Lending Club – Loan Repayment Prediction – Literature Review

## Introduction

The goal of each and every business is to make profit. For a lender profit depends on whether or not the borrower pays the interest and the principal. Without repayment the lender will incur a loss and that loss can even potentially be greater than the initial loan when lawyer, court and collection fees are taken into consideration. For these reasons it is critically important for a lender to be able to identify whether a potential borrower can and will make all of his or her loan payments. What I intend to do is identify the characteristics (limited to those found in the dataset) of persons, as well as the causes for which people default on their loans and use this information to predict whether a potential borrower would or would not make all of his or her own payments. In order to make these predictions, I will determine the most significant features in the Learning Club dataset and pass these features along with the corresponding labels (default/no-default) into the machine learning classifiers I select, creating predictive models for each. For each model I will test with the same test data set and compare the test labels to the actual labels to measure accuracy.

## Literature Review

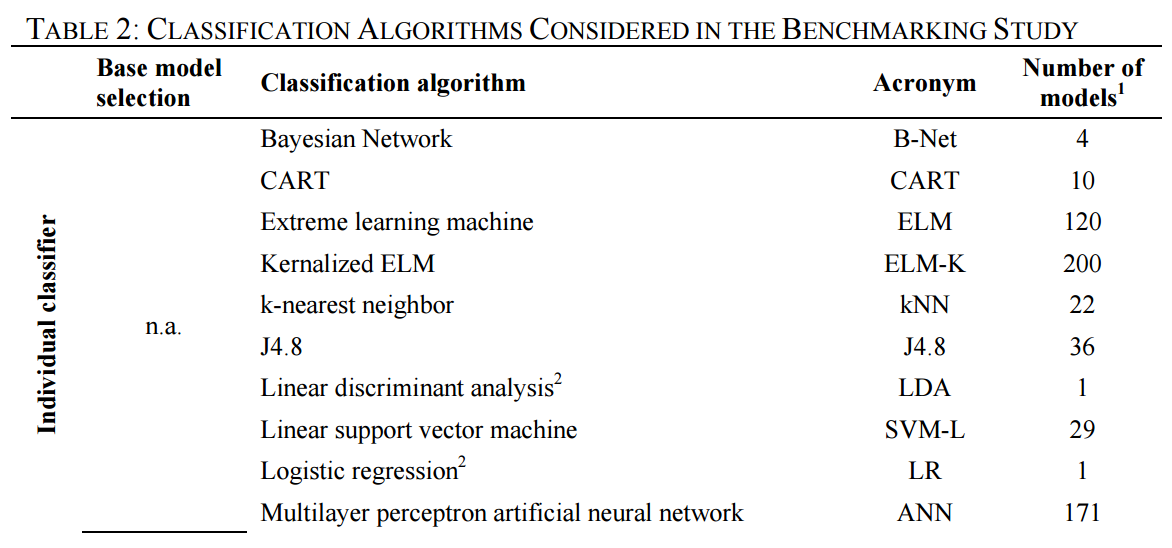
**Benchmarking state-of-the-art classification algorithms for credit scoring: A ten-year update**

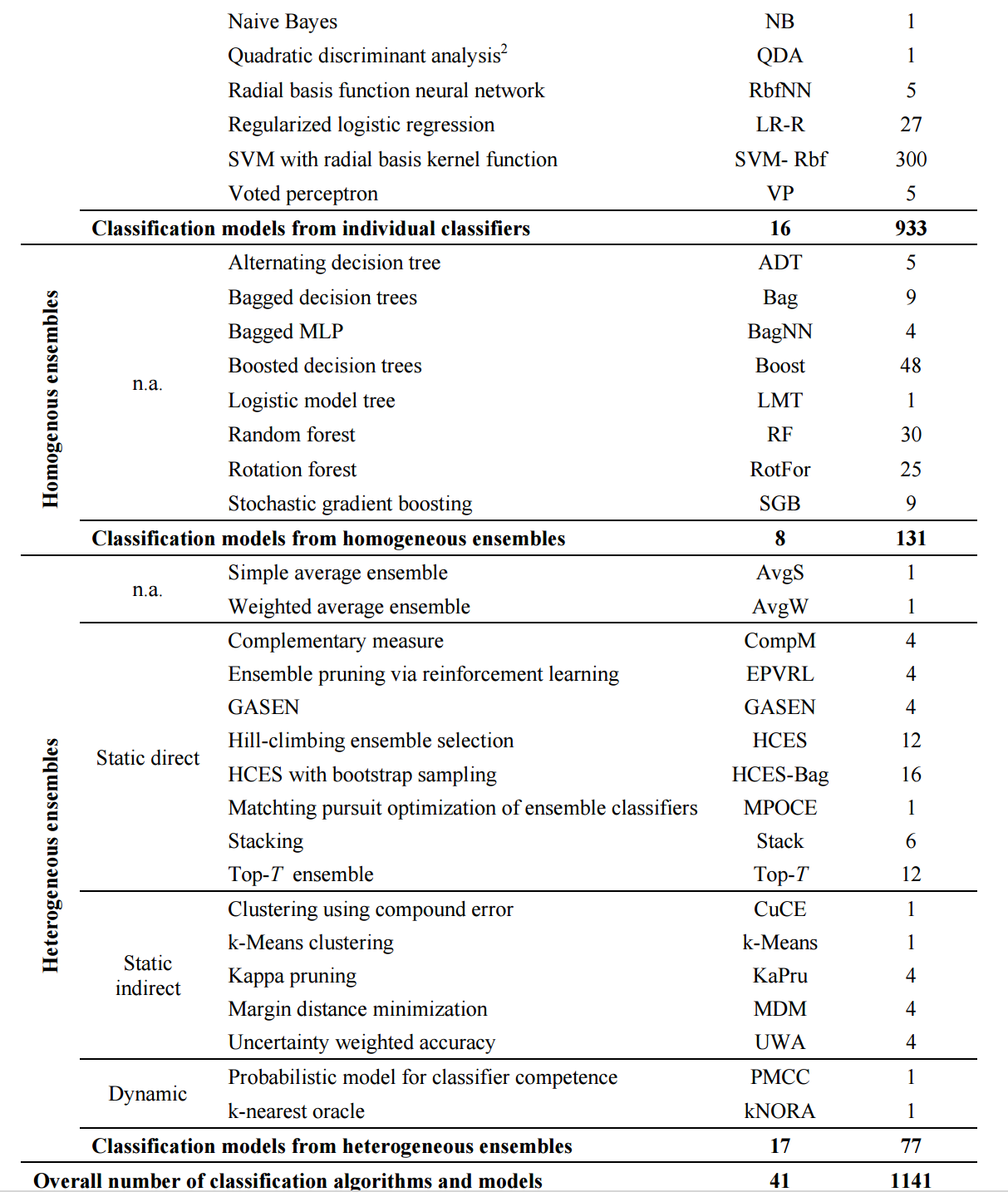
<http://www.business-school.ed.ac.uk/waf/crc_archive/2013/42.pdf>

