Prediction of Default Rates in Peer-to-Peer Lending Companies

Project Report

Group 3:

Sammi Chang

Xueting Li

Jialu Li

Junguo He

Executive summary.

This research will explore the types of customers that default on loans in peer-to-peer lending. This is relevant because on such platforms, lenders are normally at a disadvantage due to information asymmetry: usually, borrowers are more knowledgeable about the risks of the platform than lenders. If borrowers default, lenders, as an individual, bear the consequence and the financial loss themselves, unlike financial institutions, which have procedures and operations in place to mitigate loss. When lenders screen applicants for loan bidding, it is important that they consider factors that affect default rate to make a meaningful and reliable investment. The purpose of this research is to determine the relationships between the default rate of a loan and several customer factors, such as credit score, employment, income, home ownership, and indebtedness. The study will also consider the loan purpose, loan grade and various other pertinent factors.

Business Idea.

P2P lending companies provide an online platform that connects willing investors/lenders to people who need to borrow money. This type of financial product is for borrowers who have difficulty getting a loan from a bank due to restrictive loan policies or poor credit histories, and it is attractive for investors who desire a high return rate. Therefore, P2P financial products are high risk, high return. Borrowers have higher default rates than those who borrow from a bank, so it is important to predict the nature of defaults for P2P companies. In general, a loan is considered defaulted when payments have not been made for an extended period of time, and the loans become charged off when the loan is 120 days past due and there is no reasonable expectation of payment [1].

On the webpage for the Lending Club, investors are able to see the information about the purpose of the loan, the description of the loan, a loan grade, and the borrower's FICO score. The loan grade is a grade calculated by Lending Club for each borrower. A grade of A is the highest and is associated with a high credit score, while a grade of C is acceptable, and a grade of G indicates a risky borrower. The grade is assigned depending on credit history and may also decrease if the borrower requests for a loan that exceeds their borrowing amount [2]. After the grade is determined, the Lending Club designates an interest rate for the loan. In our dataset, the interest rate ranges from 5.32% to 28.99%.

Data Cleaning.

The dataset from Lending Club, a P2P lending company, features customer and loan data for loans issued from 2013-2017. It has 74 variables and 887,379 rows. Data from 2013-2016 was used as training data, while data from 2016-2017 was used as testing data. As such, some of 2016-2017 data was randomly sampled to partake in the training data, and some of it was randomly sampled to partake in the testing data. The data was split as such due to the huge volume of loans issued in 2016-2017, which was almost 50% as much as those issued in 2013-2015.

The dataset from Lending Club contained information pertaining to joint applicant borrowers. Since less than .001% of loans belonged to joint borrowers, the 4 columns of data describing these borrowers (including the joint dti, joint income, and the verification status of secondary applicant's employment) contained mostly missing data and were thus removed. Next, we removed 14 columns that contained all missing data. These included the number of revolving trades opened in the last 24 months, the number of personal financial inquiries, and the total current balance of all installment accounts. After columns of missing data were removed, we were left with 55 columns.

Other data was simply removed due to their non-predictive power. Attributes like the webpage url of the loan listing, the title of the loan's web listing, the description of the loan, zip code (where last two digits are censored), member id, loan id, and, lastly, policy code, as it contained the same value for all rows. After removing useless variables, we were left with 48 columns to work with.

Interestingly, the dataset contained a lot of leakage data that described either customer payment behavior after the loan was issued or data describing various actions of Lending Club itself. Since we want to predict loan default before the loan was issued, these variables should not be used in the prediction. For example, variables like the total late fees received to date, the total principal received to date, the total amount of interest received to date, the amount of the last payment, and the amount of recoveries were excluded from analysis as these describe customer payment behavior and would have impacted our prediction of customer solvency/default. The variable describing the date of Lending Club's last credit pull was also removed as it is again not an event in our desired timeframe. After all leaky data was removed, the final set of attributes for analysis was reduced to 30 variables, which is still a substantial amount of data to work with. The training data consisted of 107,391 rows, and the testing data consisted of 60,408 rows. Lastly, rows in which any variable had a value that was less than Q1-1.5*IQR or greater than Q3+1.5*IQR was removed from the dataset.

Feature Engineering.

Firstly, the nature of defaulting on a loan was defined as a variable. Our variable, *Default*, has a value of "1" if the loan defaulted or a value of "0" if the loan was fully paid. Many loans in the data were current; hence, all row data related to those loans had to be removed as it would not make sense to analyze current loans. In other words, any loans that had a *loan_status* value of "Current," "Late (16-30 days) ," "Late (31-120 days)", or "Issued" were removed. In addition, loans showing statuses of "'Does not meet the credit policy. Status:Fully Paid" and "Does not meet the credit policy. Status:Charged Off" were also removed from the data due to its ambiguity in meaning and lack of clear explanation for these values.

Our dataset also included categorical variables that were made into dummy variables, including: *grade*, *purpose*, *home_ownership*, *verification_status*, and *term*. The grade of the loan had six categories, so it was split into five dummy variables: *grade_a*, *grade_b*, *grade_c*, *grade_e*, and *grade_f*. The purpose for the loan that was provided by the borrower upon application, *purpose*,

was split into: *car, credit_card*, *small_business*, *other*, *wedding, debt_consolidation*, *home_improvement*, *major_purchase*, *medical*, *moving*, *vacation*, *house*, and *renewable* (for renewable energy). The verification status, which indicated if the employment was verified and contained three levels, was split into: *verified* and *not_verified.* Lastly, the term, which has two categories: 36 months, or 60 months, was made into a dummy variable, *thirtysix_mo.*

