**Assignment 1 IDS 572**

Anchit Jhingan | 663246540

Dushyant Singh Khinchi | 662632496

Pallavi Srikanth | 664941288

1. LendingClub is the world’s largest online peer-to-peer lending company. Founded in 2007, LendingClub was initially launched on Facebook to be a social networking platform to create a community where members could borrow and lend money at better interest rates.

Lending club operates by having two major stakeholders – the borrowers and the investors. The company offers 4 different types of loans – Personal Loans, Business Loans, Auto Refinancing and Patient Solutions, with a standard loan period of 3 years. Borrowers can apply for the loan they want through the LendingClub website by specifying details on the loan amount, purpose, credit score, credit history, employment status and debt to income ratio. Based on all these factors, different monthly payment plan - interest rates are provided, and the borrower can select the combination that suits them the best. Once approved, the loan is then deposited in their bank account within a week. The loans that are assigned have different grades – A (highest grade) to G (lowest grade), with 5 subgrades from 1-5 in each of them. Depending on the grade of the loan, the interest rate also varies from 6.03% - 26.06%. Peer to peer lending offers quite some advantages to the borrowers such as:

* A faster experience that happens completely online
* Generally lower rates of interest if you have a good credit score
* No prepayment penalties
* Fixed monthly payments

What sets LendingClub apart from traditional money lending solutions is that individuals can become loan investors by browsing through all the loan listings on the LendingClub website. After viewing the loan purpose, amount and grade, investors can choose which loan to invest in have to invest a minimum of $25 (to a maximum of the entire borrowed amount) in the loans they choose, and they make their money from the interest charged on the loan amount. Investors typically enjoy advantages such as:

* High returns which are several percentage points above those achieved by banks and are even higher than savings or CD accounts
* The ability to diversify your loan portfolio
* Getting a sense of community since investors get to personally choose who to back financially

LendingClub itself makes money by charging a fee to borrowers and commanding a percentage of the interest earned on all loans from the investors. To the borrowers, they charge a one-time origination fee of 1.11%-5% on the original principal amount and take a 1% fee of each payment amount from the investor. If a borrower makes a $200 payment on a loan, LendingClub would take $2 before passing the payment on to investors.1 Investors also have the option to sell the loan ‘notes’ before reaching maturity. LendingClub offers this service along with in partnership with FOLIOfn Investments which charges a 1% fee on any note sale. Thus, LendingClub was the first peer-to-peer lending company to offer a second market for peer-to-peer loans.

1. (a)
2. Through the loan\_status column, we define default as a ‘Charged Off’ loan and a non-default as a ‘Fully Paid’ loan. With a total 120,790 default and non-default loans, a total of 18,269 loans have been defaulted which is 15.12% of all loans. The default rate generally increases as we go from Grade A loans to Grade G loans and the same trend is seen within the sub-loans i.e. default rates increase from sub-loan type 1 to sub-loan type 5 within all loan categories. This is inline with the fact that Grade A loans are given to those individuals who have a lower risk and volatility (as these individuals have better credit scores, employment history etc), which increases as you go to the Grade G loans (these individuals have lower credit scores, higher debt-to-income ratio etc). The same applies within the sub-loans with type 1 having less risk and volatility that type 5 sub-loans.

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| **Grade** | **# Loans** | **Max Loan Amount ($)** | **Min Loan Amount ($)** | **Average Loan Amount ($)** | **Max Interest Rate (%)** | **Min Interest Rate (%)** | **Average Interest Rate (%)** |
| A | 30,625 | 35,000 | 1,000 | 14,323 | 8.19 | 5.32 | 7.03 |
| B | 35,746 | 35,000 | 1,000 | 12,589 | 11.99 | 6.00 | 10.16 |
| C | 33,681 | 35,000 | 1,000 | 11,722 | 14.99 | 6.00 | 13.33 |
| D | 15,389 | 35,000 | 1,000 | 12,388 | 17.86 | 6.00 | 16.63 |
| E | 4,626 | 35,000 | 1,000 | 13,506 | 21.99 | 18.25 | 19.26 |
| F | 628 | 35,000 | 1,000 | 11,088 | 25.78 | 21.99 | 23.77 |
| G | 95 | 35,000 | 2,075 | 10,012 | 28.49 | 25.80 | 26.10 |

The number of loans within each loan grade is given in the table above. While the maximum and minimum borrowing amount is the almost the same across all loan grades, the average borrowing amount decreases from grade A loans to grade G loans.

