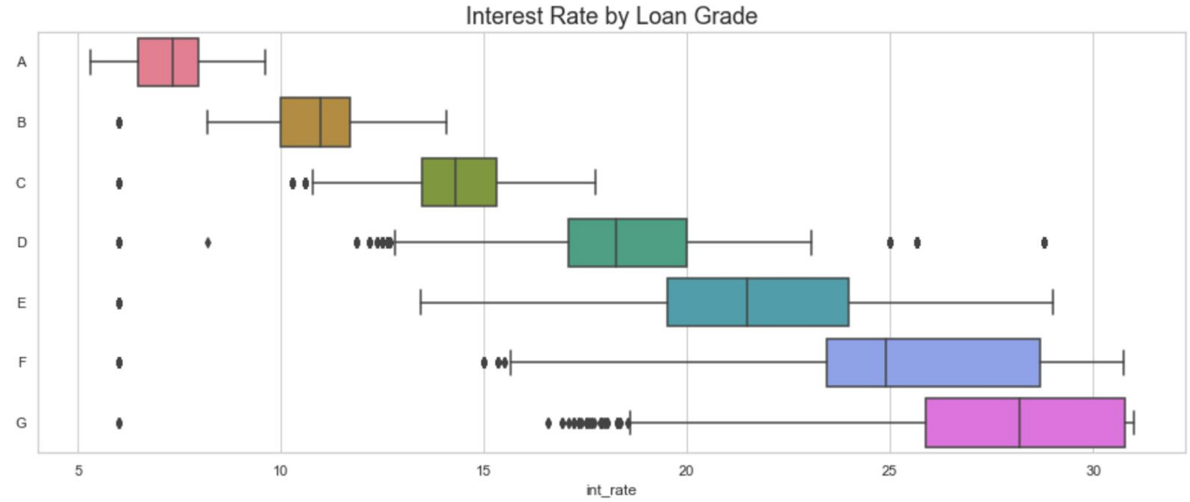


Figure 8.

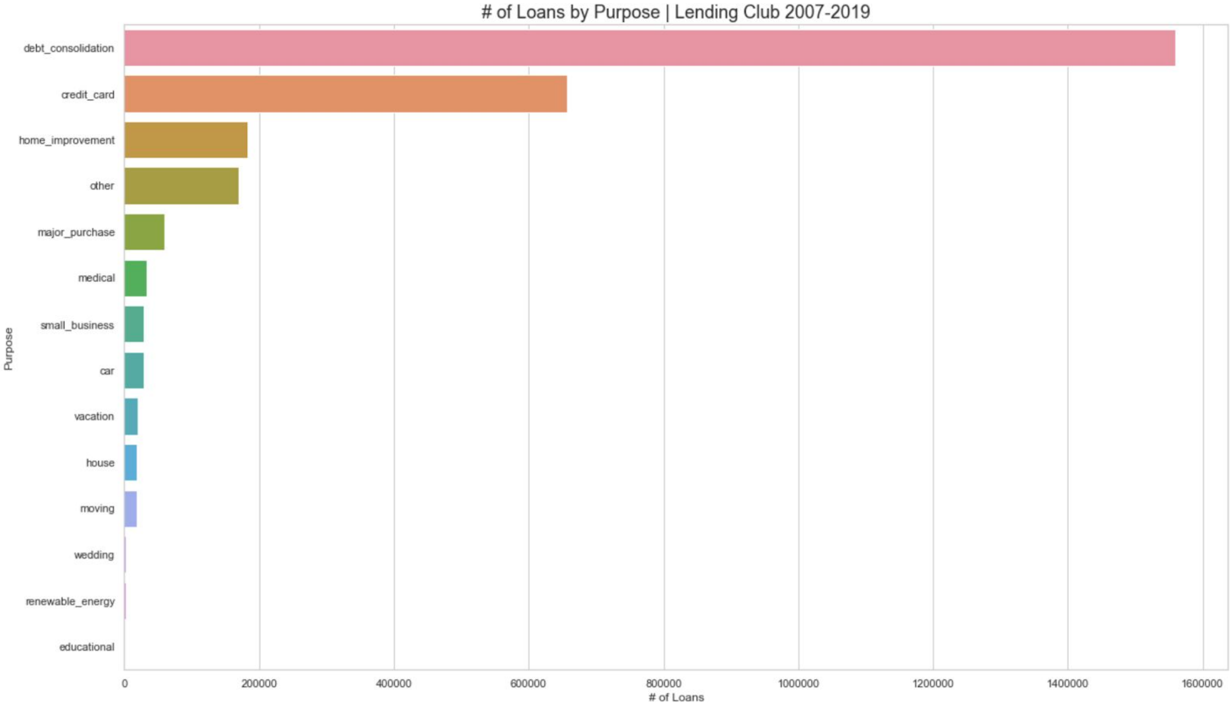
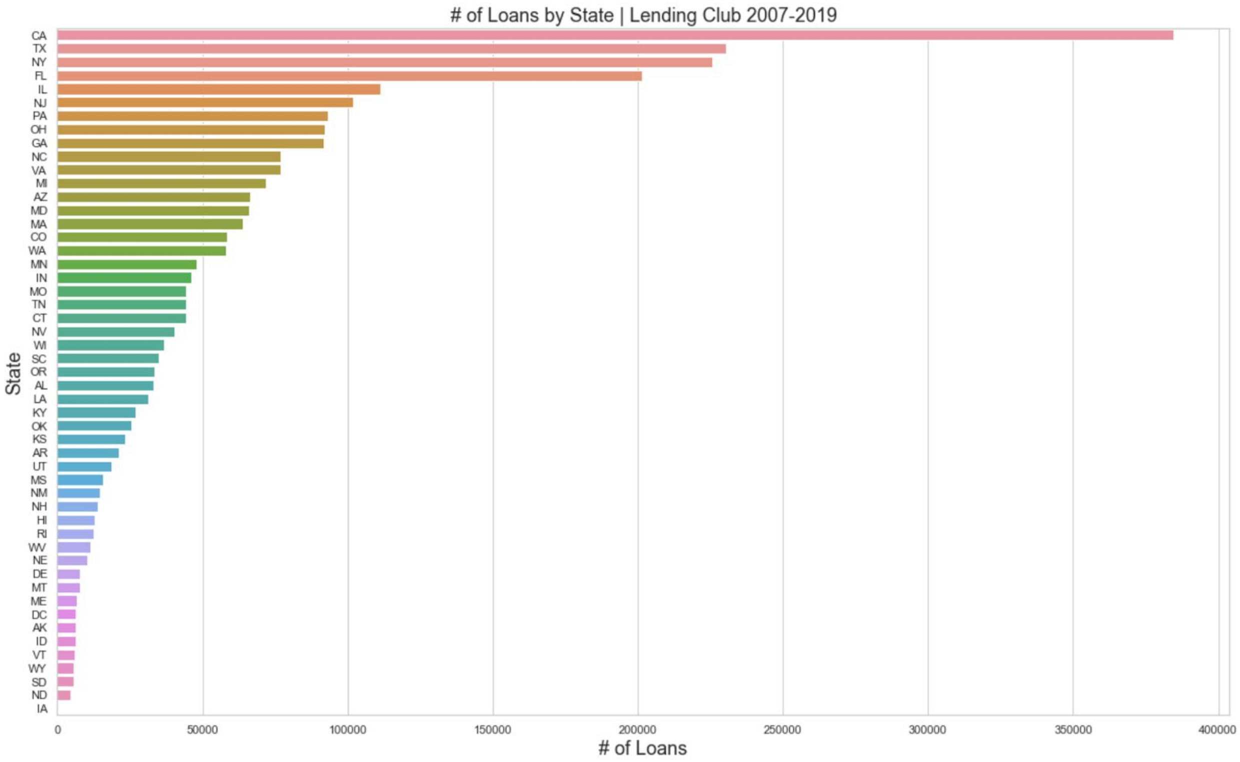
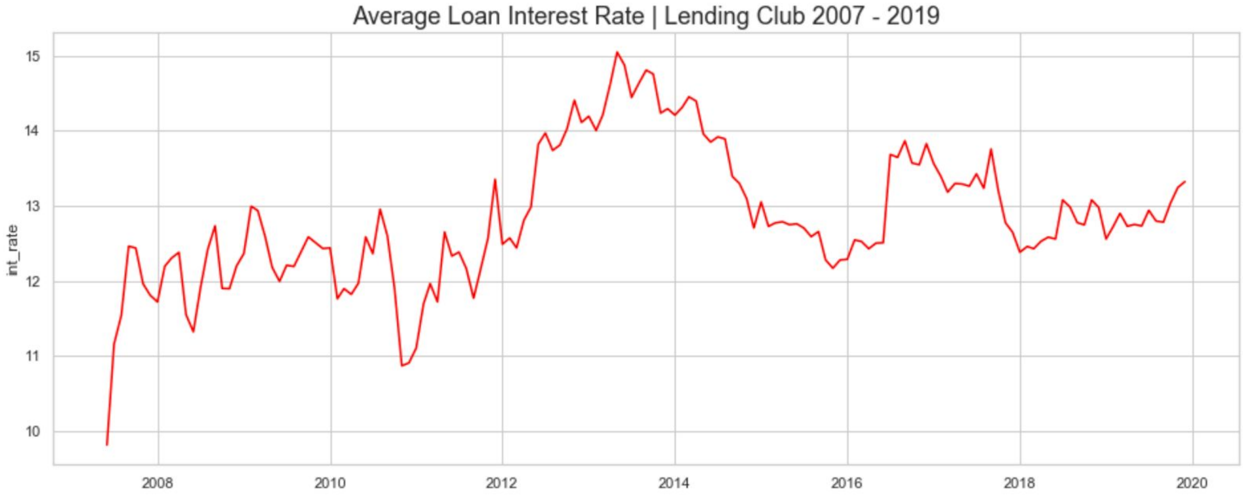


Figure 9.



Lending Club's loan interest rates take into account credit risk and market conditions. The time series of the average loan interest rate can be found in Figure 10.

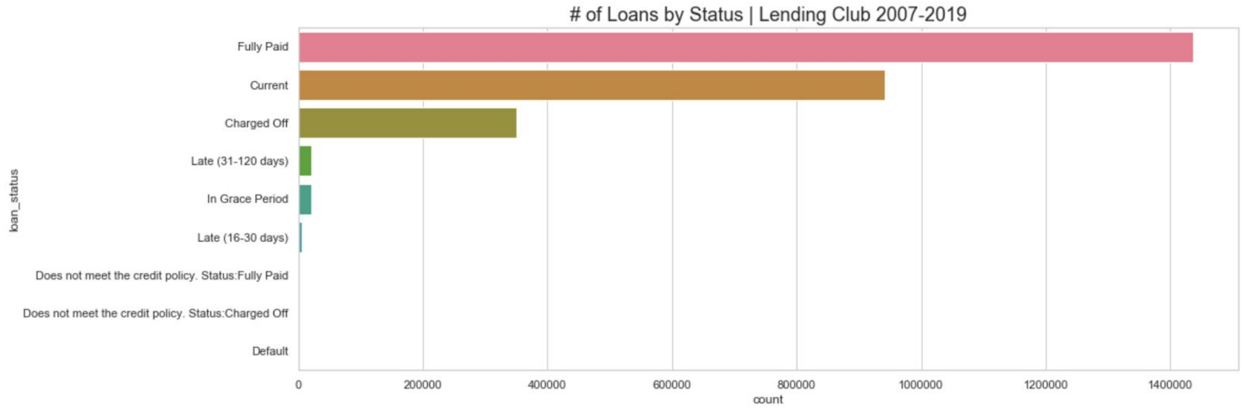
Figure 10.



There are 8 types of loans as shown in Figure 11 and we need to pay extra attention here since **we only want to include loans that are not current in the study** because we don't know if loans that are still current (including loans that are in late status) would end up as paid off or charged off in the end yet. Many previous works incorrectly included the current loans which leads to unrealistic results.

Note that loan type “default” are loans past due 121-150 days and thus are considered as “charged off”. Therefore, only loan types “paid off”, “charged off” and “default” were used in the analysis and the others were removed. **In the end there were around 1.7 million loans in the data with 20% charged off and 80% paid off**.

Figure 11.



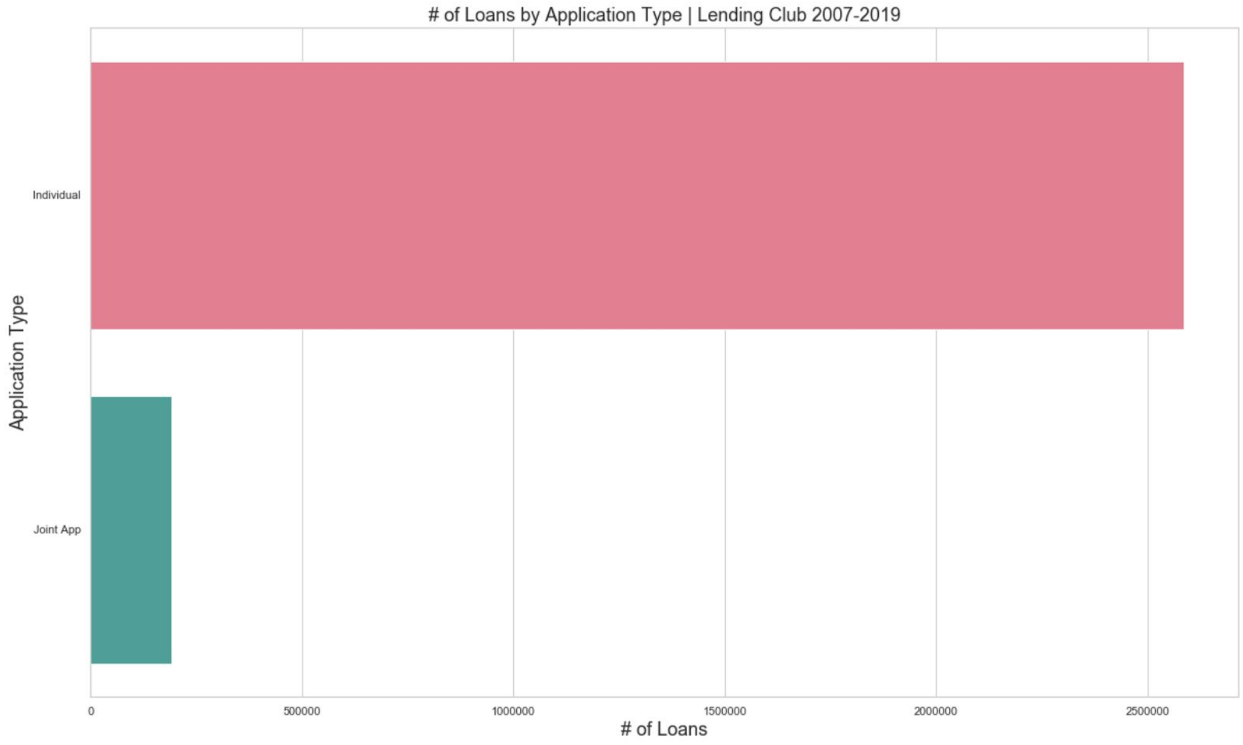
Data Cleaning

To prevent **data leakage** (variables that can not be seen at the time of making the prediction), the following features were removed: 'collection_recovery_fee', 'last_credit_pull_d', 'last_fico_range_high', 'last_fico_range_low', 'last_pymnt_amnt', 'last_pymnt_d', 'next_pymnt_d', 'num_tl_120dpd_2m', 'num_tl_30dpd', 'out_prncp', 'out_prncp_inv', 'recoveries', 'total_pymnt', 'total_pymnt_inv', 'total_rec_int','total_rec_late_fee', 'total_rec_prncp', 'debt_settlement_flag', 'debt_settlement_flag_date','settlement_status','settlement_date', 'settlement_amount', 'settlement_percentage', and 'settlement_term'.

Also, the following features were removed because of overlapping (but inferior) information or information cannot be easily extracted: 'emp_title', 'loan_amnt', 'funded_amnt_inv', 'grade', 'id','member_id', 'title','url', 'zip_code', 'pymnt_plan', 'policy_code'.