Besides coding in dummy variables, four other time-related variables given in the dataset required further manipulation, primarily: *earliest_cr_line*, *mths_since_last_delinq, mths_since_last_major_derog, and mths_since_last_record, and emp_length,* Firstly, *earliest_cr_line* only contained the month and year that the earliest credit line was opened. To obtain the age of credit, or in other words, the amount of months elapsed between the time of the first credit line and the issue date of the loan, *earliest_cr_line* and *issue_d* were first converted to a POSIXIt date with R. Then, the new variable, *cr_age* is defined:

$$cr\_age = issue\_date - earliest\_cr\_line$$

Since the variables *mths_since_last_delinq, mths_since_last_major_derog,* and *mths_since_last_record* contained many missing data, new variables were created for these, with the missing data substituted with the median. The new variables are, respectively, *Deling*, *Derog*, and *Last_Record*. Lastly, several dummy variables were made to bin the employment length variable, *emp_length*, namely: *emp_length_lessthan1*, *emp_length_1to5*, *emp_length_6to9*, *emp_length_10plus.* Respectively, these new attributes indicate an employment length of: < 1 year, 1 to 5 years, 6 to 9 years, and 10 plus years. The employment length was binned in this fashion due to their similarity in default rate; for example, those who work worked 1, 2, 3, 4, or 5 years, had similar default rates. Because default rates were unusually high for loans that presented missing data for employment length, another dummy variable was made to indicate if the value for employment length was missing: *emp_length_is_NA;* however, this variable was removed in further analysis of information gain during feature selection.

Several studies have aimed to engineer new variables from the existing data to improve accuracy rate. Namvar, Siami, Rabhi, and Naderpour [3] defined a new ratio named New DTI, which is the new debt-to-income ratio that considers the financial impact on the new lending club loan on the borrower's solvency. This is calculated from the borrower's dti and annual income, which are already attributes provided in the data. The new dti is defined as:

$$newDTI = \frac{(\frac{annual\ income}{12}) * dti + monthly\ installment)}{\frac{annual\ income}{12}}$$

The new dti does indeed record a higher correlation with default than *dti*,* but to better account for the impact of the new lending club loan on the borrower's solvency, we calculated a new variable, the incremental difference of dti, which to the best of our knowledge has not been used in previous research. The variable, *dti_diff*, is calculated by subtracting the old dti (of the given data) from the new dti. The formula for our new, engineered ratio is:

$$dti\_diff = newDTI - dti$$

Where newDTI = $\frac{(\frac{annual\ income}{12}) * dti + monthly\ installment)}{\frac{annual\ income}{12}}$

*italicized variables are from the original dataset.

It is notable that our variable, *dti_diff,* shared a higher correlation with *default* than the variable, *newDTI.* In modeling, only *dti* and *dti_diff* will be used as attributes; *newDTI* will be dropped to avoid multicollinearity.

Below is chart of the variables selected for modeling and their description. Unless otherwise stated, the stated variable is a numerical variable.

Attributes		
Type	**Attribute**	**Attribute Description**
Loan Data	loan_amnt	The amount of loan requested
	int_rate	Interest rate on the loan
	installment	Monthly payment owed by the borrower
	Grade	Grade of the loan assigned by LC. Categorical Variable (nominal): The possible grades are: a, b, c, d, e, and f.
	purpose	A category provided by the borrower upon loan request. Categorical variable (nominal): the possible values are Car, Small Business, Credit Card, Wedding, Debt Consolidation, Home Improvement, Major Purchase, Medical, Moving, Vacation, , House, Renewable Energy, and other
	thirtysix_mo	Indicates the loan term. Categorical variable(nominal): The possible terms are 36 months or 60 months
Borrower Income Data		
	emp_length_num	Employment length in years. Categorical variable(nominal): Dummy variables are made to indicate the employment length: <1 year, 2 to 5 years, 6 to 9 years, 10+ years, and NA(missing).
	annual_inc	The self-reported annual income provided by the borrower during registration.
	verification_status	Categorical variable (nominal): Indicates if income was verified by LC, not verified, or if the income source was verified
Borrower Credit Data		
	dti	The Debt-to-income ratio of the borrower
	delinq_2yrs	The number of 30+ days past-due incidences of delinquency in the borrower's credit file for the past 2 years
	inq_last_6mths	The number of inquiries in past 6 months (excluding auto and mortgage inquiries)
	open_acc	The number of open credit lines in the borrower's credit file.
	pub_rec	Number of derogatory public records
	revol_bal	Total credit revolving balance of borrower
	total_acc	The total number of credit lines currently in the borrower's credit file
	collections_12_mths_ex_med	Number of collections in 12 months excluding medical collections
	acc_now_delinq	The number of accounts on which the borrower is now delinquent.
	tot_coll_amt	Total collection amounts ever owed
	tot_cur_bal	Total current balance of all accounts
	total_rev_hi_lim	Total revolving high credit/credit limit
	credit_age	The amount of months since first credit line was opened
	Delinq	Months since last delinquency
	Derog	Months since last major derogatory record
	Last_Record	Months since last public record
	new_DTI	The new DTI that accounts for the impact of the new loan
	dti_diff	The difference between new DTI and DTI that records the incremental impact of the loan

Exploratory Data Analysis.

The default rate of the loans in the training data sat at 22%. A graph was created to visualize each state and their respective default rates. Although California and Texas are the most populous states, they did not have the greatest default rates. In fact, most states hovered around an 18% default rate, and none presented an unusually high or low default rate. Because of this and the fact that there were so many states, the state attribute was removed as it would not make a good predictor.

The chart below shows a visualization of the default rate per state:

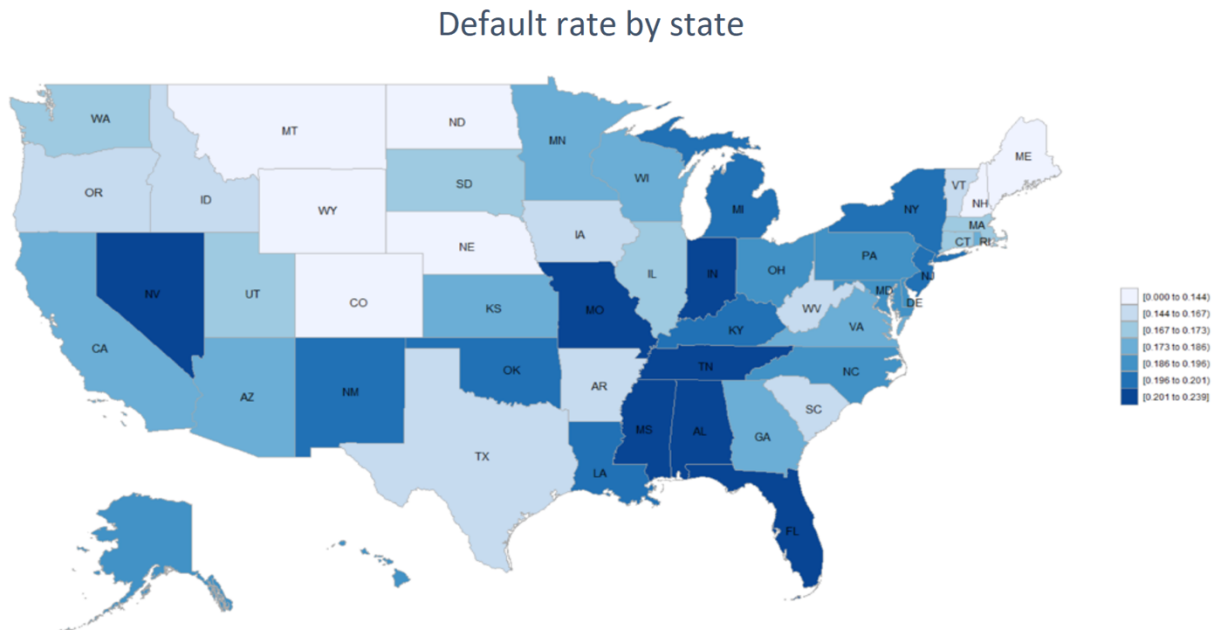


Figure 1 This chart shows the default rate per state. Darker colors indicate a higher average default rate. California sits in the middle, with Nevada, Florida, Alabama, Missouri, Indiana, Tennessee, and Mississippi holding the biggest default rates.

Below, Figure 2 shows the relationship between the interest rate and grade of the loan. Clearly, the interest rate increases as the grade of the loan degenerates from A to G. Thus, higher risk loans result in higher returns for investors, and this makes sense financially.

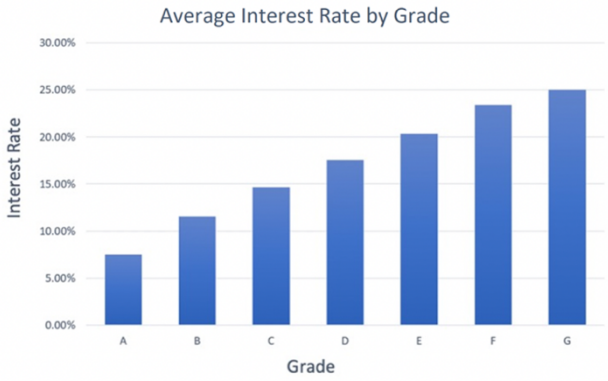


Figure 2 The mean default rate for interest rates between 0%-10%, 10%-20%, 15%-20%, and 20%-25%. The default rate increases as the interest rate increases.

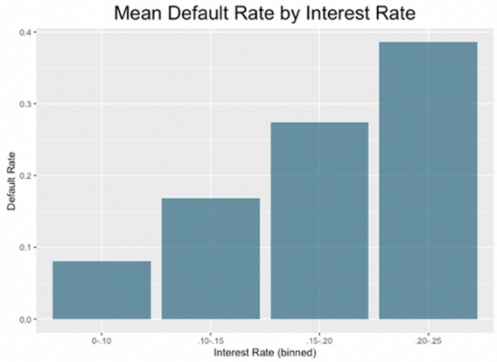


Figure 3 The mean default rate for interest rates between 0%-10%, 10%-20%, 15%-20%, and 20%-25%. The default rate increases as the interest rate increases.

Measuring default rate by interest rate gives us another interesting visual in Figure 3. As there were no interest rates that exceeded 25%, the interest rates were binned from [0%-10%], [10%-15%], [15%-20%], and [20%-25%]. The mean default rate for each interest rate level was then graphed. We can see that the default rate increases from 10% to 40% as the interest rate rises. This is likely because borrowers who take out loans with higher interest rates sustain a weightier financial burden, paying high installments each month. Thus, they are more likely to default. Interest rate is one of the more powerful predictors of default rate, as shown later in the correlation table and in the feature selection section.

The correlation table shows that most variables are not highly correlated with the default rate. Loan amount shows perfect correlation with installment, which makes sense because higher loan amounts should be associated with higher monthly payments. Because of the perfect correlation, loan_amnt will be removed from analysis to avoid multicollinearity. Furthermore, since the variable, newDTI is calculated from dti, it is evident that these two variables would show perfect correlation. Here, too, dti will be removed from our variable set and only newDTI will be used. The other variables, revol_bal and total_rev_hi_lim, also show a high positive correlation as a result of the fact that higher monthly revolving balance usually results in a higher total credit limit, but because this is not perfect correlation, both of these variables will be kept. On the other hand, our engineered variable, *dti_diff* shows high positive correlation with *loan_amnt* and *installment*; however this is simply because dti_diff is derived from those variables.

