Risk Assessment for Loan Investment

Group 7

Sheetal Chowdhary

Maaz Kamal

Rohit R Patil

Table of Contents

1. Abstract

LendingClub is a peer-to-peer loan lending platform. It enables borrowers to get unsecured loans between $1000 and $4000. The investors can then search and lend money to borrowers based on the information about the borrower, the loan amount and the purpose of the loan.

For evaluating the creditworthiness of their borrowers, Lending Club relies on many factors related to borrowers such as credit history, employment, income, ratings etc. Lending club then assigns rating/sub-rating to their borrowers based on their credit-history. This rating information is then made available to investors who fund the loan requests. Investors use this information to analyze loan request and adjudicate the approved funded amount. In addition to the grade information, Lending Club provides historical loan performance data to investors for more comprehensive analysis. Borrowers with higher credit score get lower interest rate, whereas borrowers with low credit score get higher interest rates. From the investors perspective lending to borrowers with high interest rate seems more profitable as it will give a higher Return on Investment.

But at the same time there is a risk of the loan not being returned. As per the recent studies, 3-4% of the total loans defaults every year. There is a huge risk for the investors who is funding the loans. Investors require more comprehensive assessment of these borrowers than what is presented by Lending Club to make a smart business decision. Data mining techniques and Machine Learning modelling/analysis could help predicting the loan default likelihood which may allow investors to avoid loan defaults thus limiting the risk of their investments. The goal of our project is to assist such investors in predicting which high interest loans are more likely to be returned and help them in finding worthy borrowers to lend their money.

The focus of this report was to estimate the probability that a borrower will default the loan using the LendingClub data. We also tried to predict the fraction of the loan that a borrower will pay back. This will help an investor get an idea of how much money can he expect back in each time frame. We also build a model to compute the return on investment for the investor which will help the investor to find profitable loan requests.

For predicting the loan status, we find that the purpose of the loan is a very crucial factor. To be specific if the loan is for debt consolidation or credit card, it has a higher probability of defaulting. Other loan purpose with higher probability of defaulting are home purchase, medical, small business, car, etc. Also, LendingClub categorize each loan into loan grades according to the interest rates. These grades also affect the loan status. The loan with higher interest rate is more likely to default compared to a similar loan with low interest rate. Also, a short-term loan is less likely to default.

To predict the balance payment amount the most significant predictors were the loan amount and installment. Also number of public bankruptcies records and public derogatory records affects the borrower paying the loan amount on time.

To predict the return on investment we see that loans with higher interest rates and lower term have a higher return on investment which is obvious. But our analysis also showed that borrowers whose verification is completed and who pays the loan through Direct Pay rather than cash has a higher return on investment too.

Our analysis could give some useful insights which when available to an investor she/he can take an informed decision about whether to invest in a particular loan request or not.

2. Data Collection/ Cleaning / Exploration

We are using Lending Club dataset provided by Nathan George from Kaggle. The dataset contains 151 columns and 2260701 rows which are a mix of categorical and numerical type. This dataset contains complete loan data for all loans issued through the 2007-2018, including the current loan status (Current, Late, Fully Paid, etc.) and latest payment information. The file contains complete loan data for all loans issued through the previous completed calendar quarter. Additional features include credit scores, number of finance inquiries, address including zip codes, and state, and collections, among others. We are using subset of this dataset and our dataset contains all 151 columns but only 1048575 rows. W

To deal with null values we have removed the columns with more than 30% null values, because using these columns would lead to a significant loss in the information we could use for our analysis. It might affect our analysis in a negative way. Next we have checked the distribution of some of the important variables like loan amount, loan status, grade, etc. Some of the visualization are depicted in **Error! Reference source not found.**, Figure 2 and Figure 3