There are two types of loan applications: single applicant or joint applicants with the former the majority (see Figure 12). If it’s joint applications, their total annual income, dti, verification status, revolving balance and other credit related features were substituted with the joint numbers. This is another often overlooked data cleaning step and is addressed in this study.

Figure 12.



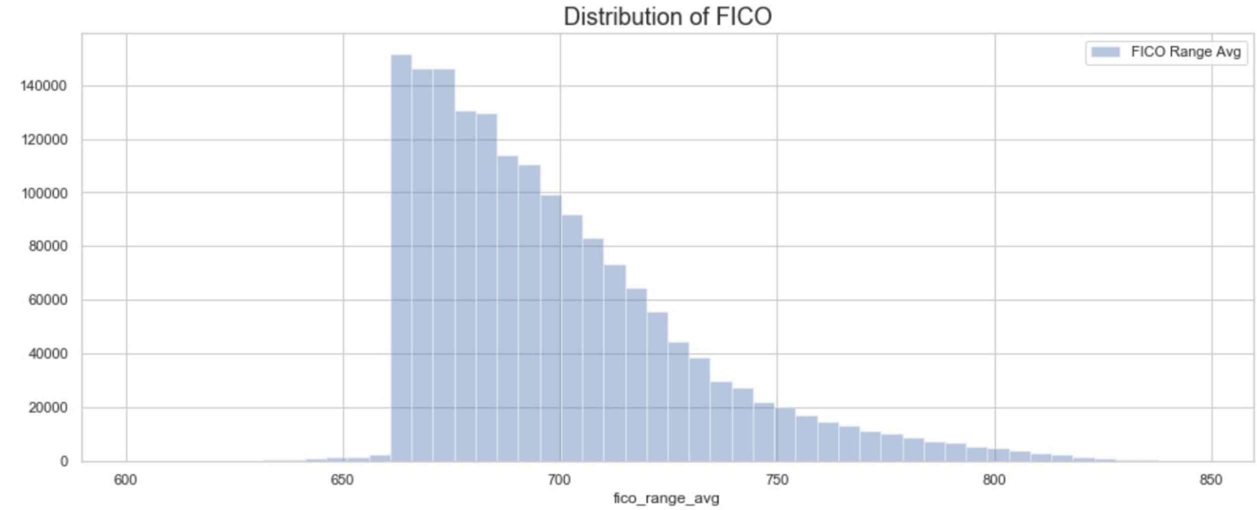
Feature Engineering

Dummy variables were created for the following categorical features: term, sub_grade, home_ownership, verification_status, loan purpose, state, initial_list_status, and application type. For home ownership types, “ANY” and “MORE” were grouped into “OTHER”. Employment length was converted to numerical values by using the corresponding number of years with the exceptions of “<1 year” assigned to 0.5 and “10+ years” assigned to 15. Also, state dummy variables with less than 10000 loans were dropped.

A few new features were created: 1) the length of the loan description provided by the applicants; 2) the history of credit line (the # of months between the earliest credit line date and the loan initiation date); 3) the 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. This is a variant of the DTI but focuses on this particular LC loan.

Figure 13 shows the distribution of the average FICO score and suggests there might be some hard limits (at around 620) used by LC for loan approvals.

Figure 13.



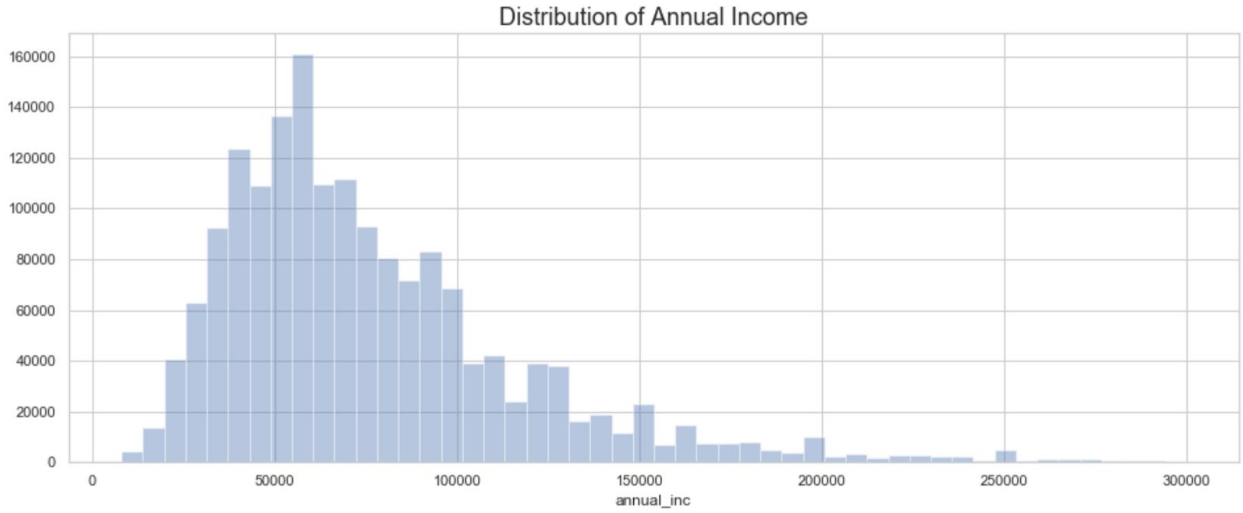
Missing Values and Outliers

For variables with missing values, it's important to learn the underlying reason and what it means by "NA" for the feature. This is another step that many previous works missed. The NAs for features like mths_since_recent_inq means there was never a delinquency and thus is replaced with 2 * maximum (for this feature, the higher the better). Some previous works mistakenly removed these valuable features or replaced them with 0 (which is the opposite of reality).

For other features like dti, employment length etc. the missing values were filled with the median. There are also some new features that were added after 2015 by LC for example open_acc_6m and thus were discarded because filling pre-2015 NAs with post-2015 values doesn't make much sense. I did a sanity check in a separate sample just using 2015-2019 data and these new variables didn't add much value so it's safe to remove them entirely in this study.

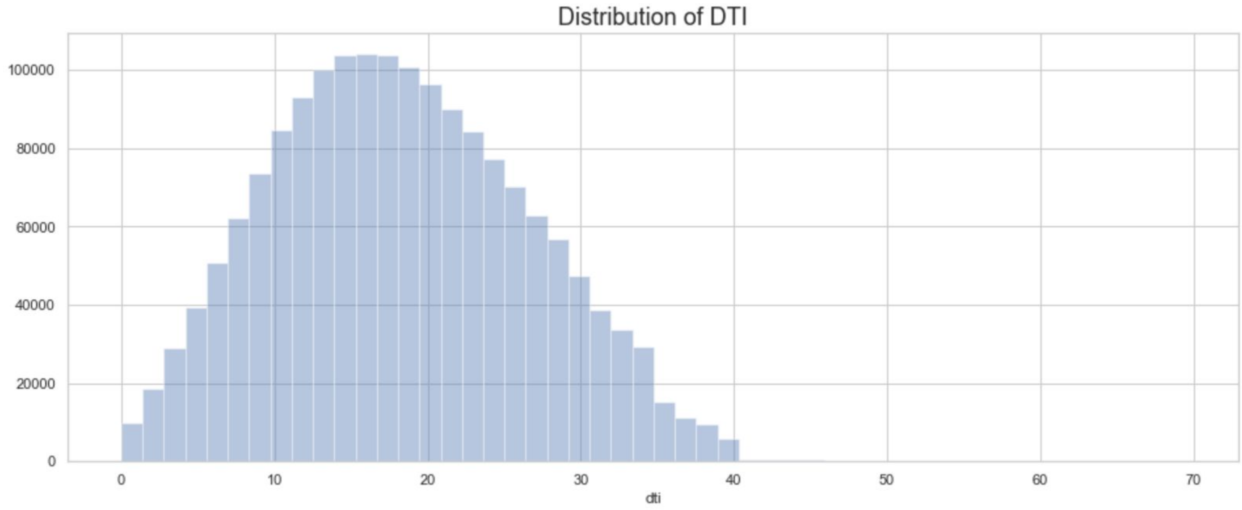
There were some extreme values for annual income with the lowest being $2000 (per year) which might be because applicants accidentally thought it's monthly income. The highest income was $110 million which you may wonder why billionaires need to borrow micro loans on Lending Club and so it's very likely the numbers were made up to get the loan approved and LC didn't catch and/or correct them. I removed outliers with annual income values below $8000 or above $300K which accounts for less than 1% of the total sample size. The distribution of annual income can be seen in Figure 14 with the median around $65K and is skewed to the right.

Figure 14.



Also, observations with negative DTI were also removed and the distribution can be seen in Figure 15 with most DTIs from 0 to 40%.

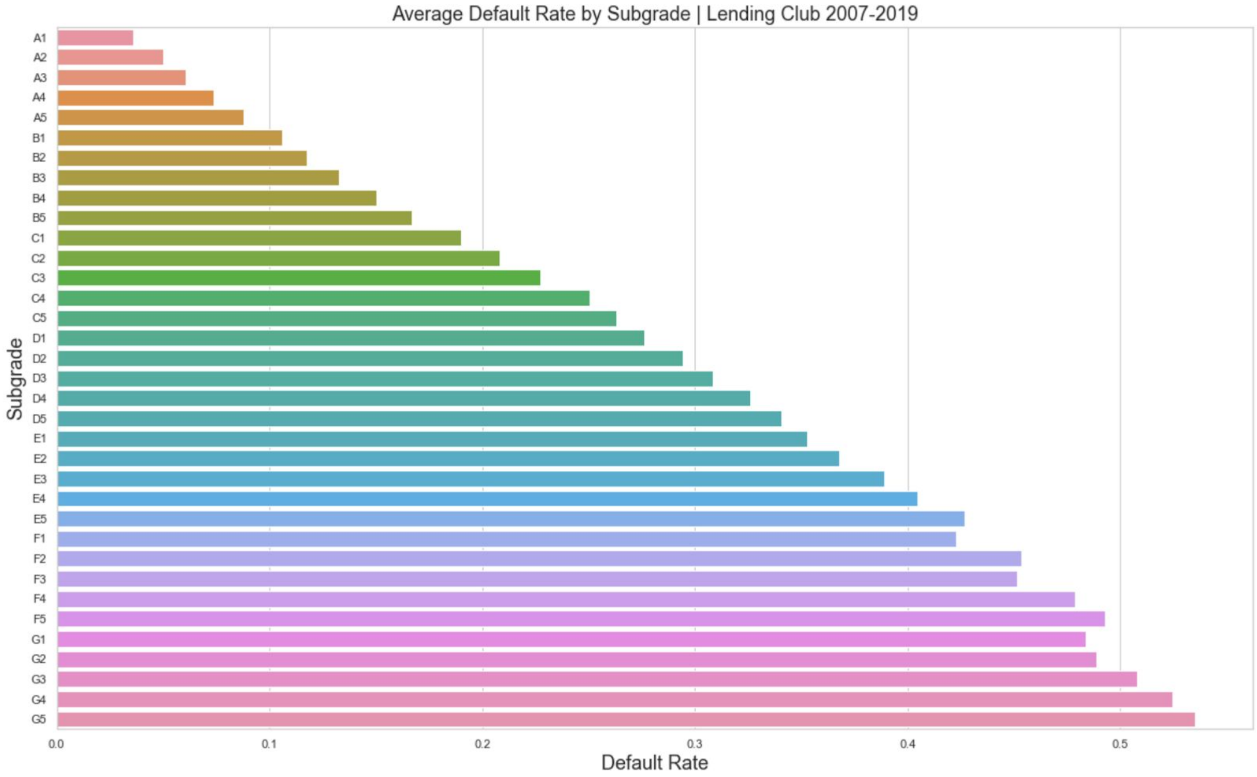
Figure 15.



Indepth EDA Pt.1: Default Prediction

After cleaning the data, let’s see the default rate grouped by the sub grade as in Figure 16. Generally the default rate increases from 4% for A1 to 55% for G5. Big differences here! With the default rate more than 50%, the actual realized returns for the “marketed” high yields of the low grade junk bonds could be easily cut in half.

Figure 16.



The higher default rate for the 60 month term loans is well in line with the higher interest rate and higher risk for longer maturity (see Figure 17). So far so good!

Interestingly, as shown in Figure 18, the average default rate for join applicants is higher than single applicants probably due to self selection reasons - applicants with lower credit tend to file joint applications to increase the chance of approval.

Figure 17.

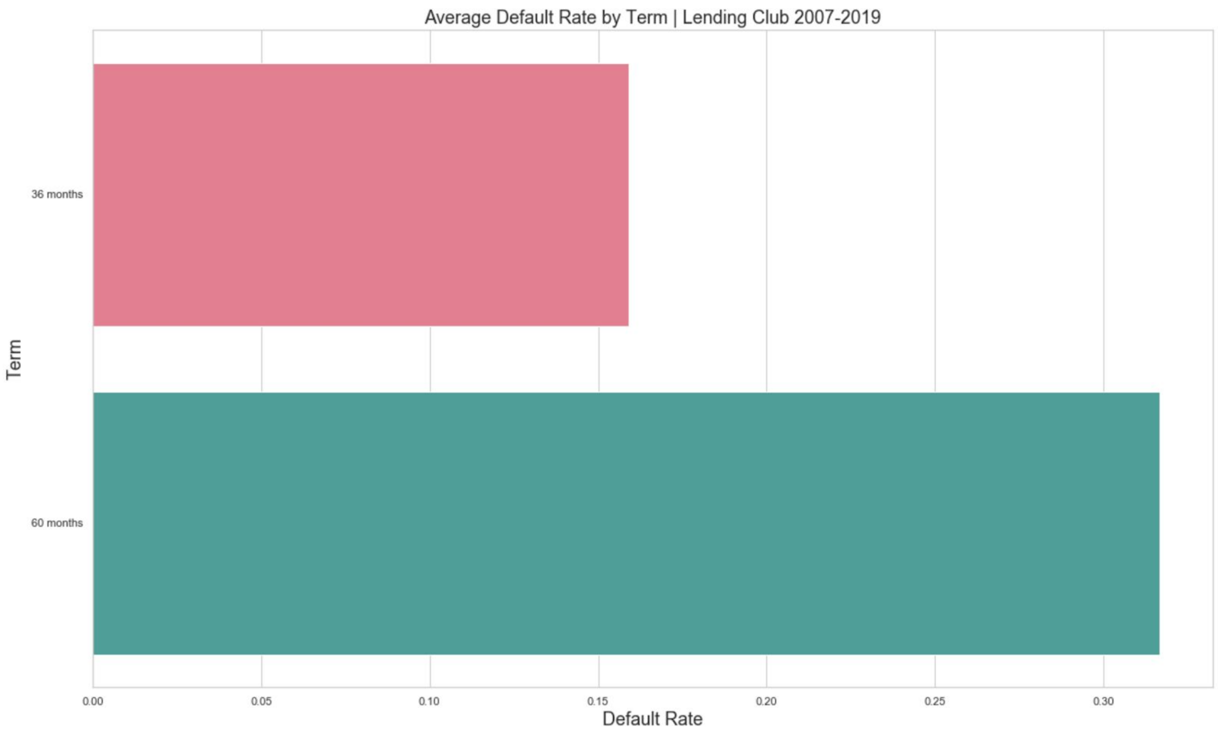


Figure 18.