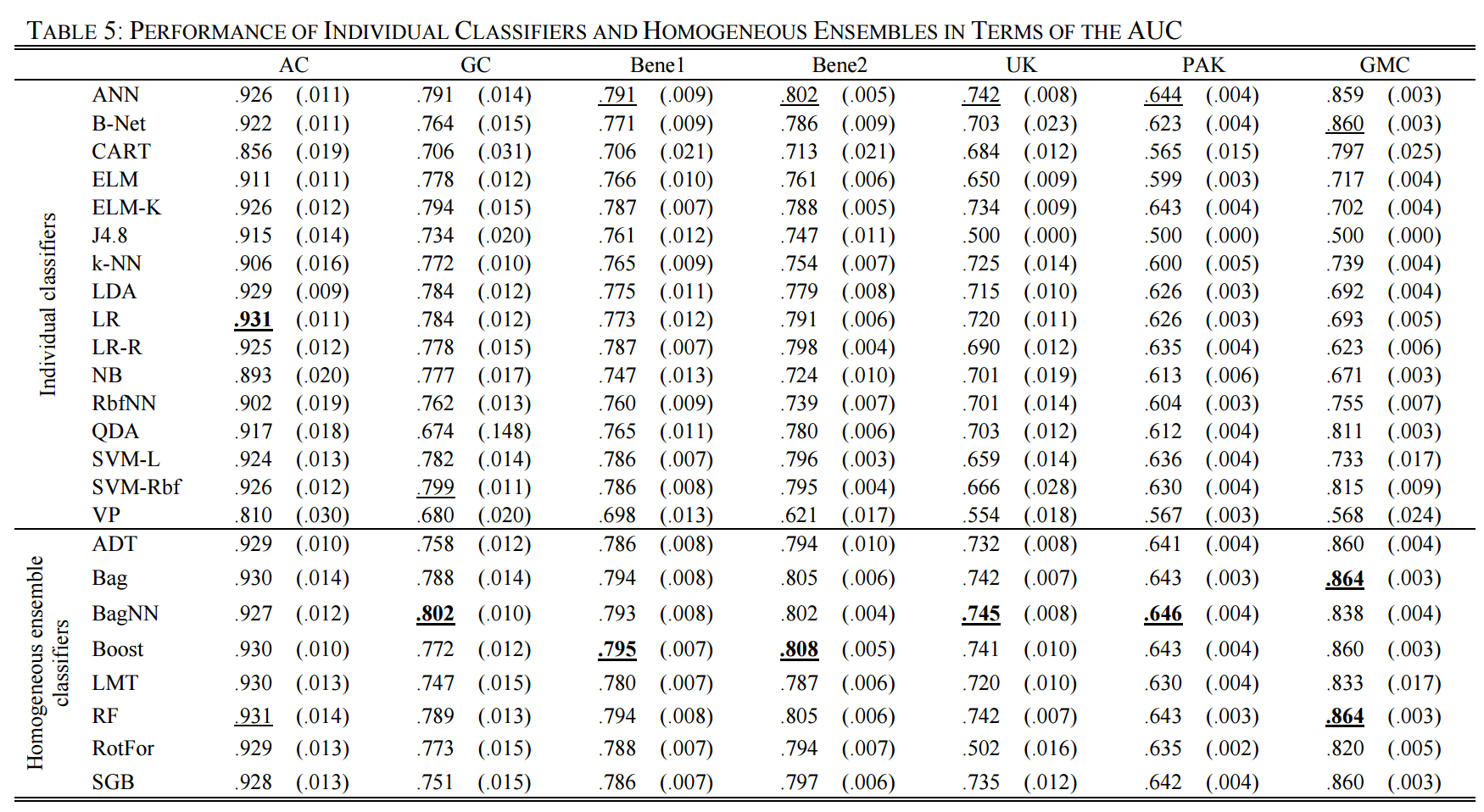
Essentially this paper is a modern update to the landmark benchmarking study of classification algorithms for credit scoring by Baesens, Van Gestel, Viaene, Stepanova, Suykens and Vanthienen in 2003. Since 2003 many new techniques and algorithms have been developed in predictive modeling. This paper builds upon Baesens et al. by including all of the newer state-of-the-art techniques as well as those covered in the previous study.