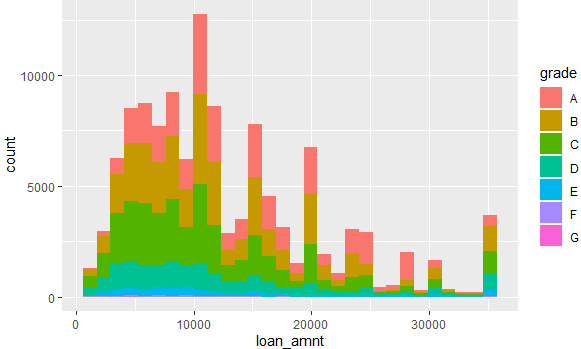
As the interest rates are calculated depending on the individual’s risk and volatility (which increase from grade A to grade G), the interest rates (maximum, minimum and average) increase while going from grade A loans to grade G loans. We see the same trend within the loan sub-grades with type 1 having the lowest interest rate and type 5 having higher interest rates.

1. There are 14 different borrowing purposes as given in the table below. The highest proportion of loans are for debt consolidation (59.6%) and the lowest number of loans are taken for wedding (2 loans) and education purposes (1 loan). The table below gives the average loan amount, average interest rates and the number and percentage of loans that were fully paid and charged off by loan purpose. We can see that the highest number of defaults are for loans taken out for the purpose of small businesses (23.11%) and renewable energy (24%), with these businesses/purposes having a higher risk involved.

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **Purpose** | **# Count** | **Count (%)** | **Avg Loan Amount ($)** | **Avg Interest Rate (%)** | **# Fully Paid** | **# Charged Off** | **Fully Paid (%)** | **Charged off (%)** |
| Debt Consolidation | 71,988 | 59.60% | 13,128 | 11.8 | 60,516 | 11,472 | 84.06% | 15.94% |
| Moving | 866 | 0.72% | 7,083 | 15.0 | 688 | 178 | 79.45% | 20.55% |
| Credit Card | 28,495 | 23.59% | 13,850 | 10.0 | 24,944 | 3,551 | 87.54% | 12.46% |
| Other | 6,031 | 4.99% | 8,628 | 13.7 | 4,997 | 1,034 | 82.86% | 17.14% |
| Home Improvement | 6,631 | 5.49% | 12,275 | 11.4 | 5,751 | 880 | 86.73% | 13.27% |
| Major Purchase | 2,046 | 1.69% | 10,428 | 11.7 | 1,739 | 307 | 85.00% | 15.00% |
| Small Business | 1,056 | 0.87% | 14,167 | 16.1 | 812 | 244 | 76.89% | 23.11% |
| Medical | 1,150 | 0.95% | 7,494 | 13.2 | 947 | 203 | 82.35% | 17.65% |
| Vacation | 857 | 0.71% | 5,753 | 13.6 | 715 | 142 | 83.43% | 16.57% |
| House | 441 | 0.37% | 12,181 | 16.5 | 350 | 91 | 79.37% | 20.63% |
| Car | 1,151 | 0.95% | 8,144 | 11.2 | 1,002 | 149 | 87.05% | 12.95% |
| Renewable Energy | 75 | 0.06% | 7,894 | 15.9 | 57 | 18 | 76.00% | 24.00% |
| Wedding | 2 | 0.00% | 3,600 | 11.8 | 2 | 0 | 100.00% | 0.00% |
| Educational | 1 | 0.00% | 2,200 | 11.5 | 1 | 0 | 100.00% | 0.00% |

For each loan purpose, the number of loans decreases form grade A to grade G loans as given in the table below.