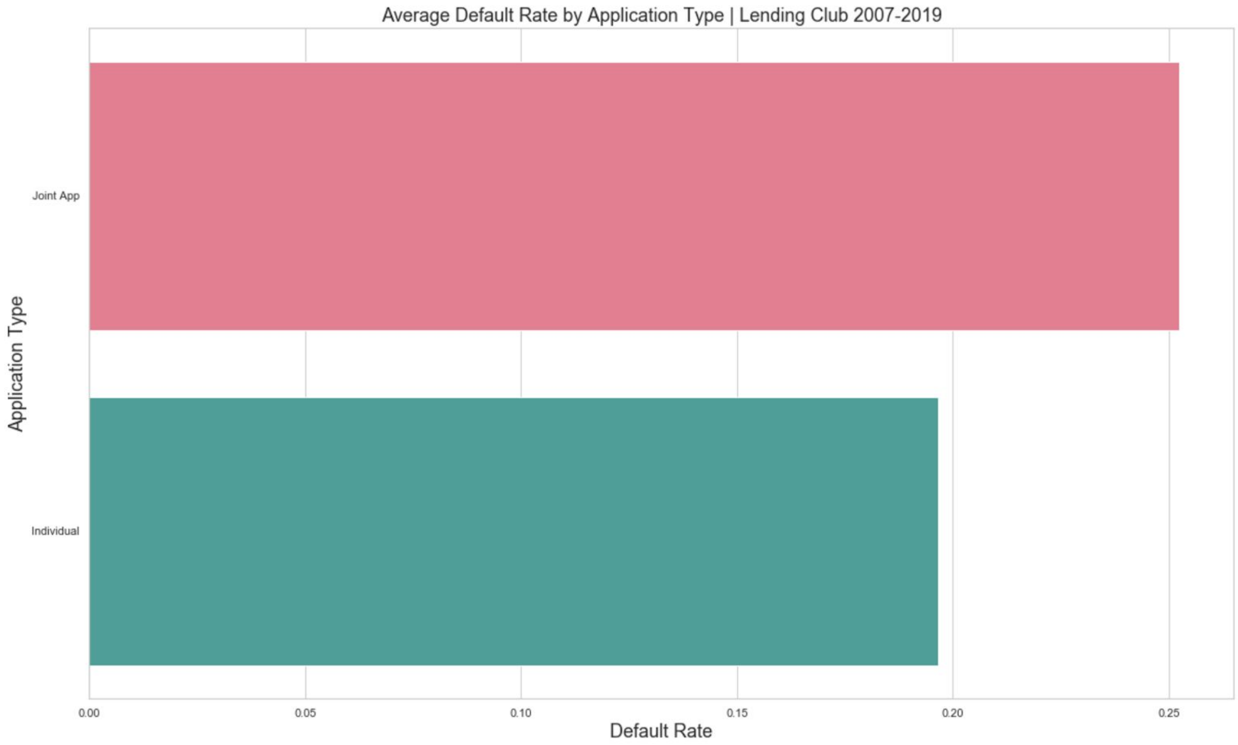


Figure 19 is another interesting one. The default rate for applicants with their income verified by LC is higher than not verified. This is probably exactly why LC verified them in the first place!

Figure 19.

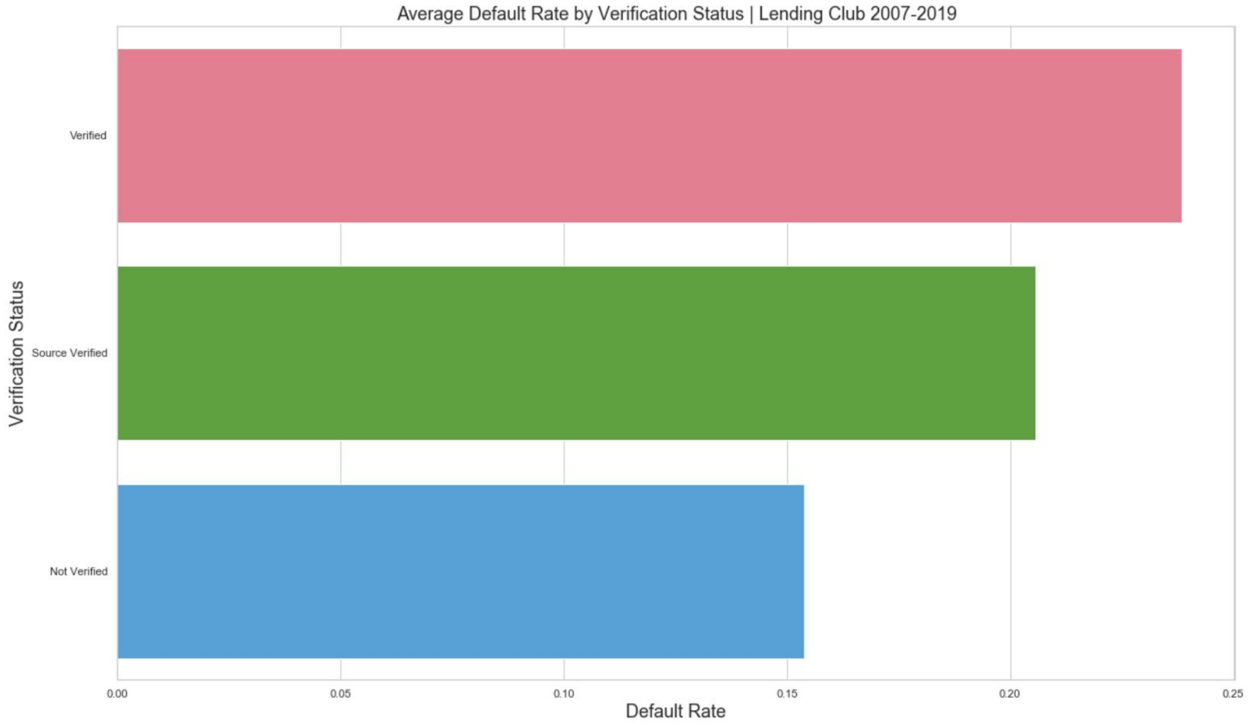


Figure 20 sorts the average default rate by loan purposes. It suggests starting up small businesses are indeed the riskiest investment while wedding and education tend to have lower default rates! Note that the majority loan purposes are debt consolidation and credit cards so it might be hard to implement what we learned from this chart in practice.

Figure 20.

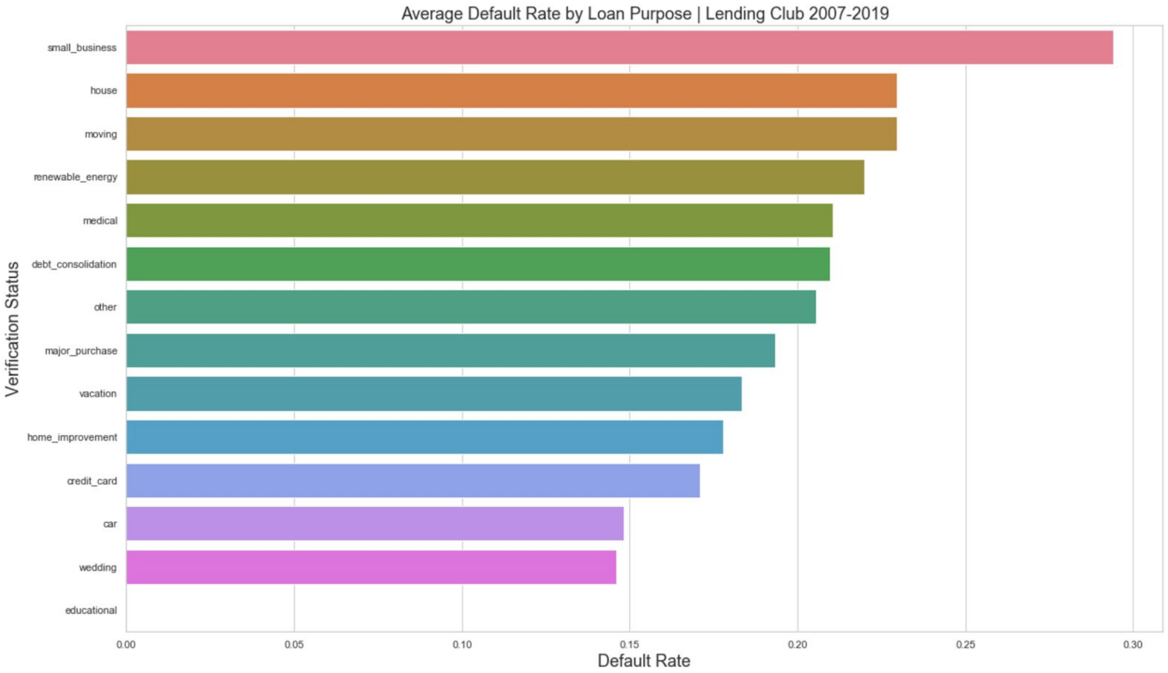


Figure 21 suggests the default rate for people who rent is on average higher than people who own which is highly correlated with income and credit history.

Figure 21.

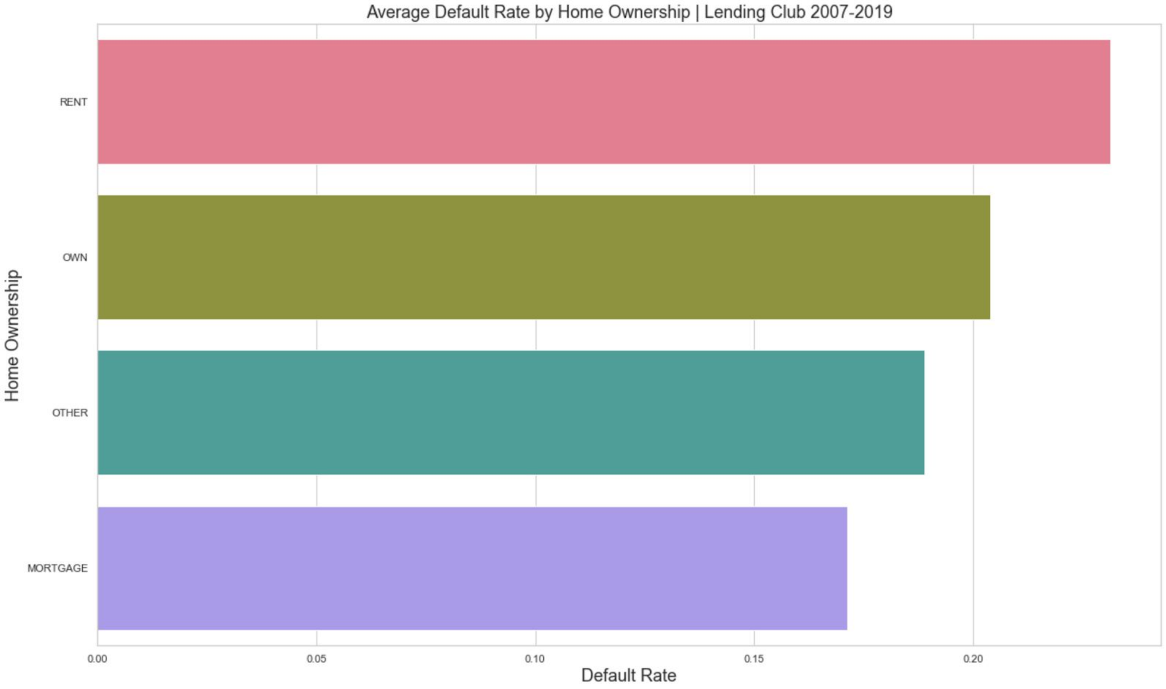
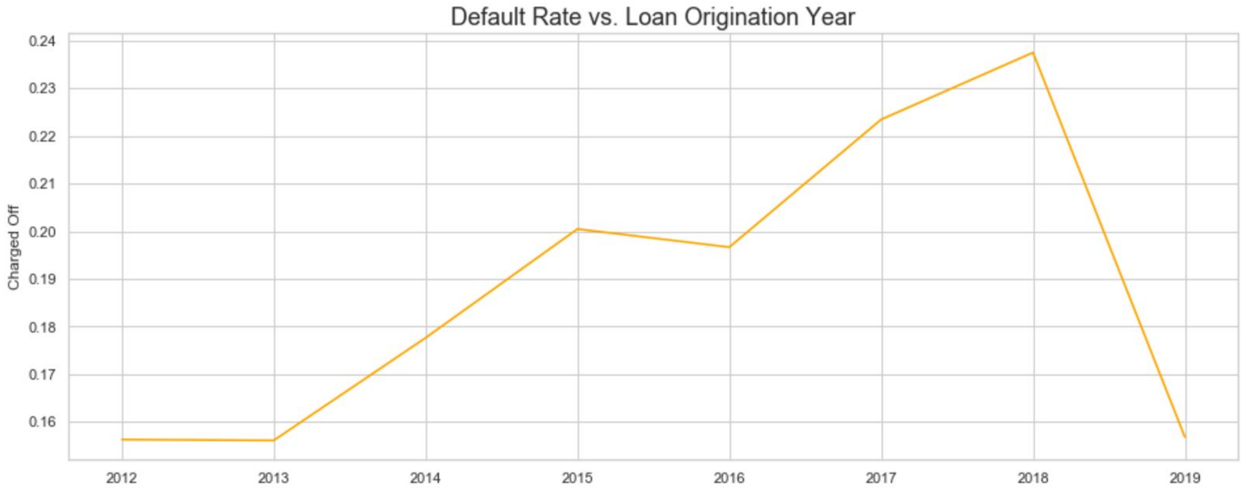


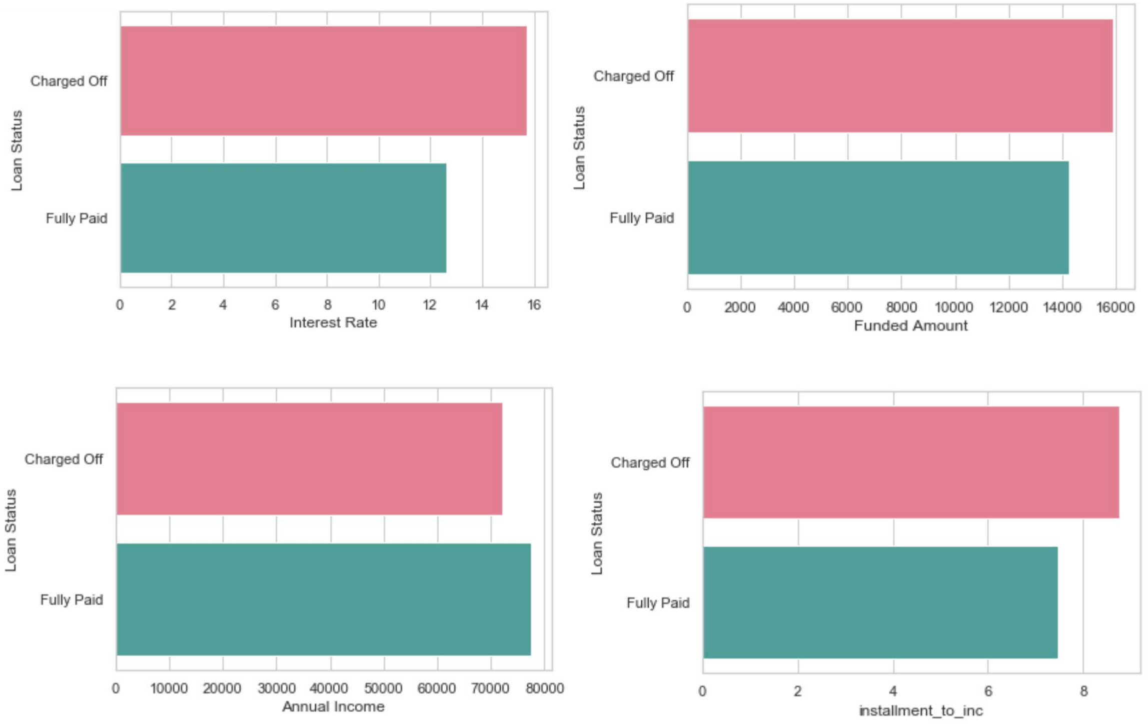
Figure 22 shows the default rate in time series from 2012 to 2019. Notice the soaring default rate around the scandal year of 2016.

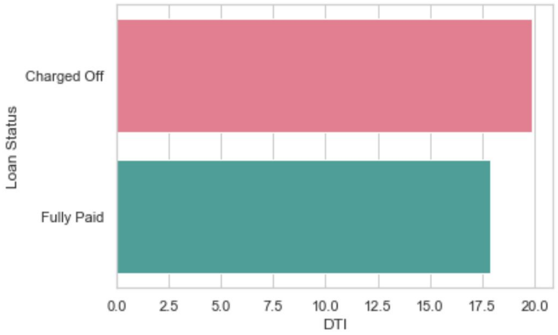
Figure 22.