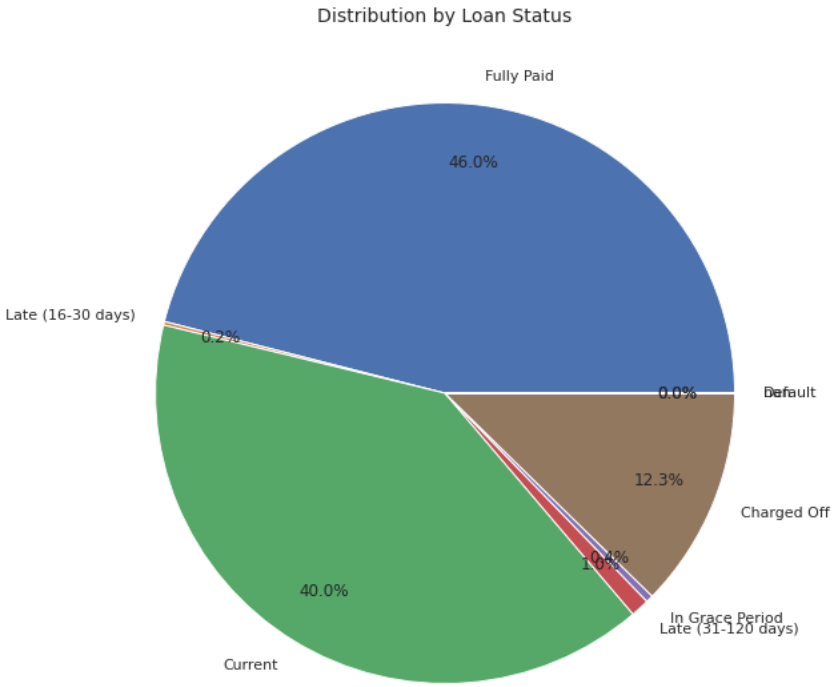

Figure 1: PieChart of Loan-status distribution

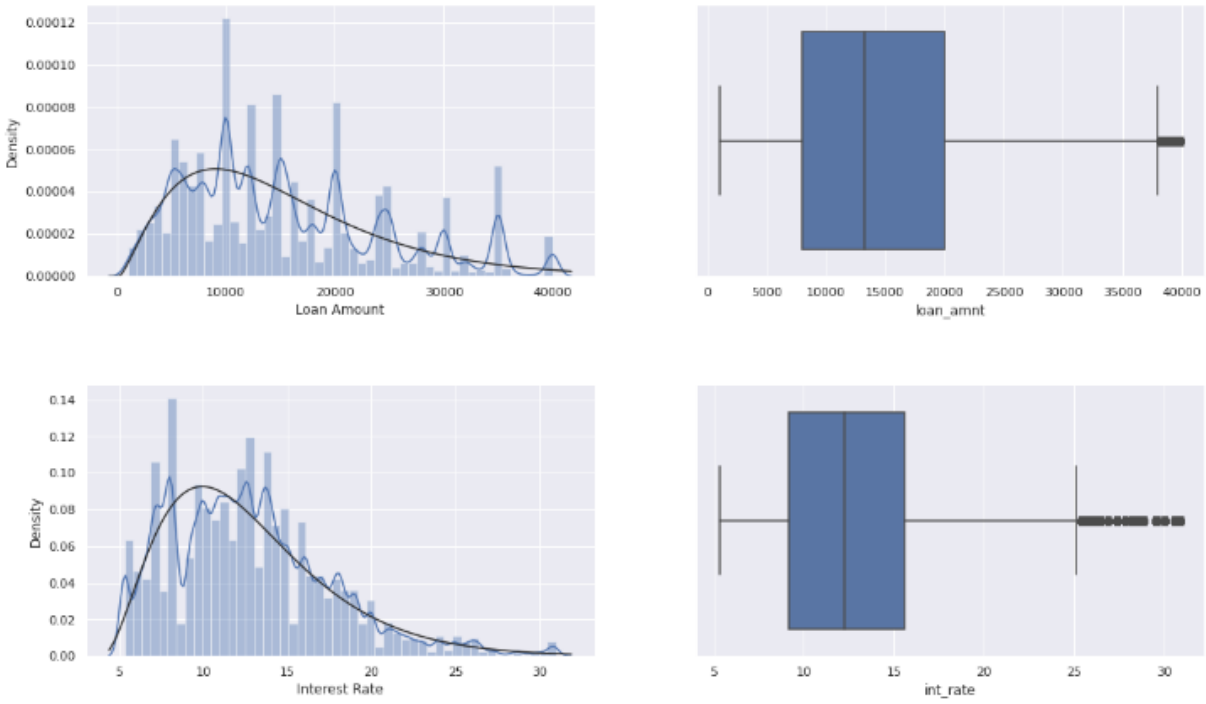

Figure 2: Histogram and box-plot of Interest Rate and Loan Amount

After this initial cleaning of null-value columns we have 93 columns left. We further removed some columns like url, zipcode, emp_title which were descriptive in nature and didn't have any relevance toward the prediction. Next, we meticulously analyzed what all the other variables mean by referring to the Data Dictionary for this data and doing some research on our own. After careful consideration we dropped the ones which we felt weren't relevant in predicting whether a loan will get default or not. Some fields are for internal use and not necessarily available to the investor at the time of investment. Some are current loan values like pymt_plan. It indicates that the loan is in jeopardy and that the borrower has been placed on a payment plan. So we dropped such fields.

Looking at the pair-plots in **Error! Reference source not found.Error! Reference source not found.**, it is clear that other than a few predicting variables, there is no real correlation which might be helpful when using linear regression.

Next we have removed the null values in some of the variables and imputed them with mean values for e.g. dti, revol_util, bc_util. The column emp_length had 69949 null values. Null values in emp_length meant no employment so we created a new column emp_status in which 0 means

non-employed and 1 means employed. So, we changed all 69949 null values in emp_length as 0 and corresponding emp_status as 0. And converted some of the categorical variables into numeric by label encoder or one hot encoding. Converted all the numeric columns which were in string to integer or float type as applicable. To make sure that we don't use any highly correlated variables to build our models we have computed the correlation matrix for all the numeric variables and then removed the variables which have a high correlation value. This is to make sure that we don't have erroneous results due to these correlated variables.

"fico_range_low" & "fico_range_high" scores had high correlation so we took an average of these columns and created a new column fico_range_avg and dropped the previous two columns. We also dropped "earliest_cr_line" column because it affects the "fico_range" so we thought of going ahead with "fico_range_avg".

After this cleaning we are left with about 40 variables. We have used these variables for all of our future model building process.