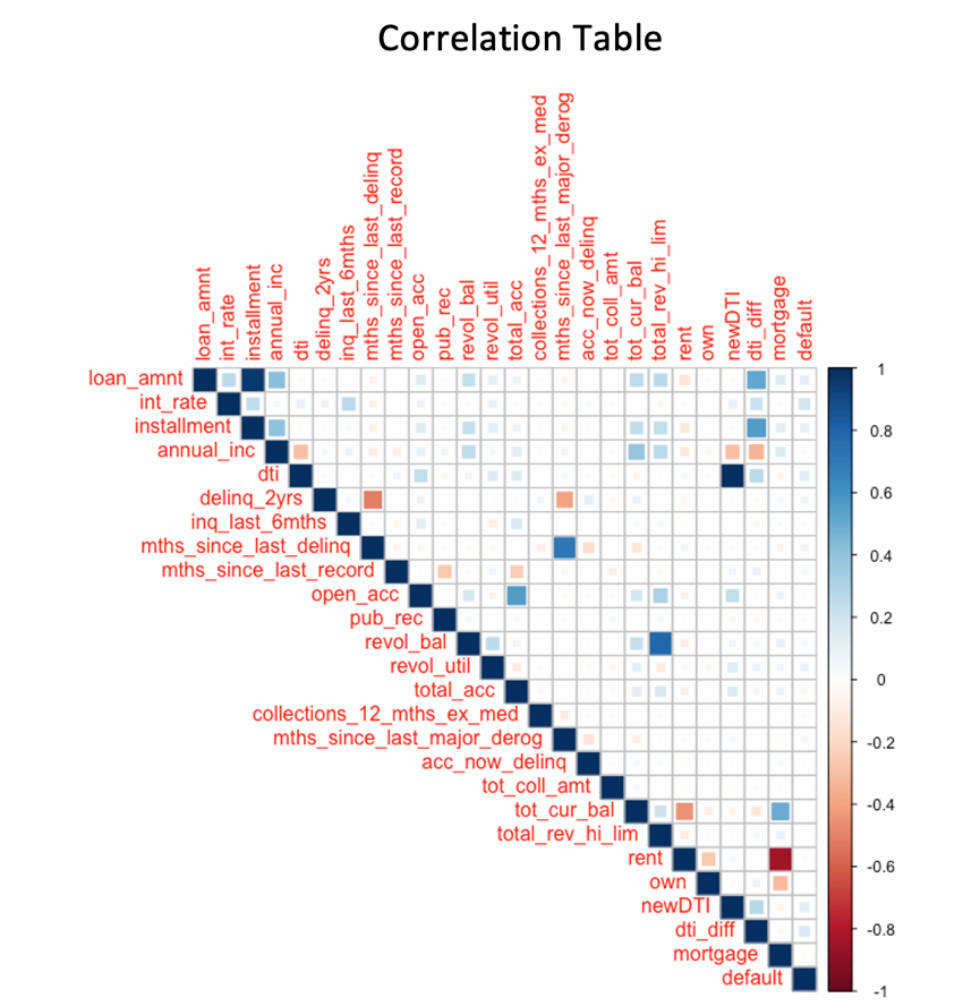


Figure 2 The correlation table shows the correlation between different variables.

On the other hand, there are other variables that also show high negative correlation. The variable, *rent*, is negatively correlated with *mortgage* and *own.* This is rather superficial as they are dummy variables for the same categorical variable. Another pair of variables in the same situation is *dti_diff* and *annual_inc*, where the negative correlation is a result of dti_diff being derived from dividing by *annual_inc*. In contrast, the pair, *rent* and *tot_cur_bal* share a non-superficial negative correlation—the correlation here provides insight into the nature of current balance accounts in borrowers who rent their places – renters are more likely to have a lesser amount of current balance of debt. This makes sense because mortgagers are indeed more likely than renters to incur more debt in their current accounts.

Perhaps the most important object of interest in the correlation table is the column for *default*. Most variables are not very correlated with our class variable, *default*, and most correlation is positive correlation. Indeed, the only variable that holds an (although weak) negative correlation with the class variable is mths_since_last_delinq, which measures the number of months since the borrower's last delinquency, so it seems to be that if a borrower was recently late for a

payment, he or she is more likely to default on the loan. However, this correlation is weak at best. Variables that share a positive correlation with default are *loan_amnt*, *int_rate*, *installment*, *dti*, *newDTI*, and *dti_diff*. These variables are the most important predictors for default. It is notable that our engineered variable, *dti_diff,* shares a greater correlation with *default* than *dti*.

On the other hand, the loan amount tends to increase as the borrower's grade decreases. In the below graph, loan amounts of upwards of $20,000 are most prevalent in G-grade loans. E- and F-grade loans also tend to borrow loan amounts of upwards of $10,000. Interestingly, A-, B-, and C- grade loans tend to borrow less with a slight uptick at $35,000, which is the maximum amount to borrow. Amongst all loans, A-grade loans are most likely to borrow the least amounts (<$10,000). This makes sense because a borrower's loan grade may decrease if the borrower requests a loan over their personal maximum. The wave shape of each graph is a result of borrowers requesting similar loan amounts: for example, in the curve of the D-grade loans the graph holds local maximums at 10,000, 15,000, and 20,000, indicating that these are popular loan amounts to request amongst borrowers.

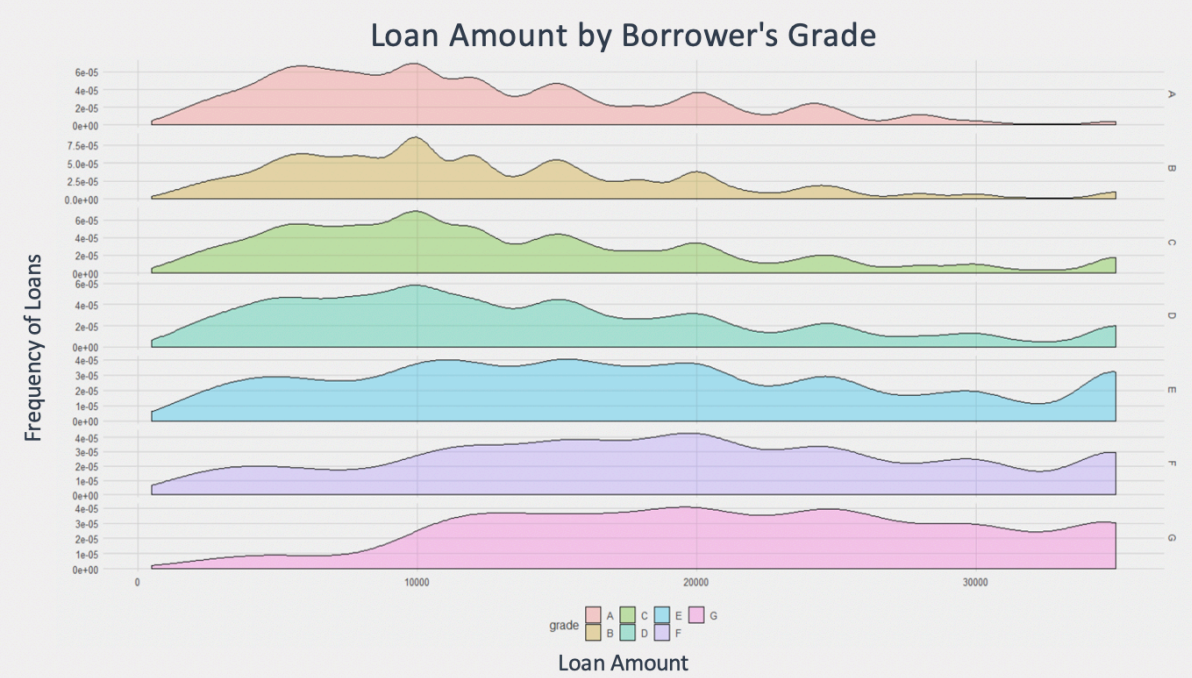


Figure 6 shows the loan amount per borrower's grade. The lower the grade, the higher amount of borrowing. The y-axis shows the frequency of such loans, and the x-axis shows the loan amount.