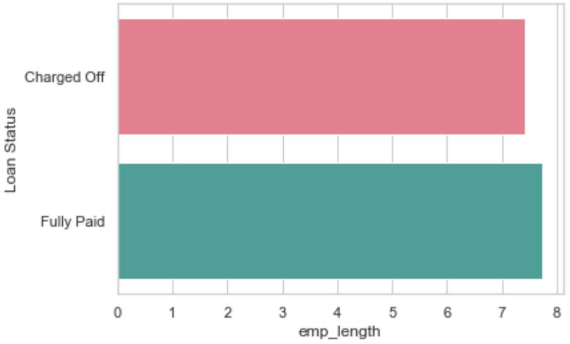


What about features with numeric values? Figure 23 suggests the average default rate is higher for higher interest rate, higher loan amount, lower annual income, higher installment to income ratio, higher DTI, and lower employment length. All makes sense.

Figure 23.







x`

Machine Learning Pt.1: Default Prediction

Goal

The goal of the in-depth analysis is, given the features of loan characteristics (for example terms, interest rate, sub-grades) and features of applicant's profiles (income, DTI, employment length, FICO etc.) to predict the loan default probability.

The target variable is the loan status for all non-current loans with 1 being charged off and 0 being paid off.

Data Preparation

Multicollinearity would be a big issue for logistic regression. It's not necessarily a problem for a random forest model but still reduces the feature importance interpretations. To address this, a correlation matrix was obtained and features with correlations more than 95% were removed including: 'num_sats', 'num_rev_tl_bal_gt_0', 'tot_hi_cred_lim', 'installment'.

Since only 20% of the loans are charged off, this is a typical **imbalanced data** and resampling was done to reduce the imbalance. This is a crucial step, otherwise the predicted default probability will be biased towards 0 and using a threshold of 0.5 will be too high. Two methods were tested and yielded similar results: 1) under-sampling (randomly selecting paid off loans so the number of paid off loans is the same as charged off loans) and 2) over-sampling (randomly sampling with replacement in the charged off loans to have the same number of charged off loans as the paid off loans). For simplicity, under-sampling results are shown here with 676,734 loans in total and 50% paid off and 50% charged off loans divided equally.

A **70%/30% train and test split with stratification** approach was used for model training and testing. I also tested a different approach which was using the data from 2012 to 2017 for training and then testing on 2018-2019 data which is more in line with reality. The results were similar to the 70%/30% split approach. And since the 2018-2019 period is closer to the current date and many 60m loans are still current as of now, the 2nd approach will have the testing data dominated by shorter term loans and thus was not used in the end.

Before running logistic regressions, one of the categories need to be dropped for each category to avoid perfect correlation of features. Also, features were scaled for easier coefficient interpretations which captures the effect of one unit change in the feature on the log odds of “success”.

Logistic Regression (with regularization)

Since we have 152 features, we need to do some feature selections to reduce the list and thus **recursive feature elimination (RFE)** was used as the first step. It’s based on the idea to repeatedly construct a model to choose the best features recursively. Top 30 features were selected by the ranking to run a first round of logistic regression. And then features with p-values higher than 0.05 were removed to further reduce the list of features.

Figure 24 lists the features selected in the end and their coefficients. To interpret the coefficients, **the 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. It’s a good start.

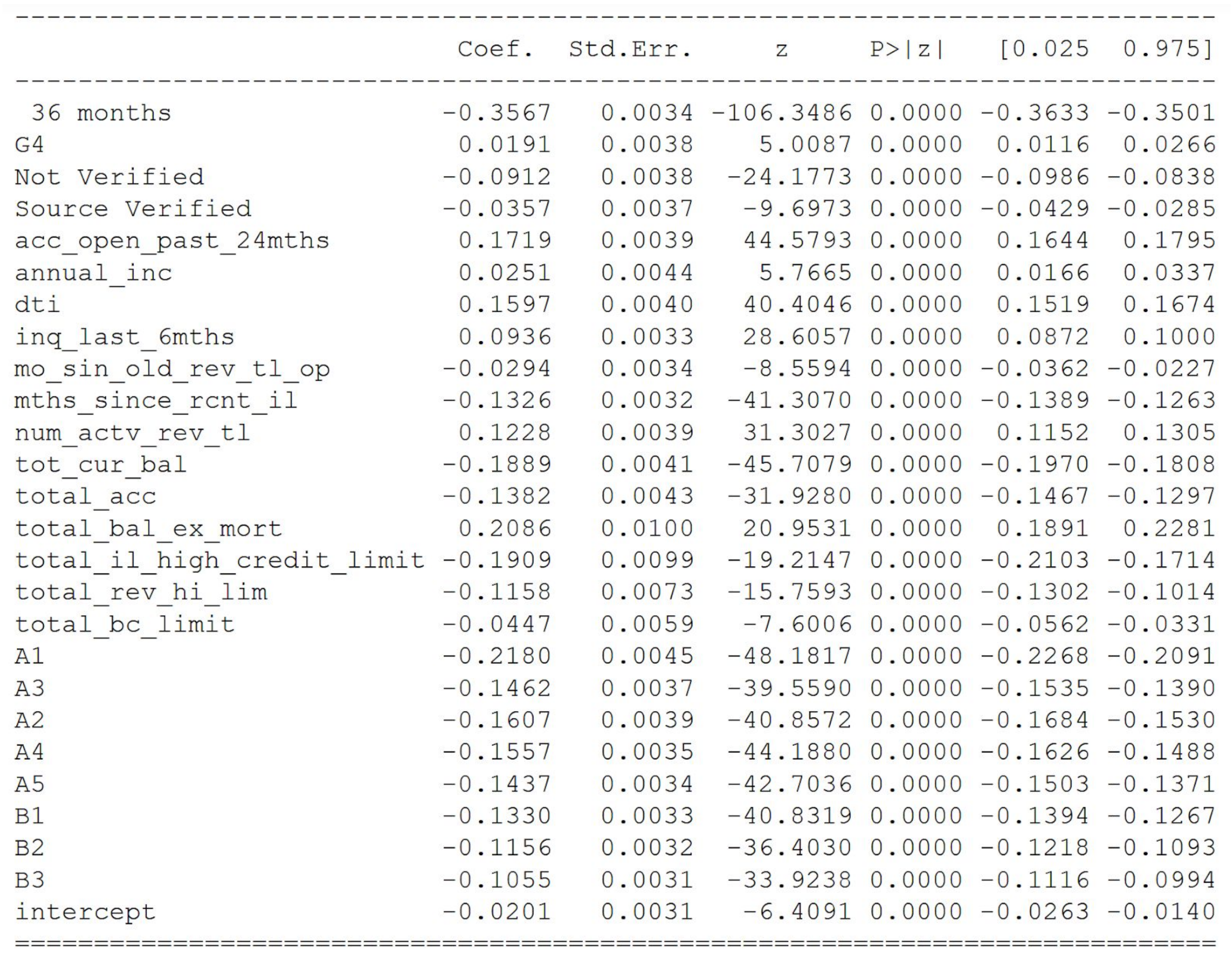
Decision Tree and Random Forest

A random forest classifier model with n_estimator of 200 and max_depth of 10 was trained and tested with an AUC of 0.66 which is marginally better than the logistic regression. The most important features include: interest rate, loan terms, FICO, installment to income ratio, etc which are quite reasonable. The decision tree model, on the other hand, did poorly with an AUC of 0.57. This suggests the wisdom of the crowd nature of the random forest model helps reduce the variance tremendously.

Boosting (AdaBoost and XGBoost)

Boosting tries to improve the prediction power by training a sequence of weak models into a strong one. There are two types of boosting. AdaBoost identifies miss-classified data points and increases their weights so the next classifier will pay extra attention. Gradient boosting fits the new predictor to the residual error made by the previous predictor and has strong prediction power. The drawback is it’s hard to parallelize it because it’s sequential. And this is where XGBoost comes and shines! It’s scalable, portable, and distributed.

Figure 24.



```
-------------------------------------------------------------------------------
                             Coef.  Std.Err.      z      P>|z|   [0.025  0.975]
-------------------------------------------------------------------------------
 36 months                  -0.3567   0.0034 -106.3486 0.0000 -0.3633 -0.3501
G4                           0.0191   0.0038    5.0087 0.0000  0.0116  0.0266
Not Verified                -0.0912   0.0038  -24.1773 0.0000 -0.0986 -0.0838
Source Verified             -0.0357   0.0037   -9.6973 0.0000 -0.0429 -0.0285
acc_open_past_24mths         0.1719   0.0039   44.5793 0.0000  0.1644  0.1795
annual_inc                   0.0251   0.0044    5.7665 0.0000  0.0166  0.0337
dti                          0.1597   0.0040   40.4046 0.0000  0.1519  0.1674
inq_last_6mths               0.0936   0.0033   28.6057 0.0000  0.0872  0.1000
mo_sin_old_rev_tl_op        -0.0294   0.0034   -8.5594 0.0000 -0.0362 -0.0227
mths_since_rcnt_il          -0.1326   0.0032  -41.3070 0.0000 -0.1389 -0.1263
num_actv_rev_tl              0.1228   0.0039   31.3027 0.0000  0.1152  0.1305
tot_cur_bal                 -0.1889   0.0041  -45.7079 0.0000 -0.1970 -0.1808
total_acc                   -0.1382   0.0043  -31.9280 0.0000 -0.1467 -0.1297
total_bal_ex_mort            0.2086   0.0100   20.9531 0.0000  0.1891  0.2281
total_il_high_credit_limit -0.1909   0.0099  -19.2147 0.0000 -0.2103 -0.1714
total_rev_hi_lim            -0.1158   0.0073  -15.7593 0.0000 -0.1302 -0.1014
total_bc_limit              -0.0447   0.0059   -7.6006 0.0000 -0.0562 -0.0331
A1                          -0.2180   0.0045  -48.1817 0.0000 -0.2268 -0.2091
A3                          -0.1462   0.0037  -39.5590 0.0000 -0.1535 -0.1390
A2                          -0.1607   0.0039  -40.8572 0.0000 -0.1684 -0.1530
A4                          -0.1557   0.0035  -44.1880 0.0000 -0.1626 -0.1488
A5                          -0.1437   0.0034  -42.7036 0.0000 -0.1503 -0.1371
B1                          -0.1330   0.0033  -40.8319 0.0000 -0.1394 -0.1267
B2                          -0.1156   0.0032  -36.4030 0.0000 -0.1218 -0.1093
B3                          -0.1055   0.0031  -33.9238 0.0000 -0.1116 -0.0994
intercept                   -0.0201   0.0031   -6.4091 0.0000 -0.0263 -0.0140
===============================================================================
```

AUC is 0.66 for AdaBoost. AUC is 0.67 for XGBoost with n_estimator of 200 and max_depth of 3. This is a winner! Figure 25 compares the AUCs for different models. Figure 26 lists **the most important features in the XGBoost model including interest rates, DTI, funded amount, installment to income, revolving balance** etc. which makes sense.

The XGBoost model yields an accuracy score of 0.67, precision 0.66, recall 0.68, and F1 score of 0.67. The confusion matrix can be found in Figure 27. Randomized search on hyper parameters with cross validation was tested and the results didn’t improve much and thus are not shown in here.

Figure 25.

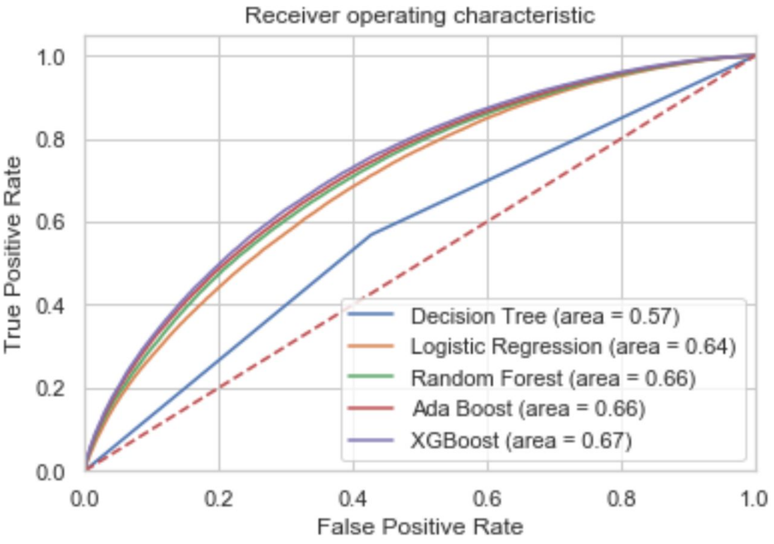


Figure 26.

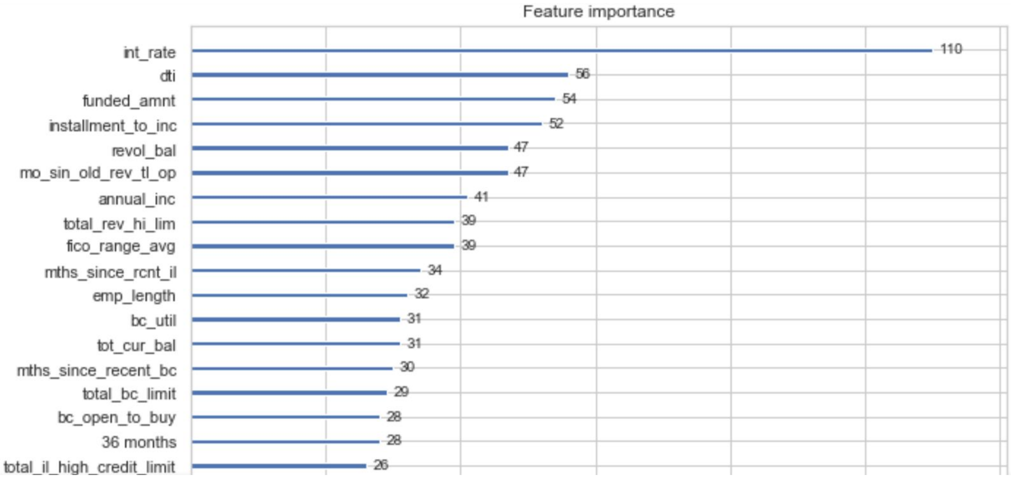
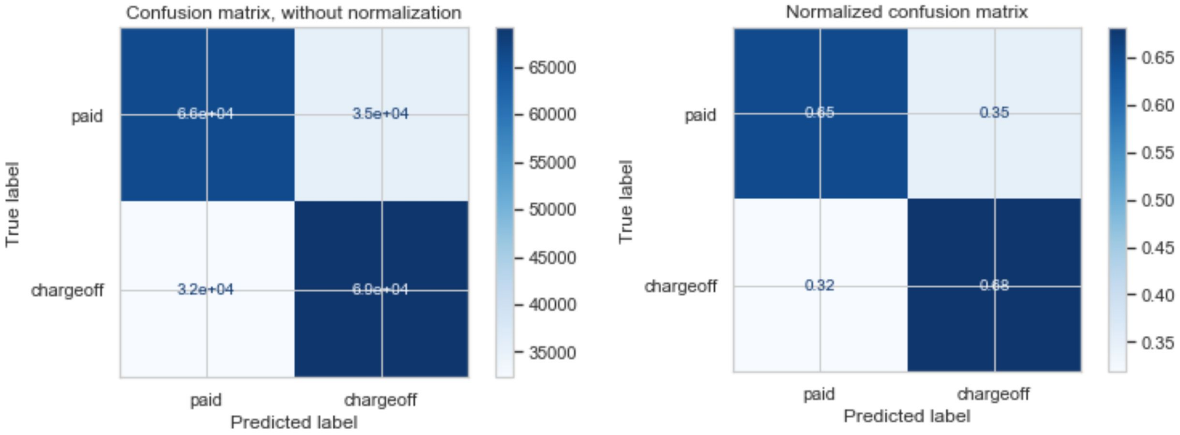


Figure 27.