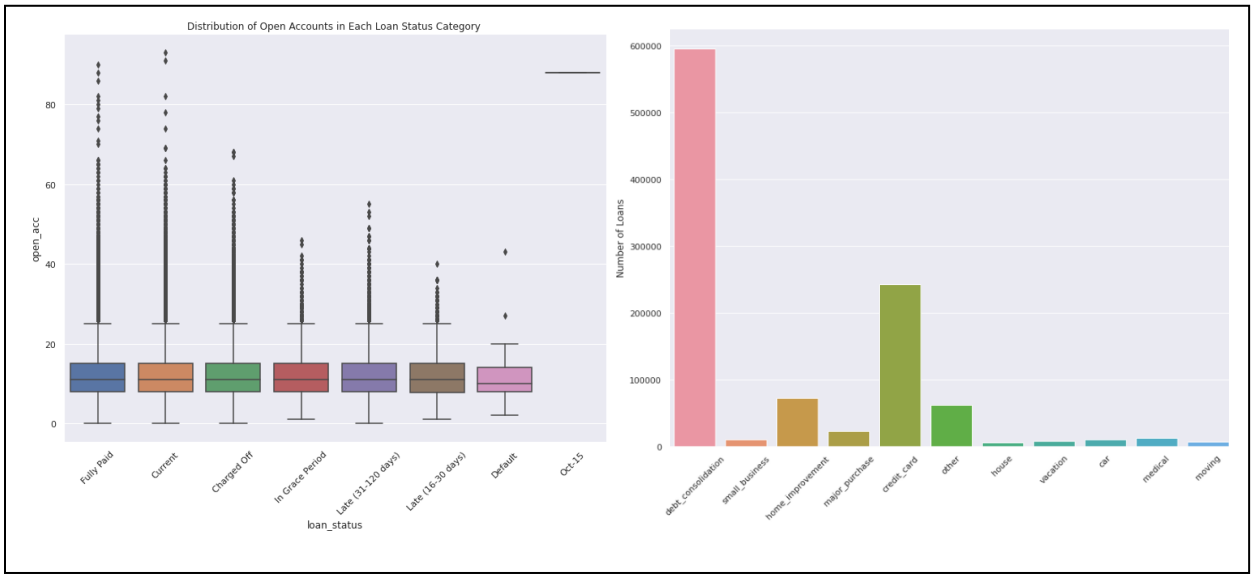

Figure 3: Distribution of OpenAccounts in each loan status (left). Histogram of purpose of loans

We see that the loans labelled as default are very less (12 records - almost 0%). We have included the loans in the category of charged off, late and in grace-period as default too. For this we

created a new column “isDefault” where 0 means defaulted and 1 means not defaulted. Even after this transformation we have only 20% of data that are default, therefore we have computed a weight factor which is the ratio of number of default rows to number of non-default rows and applied it in all our models.

3. Methodology

We had three different set of predictors to work on. For each of these we have used a number of algorithms like Logistic Regression, Random Forest Regression, Linear Regression and Gradient Boosted Trees Regressor to predict our outcome variable. The general flow of our pipeline in building these models is as shown in Figure 4

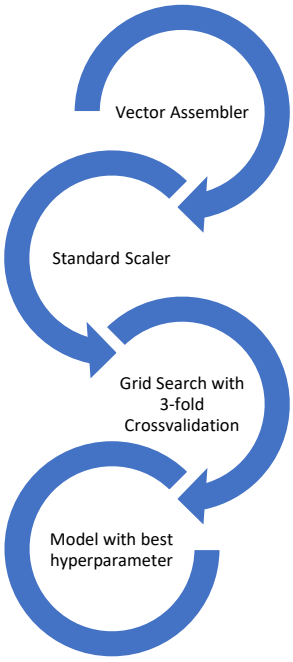

Figure 4: Flow Diagram of the Processes to build a Regression Model

4. Models

4.1. Probability that a given borrower will default

For all the models we have used a set of 43 independent variables to predict the outcome of whether the loan defaults or not.

4.1.1. Logistic Regression Model

For Logistic regression we have computed the MSE, ROC and AUC for this model.

We find that the purpose of the loan is a very crucial factor. To be specific if the loan is for debt consolidation or credit card, it has a higher probability of defaulting. Other loan purposes with higher probability of defaulting are home purchase, medical, small business, car, etc.

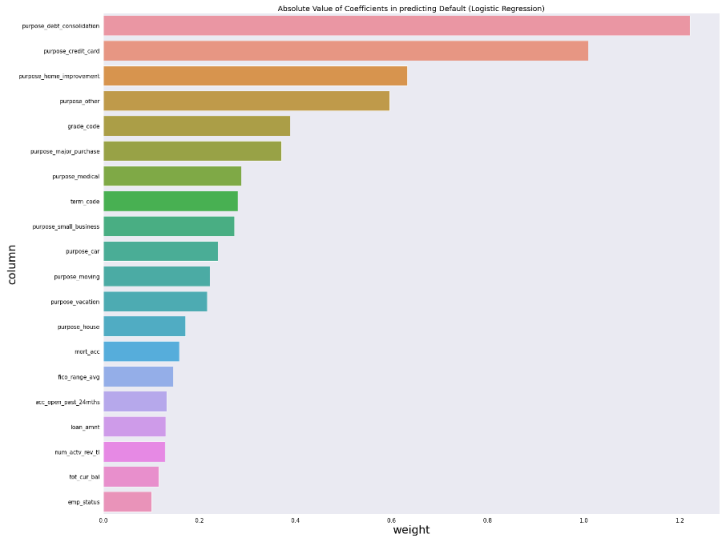

Figure 5: Feature Importance of logistic regression to predict loan status

4.1.2. Random Forest Regression

According to this model the loan grade and the term of the loan is a crucial factor in prediction of the loan status.