I chose this paper because of its goal of comparing of classification algorithms for credit scoring is highly related to the problem I wish to solve in this project. The paper describes a vector x of m dimensions with each dimension as a feature characterizing an application for a credit products such as a loan. It then goes on to discuss a binary response variable which indicates the existence or non-existence of a default event. The probability of a default event given x is classification problem being addressed in the study. Finally, a decision maker will take this probability and if it falls under a given threshold the application will be accepted, otherwise it will be rejected. This is essentially the approach I am taking for this capstone project. The Lending Club data set I am using for this project contains over 100 features characterizing the borrower, my vector x. The data set also contains a feature stating the current status of the loan with various possible values that can easily be grouped into default or non-default statuses. This is effectively my binary response variable described in the study. Finally my goal is to estimate the probability of default given a set of borrower characteristics and use that to determine whether they are likely or unlikely to default. This is parallel to the study.

The study considered the following classification algorithms.



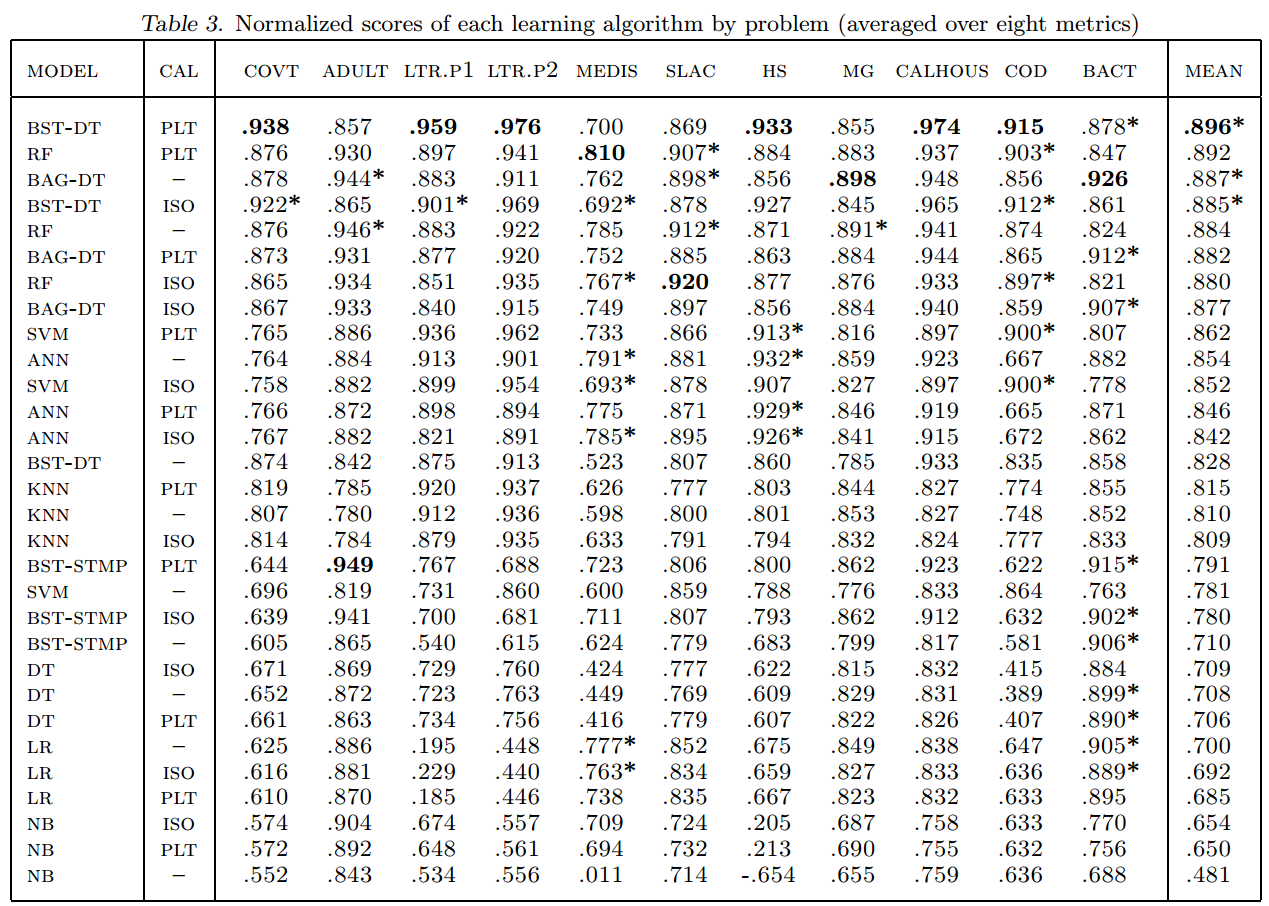


The following table measures the performance of each algorithm in credit scoring classification using Area Under a ROC Curve (AUC). According to the study across all performance measures the top three most accurate classifiers are Random Forests, Bagged (MLP) Neural Networks, and Bagged Decision Trees. 

**An Empirical Comparison of Supervised Learning Algorithms**

<https://www.cs.cornell.edu/~caruana/ctp/ct.papers/caruana.icml06.pdf>

This study compares the performance of eight machine learning algorithms namely, SVMs, neural nets, logistic regression, naïve bayes, memory based learning, random forests, decision trees, bagged trees, boosted trees, and boosted stumps. The performance metrics used are, accuracy, F-score, Lift, ROC Area, average precision, squared error and cross entropy. The study concludes that bagged trees, random forests and neural nets have the best average performance (prior to calibration) over all the metrics and over all the problems. When calibration is taken into account, the overall best performing algorithm is boosted decision trees (calibrated). In close second is Random forests, followed by bagged decision trees (uncalibrated).