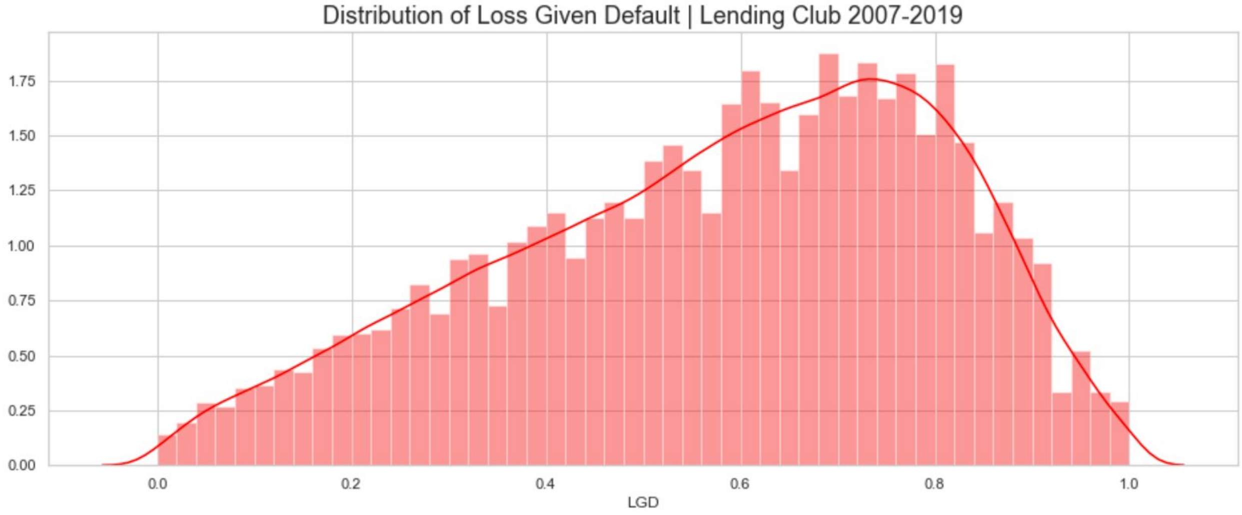


Indepth EDA Pt.2: Loss Given Default (LGD) Prediction

In Pt.1 we analyze and predict the default probability of any loan. If a loan defaults, it would be useful to know the magnitude of the loss and that's what we're trying to predict in this part. Loss given default can also be written as 1 - recovery rate. If we're able to predict the default probability and LGD then we can predict the expected loss of a loan which is Default Probability * LGD * Loan Explosures.

LGD prediction was very rarely looked at in previous works and this study is an attempt to get insights and build a model to predict it. LGD is defined as the total payments made by borrowers (including any recoveries) divided by the total amount of principal and interests which is installment * term. Figure 28 shows the distribution of LGD which ranges from 0 to 100% with the mean LGD of 58% and standard deviation of 23%.

Figure 28.



Figures 29 and 30 compare the LGD across different grades and LGD is generally higher for lower graded loans which makes sense. Figure 31 suggests that LGD is on average higher for longer term loans which is in line with their higher risks.

Figure 29.

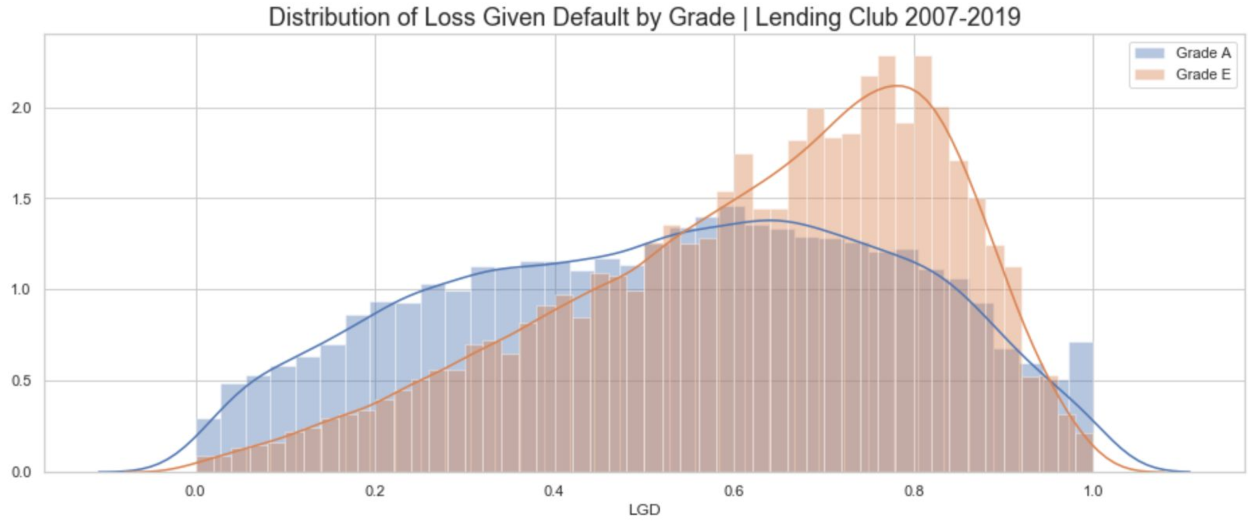


Figure 30.

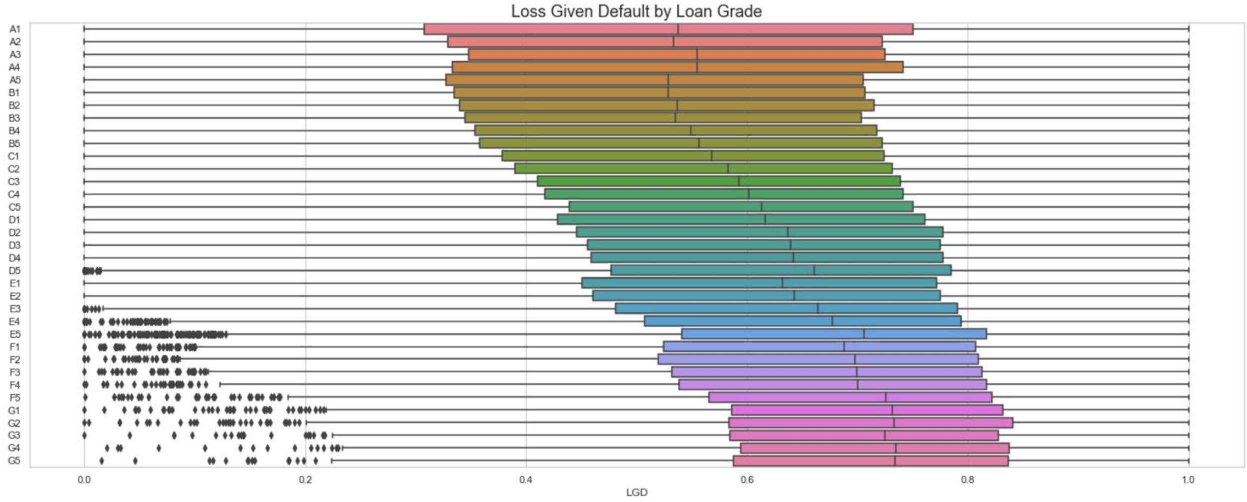
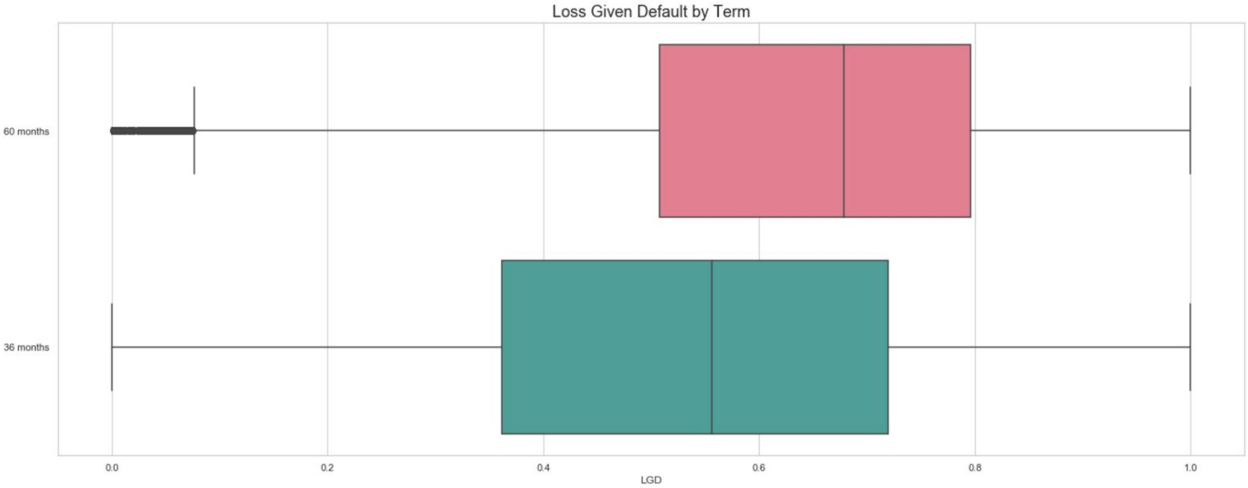


Figure 31.



Machine Learning Pt.2: LGD Prediction

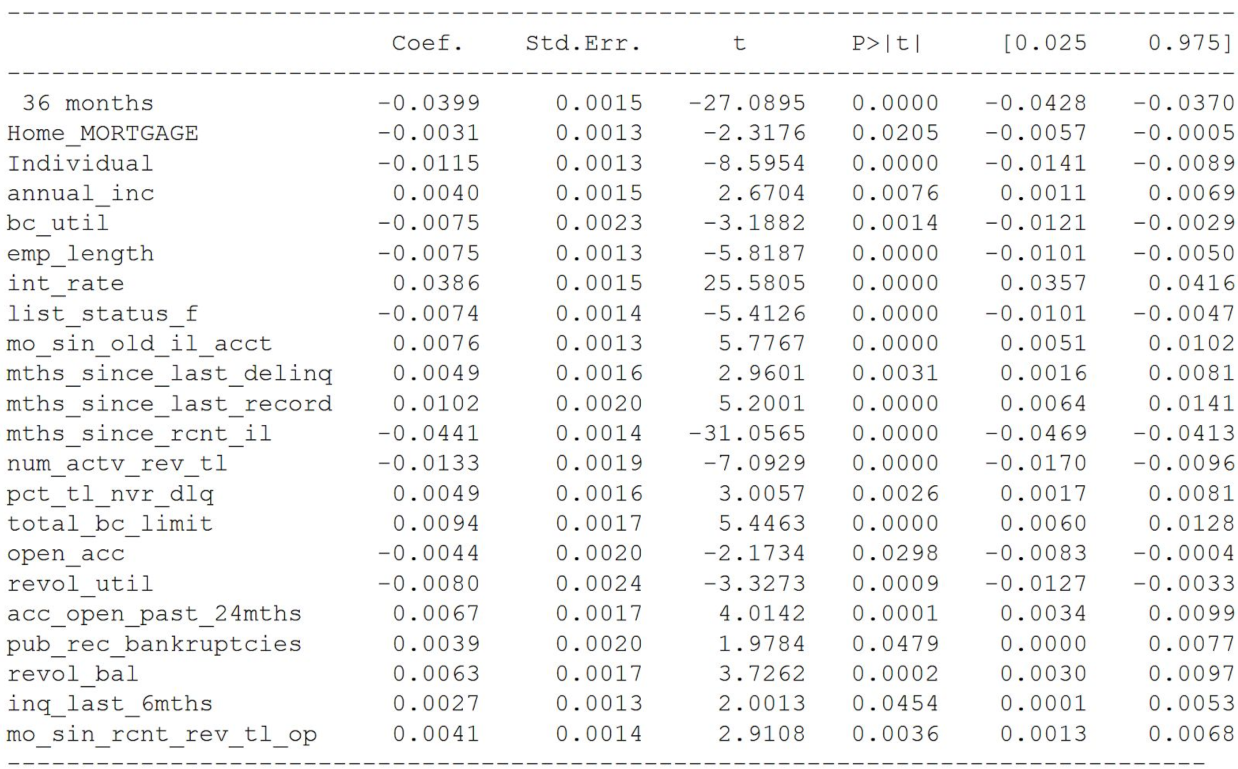
The data preparation, feature engineering and transformation steps are exactly the same as the default prediction process except now LGD is the target variable instead of loan status and since it's a numeric value linear regression is used.

Linear Regression

Like what we did for default prediction, recursive feature elimination (RFE) was used for feature selections. Top 30 features were selected then features with p-values higher than 0.05 were removed to further reduce the list of features. The R^2 is 15% and RMSE is 0.2.

Figure 32 lists the features selected in the end and their coefficients. To interpret it, LGD is lower for a shorter term loan, and for borrowers who didn't open installment accounts recently, and for a lower interest rate.

Figure 32.



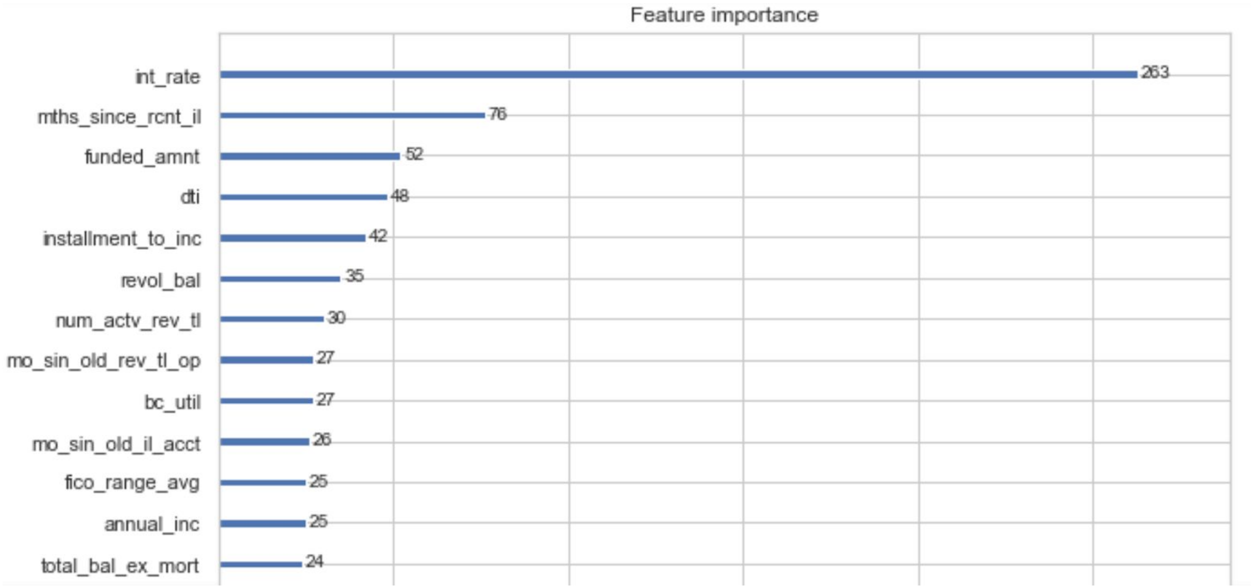
```
---------------------------------------------------------------------------------
                          Coef.    Std.Err.       t       P>|t|     [0.025    0.975]
---------------------------------------------------------------------------------
 36 months              -0.0399     0.0015   -27.0895   0.0000   -0.0428   -0.0370
Home_MORTGAGE           -0.0031     0.0013    -2.3176   0.0205   -0.0057   -0.0005
Individual              -0.0115     0.0013    -8.5954   0.0000   -0.0141   -0.0089
annual_inc               0.0040     0.0015     2.6704   0.0076    0.0011    0.0069
bc_util                 -0.0075     0.0023    -3.1882   0.0014   -0.0121   -0.0029
emp_length              -0.0075     0.0013    -5.8187   0.0000   -0.0101   -0.0050
int_rate                 0.0386     0.0015    25.5805   0.0000    0.0357    0.0416
list_status_f           -0.0074     0.0014    -5.4126   0.0000   -0.0101   -0.0047
mo_sin_old_il_acct       0.0076     0.0013     5.7767   0.0000    0.0051    0.0102
mths_since_last_delinq   0.0049     0.0016     2.9601   0.0031    0.0016    0.0081
mths_since_last_record   0.0102     0.0020     5.2001   0.0000    0.0064    0.0141
mths_since_rcnt_il      -0.0441     0.0014   -31.0565   0.0000   -0.0469   -0.0413
num_actv_rev_tl         -0.0133     0.0019    -7.0929   0.0000   -0.0170   -0.0096
pct_tl_nvr_dlq           0.0049     0.0016     3.0057   0.0026    0.0017    0.0081
total_bc_limit           0.0094     0.0017     5.4463   0.0000    0.0060    0.0128
open_acc                -0.0044     0.0020    -2.1734   0.0298   -0.0083   -0.0004
revol_util              -0.0080     0.0024    -3.3273   0.0009   -0.0127   -0.0033
acc_open_past_24mths     0.0067     0.0017     4.0142   0.0001    0.0034    0.0099
pub_rec_bankruptcies     0.0039     0.0020     1.9784   0.0479    0.0000    0.0077
revol_bal                0.0063     0.0017     3.7262   0.0002    0.0030    0.0097
inq_last_6mths           0.0027     0.0013     2.0013   0.0454    0.0001    0.0053
mo_sin_rcnt_rev_tl_op    0.0041     0.0014     2.9108   0.0036    0.0013    0.0068
-------------------------------------------------------------------------------
```

XGBoost

XGBoost is still the winner here! It achieves the highest R^2 of 21% among all tested models including Random Forest (with R^2 = 19%). RMSE is 0.2 for XGBoost.