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| **Purpose** | **A** | **B** | **C** | **D** | **E** | **F** | **G** |
| Debt Consolidation | 15,935 | 21,491 | 21,413 | 9,950 | 2,839 | 332 | 28 |
| Moving | 22 | 94 | 353 | 281 | 98 | 13 | 5 |
| Credit Card | 11,328 | 9,403 | 5,789 | 1,588 | 343 | 39 | 5 |
| Other | 509 | 1,314 | 2,229 | 1,379 | 494 | 86 | 20 |
| Home Improvement | 1,798 | 1,923 | 1,761 | 848 | 272 | 29 | 0 |
| Major Purchase | 516 | 556 | 557 | 306 | 96 | 12 | 3 |
| Small Business | 26 | 98 | 314 | 337 | 210 | 54 | 17 |
| Medical | 98 | 282 | 450 | 242 | 69 | 7 | 2 |
| Vacation | 55 | 190 | 353 | 193 | 53 | 12 | 1 |
| House | 12 | 44 | 118 | 125 | 94 | 35 | 13 |
| Car | 324 | 342 | 317 | 119 | 43 | 5 | 1 |
| Renewable Energy | 2 | 7 | 26 | 21 | 15 | 4 | 0 |
| Wedding | 0 | 1 | 1 | 0 | 0 | 0 | 0 |
| Educational | 0 | 1 | 0 | 0 | 0 | 0 | 0 |



1. Annual return is calculated for non-default loans. The value is calculated as given below:

where loan duration is calculated as:

We can observe that the average annual return rate is less than the average interest rate. This is due to the fact that most individuals pay back the loan amounts before the loan period of 3 years. Due to this, the return observed by the investor is less than the overall interest rate charged on the loan.

We can observe the highest returns among the A, B and C grade loans and lower returns from loans with grade D, E and F. We are observing a difference in the average return rate between the F and G loans which can be attributed to the higher default rate in F loans as compared to G loans. Overall, the Annual Return rate reduces as we go from A1 sub-loans to G5 sub-loans.

|  |  |
| --- | --- |
| **Grade** | **Average Return** |
| A | 2.26% |
| B | 2.49% |
| C | 2.32% |
| D | 2.01% |
| E | 1.29% |
| F | 0.39% |
| G | 2.22% |

1. Some derived attributes that may be useful for predicting default are:
   1. Ratio of open accounts to total accounts
   2. Ratio of satisfactory bank card accounts to total bank card accounts

(b) There are a 19 variables with missing values in the data set and the proportion of missing values among these variables ranges from 0.0033% to 100% (excluding variables who have more than 99% missing values). These variables have been listed below along with the proportion of missing values:

|  |  |
| --- | --- |
| **Variable** | **% Missing Values** |
| emp\_title | 6.25% |
| mths\_since\_last\_delinq | 47.91% |
| mths\_since\_last\_record | 81.29% |
| revol\_util | 0.04% |
| last\_pymnt\_d | 0.06% |
| next\_pymnt\_d | 100.00% |
| last\_credit\_pull\_d | 0.0033% |
| mths\_since\_last\_major\_derog | 70.00% |
| bc\_open\_to\_buy | 1.02% |
| bc\_util | 1.09% |
| mo\_sin\_old\_il\_acct | 3.67% |
| mths\_since\_recent\_bc | 0.97% |
| mths\_since\_recent\_bc\_dlq | 73.46% |
| mths\_since\_recent\_inq | 10.65% |
| mths\_since\_recent\_revol\_delinq | 63.26% |
| num\_tl\_120dpd\_2m | 3.54% |
| percent\_bc\_gt\_75 | 1.07% |
| hardship\_dpd | 99.91% |
| settlement\_term | 99.56% |

We have employed a method of replacing missing values of a particular variable with the median value from the data of the variable so that our model has enough information to complete the learning process.