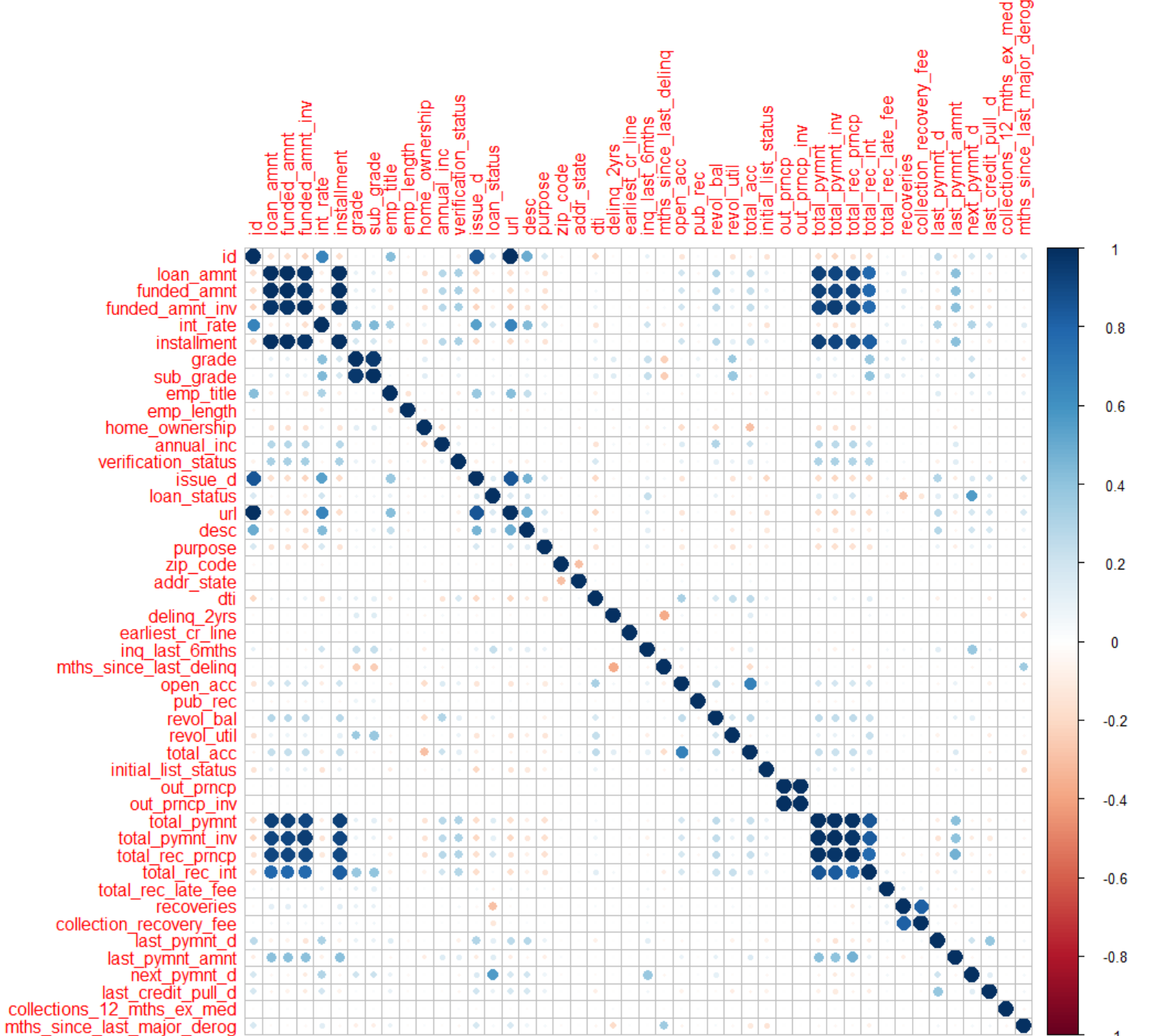
**Conclusion**

I was fascinated to discover that from both of the studies I researched, ensemble methods were generally the best performers. I was also pleasantly surprised to find that in both studies, the top three performing algorithms were almost identical. According to **Benchmarking state-of-the-art classification algorithms for credit scoring: A ten-year update**, the top three classifiers were Random Forests, Bagged (MLP) Neural Networks, and Bagged Decision Trees. Meanwhile according to **An Empirical Comparison of Supervised Learning Algorithms**, the top three are Boosted Decision Trees, Random Forests and Bagged Decision Trees. In both studies Random Forests and Bagged Decision Trees come out on top. It’s important to note that **An Empirical Comparison of Supervised Learning Algorithms** did not include Bagged Neural Networks. Given the agreement among both studies I have decided to use Random Forests, Bagged (MLP) Neural Networks, and Bagged Decision Trees for this project. In addition to selecting classification algorithms, metrics for measuring the performance of said classifiers are required. Given the similarities of the goals of my project and the aims of **Benchmarking state-of-the-art classification algorithms for credit scoring: A ten-year update**, I have decided to use its performance measures. Specifically, the performance metrics for this project are Percentage Correctly Classified (PCC), the area under a ROC curve (AUC), H-measure, and Brier Score (BS).