Figure 33 shows the feature importances. It’s interesting to see the mths_since_rcnt_il (Months since most recent installment accounts opened) is the 2nd most important feature here. Based on Figure 32, it suggests that if someone recently opened installment accounts, then they’re more hungry of seeking credit and thus LGD is higher.

Figure 33.



Caveats for Default and LGD Predictions

There are apparently important missing features for example the macroeconomic conditions: the yield curve, inflation, unemployment and GDP that could cause **systematic default risks rise or fall across all applicants**. The only feature used in this study that might be related is the loan interest rate imposed by LC which incorporates both the applicant’s credit risk and market risk. However it only reflects the market risk at the time of loan origination. If the economy turns south during the loan payment periods, the default rate and LGD are likely to rise. Since most of the data spans 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.

Similarly, the loan applicant’s own credit condition might change from time to time and cause defaults. It’s possible to run predictions continuously using their most up to date data like FICO, DTI, employment, income, assets and liabilities. Note that this is why holding a diversified portfolio of loans is important because the idiosyncratic risks can be diversified away and the portfolio is only subject to systematic risks.

Another shortfall is when we calculated LGD, cash flows were not discounted to the present value due to the lack of the paths of cash flows and the corresponding dates in this data.

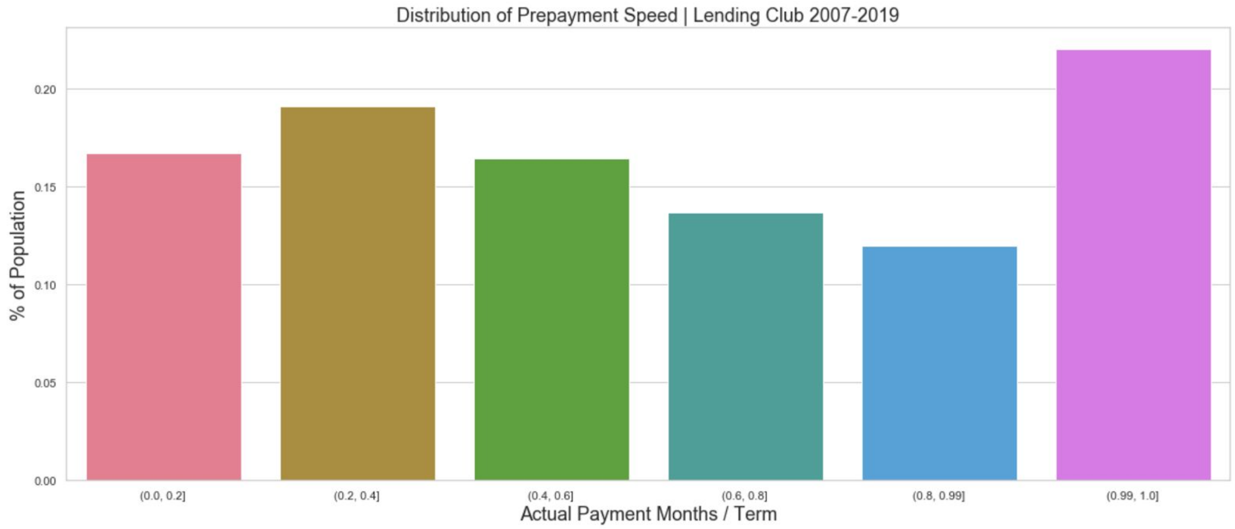
Indepth EDA Pt.3: Prepayment Speed Prediction

Part 1 and 2 focus on the default risk. There is another risk that's also important to investor's realized returns which is the **prepayment risk**. If the prevailing interest rates fall or the borrowers' credit improves, then they tend to prepay the loan and borrow at a lower rate for example refinancing a mortgage. From an investor's point of view, in this case it's good that the loans didn't default but the bad news is they had to reinvest the proceeds usually in a lower interest rate and thus lower returns.

Since this data lack the information of the paths of prepayment and we only know the last payment date and total payment made, we will ignore the reinvestment part and **define prepayment speed as the actual # of months pay off / # of months in the loan term**. So if a 36 month loan is paid off in 18 months then its prepayment speed is 50%.

We categorize the prepayment speed into 6 buckets (0, 0.2], (0.2, 0.4], (0.4, 0.6], (0.6, 0.8], (0.8, 0.99], and (0.99, 1]. The last one is the group of loans that were paid off exactly till the end of the term. Figure 34 shows the percentage of population for each bucket across all loans. Interestingly, **only around 25% of total loans were not prepaid**! **And 75% of the loans were prepaid at different speeds.** This is probably due to the ease of closing a loan and opening another one on the LC platform whenever the borrower wants to lower the borrowing rate. Some may even take another loan first and repay the previous loans (refinancing).

Figure 34.



The prepayment speed for longer term loans (60m) is faster than 36m loans as seen in Figure 35 which reflects the higher prepayment risk for longer duration loans. Figure 36 shows this trend from a different perspective.

Figure 35.

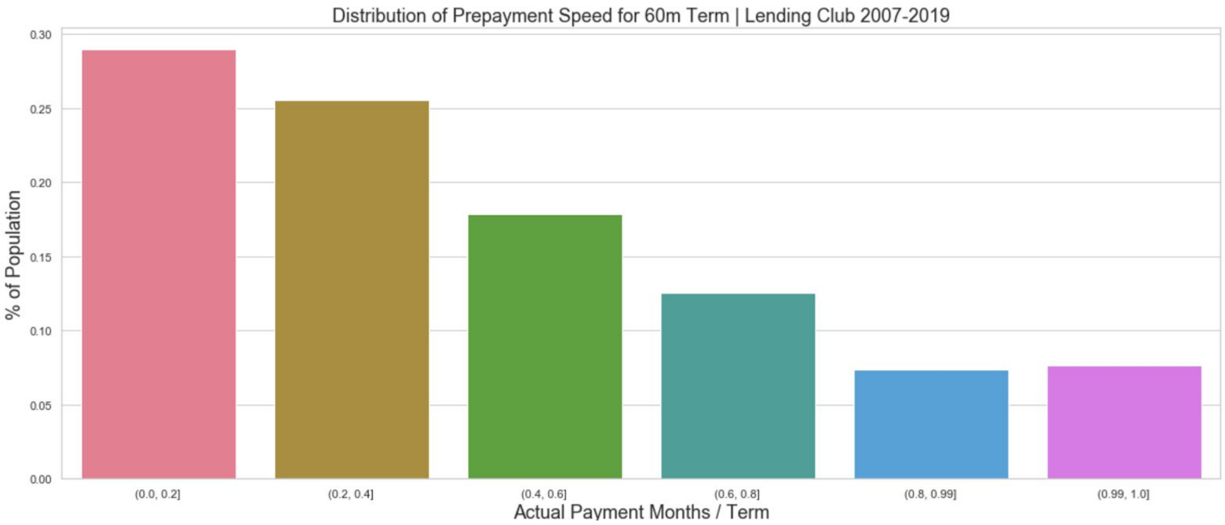


Figure 36.

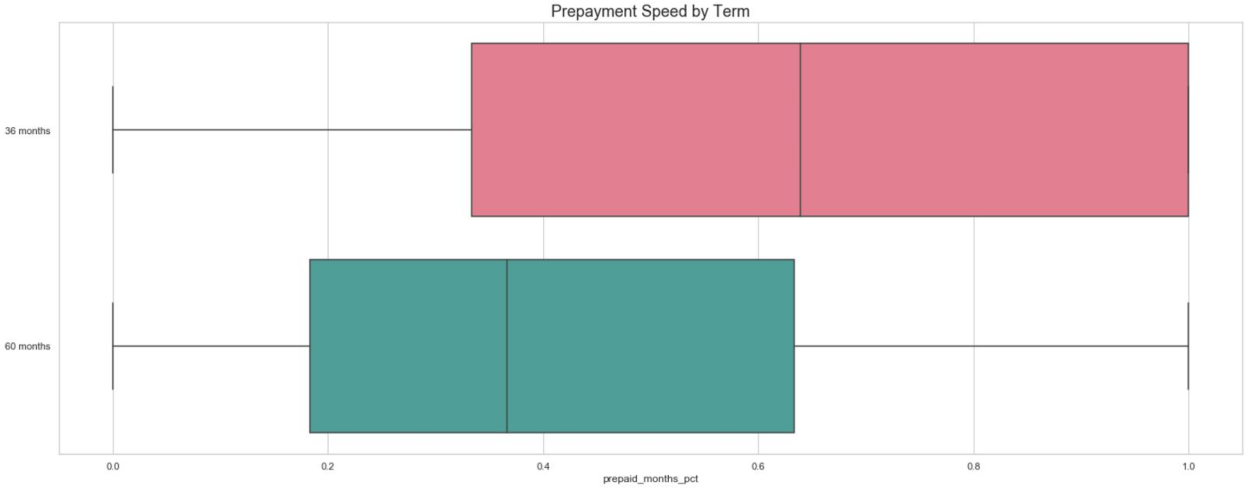


Figure 37 looks at the prepayment speed by loan grades. It seems borrowers pay off lower grade loans faster than the higher grade ones in general. This is probably because of the higher interest rates - lower grade loan borrowers have more motivation to pay off (if they can) and refinance at a lower rate.

Having the motivation to prepay is one story, having the ability to prepay is another story. This can be seen in Figure 38 - the lower the debt to income ratio, the faster the prepayment happens on average. This is even true if the borrower tried to open another loan to pay off this one since the loan approval and interest rates are related to borrower's credit information.

Figure 37.

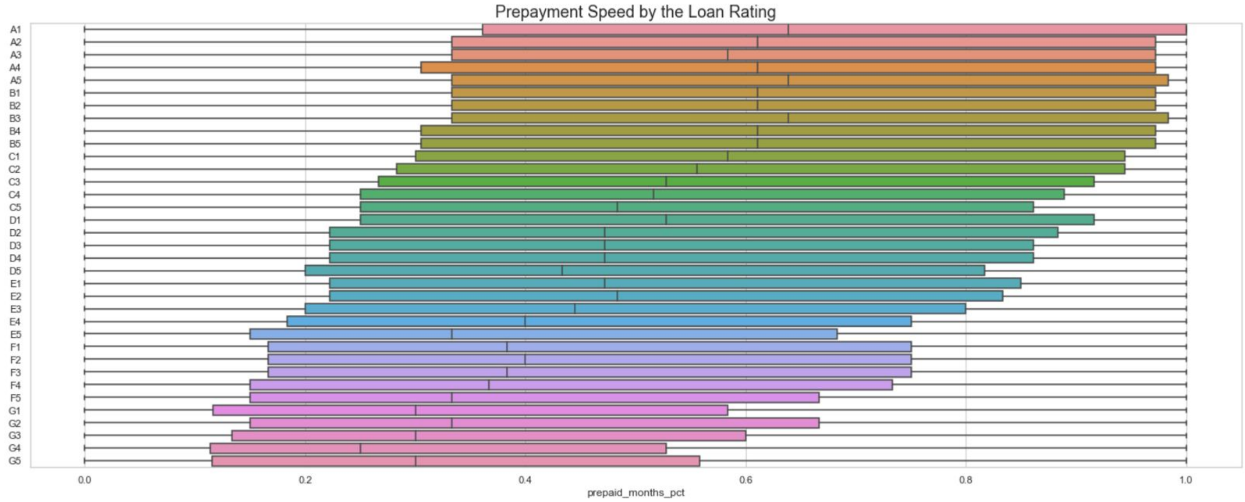
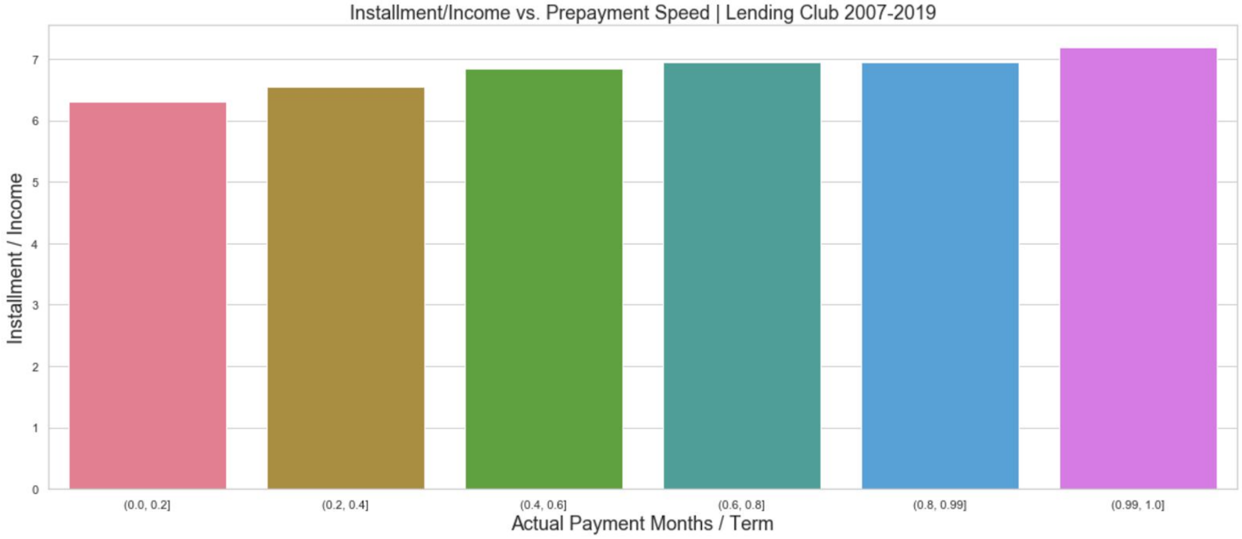


Figure 38.