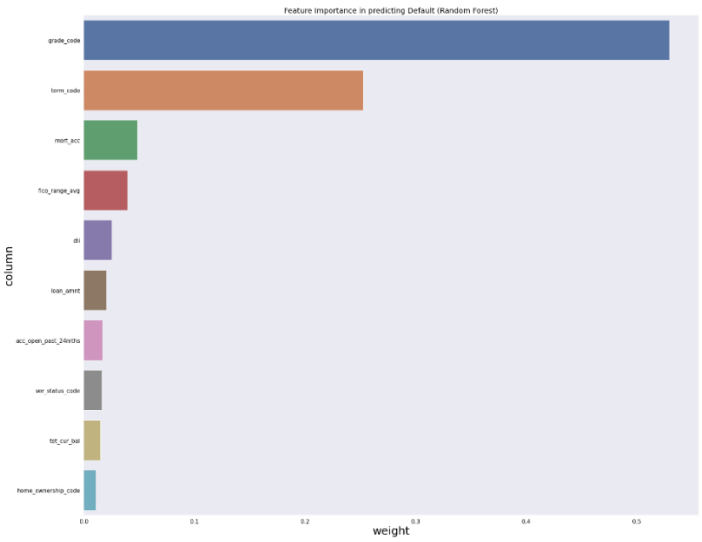

Figure 6: Feature importance of Random Forest to predict loan status

4.1.3. Gradient Boosted Tree Regression

According to this model too, the loan grade and the term of the loan are crucial factors in prediction of the loan status which is similar to the result from Random Forest Regression.

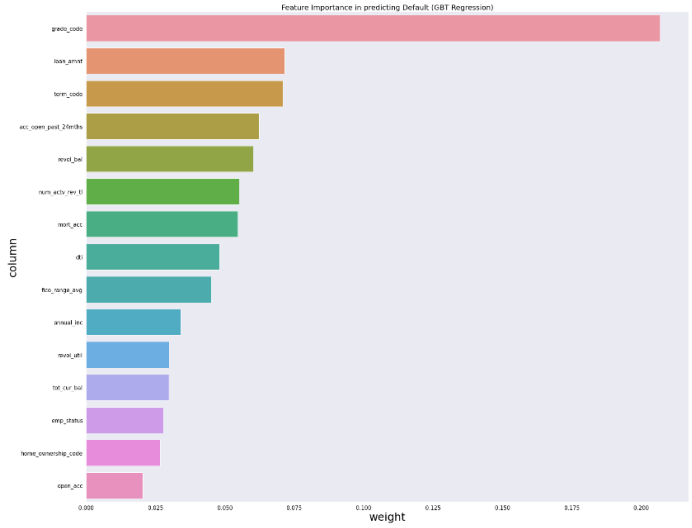

Figure 7: Feature importance for GBT to predict loan status

4.2. Fraction of amount a borrower will return

In this section we have tried to build a model to predict how much amount can an investor expect to be returned by a borrower before the term expires. We have used the same input independent variables as in the previous section. The outcome variable here is the balance of principal amount yet to be paid. Also, we have used a subset of the data here. We filtered the data keeping only the loans qualifying as 'current'.

4.2.1. Linear Regression

According to this model the type of applicant (individual or joint) is the most significant predictor. Also, public recorded bankruptcies and number of public derogatory records also affect the amount of balance left to be paid by the borrower.