The table below charts the loan amount by employment length. People who have worked for 10+ years tend to borrow more money, and those who have worked for 9 years tend to borrow the least. Because of this, there seems to be no correlation between employment length and loan amount. However, since borrowers who have worked for more than 10 years tend to request more loan money, the dummy variable indicating an employment length of 10+ years may be able to hold some predictive power.

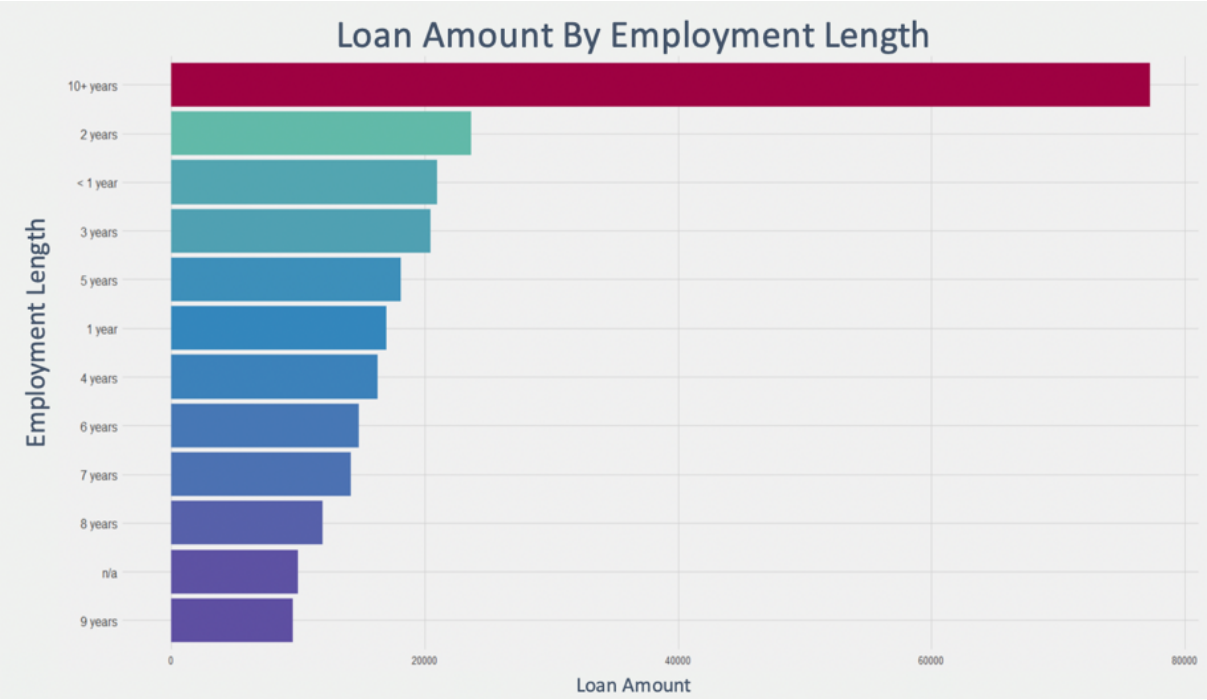


Figure 3 The loan amount by employment length, we can tell people with more than 10 years employment will tend to borrow more money or qualify to borrow more.

The density plots for each variable plot the distribution against the defaulters and the non-defaulters data. Several notable features of the data arise. Firstly, the debt-to-income ratio, or dti, which measures the indebtedness of the borrower, is more skewed right for non-defaulters. The difference of dti, *dti_diff,* further emphasizes this skew. The curve for non-defaulters obtains a greater right-skew, and some abnormalities in the defaulters curve are emphasized. This is perhaps indicative that our engineered variable, *dti_diff* is able to capture more abnormalities in the data, compared to *newDTI*, which retains the same distribution as *dti.* In particular, we notice that the distributions for most of the variables are similar for both the defaulters and non-defaulters, which reflects the general low correlation that the independent variables have with *default.* Most notably, the density plot for total_acc shows very close normally distributed curves for both defaulters and non-defaulters, but because the blue curve extends a little farther than the red, total_acc may still be a good predictor.