Machine Learning Pt.3: Prepayment Speed Prediction

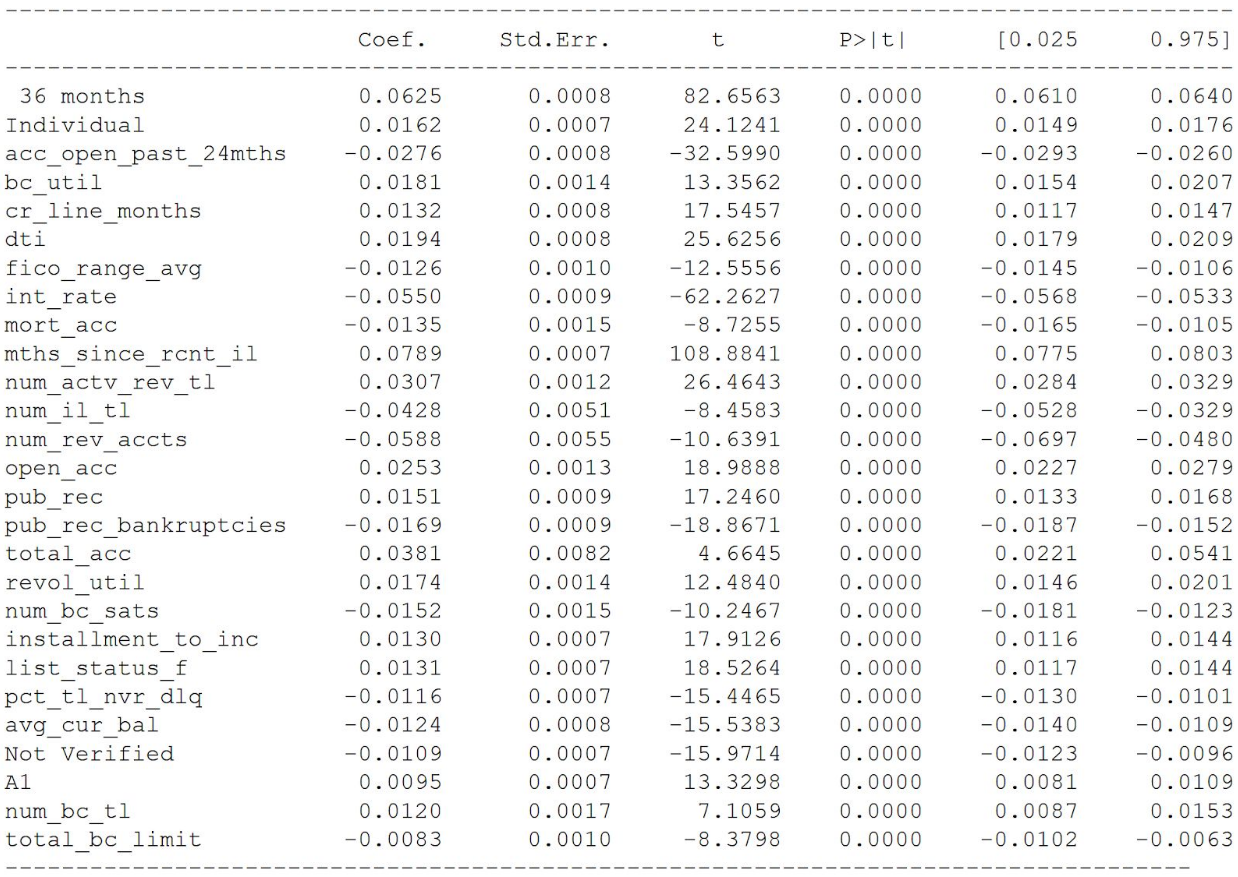
The data preparation, feature engineering and transformation steps are the same as the previous process except now Prepayment Speed is the target variable.

Linear Regression

Recursive feature elimination (RFE) was again used for feature selections. Top 30 features were selected by the ranking to run a first round of linear regression. And then features with p-values higher than 0.05 were removed to further reduce the list of features.

I was able to achieve 21.5% R^2 and 0.28 RMSE using the linear regression model. Figure 39 shows the significant features. Intuitively, prepayment is faster if loans have 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).

Figure 39.



```
-------------------------------------------------------------------------------------
                          Coef.    Std.Err.       t        P>|t|     [0.025     0.975]
-------------------------------------------------------------------------------------
 36 months               0.0625    0.0008     82.6563    0.0000    0.0610    0.0640
Individual               0.0162    0.0007     24.1241    0.0000    0.0149    0.0176
acc_open_past_24mths    -0.0276    0.0008    -32.5990    0.0000   -0.0293   -0.0260
bc_util                  0.0181    0.0014     13.3562    0.0000    0.0154    0.0207
cr_line_months           0.0132    0.0008     17.5457    0.0000    0.0117    0.0147
dti                      0.0194    0.0008     25.6256    0.0000    0.0179    0.0209
fico_range_avg          -0.0126    0.0010    -12.5556    0.0000   -0.0145   -0.0106
int_rate                -0.0550    0.0009    -62.2627    0.0000   -0.0568   -0.0533
mort_acc                -0.0135    0.0015     -8.7255    0.0000   -0.0165   -0.0105
mths_since_rcnt_il       0.0789    0.0007    108.8841    0.0000    0.0775    0.0803
num_actv_rev_tl          0.0307    0.0012     26.4643    0.0000    0.0284    0.0329
num_il_tl               -0.0428    0.0051     -8.4583    0.0000   -0.0528   -0.0329
num_rev_accts           -0.0588    0.0055    -10.6391    0.0000   -0.0697   -0.0480
open_acc                 0.0253    0.0013     18.9888    0.0000    0.0227    0.0279
pub_rec                  0.0151    0.0009     17.2460    0.0000    0.0133    0.0168
pub_rec_bankruptcies    -0.0169    0.0009    -18.8671    0.0000   -0.0187   -0.0152
total_acc                0.0381    0.0082      4.6645    0.0000    0.0221    0.0541
revol_util               0.0174    0.0014     12.4840    0.0000    0.0146    0.0201
num_bc_sats             -0.0152    0.0015    -10.2467    0.0000   -0.0181   -0.0123
installment_to_inc       0.0130    0.0007     17.9126    0.0000    0.0116    0.0144
list_status_f            0.0131    0.0007     18.5264    0.0000    0.0117    0.0144
pct_tl_nvr_dlq          -0.0116    0.0007    -15.4465    0.0000   -0.0130   -0.0101
avg_cur_bal             -0.0124    0.0008    -15.5383    0.0000   -0.0140   -0.0109
Not Verified            -0.0109    0.0007    -15.9714    0.0000   -0.0123   -0.0096
A1                       0.0095    0.0007     13.3298    0.0000    0.0081    0.0109
num_bc_tl                0.0120    0.0017      7.1059    0.0000    0.0087    0.0153
total_bc_limit          -0.0083    0.0010     -8.3798    0.0000   -0.0102   -0.0063
==================================================================================
```

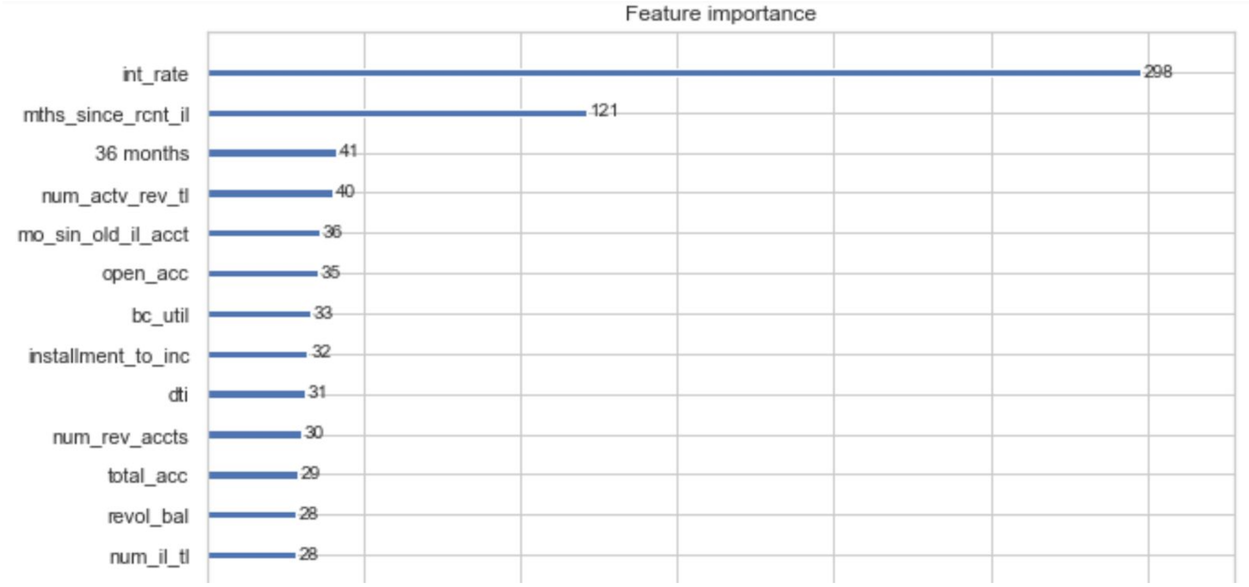
XGBoost

Again, XGBoost beat all the other models including random forest and delivered a performance of 30% R^2 and 0.28 RMSE. Some of the most important features can be found in Figure 40 with interest rate, months since recent installment accounts opened, and term being on top.

Caveats

The most often used metrics for prepayment is Single Monthly Mortality (SMM) and Conditional Prepayment Rate (CPR) which is SMM’s annualized version. However the calculation requires payment paths in each period which is unknown in this data. That’s why Prepayment Speed is used in this study. 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.

Figure 40.



Indepth EDA Pt.4: Return Prediction

As mentioned before, both default and prepayment hurt investors' returns. In this part, I try to **predict returns directly which incorporates the loans' default risk, LGD, and prepayment risk**. This is the "Holy Grail" for investors and let's see if our model can beat the benchmark and generate Alpha.

The 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 was defaulted or prepaid which might not be realistic. But given that the data lack the information of when and what reinvestment happened, this is the best we can do. It still generates invaluable insights and the investment performance is quite strong.

This time we include all the non-current loans in the study and the 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). See Figure 41 for the distribution with the 1st percentile and 99th percentile -43% and 10.3% respectively.

How does the distribution compare to the marketed expected returns on LC's website (see Figure 42)? It claims the expected return is 26% for loans with a grade of G, if they don't default (or prepaid). That's a big "if" and Figure 43 shows the actual median returns across grades. The realized returns exhibit a bell curve. The returns increase at first and peak at around grade C then start to drop as grade worsens. In fact, the median return for grade G is around 1%, not the

26% in the previous chart! This is contrary to the expectation that higher risks are compensated with higher returns.

Figure 41.

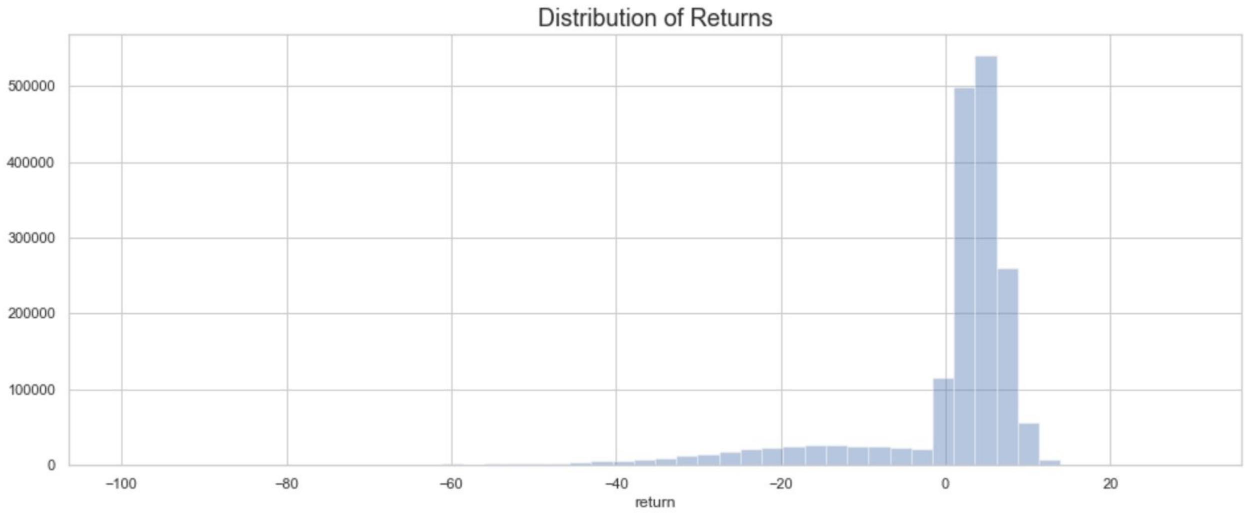
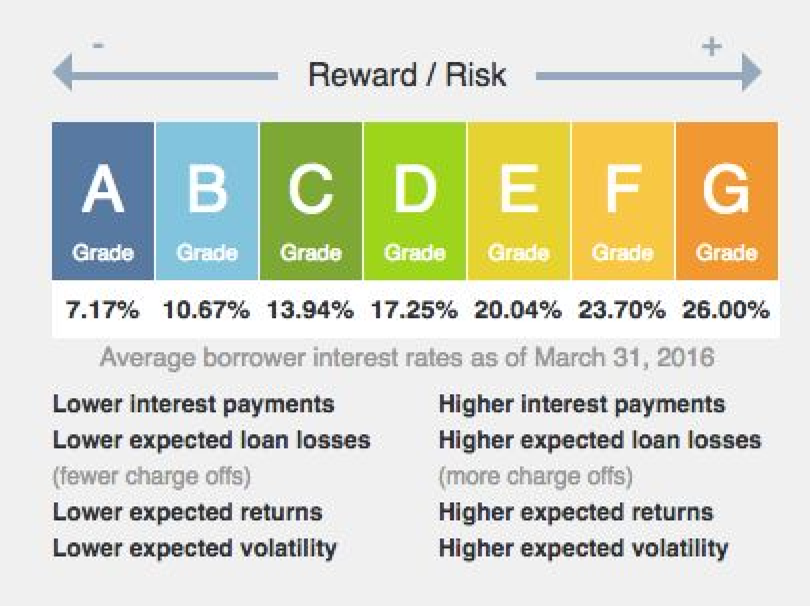


Figure 44 is super interesting in the sense that not only it shows the bell curve shaped interest rates but also you can observe the **different levels of volatility among loans with different grades**. The higher the grade, the less spread out the returns and less risks for investors. As grade gets lower, returns are more spread out and investors start bearing more risks (but not necessarily getting more returns!) This suggests randomly selecting loans in your portfolio could be harmful to total returns. We need a model that can predict returns and select the ones ranked the highest by returns.

Figure 42.



A Grade	B Grade	C Grade	D Grade	E Grade	F Grade	G Grade
7.17%	10.67%	13.94%	17.25%	20.04%	23.70%	26.00%

Average borrower interest rates as of March 31, 2016

Lower interest payments
Lower expected loan losses
(fewer charge offs)
Lower expected returns
Lower expected volatility

Higher interest payments
Higher expected loan losses
(more charge offs)
Higher expected returns
Higher expected volatility

Figure 43.