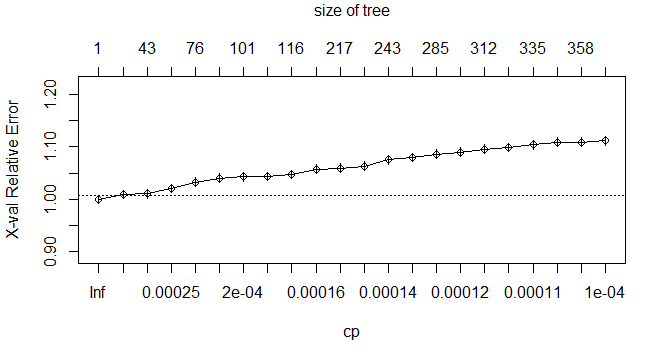
1. There are three kinds of variables that we have chosen to exclude while building our models:
   1. Variables that have more than 99% missing values
   2. Variables that we think have no relation to whether an individual will default or fully pay a loan
   3. Variables whose values have been computed after/during the loan period has been completed

By excluding these variables, we are ensuring that our models are created based only on pertinent information which is present at the time that an individual is buying the loan and the learning algorithm has not been trained on information that would not be present in a real world scenario. The variables that have been excluded from the model have been listed below:

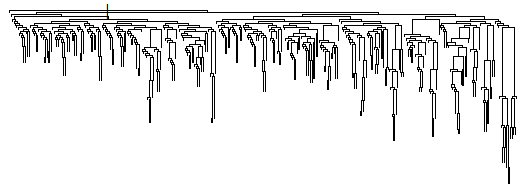
|  |  |
| --- | --- |
| **Variable** | **Reason from excluding from model** |
| mths\_since\_last\_record | Information collected after/during loan period |
| out\_prncp | Information collected after/during loan period |
| out\_prncp\_inv | Information collected after/during loan period |
| total\_pymnt | Information collected after/during loan period |
| total\_pymnt\_inv | Information collected after/during loan period |
| total\_rec\_prncp | Information collected after/during loan period |
| total\_rec\_int | Information collected after/during loan period |
| total\_rec\_late\_fee | Information collected after/during loan period |
| Recoveries | Information collected after/during loan period |
| collection\_recovery\_fee | Information collected after/during loan period |
| last\_pymnt\_d | Information collected after/during loan period |
| Id | More than 99% missing values |
| member\_id | More than 99% missing values |
| url | More than 99% missing values |
| Desc | More than 99% missing values |
| annual\_inc\_joint | More than 99% missing values |
| dti\_joint | More than 99% missing values |
| verification\_status\_joint | More than 99% missing values |
| open\_acc\_6m | More than 99% missing values |
| open\_act\_il | More than 99% missing values |
| open\_il\_12m | More than 99% missing values |
| open\_il\_24m | More than 99% missing values |
| mths\_since\_rcnt\_il | More than 99% missing values |
| total\_bal\_il | More than 99% missing values |
| il\_util | More than 99% missing values |
| open\_rv\_12m | More than 99% missing values |
| open\_rv\_24m | More than 99% missing values |
| max\_bal\_bc | More than 99% missing values |
| all\_util | More than 99% missing values |
| inq\_fi | More than 99% missing values |
| total\_cu\_tl | More than 99% missing values |
| inq\_last\_12m | More than 99% missing values |
| revol\_bal\_joint | More than 99% missing values |
| sec\_app\_earliest\_cr\_line | More than 99% missing values |
| sec\_app\_inq\_last\_6mths | More than 99% missing values |
| sec\_app\_mort\_acc | More than 99% missing values |
| sec\_app\_open\_acc | More than 99% missing values |
| sec\_app\_revol\_util | More than 99% missing values |
| sec\_app\_open\_act\_il | More than 99% missing values |
| sec\_app\_num\_rev\_accts | More than 99% missing values |
| sec\_app\_chargeoff\_within\_12\_mths | More than 99% missing values |
| sec\_app\_collections\_12\_mths\_ex\_med | More than 99% missing values |
| sec\_app\_mths\_since\_last\_major\_derog | More than 99% missing values |
| hardship\_type | More than 99% missing values |
| hardship\_reason | More than 99% missing values |
| hardship\_status | More than 99% missing values |
| deferral\_term | More than 99% missing values |
| hardship\_amount | More than 99% missing values |
| hardship\_start\_date | More than 99% missing values |
| hardship\_end\_date | More than 99% missing values |
| payment\_plan\_start\_date | More than 99% missing values |
| hardship\_length | More than 99% missing values |
| hardship\_loan\_status | More than 99% missing values |
| orig\_projected\_additional\_accrued\_interest | More than 99% missing values |
| hardship\_payoff\_balance\_amount | More than 99% missing values |
| hardship\_last\_payment\_amount | More than 99% missing values |
| debt\_settlement\_flag\_date | More than 99% missing values |
| settlement\_status | More than 99% missing values |
| settlement\_date | More than 99% missing values |
| settlement\_amount | More than 99% missing values |
| settlement\_percentage | More than 99% missing values |
| next\_pymnt\_d | More than 99% missing values |
| hardship\_dpd | More than 99% missing values |
| settlement\_term | More than 99% missing values |
| emp\_title | Information not relevant |
| issue\_d | Information not relevant |
| zip\_code | Information not relevant |
| addr\_state | Information not relevant |
| last\_pymnt\_amnt | Irrelevant information |
| last\_credit\_pull\_d | Irrelevant information |
| mths\_since\_last\_major\_derog | Irrelevant information |
| tot\_cur\_bal | Irrelevant information |
| avg\_cur\_bal | Irrelevant information |
| mths\_since\_recent\_bc\_dlq | Irrelevant information |
| mths\_since\_recent\_revol\_delinq | Irrelevant information |
| num\_il\_tl | Irrelevant information |
| total\_il\_high\_credit\_limit | Irrelevant information |
| debt\_settlement\_flag | Irrelevant information |