## Dataset

The dataset for this project is the Loan Data dataset from the Lending Club (<https://www.lendingclub.com/info/download-data.action>). Only 36 month term data from 2007 to February 2013 is being used because it is important that all loans examined are well past their due date. A borrower which defaults on a current loan may do so due to a temporary job loss or a myriad of other reasons. The borrower may then recover and return to good standing once again. Using 36 month term data up until February 2013, ensures that labelling potentially temporary defaults as permanent does not occur.

The original dataset has 111 features. The first step was to review the data and remove any features consisted of majority NULL values. Once those features were removed a correlation analysis was performed. Of each set of highly correlated features, one was removed. The image below displays the correlation plot before the removal of highly correlated features.



The next step was to determine which of the remaining features were significant. Two methods were employed: the Boruta algorithm and the Recursive Feature Elimination algorithm (RFE).

**Boruta:**

After running Boruta four separate times each on a unique set of random records from the dataset, the following results were found. Only features/attributes which were deemed important in 50% or more of the Boruta runs were ultimately selected as important.

22 features/attributes deemed important

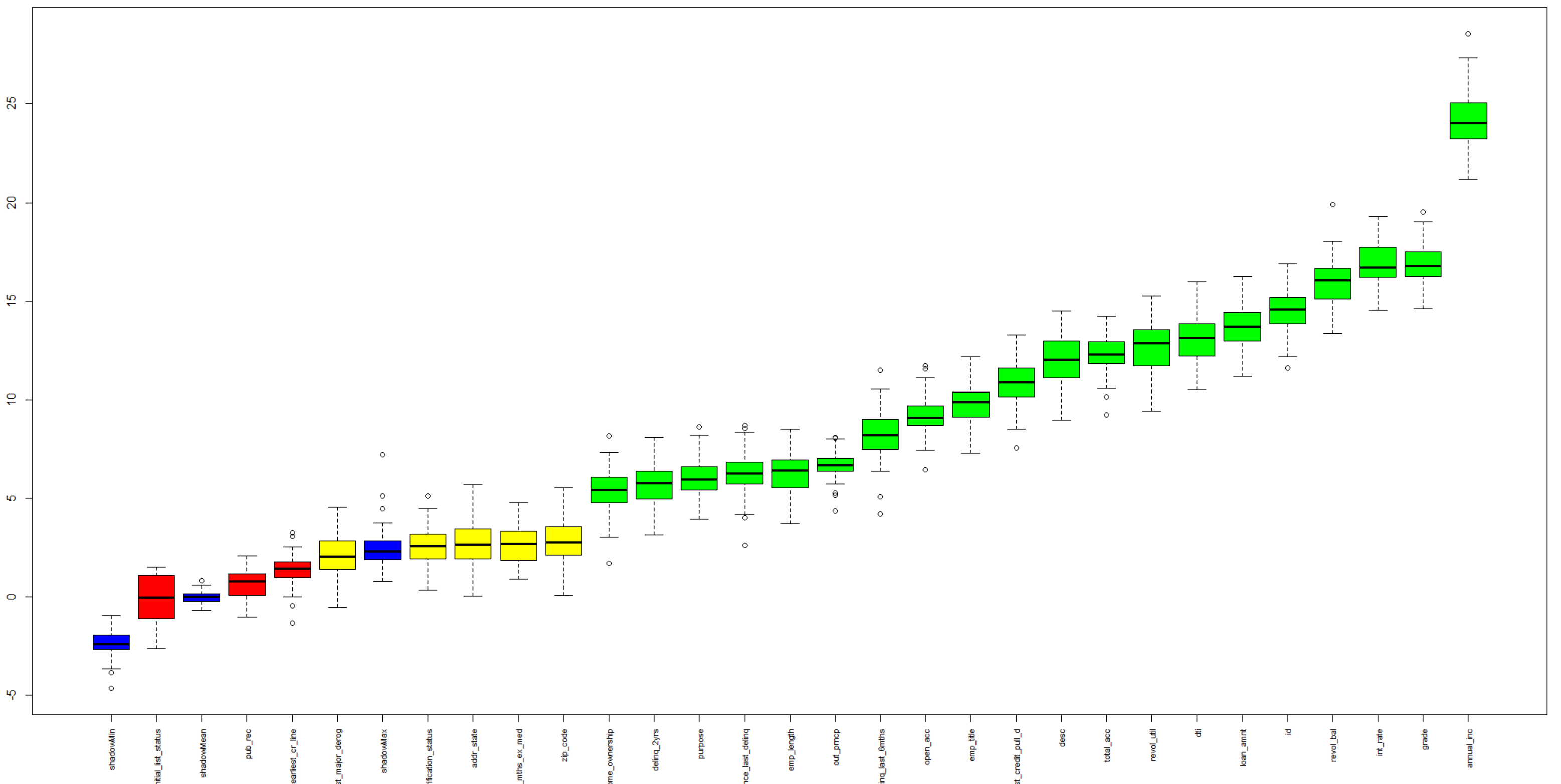
2 features/attributes deemed unimportant

4 features/attributes with tentative importance

The following features/attributes were found to be significantly important in 50% or more of the Boruta analyses:

1. id
2. loan\_amnt
3. int\_rate
4. grade
5. emp\_title
6. emp\_length
7. home\_ownership
8. annual\_inc
9. verification\_status
10. desc
11. purpose
12. dti
13. zip\_code
14. delinq\_2yrs
15. inq\_last\_6mths
16. mths\_since\_last\_delinq
17. open\_acc
18. revol\_bal
19. revol\_util
20. total\_acc
21. out\_prncp
22. last\_credit\_pull\_d

The graph below displays the importance of each of the features. It was taken from one of the four iterations of the algorithm and generally reflects the overall result. Green = Significant importance, Yellow = Tentative importance and Red = Insignificant importance



**Recursive Feature Elimination (RFE):**

According to the RFE algorithm, there are 17 features/attributes of significant importance. These features are listed below along with their overall significance.