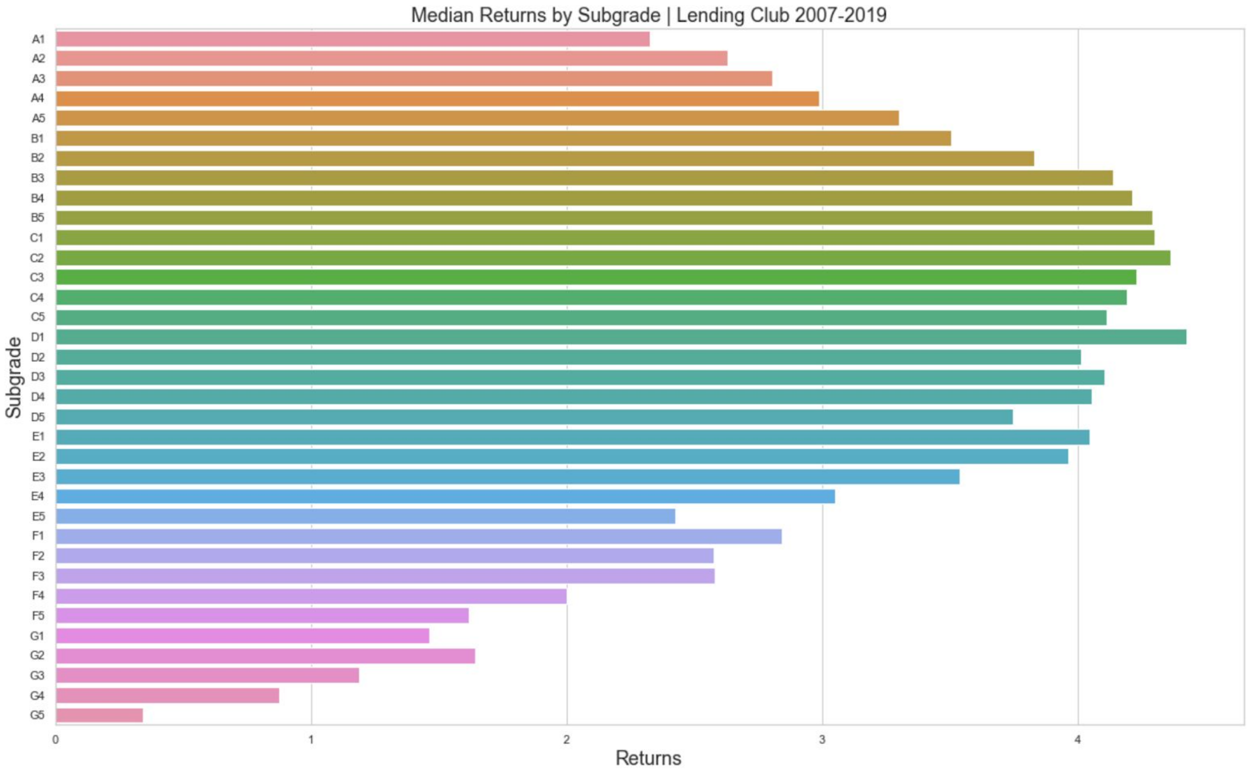
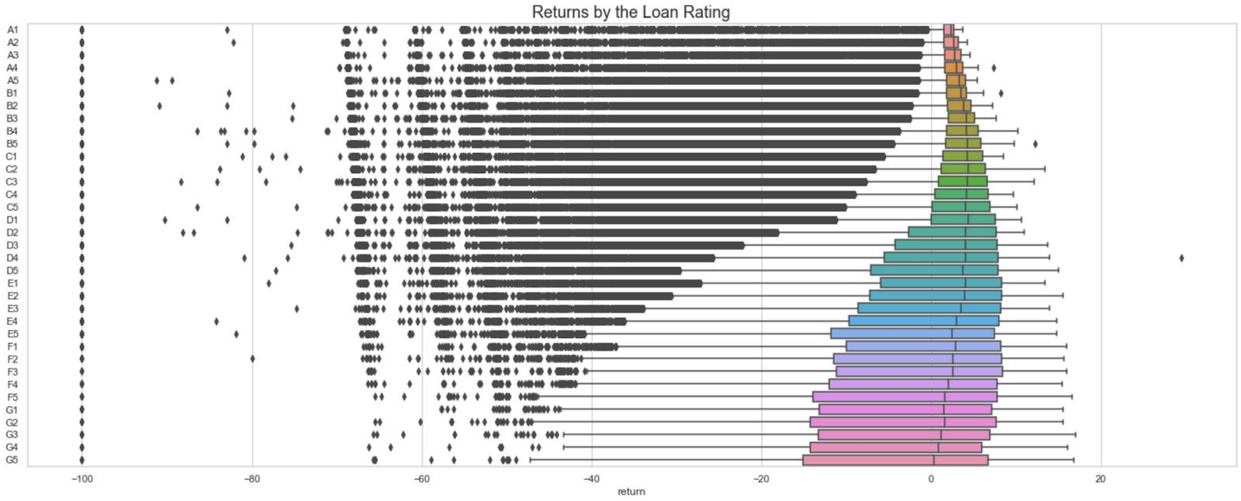
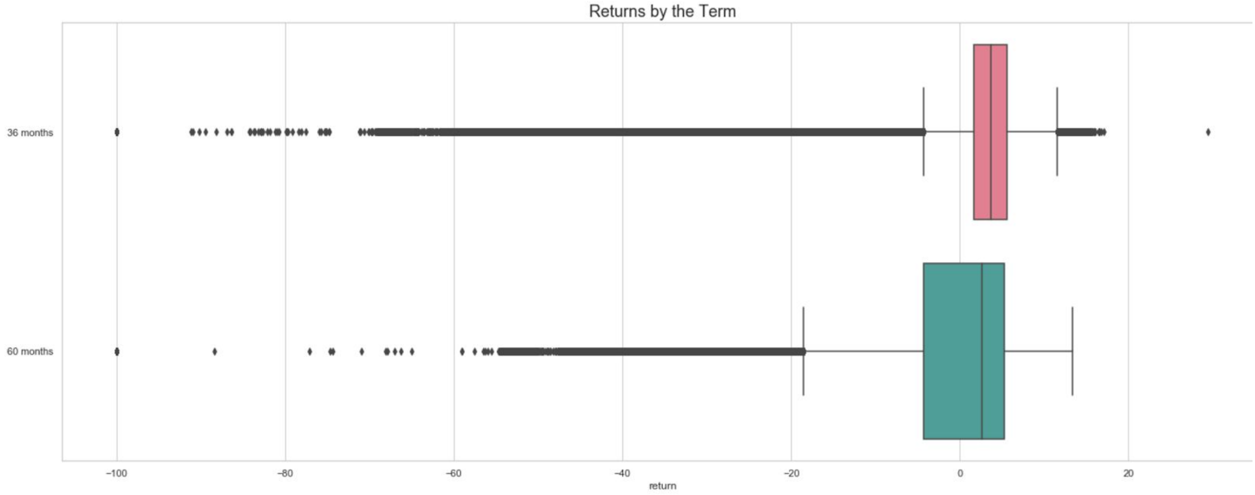


Figure 44.



In theory, there are more risks of holding longer term loans and investors are supposed to be compensated for this risk. But notice that Figure 45 suggests that's not necessarily true. If we remember the previous results, default probability, LGD, and prepayment all increase with longer duration loans and thus returns hurt by the combinations of three.

Figure 45.



Machine Learning Pt.4: Return Prediction

The data preparation, feature engineering and transformation steps are the same as the previous process except now Return is the target variable.

Linear Regression

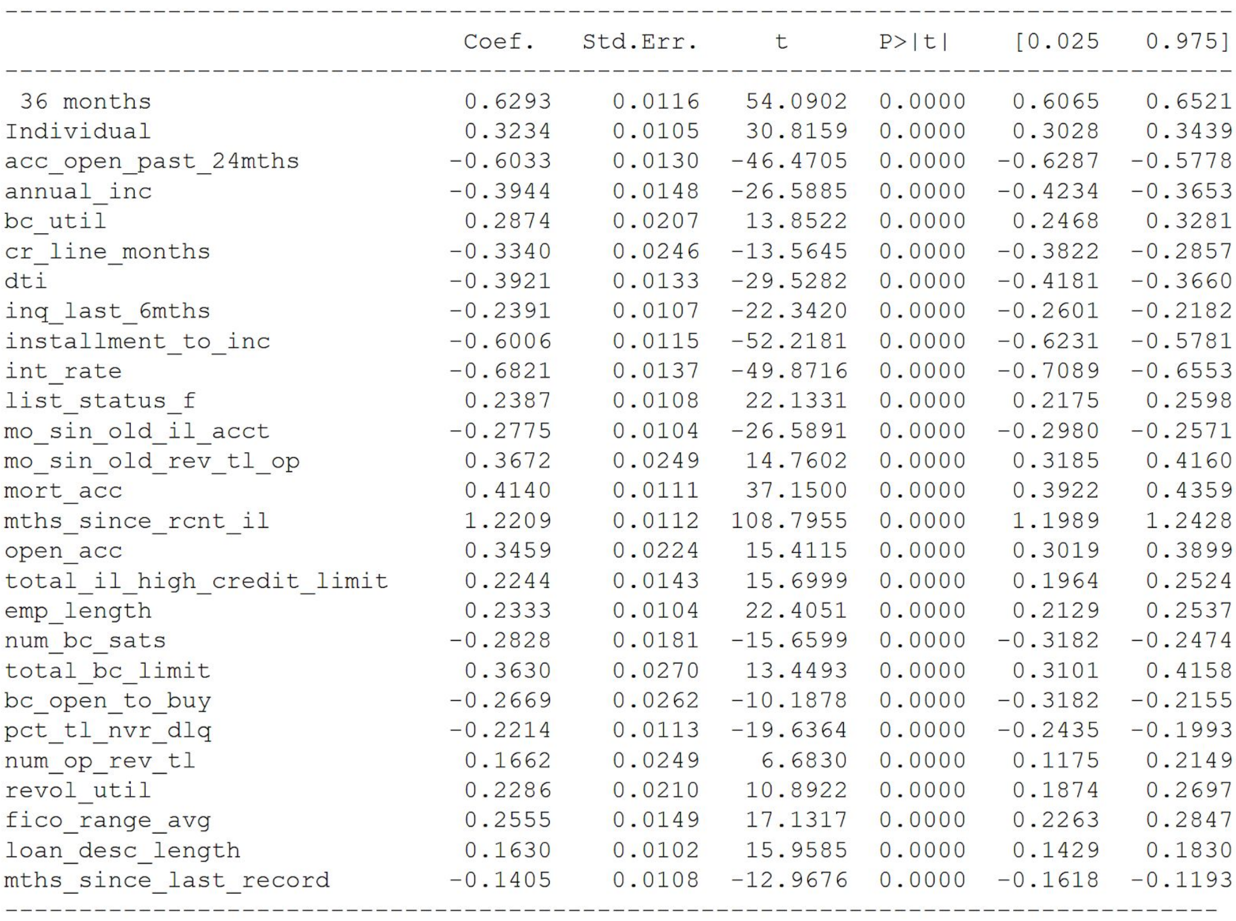
Recursive feature elimination (RFE) was used for feature selections. Top 30 features were selected by the ranking to run a first round of linear regression. And then features with p-values higher than 0.05 were removed to further reduce the list of features.

The R^2 is 4.4% and RMSE is 0.11 with the significant features and coefficients as seen in Figure 46. To interpret the signs of the coefficients, the expected returns are higher for shorter term loans, lower DTI, and less accounts opened in the last 2 years and less recent installment accounts opened (not hungry for credit). As we have seen the relationship between returns and the interest rates is not linear so the negative coefficient here is more a reflection of the lower grade loans.

XGBoost

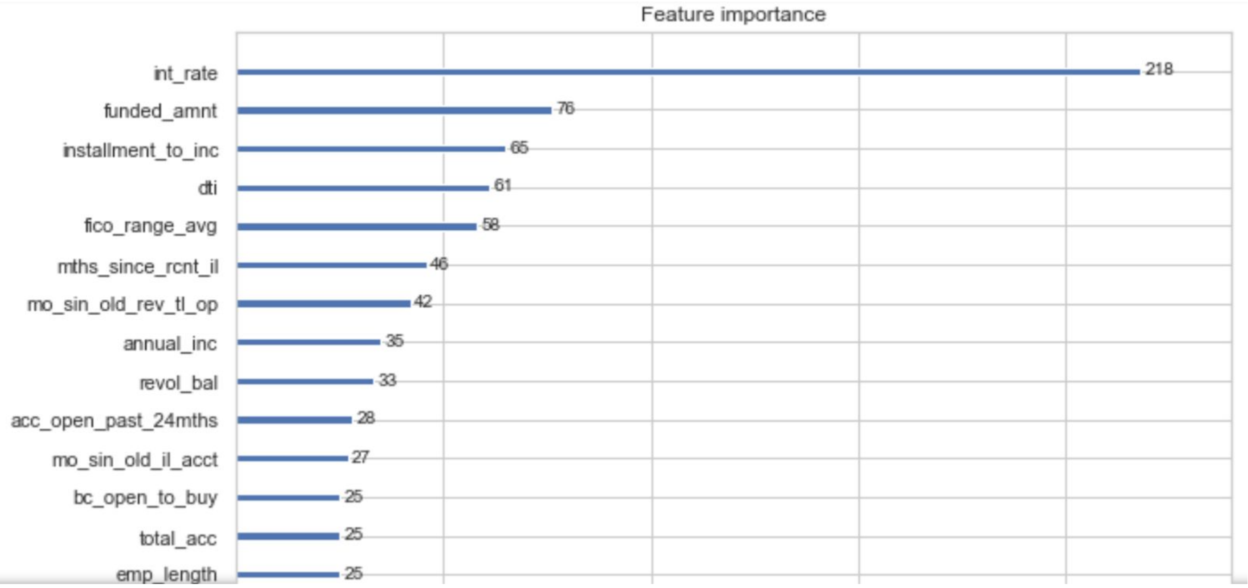
Once again, XGBoost is the best performing model, delivering a R^2 of 7.2% (vs. 5.9% using Random Forest) and RMSE of 0.11. The most important features can be seen in Figure 47 with the interest rate, funded amount, installment to income, dti, fico, and months since most recent installment accounts opened on top.

Figure 46.



```
---------------------------------------------------------------------------------------
                                      Coef.   Std.Err.      t      P>|t|    [0.025   0.975]
---------------------------------------------------------------------------------------
 36 months                           0.6293   0.0116   54.0902  0.0000   0.6065   0.6521
Individual                           0.3234   0.0105   30.8159  0.0000   0.3028   0.3439
acc_open_past_24mths                -0.6033   0.0130  -46.4705  0.0000  -0.6287  -0.5778
annual_inc                          -0.3944   0.0148  -26.5885  0.0000  -0.4234  -0.3653
bc_util                              0.2874   0.0207   13.8522  0.0000   0.2468   0.3281
cr_line_months                      -0.3340   0.0246  -13.5645  0.0000  -0.3822  -0.2857
dti                                 -0.3921   0.0133  -29.5282  0.0000  -0.4181  -0.3660
inq_last_6mths                      -0.2391   0.0107  -22.3420  0.0000  -0.2601  -0.2182
installment_to_inc                  -0.6006   0.0115  -52.2181  0.0000  -0.6231  -0.5781
int_rate                            -0.6821   0.0137  -49.8716  0.0000  -0.7089  -0.6553
list_status_f                        0.2387   0.0108   22.1331  0.0000   0.2175   0.2598
mo_sin_old_il_acct                  -0.2775   0.0104  -26.5891  0.0000  -0.2980  -0.2571
mo_sin_old_rev_tl_op                 0.3672   0.0249   14.7602  0.0000   0.3185   0.4160
mort_acc                             0.4140   0.0111   37.1500  0.0000   0.3922   0.4359
mths_since_rcnt_il                   1.2209   0.0112  108.7955  0.0000   1.1989   1.2428
open_acc                             0.3459   0.0224   15.4115  0.0000   0.3019   0.3899
total_il_high_credit_limit           0.2244   0.0143   15.6999  0.0000   0.1964   0.2524
emp_length                           0.2333   0.0104   22.4051  0.0000   0.2129   0.2537
num_bc_sats                         -0.2828   0.0181  -15.6599  0.0000  -0.3182  -0.2474
total_bc_limit                       0.3630   0.0270   13.4493  0.0000   0.3101   0.4158
bc_open_to_buy                      -0.2669   0.0262  -10.1878  0.0000  -0.3182  -0.2155
pct_tl_nvr_dlq                      -0.2214   0.0113  -19.6364  0.0000  -0.2435  -0.1993
num_op_rev_tl                        0.1662   0.0249    6.6830  0.0000   0.1175   0.2149
revol_util                           0.2286   0.0210   10.8922  0.0000   0.1874   0.2697
fico_range_avg                       0.2555   0.0149   17.1317  0.0000   0.2263   0.2847
loan_desc_length                     0.1630   0.0102   15.9585  0.0000   0.1429   0.1830
mths_since_last_record              -0.1405   0.0108  -12.9676  0.0000  -0.1618  -0.1193
-------------------------------------------------------------------------------------
```

Figure 47.



Portfolio Returns

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%. Yes, 0% if you don't have a clue of which loans to invest on LC!

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

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

Caveats

In reality, there are two frictions that could prevent us from realizing this return. First, we might not know all the feature values on the LC platform although we might be able to design another model with the available features. Second, good loans were taken fast due to many automated investors and it's hard for normal individual investors to seize the opportunities.

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 for example incorrectly including current loans, and/or data leakage and imbalance issues.

What's more, I explored three important and often overlooked subjects: LGD, prepayment risk, and loan 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.

The next steps could include periodically updated borrowers and economic features and running a cohort study to predict the default risk and prepayment risk in a more precise and timely manner. For investors, we could 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 tolerance. Last but not least, we can analyze and predict the default risk, prepayment risk, and returns for different tranches of the loans by incorporating the joint default correlation of loans into the model. The sky's the limit.