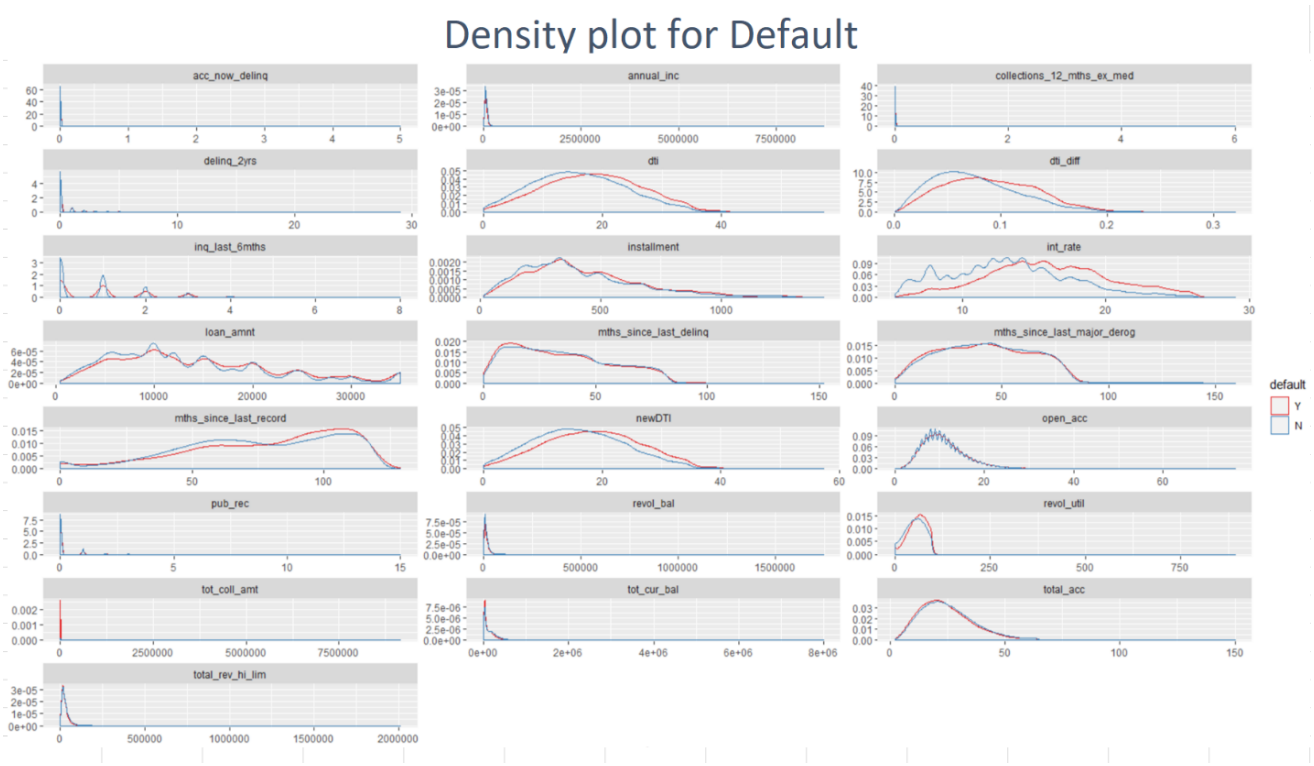


Figure 4 Shows the density plot for several variables. The red curve represents the variable's density plot represented by the defaulter data.

Below is chart visualizing the default rate by term. The y-axis counts the number of defaults. The data contains mostly 36-term loans. Eighty-five percent of loans that have a term of 36 months reach maturity and do not default; only 15% of these default. On the other hand, 29% percent of 60 month loans eventually default, and 71% reach eventual fully paid status. This shows that 60-month term loans are more likely to default; this makes sense because the higher the loan amount, the higher the term. Thus, there exists some multicollinearity between these two variables.

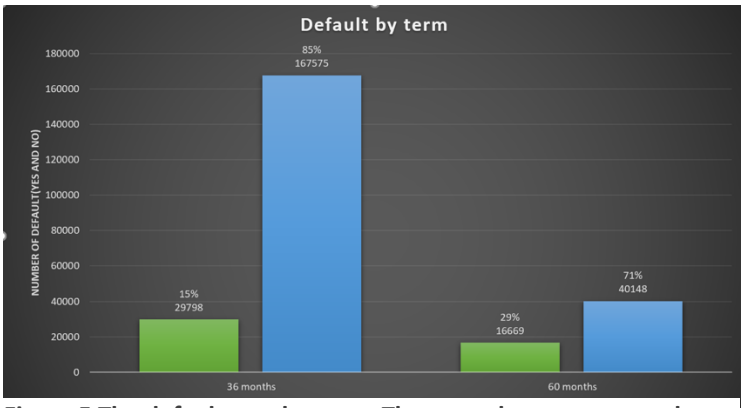


Figure 5 The default rate by term. The green bar represents the default rate, and the numbers on top of the bars represent the number of defaults in that category.

Feature Selection.

Ran with Weka's information gain evaluator and the Ranker search method with 10-fold cross validation, all features with an information gain close to 0 were removed from the analysis. These attributes were also cross-checked with the correlation table to ensure that they indeed would not be useful towards the prediction of default.

The final attributes selected for modeling are these 22 attributes: *int_rate*, the loan grades: (*grade_a, grade_b, grade_e, grade_f, grade_g*), *dti_diff, dti, thirtysix_mo, revol_util, annual_inc, tot_cur_bal, loan_amnt, verification_status*, the home ownership statuses: (*mortgage, rent*), *installment, total_rev_hi_lim, inq_last_6mths, credit_age, total_acc, Last_Record,* the loan purpose: *Small_Business*, the employment length: *emp_length_10plus, Derog, and Delinq.*

Features With The Most Information Gain

```
=== Attribute selection 10 fold cross-validation (stratified), seed: 1 ===

average merit      average rank  attribute
 0.05  +- 0          1   +- 0        2 int_rate
 0.017 +- 0          2.2 +- 0.6     29 grade_a
 0.016 +- 0          3.3 +- 0.64    57 dti_diff
 0.013 +- 0          6   +- 0        6 dti
 0.011 +- 0          7   +- 0       52 thirtysix_mo
 0.009 +- 0          8   +- 0       30 grade_b
 0.008 +- 0          9   +- 0       32 grade_e
 0.007 +- 0         10.4 +- 0.49    12 revol_util
 0.007 +- 0         10.6 +- 0.49     4 annual_inc
 0.006 +- 0         12.2 +- 0.4     17 tot_cur_bal
 0.006 +- 0         12.9 +- 0.54    33 grade_f
 0.005 +- 0         13.9 +- 0.3      1 loan_amnt
 0.005 +- 0         15   +- 0        5 verification_status
 0.004 +- 0         16.5 +- 0.5     22 mortgage _
 0.004 +- 0         18.1 +- 0.54    20 rent
 0.004 +- 0         18.5 +- 1.02     3 installment
 0.003 +- 0         20   +- 0       18 total_rev_hi_lim
 0.002 +- 0         21   +- 0        8 inq_last_6mths
 0.002 +- 0         22   +- 0       34 grade_g
 0.001 +- 0         23   +- 0       48 credit_age
 0.001 +- 0         24   +- 0       13 total_acc
 0.001 +- 0         25   +- 0       51 application_type_dummy
 0.001 +- 0         26.4 +- 0.49    53 Last_Record
 0.001 +- 0         27.2 +- 0.87    37 Small_Business
 0.001 +- 0         27.4 +- 0.66    27 emp_length_10plus
 0.001 +- 0         29   +- 0       50 Derog
 0.001 +- 0         30.2 +- 0.4     49 Delinq
```

These features are kept in the final analysis.