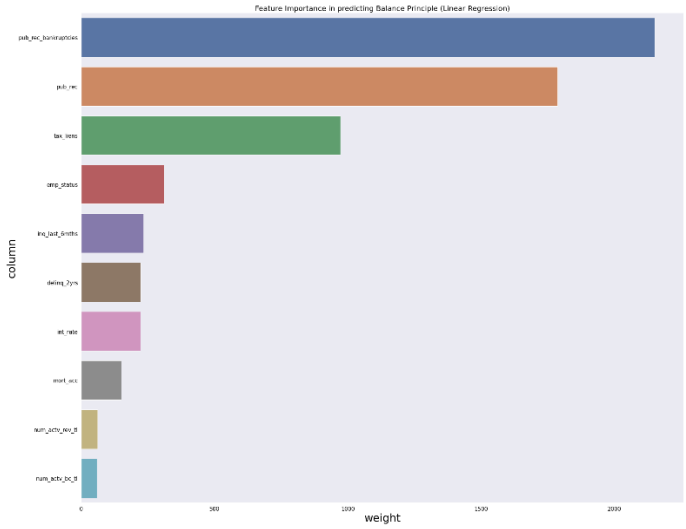

Figure 8: Feature Importance for Linear Regression to predict balance payment

4.2.2. Random Forest Regression

According to this model loan amount is the most significant predictor. This is an obvious predictor though since higher the loan amount more would be the balance remaining in a given time. But apart from that interest rate and term length of the loan is also a significant predictor.

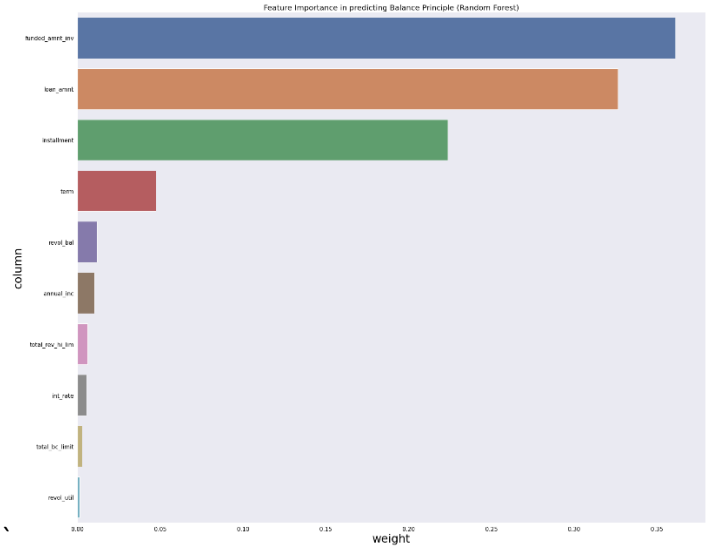

Figure 9: Feature Importance of Random Forest to predict balance payment

4.2.3. Gradient Boosted Tree Regression

According to this model too the loan amount, installment and interest rate affect the balance payment by a borrower.

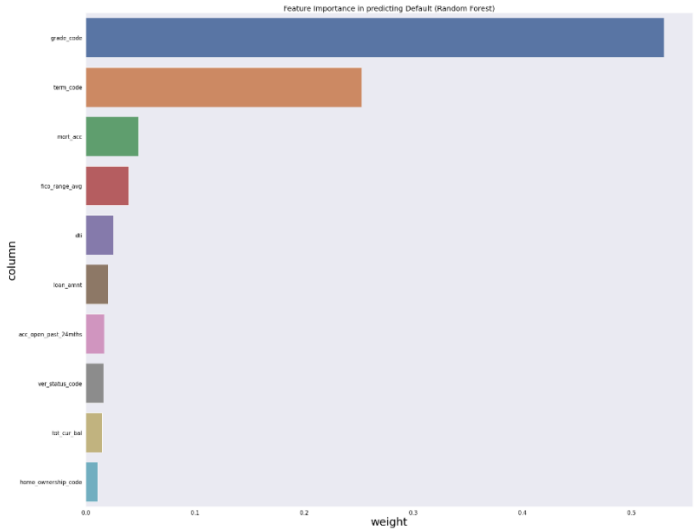

Figure 10:Feature Importance for GBT to predict balance payment

4.3. Total Return on Investment

In this section we have tried to build a model to predict what would be the expected return on investment for the investor. We have used a few different predicting variables compared to the previous sections. The outcome variable here is the “return_inv” (return on Investment) which was computed using the variables “total_pymnt_inv” and “funded_amnt_inv”.

4.3.1. Linear Regression

According to this model “disbursement_method” and “grade” affect ROI (return on investment) the most. When mode of disbursement changes from "Cash” to “Direct Pay”, the ROI increase decreases by .4%. When grade of the investment changes from “A” to “B”, the ROI increases by .2%.