Overall

annual\_inc 20.766586

revol\_bal 13.414077

int\_rate 12.425477

id 12.326874

emp\_title 11.834513

loan\_amnt 9.727355

grade 9.581749

total\_acc 7.796833

revol\_util 6.757664

open\_acc 6.715921

purpose 6.253235

dti 6.137568

last\_credit\_pull\_d 5.895808

emp\_length 5.670145

inq\_last\_6mths 5.322751

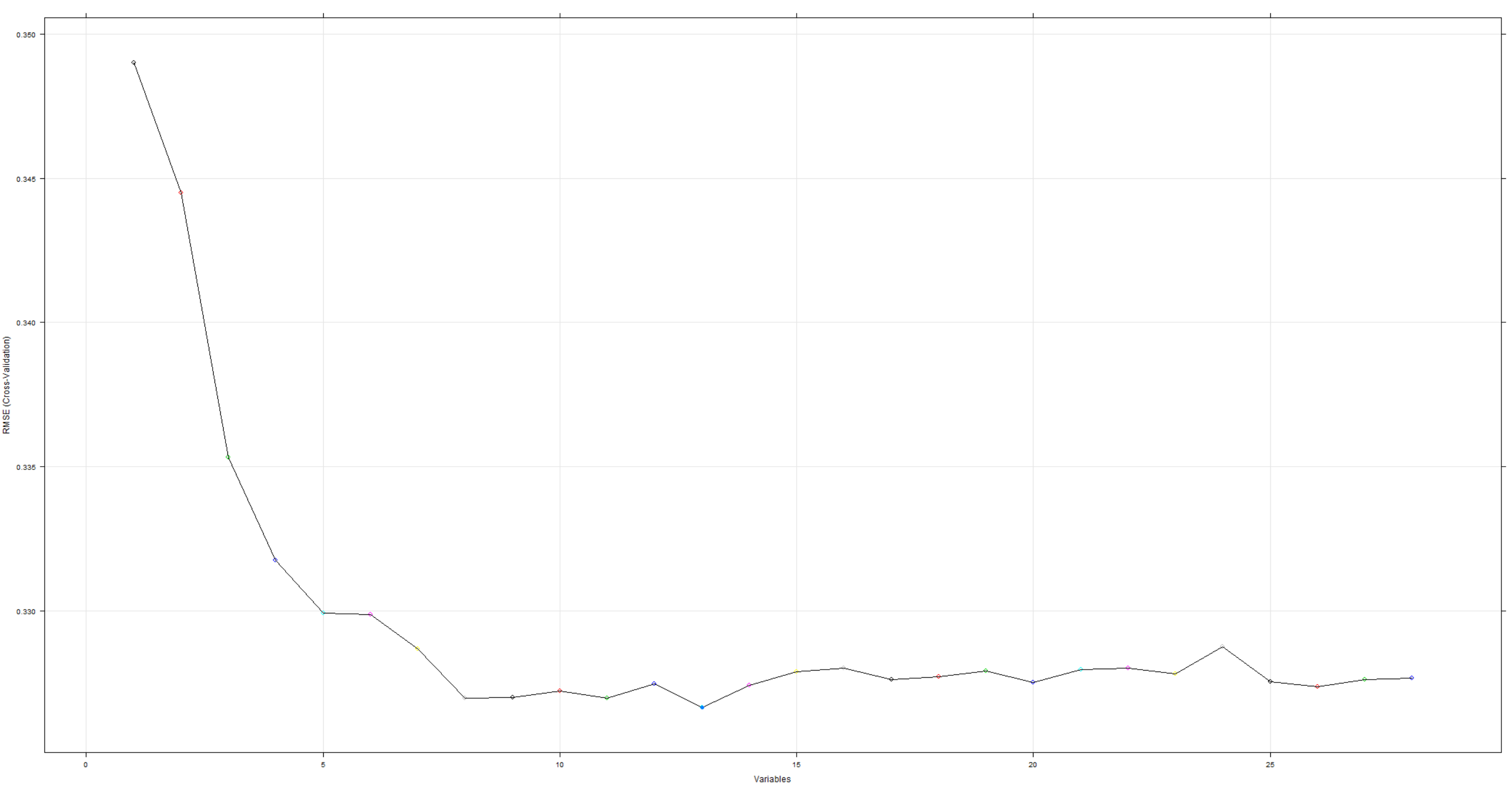
home\_ownership 4.888201

desc 4.841160

It is important to note that all of the significant features selected by RFE were also deemed as significant by the Boruta algorithm. Of the 16 deemed important, the RFE algorithm selected an optimal subset of 13 features.

1. annual\_inc
2. revol\_bal
3. int\_rate
4. id
5. emp\_title
6. loan\_amnt
7. grade
8. total\_acc
9. revol\_util
10. dti
11. open\_acc
12. last\_credit\_pull\_d
13. purpose

According to the RFE algorithm, the top feature/variable is **annual\_inc**. This corresponds to the Boruta analysis. The diagram below shows the relationship between the number of variables/features to and Root Mean Square Error (RMSE).



In conclusion, both the Boruta and RFE algorithms found a similar set of significant set of features/variables. The main difference being that Boruta tended to find more features/variables significant than RFE. One advantage of RFE was the ability to determine the optimal subset of features/variables which would minimize the RMSE. As such, I will be using the set of 13 optimal features/variables as listed above.