1. From the final data set, we have considered a 70-30 split, to split the data into a training data set and a testing data set. This split ensures that we have enough data to train the model and test it for accuracy.

Using rpart we have created a decision tree with 68 variables (after removing the variables given in the table above) and a minimum split of 30. The following CP plot has been obtained:



We have thus used a CP value of 0.0001 to prune the tree. The structure of the final tree we have obtained has been given below:



**Performance Evaluation on Training data:**

Accuracy: 86.15%

Confusion Matrix:

|  |  |  |  |
| --- | --- | --- | --- |
| **Prediction** | **Actual** | | |
|  | Fully Paid | Charged Off |
| Fully Paid | 70,535 | 10,493 |
| Charged Off | 1,218 | 2,308 |

**Performance Evaluation on Testing data:**

Accuracy: 83.20%

Confusion Matrix:

|  |  |  |  |
| --- | --- | --- | --- |
| **Prediction** | **Actual** | | |
|  | Fully Paid | Charged Off |
| Fully Paid | 29,678 | 4,999 |
| Charged Off | 1,090 | 469 |

Model Characteristics

Our pruned tree has a depth of 22

Model variable importance:

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Variable** | **Importance** |  | **Variable** | **Importance** |
| int\_rate | 2110.553303 |  | revol\_util | 130.4918755 |
| sub\_grade | 2103.92424 |  | mths\_since\_recent\_inq | 127.7629619 |
| Grade | 1801.657561 |  | pct\_tl\_nvr\_dlq | 118.4012658 |
| acc\_open\_past\_24mths | 717.9484869 |  | num\_actv\_rev\_tl | 110.554667 |
| total\_bc\_limit | 493.4053443 |  | mo\_sin\_rcnt\_rev\_tl\_op | 108.186396 |
| tot\_hi\_cred\_lim | 462.5204523 |  | num\_rev\_tl\_bal\_gt\_0 | 105.7068992 |
| emp\_length | 459.7406049 |  | mths\_since\_last\_delinq | 103.0216801 |
| installment | 380.5536332 |  | total\_acc | 99.92748692 |
| loan\_amnt | 361.8351499 |  | num\_sats | 91.88547553 |
| funded\_amnt | 317.4235741 |  | num\_op\_rev\_tl | 82.05418226 |
| Dti | 303.3688815 |  | open\_acc | 81.81462019 |
| funded\_amnt\_inv | 286.0251901 |  | num\_bc\_tl | 80.1136277 |
| mo\_sin\_old\_rev\_tl\_op | 248.2323131 |  | num\_rev\_accts | 70.81987882 |
| revol\_bal | 244.915497 |  | num\_actv\_bc\_tl | 59.06401228 |
| bc\_util | 216.1101487 |  | num\_tl\_op\_past\_12m | 57.03136047 |
| earliest\_cr\_line | 212.6363808 |  | percent\_bc\_gt\_75 | 56.61222849 |
| mths\_since\_recent\_bc | 210.4037006 |  | mo\_sin\_rcnt\_tl | 48.83010345 |