Features With The Least Information Gain

```
0     +- 0         31   +- 0.63    35 Credit_Card
0     +- 0         31.8 +- 0.4     11 revol_bal
0     +- 0         33.2 +- 0.4     16 tot_coll_amt
0     +- 0         34   +- 0.63     7 delinq_2yrs
0     +- 0         34.8 +- 0.4     41 Home_Improvement
0     +- 0         36.8 +- 0.98     9 open_acc
0     +- 0         37   +- 1       36 Car
0     +- 0         38.1 +- 0.83    31 grade_c
0     +- 0         38.1 +- 0.94    14 collections_12_mths_ex_med
0     +- 0         40   +- 0       38 Purpose_Other
0     +- 0         41.2 +- 0.6     44 Moving
0     +- 0         42.5 +- 0.92    43 Medical
0     +- 0         43.1 +- 0.83    10 pub_rec
0     +- 0         43.2 +- 0.75    24 emp_length_lessthan1
0     +- 0         45.2 +- 0.4     25 emp_length_1to5
0     +- 0         45.8 +- 0.4     42 Major_Purchase
0     +- 0         49   +- 0.63    47 Renewable
0     +- 0         49.9 +- 2.66    21 own
0     +- 0         50.4 +- 3.38    46 House
0     +- 0         50.8 +- 2.09    26 emp_length_6to9
0     +- 0         51.2 +- 2.75    40 Debt_Consolidation
0     +- 0         52.2 +- 1.78    45 Vacation
0     +- 0         52.5 +- 3.64    28 emp_length_is_NA
0     +- 0         52.8 +- 3.16    15 acc_now_delinq
0     +- 0         53   +- 2.37    39 Wedding
0     +- 0         53.2 +- 0.87    23 other
```

These features are removed in the final analysis.

Modeling.

Precision is equal to the number of true positives over the number of true positives plus false positives. On the other hand, recall measures the stratified accuracy: the number of true positives over the number of false positives. The goal here is to find a model with a high weighted accuracy and also relatively desirable recall and precision rates. However, the ROC area under the curve will be the primary measure to determine the best model for default rate. The data will be oversampled to create a 1:1 ratio between the minority and majority class and remedy the problem of class imbalance. The models are also run without re-sampling to create a baseline accuracy for comparison.

Naïve Bayes.

Naïve Bayes gave an accuracy of 63.93%, a recall rate of .636 for class 0 and a recall rate of .651 for class 1. The ROC Area is .695, which is one of the highest out of all models. The precision was similar to that of all the other models; in general, the precision does not improve with any model.

```
=== Evaluation on test set ===

Time taken to test model on supplied test set: 31.83 seconds

=== Summary ===

Correctly Classified Instances       38618               63.9286 %
Incorrectly Classified Instances     21790               36.0714 %
Kappa statistic                          0.2156
Mean absolute error                      0.3853
Root mean squared error                  0.5078
Relative absolute error                 77.0693 %
Root relative squared error            101.5577 %
Total Number of Instances            60408

=== Detailed Accuracy By Class ===

                 TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
                 0.636    0.349    0.865      0.636    0.733      0.241    0.695     0.877     N
                 0.651    0.364    0.338      0.651    0.445      0.241    0.695     0.379     Y
Weighted Avg.    0.639    0.352    0.748      0.639    0.669      0.241    0.695     0.766

=== Confusion Matrix ===

     a     b   <-- classified as
 29884 17109 |     a = N
  4681  8734 |     b = Y
```

Logistic Regression.

Using *dti_diff*, the accuracy rate rose to 65.41%. The Recall and Precision rates are also the best out of all other models.

The Logistic Regression Output:

```
=== Evaluation on test set ===

Time taken to test model on supplied test set: 31.05 seconds

=== Summary ===

Correctly Classified Instances       39514               65.4119 %
Incorrectly Classified Instances     20894               34.5881 %
Kappa statistic                          0.2281
Mean absolute error                      0.4295
Root mean squared error                  0.4645
Relative absolute error                 85.9055 %
Root relative squared error             92.8914 %
Total Number of Instances            60408

=== Detailed Accuracy By Class ===

                 TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
                 0.659    0.363    0.864      0.659    0.748      0.250    0.706     0.888     N
                 0.637    0.341    0.348      0.637    0.450      0.250    0.706     0.387     Y
Weighted Avg.    0.654    0.358    0.749      0.654    0.682      0.250    0.706     0.777

=== Confusion Matrix ===

     a     b   <-- classified as
 30972 16021 |     a = N
  4873  8542 |     b = Y
```

Decision Tree.

Multiple models of J48 decision tree was run in Weka to determine the optimum number of trees for the best recall rates and ROC. When a size of 93 was tried, the overall accuracy was 92% and the stratified accuracy for classes 0(no default) and 1(default) were .605 and .672 respectively, with an ROC AUC of .683. Although the stratified accuracy for class 1 was relatively high, the accuracy for class 0 was only 60%. Similarly, when a tree size of 109 was tried, the overall accuracy rose to 62.7%, with recall rates for classes 0 and 1, .617 and .664 respectively, and an ROC of .683. Using a tree size of 127 gave optimal results. The accuracy rate is 63.13% and ROC AUC is .684. Although the area under the ROC curve is not much different from the other decision tree results, it nevertheless shows that a tree size of about 127 is optimal. Lastly, although a tree size of 155 provided a higher AUC of .688, the weighted accuracy was about .10 lower, and the recall rates were similar to that of the size 93 tree. Further details of the decision tree of size 127 are given below in the Weka output.