Table 1:Linear Regression Coefficients for predicting Return on Investment

coefficient	Values
funded_amnt_inv	-0.00001
int_rate	-0.07109
installment	0.00026
annual_inc	0

dti	-0.00085
fico_range_low	-0.00099
open_acc	-0.00354
total_acc	0.00285
tot_cur_bal	0
term	-0.04447
grade	0.25959
emp_length	0.00077
home_ownership	0.00724
verification_status	-0.02849
purpose	-0.00306
disbursement_method	-0.46334

4.3.2. Random Forest Regression

According to this model Disbursement Method is the most significant predictor. Which was already concluded from the Linear Regression Model. Interestingly, however, term of investment turns out to be the second most significant predictor.

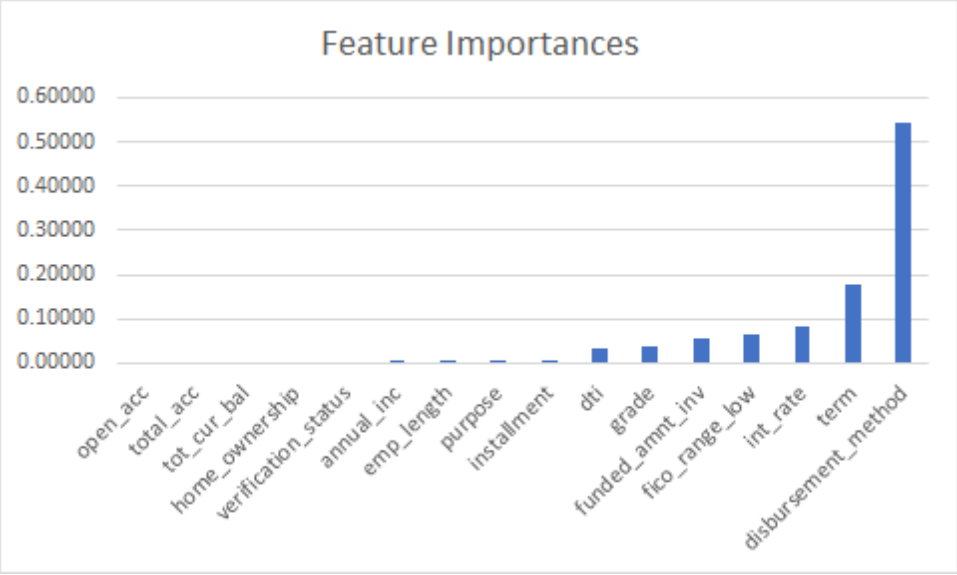

Figure 11: Feature Importance for Random forest to predict ROI

5. Conclusion

The following tables summarizes the result of our analysis.

Table 2: Model Comparison for Predicting Loan default

	Best Features	Accuracy	RMSE
Logistic Regression	Loan Purpose (debt consolidation or credit card)	77.2931%	0.478024
Random Forest	Loan grade and Term length	77.2703%	0.429977
Gradiant Boosted Tree	Loan grade and Term length	77.3327%	0.424205

Table 3: Model Comparison for predicting balance principal

	Best Features	RMSE
Linear Regression	Public record bankruptcies, number of derogatory public records	4312.49
Random Forest	Loan Amount, installment	4275.11
Gradiant Boosted Tree	Loan Amount, interest rate, installment	3739.77

Table 4: Model Comparison for predicting Return on Investment

	Best Features	MSE
Linear Regression	Disbursement Method, Grade, Interest Rate	0.116
Random Forest	Annual Income, Installment, Funded Amount	0.126

In conclusion, we can say that an investor looking for loans which will not default should choose from the loans with lower grade or longer term, or whose purpose of loan is something other than debt consolidation or credit card dept.

An investor expecting his loan to be returned on time should choose higher loan amounts having higher interest rate and installment amounts.

Finally, the highest return on investment is expected for loans with higher rate of interest, and loans with direct pay as disbursement method rather than cash and the borrowers should have higher annual income.