Below is the Data Dictionary, describing each feature and whether or not it will be used for analysis. Note: Only those features labeled as ‘**Significant and part of optimal subset**’ will be used for analysis.

**Data Dictionary:**

|  |  |
| --- | --- |
| **Feature** | **Description** |
| acc\_now\_delinq | The number of accounts on which the borrower is now delinquent. |
| acc\_open\_past\_24mths | Number of trades opened in past 24 months. |
| addr\_state | The state provided by the borrower in the loan application |
| all\_util | Balance to credit limit on all trades |
| annual\_inc | The self-reported annual income provided by the borrower during registration. |
| annual\_inc\_joint | The combined self-reported annual income provided by the co-borrowers during registration |
| application\_type | Indicates whether the loan is an individual application or a joint application with two co-borrowers |
| avg\_cur\_bal | Average current balance of all accounts |
| bc\_open\_to\_buy | Total open to buy on revolving bankcards. |
| bc\_util | Ratio of total current balance to high credit/credit limit for all bankcard accounts. |
| chargeoff\_within\_12\_mths | Number of charge-offs within 12 months |
| collection\_recovery\_fee | post charge off collection fee |
| collections\_12\_mths\_ex\_med | Number of collections in 12 months excluding medical collections |
| delinq\_2yrs | The number of 30+ days past-due incidences of delinquency in the borrower's credit file for the past 2 years |
| delinq\_amnt | The past-due amount owed for the accounts on which the borrower is now delinquent. |
| Desc | Loan description provided by the borrower |
| Dti | A ratio calculated using the borrower’s total monthly debt payments on the total debt obligations, excluding mortgage and the requested LC loan, divided by the borrower’s self-reported monthly income. |
| dti\_joint | A ratio calculated using the co-borrowers' total monthly payments on the total debt obligations, excluding mortgages and the requested LC loan, divided by the co-borrowers' combined self-reported monthly income |
| earliest\_cr\_line | The month the borrower's earliest reported credit line was opened |
| emp\_length | Employment length in years. Possible values are between 0 and 10 where 0 means less than one year and 10 means ten or more years. |
| emp\_title | The job title supplied by the Borrower when applying for the loan.\* |
| fico\_range\_high | The upper boundary range the borrower’s FICO at loan origination belongs to. |
| fico\_range\_low | The lower boundary range the borrower’s FICO at loan origination belongs to. |
| funded\_amnt | The total amount committed to that loan at that point in time. |
| funded\_amnt\_inv | The total amount committed by investors for that loan at that point in time. |
| Grade | LC assigned loan grade |
| home\_ownership | The home ownership status provided by the borrower during registration. Our values are: RENT, OWN, MORTGAGE, OTHER. |
| Id | A unique LC assigned ID for the loan listing. |
| il\_util | Ratio of total current balance to high credit/credit limit on all install acct |
| initial\_list\_status | The initial listing status of the loan. Possible values are – W, F |
| inq\_fi | Number of personal finance inquiries |
| inq\_last\_12m | Number of credit inquiries in past 12 months |
| inq\_last\_6mths | The number of inquiries in past 6 months (excluding auto and mortgage inquiries) |
| Installment | The monthly payment owed by the borrower if the loan originates. |
| int\_rate | Interest Rate on the loan |
| issue\_d | The month which the loan was funded |
| last\_credit\_pull\_d | The most recent month LC pulled credit for this loan |
| last\_fico\_range\_high | The upper boundary range the borrower’s last FICO pulled belongs to. |
| last\_fico\_range\_low | The lower boundary range the borrower’s last FICO pulled belongs to. |
| last\_pymnt\_amnt | Last total payment amount received |
| last\_pymnt\_d | Last month payment was received |
| loan\_amnt | The listed amount of the loan applied for by the borrower. If at some point in time, the credit department reduces the loan amount, then it will be reflected in this value. |
| loan\_status | Current status of the loan |
| max\_bal\_bc | Maximum current balance owed on all revolving accounts |
| member\_id | A unique LC assigned Id for the borrower member. |
| mo\_sin\_old\_il\_acct | Months since oldest bank installment account opened |
| mo\_sin\_old\_rev\_tl\_op | Months since oldest revolving account opened |
| mo\_sin\_rcnt\_rev\_tl\_op | Months since most recent revolving account opened |
| mo\_sin\_rcnt\_tl | Months since most recent account opened |
| mort\_acc | Number of mortgage accounts. |
| mths\_since\_last\_delinq | The number of months since the borrower's last delinquency. |
| mths\_since\_last\_major\_derog | Months since most recent 90-day or worse rating |
| mths\_since\_last\_record | The number of months since the last public record. |
| mths\_since\_rcnt\_il | Months since most recent installment accounts opened |
| mths\_since\_recent\_bc | Months since most recent bankcard account opened. |
| mths\_since\_recent\_bc\_dlq | Months since most recent bankcard delinquency |
| mths\_since\_recent\_inq | Months since most recent inquiry. |
| mths\_since\_recent\_revol\_delinq | Months since most recent revolving delinquency. |
| next\_pymnt\_d | Next scheduled payment date |
| num\_accts\_ever\_120\_pd | Number of accounts ever 120 or more days past due |
| num\_actv\_bc\_tl | Number of currently active bankcard accounts |
| num\_actv\_rev\_tl | Number of currently active revolving trades |
| num\_bc\_sats | Number of satisfactory bankcard accounts |
| num\_bc\_tl | Number of bankcard accounts |
| num\_il\_tl | Number of installment accounts |
| num\_op\_rev\_tl | Number of open revolving accounts |
| num\_rev\_accts | Number of revolving accounts |
| num\_rev\_tl\_bal\_gt\_0 | Number of revolving trades with balance >0 |
| num\_sats | Number of satisfactory accounts |
| num\_tl\_120dpd\_2m | Number of accounts currently 120 days past due (updated in past 2 months) |
| num\_tl\_30dpd | Number of accounts currently 30 days past due (updated in past 2 months) |
| num\_tl\_90g\_dpd\_24m | Number of accounts 90 or more days past due in last 24 months |
| num\_tl\_op\_past\_12m | Number of accounts opened in past 12 months |
| open\_acc | The number of open credit lines in the borrower's credit file. |
| open\_acc\_6m | Number of open trades in last 6 months |
| open\_il\_12m | Number of installment accounts opened in past 12 months |
| open\_il\_24m | Number of installment accounts opened in past 24 months |
| open\_il\_6m | Number of currently active installment trades |
| open\_rv\_12m | Number of revolving trades opened in past 12 months |
| open\_rv\_24m | Number of revolving trades opened in past 24 months |
| out\_prncp | Remaining outstanding principal for total amount funded |
| out\_prncp\_inv | Remaining outstanding principal for portion of total amount funded by investors |
| pct\_tl\_nvr\_dlq | Percent of trades never delinquent |
| percent\_bc\_gt\_75 | Percentage of all bankcard accounts > 75% of limit. |
| policy\_code | publicly available policy\_code=1 |
| new products not publicly available policy\_code=2 |
| pub\_rec | Number of derogatory public records |
| pub\_rec\_bankruptcies | Number of public record bankruptcies |
| Purpose | A category provided by the borrower for the loan request. |
| pymnt\_plan | Indicates if a payment plan has been put in place for the loan |
| Recoveries | post charge off gross recovery |
| revol\_bal | Total credit revolving balance |
| revol\_util | Revolving line utilization rate, or the amount of credit the borrower is using relative to all available revolving credit. |
| sub\_grade | LC assigned loan subgrade |
| tax\_liens | Number of tax liens |
| Term | The number of payments on the loan. Values are in months and can be either 36 or 60. |
| Title | The loan title provided by the borrower |
| tot\_coll\_amt | Total collection amounts ever owed |
| tot\_cur\_bal | Total current balance of all accounts |
| tot\_hi\_cred\_lim | Total high credit/credit limit |
| total\_acc | The total number of credit lines currently in the borrower's credit file |
| total\_bal\_ex\_mort | Total credit balance excluding mortgagemp\_lengthe |
| total\_bal\_il | Total current balance of all installment accounts |
| total\_bc\_limit | Total bankcard high credit/credit limit |
| total\_cu\_tl | Number of finance trades |
| total\_il\_high\_credit\_limit | Total installment high credit/credit limit |
| total\_pymnt | Payments received to date for total amount funded |
| total\_pymnt\_inv | Payments received to date for portion of total amount funded by investors |
| total\_rec\_int | Interest received to date |
| total\_rec\_late\_fee | Late fees received to date |
| total\_rec\_prncp | Principal received to date |
| total\_rev\_hi\_lim | Total revolving high credit/credit limit |
| url | URL for the LC page with listing data. |
| verification\_status | Indicates if income was verified by LC, not verified, or if the income source was verified |
| verified\_status\_joint | Indicates if the co-borrowers' joint income was verified by LC, not verified, or if the income source was verified |
| zip\_code | The first 3 numbers of the zip code provided by the borrower in the loan application. |
|  |  |
| **LEGEND** |  |
|  | Removed due to high correlation |
|  | Insufficient values in data set |
|  | Removed due to irrelevance |
|  | Removed because data acquired after default |
|  | Significant relevance |
|  | Not found in dataset |
|  | Label source |
|  | Tentative relevance |
|  | Significant and part of optimal subset |