Decision Tree Output, with size 127: the recall rates look more 'smoothed out' here, with a recall rate of 62.4% for class 0 and a recall rate of 65.8% for class 1.

```
=== Summary ===

Correctly Classified Instances       38136               63.1307 %
Incorrectly Classified Instances     22272               36.8693 %
Kappa statistic                          0.2088
Mean absolute error                      0.443
Root mean squared error                  0.4736
Relative absolute error                 88.6032 %
Root relative squared error             94.7182 %
Total Number of Instances            60408

=== Detailed Accuracy By Class ===

                 TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
                 0.624    0.342    0.865      0.624    0.725      0.236    0.684     0.871     N
                 0.658    0.376    0.333      0.658    0.442      0.236    0.684     0.343     Y
Weighted Avg.    0.631    0.350    0.747      0.631    0.662      0.236    0.684     0.754

=== Confusion Matrix ===

     a     b   <-- classified as
 29309 17684 |     a = N
  4588  8827 |     b = Y
```

The confusion matrix shown above also characterizes the optimal size decision tree. Overall, the results of the decision tree algorithm is successful.

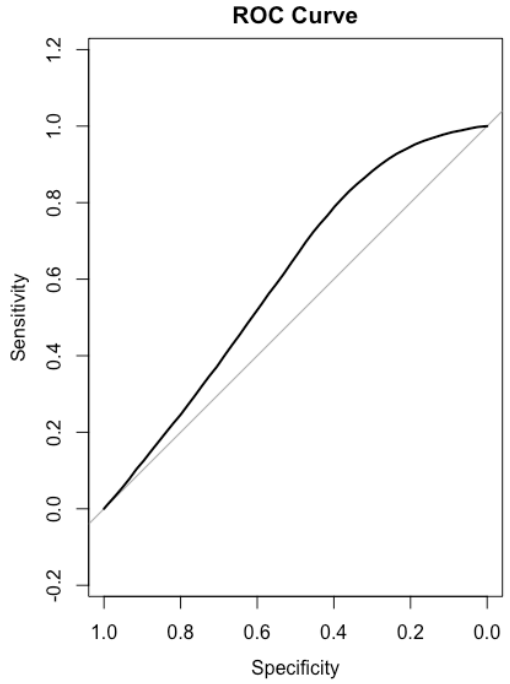
Random Forest.

Although random forest was expected to improve on the accuracy of the decision tree, the results are not better than that of the other models. The accuracy rate for our random forest model is 68.5%, which is the highest accuracy out of all the models; however, the recall rate of class 1 is fairly low, at .57. In contrast, the recall rate of class 0 is high, at .71, hence the optimistic weighted accuracy of 68.5%.

```
         predicted
observed     0     1
       0 33637 13356
       1  5666  7749
```

```
           precision    recall        f1
0 0.8558380 0.7157875 0.7795726
1 0.3671642 0.5776370 0.4489571
```

An ROC curve was plotted to show the relationship between specificity(precision) and sensitivity(recall). In general, a value of .70-.80 is acceptable and anywhere between .80-.90 is good. In the case of the random forest model, the ROC area under curve was only .396, indicating that the random forest model is not a good predictor of default rate.



The results are charted against each other in the below table. The models are run with oversampling to prevent bias while training models and are also run with and without *dti_diff.* In almost all the models, each algorithm fares better with the leverage of this variable. For example, the accuracy rate in logistic regression is raised by almost .060 with the inclusion of this variable. The ROC Area Under Curve also receives marginal gain, albeit only a small amount. Precision does not fluctuate very much in each model, but the recall, which refers to the stratified accuracy of each class, is a little different amongst all classes. For example, in the decision tree model, recall for class 0 is at .622 while recall for class 0 in the logistic regression is .649. Clearly, there is an increase for class 0, but there is also a trade off for class 1: the recall rate declines by .01.

In all cases, the models predict better with the oversampling method than with no resampling. This is because the data itself contains far more 1's than 0's: there are only 22% of loans that eventually default. With the class imbalance problem solved, the models are more likely to perform without bias and are able to capture more abnormalities in the data.

		Oversampling				No resampling	
		With *dti_diff*		Without *dti_diff*			
		Class: 1	Class: 0	Class: 1	Class: 0	Class:1	Class: 0
Logistic Regression	**precision:**	0.348	0.864	0.345	0.865	0.544	0.786
	recall:	0.637	0.659	0.647	0.649	0.058	0.986
	weighted accuracy:	65.41%		64.85%		78.00%	
	ROC Area	0.706		0.705			
J48 Decision Tree	**precision:**	0.333	0.865	0.332	0.864	0.583	0.784
	recall:	0.658	0.624	0.657	0.622	0.042	0.991
	weighted accuracy:	63.13%		63.00%		0.78%	
	ROC Area	0.684		0.684		0.641	
Random Forest	**precision:**	0.367	0.856	0.869	0.687	0.78	0.997
	recall:	0.577	0.716	0.605	0.687	0.692	0.0187
	weighted accuracy:	68.50%		66.90%		78%	
	ROC Area	0.396		0.412		0.3	
Naïve Bayes	**precision:**	0.338	0.865	0.333	0.866	0.835	0.384
	recall:	0.651	0.636	0.666	0.619	0.793	0.451
	weighted accuracy:	63.90%		62.90%		0.74%	
	ROC Area	0.695		0.693		0.69	

Conclusion.

The best model is logistic regression with an overall accuracy of 65.41 and the highest ROC AUC of .706. Our recommendations to investors, when looking for a loan to buy is to look for loans with a grade of a and a low debt-to-income ratio. Surprisingly, since lower grades of e, g, or f are not as important to the prediction as a grade of a, there is still a good chance that loans of lower grades will not default. It is ultimately up to the discernment of the investor: lower grades result in higher interest rates with a chance of high returns. On the other hand, borrowers who have a higher income are also less likely to default, and so are borrowers who have lower interest rates. In summary, the financial market is a world of tug-of-war: in a high risk, high return environment, investors must navigate the lending platform with keenness and aptitude, and most importantly, with understanding of the types of customers most likely to default.

References.

[1] Lending Club Website. https://help.lendingclub.com/hc/en-us/articles/216127747

[2] Lending Club SEC Filing.
https://www.sec.gov/Archives/edgar/data/1409970/000089161808000318/f41480orsv1.htm

[3] Anahita Namvar, Mohammad Siami , Fethi Rabhi , Mohsen Naderpour.: 'Credit risk prediction in an imbalanced social lending environment.' https://arxiv.org/pdf/1805.00801.pdf