| Title | 208.3223044 |  | tot\_coll\_amt | 47.81012216 |
| propSatisBankcardAccts | 190.2478586 |  | num\_accts\_ever\_120\_pd | 29.18210323 |
| purpose | 190.0009899 |  | initial\_list\_status | 26.21738896 |
| annual\_inc | 181.2763946 |  | verification\_status | 23.5525225 |
| bc\_open\_to\_buy | 173.2274945 |  | inq\_last\_6mths | 22.82148188 |
| mort\_acc | 171.9374463 |  | delinq\_2yrs | 22.54245543 |
| total\_rev\_hi\_lim | 165.2174623 |  | pub\_rec | 21.94118722 |
| total\_bal\_ex\_mort | 160.6599386 |  | num\_bc\_sats | 6.844486079 |
| home\_ownership | 157.7950769 |  | delinq\_amnt | 5.243755737 |
| mo\_sin\_old\_il\_acct | 153.2853968 |  | pub\_rec\_bankruptcies | 4.134666403 |
| openacc\_totalacc\_Ratio | 130.6076945 |  | num\_tl\_90g\_dpd\_24m | 2.893920856 |

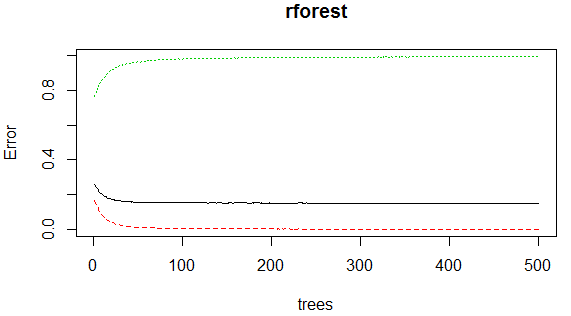
The following variables have a importance score of 0: term, pymnt\_plan, collections\_12\_mths\_ex\_med, policy\_code, application\_type, acc\_now\_delinq, chargeoff\_within\_12\_mths, num\_tl\_120dpd\_2m, num\_tl\_30dpd, tax\_liens, hardship\_flag, disbursement\_method.

A screenshot of text

Description automatically generated

Random Forest Confusion Matrix

1. We have made a random forest using the same split of data (70-30 for training and testing respectively). We have grown the model considering a total number of trees as 500. As seen in the error plot given below, the error decreases significantly with ~100 trees and continues to decrease until 500 trees.



**Performance Evaluation on Training data:**

Accuracy: 100%

Confusion Matrix:

|  |  |  |  |
| --- | --- | --- | --- |
| **Prediction** | **Actual** | | |
|  | Fully Paid | Charged Off |
| Fully Paid | 71,753 | 0 |
| Charged Off | 0 | 12,801 |

**Performance Evaluation on Testing data:**

Accuracy: 84.93%

Confusion Matrix:

|  |  |  |  |
| --- | --- | --- | --- |
| **Prediction** | **Actual** | | |
|  | Fully Paid | Charged Off |
| Fully Paid | 30,746 | 5,439 |
| Charged Off | 22 | 29 |

Accuracy comparison between Decision Tree and Random Forest:

|  |  |  |
| --- | --- | --- |
|  | **On Training Data** | **On Testing Data** |
| **Decision Tree** | 86.15% | 83.20% |
| **Random Forest** | 100.00% | 84.93% |

Thus, as compared to the decision tree model from Q4, the random tree that has been developed is better in terms of accuracy on both the training and testing data.