## Approach

Load and Cleanse

Feature Selection

Feature Engineering

Create & train model

Make Predictions

Measure Performance

Causes of Default

### Step 1: Load and Cleanse

Load dataset and perform initial cleansing. Cleansing will entail removing records with little or no information other than NULL values and potentially populating empty cells with imputed values.

### Step 2: Feature Selection

Determine which features are significant and which will be used in the predictive models. There are three sub-steps:

1. Manually remove all features which mainly consist of NULL values.
2. Using correlation analysis, remove highly correlated features from the dataset.
3. Using the Boruta algorithm and Recursive Feature Elimination, determine which features are significant. Note: The significant features will be used in the predictive models.

### Step 3: Feature Engineering

Perform further analysis of the data and determine whether features can be modified, combined or split to provide even more useful information. Please note that this may involve going back to feature selection before moving to the next step.

### Step 4: Create and Train Model

Using the pruned dataset consisting of features determined from steps 2 and 3, create a test and training dataset. Create predictive models on the training dataset for the following algorithms: Random Forests, Bagged (MLP) Neural Networks, and Bagged Decision Trees. Note: Each model is created using the same training dataset.

### Step 5: Make Predictions

Use the models on the test dataset and record the results.

### Step 6: Measure Performance

This is a sub-step of step 5. For each result set in step 5, measure the Percentage Correctly Classified (PCC), the area under a ROC curve (AUC), H-measure, and Brier Score (BS) and compare the performance of each of the classification algorithms.

### Step 7: Causes of Default

Using the knowledge gained from all of the previous steps, determine the causes of loan default.

## Code

<https://github.com/ribeiros/lending_club_default_prediction>