Variable Importance:

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Variable** | **Importance** |  | **Variable** | **Importance** |
| emp\_length | 1048.480222 |  | num\_op\_rev\_tl | 395.6394081 |
| sub\_grade | 1000.750662 |  | num\_rev\_tl\_bal\_gt\_0 | 360.3031607 |
| dti | 814.6600048 |  | num\_actv\_rev\_tl | 358.2677975 |
| tot\_hi\_cred\_lim | 807.3342984 |  | num\_bc\_sats | 351.6492363 |
| mo\_sin\_old\_rev\_tl\_op | 778.3880873 |  | num\_actv\_bc\_tl | 334.4897187 |
| total\_bal\_ex\_mort | 737.0126871 |  | num\_tl\_op\_past\_12m | 330.7713719 |
| total\_rev\_hi\_lim | 733.9356635 |  | title | 325.9560492 |
| revol\_bal | 733.0415629 |  | purpose | 318.482513 |
| total\_bc\_limit | 731.5151016 |  | grade | 289.9923353 |
| annual\_inc | 706.6832904 |  | mort\_acc | 260.8596852 |
| int\_rate | 705.5193288 |  | tot\_coll\_amt | 235.8651398 |
| installment | 627.8744764 |  | inq\_last\_6mths | 211.4163721 |
| openacc\_totalacc\_Ratio | 588.7135131 |  | num\_accts\_ever\_120\_pd | 186.7879394 |
| total\_acc | 575.4110157 |  | verification\_status | 171.067318 |
| mo\_sin\_rcnt\_rev\_tl\_op | 529.6497825 |  | delinq\_2yrs | 160.5447494 |
| mths\_since\_recent\_inq | 505.4860789 |  | home\_ownership | 151.7775206 |
| funded\_amnt\_inv | 499.7569809 |  | pub\_rec | 124.7367981 |
| num\_rev\_accts | 497.1563469 |  | pub\_rec\_bankruptcies | 88.87115425 |
| funded\_amnt | 489.3193643 |  | initial\_list\_status | 84.93286975 |
| loan\_amnt | 488.7130069 |  | num\_tl\_90g\_dpd\_24m | 72.53038297 |
| mo\_sin\_rcnt\_tl | 477.1928325 |  | tax\_liens | 64.21043712 |
| propSatisBankcardAccts | 474.7977617 |  | collections\_12\_mths\_ex\_med | 42.06532733 |
| acc\_open\_past\_24mths | 469.3096977 |  | chargeoff\_within\_12\_mths | 18.37671543 |
| num\_bc\_tl | 454.137477 |  | delinq\_amnt | 18.03112352 |
| pct\_tl\_nvr\_dlq | 429.4178842 |  | acc\_now\_delinq | 11.33458548 |
| open\_acc | 421.5859439 |  | num\_tl\_30dpd | 7.218674516 |
| num\_sats | 419.3304977 |  |  |  |

The following variables have an importance score of 0: term, pymnt\_plan, policy\_code, application\_type, hardship\_flag, disbursement\_method.

We can observe a difference in the variable importance between the decision tree model and the random forest model. The top 3 variables in the random forest model are employment length, sub grade and dti. Where as in the decision tree model, the top 3 variables by importance are interest, sub grade and grade. From the top 20 variables in the random forest model, 12 are also present in the list of top 20 variables by the decision tree model.

Overall we would prefer the random forest model over the decision tree model due to better accuracy observed on both the training and testing data.

ROC Analysis

A map of a person

Description automatically generated

Our Roc graph shows higher TP rate and low FP rate which shows better performance.

LIFTS

A close up of a map

Description automatically generated

AUC Value = 0.6401581

AUC =0.6401581, it means there is 64% chance that model will be able to distinguish between positive class and negative class.

1. We would consider the average annual return to calculate a more accurate value of profit. The average interest rate would be suitable for use if the loans that are fully paid are paid at the end of the actual loan duration (i.e. 3 years). However, since most fully paid loans are paid off before the loan period is complete, the return for the investor is less than the interest rate. The overall average annual return rate across fully paid loans is 1.86%. Results can be seen in the graph below.

A screenshot of a cell phone

Description automatically generated

7.

A close up of a map

Description automatically generated

A close up of a map

Description automatically generated

According to above graph, our gbm model is predicting a little better than random forest.