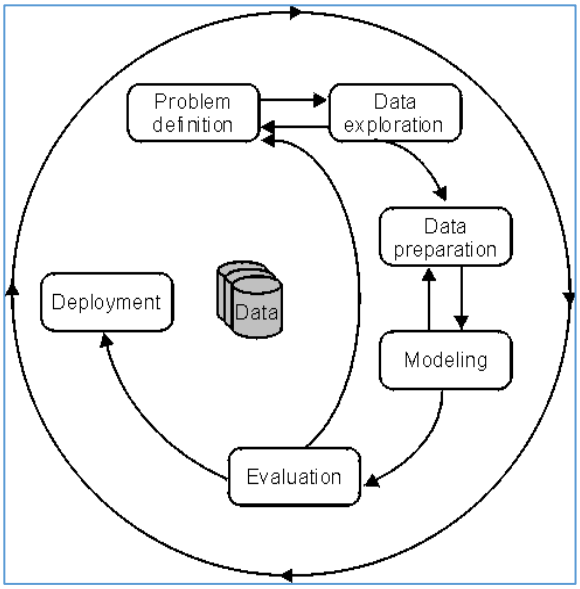
Table of Contents

Executive Summary

The project aims at addressing one of the most critical challenges faced by credit risk analysts i.e. reducing default/bad loans. Bad loans are a huge overhead to all the financial institutions. According to a survey in 2015-2016, more than 16% of all the loans in United States go default.

In this project we have tried to handle the issue of one such financial organization, Lending Club. Lending Club is the world's largest online marketplace for investors and borrowers. Any US citizen can become an investor or a borrower. However, becoming an investor entails the dilemma of extending loan to a borrower who might default.

In this project, we solve this dilemma of the investors. **We have come out with predictive models that can predict loan defaults with an accuracy as high as 75%.** We started this project by collecting historical data from Lending club. The website provided us data from 2007 to 2016. Further, we started exploring and processing the data. Various charts, plots and summary statistics provided us useful insights from the data. After processing and visualizing the data, we moved on to build predictive models. **We tried 7 different models, of which Random Forest proved the most accurate in predicting defaults.**



Along with performance measures, we also came out with various insights from the dataset. We found out that interest rate, loan amount, annual income and debt-to-income ratio were some of the most important variables. Also, grades given by lending club to each borrower/loan are a direct impact of the interest rate charged from the borrower. We should also note that FICO score, one of other criteria to measure person's creditworthiness was not used in the analysis. In future, we would like to implement this model in real world scenario and note down its performance with real loans.

Data Description

Data Source: https://www.lendingclub.com/info/download-data.action

Sample size: 42540 records.

Number of variables: 121

The data is of interest because of the unique nature of lending club's business model with investors choosing to invest in loan applicants based on the loan grade and various other factors. Also, choosing a dataset from the financial industry was of particular interest to us as most of wish to work in this sector as a consultant or a product manager and hence an opportunity to understand what gives was in fact desirable.

Loan Amount

Loan amount is the amount of loan that has been extended to the borrower. In our dataset, the amount ranged from $500 - $35,000 and the maximum number of loans were borrowed in the $7,500-$10,000 range.

```
 Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
 1000    8400   14400   15589   21000   40000
```

Term

Term is the period for which the loan is extended and is measured by the number of payments on the loan. Values are in months which are either 36 or 60. It was seen that most of the loans, 70%, were taken on 36-month payment plan.

```
   36    60
96120 37767
```

Interest Rate

Interest Rate is the rate of interest on the loan. The histogram is somewhat right skewed and even though the interest rate ranges from 5.32-28.9, a big part of the loans were extended for an interest rate up to 20%, with the average rate being 13.25%.

```
 Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
 5.32    8.49   11.99   12.48   15.31   28.99
```

Grade

Grade field is the grade assigned to each loan by Lending Club. It reflects how likely the loan is to be paid off and is determined based on the creditworthiness of the borrower. The grades in the dataset range from A-G, B and C being the most assigned and very few G grade loans extended.

```
   A    B    C    D    E    F    G
26482 40267 36777 16454  9540  3482   885
```

Employment Length

Employment length is the length in years for which the borrower has been employed. Possible values are between 0 and 10 where 0 being less than one year and 10 being ten or more years. N/a is the value assigned to the borrower who has never been in employment so far. A vast majority of the loans have been extended to people with 10 or more years in employment.

Home Ownership

The home ownership status provided by the borrower during registration. Our values include RENT, OWN, MORTGAGE, OTHER, however, as is evident from the visualization, none of the applicants selected 'Other'. Also, close to 50% of the loans were extended to people having an existing mortgage and those owning a house formed the lowest part, being 10%.

```
MORTGAGE     OWN    RENT
  66829   16194   50864
```

Annual Income

The annual income is what was provided by the borrower during registration. In our dataset, the annual income ranges from anywhere between $0- $150,000 and the distribution is somewhat right skewed with the majority of borrowers making within the $100,000 range, the average being around $67,000.

```
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
      0   49500   68000   80464   95008 9550000
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
      0   49500   68000   69911   86000  150000
```

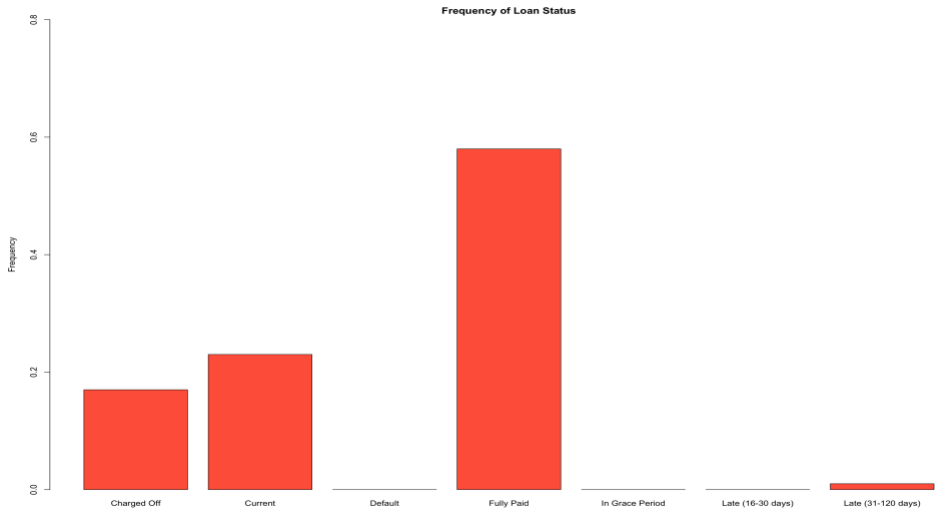
Verification Status

Indicates whether the co-borrowers' joint income has been verified by Lending Club. It is classified into 'Verified', 'Not verified' or 'Source verified' and from the frequency plot, it can be seen that the cases belonging to each category are uniformly distributed with 37% being source verified and 30% not verified.

Not Verified	Source Verified	Verified
44653	52872	36362

Loan Status

The term loan status means the current status of the borrower. There are 8 different categories a borrower can be placed under which are the following: charged off, current, default, fully paid, in grace period, issued, last (16-30 days), late (31-120 days).

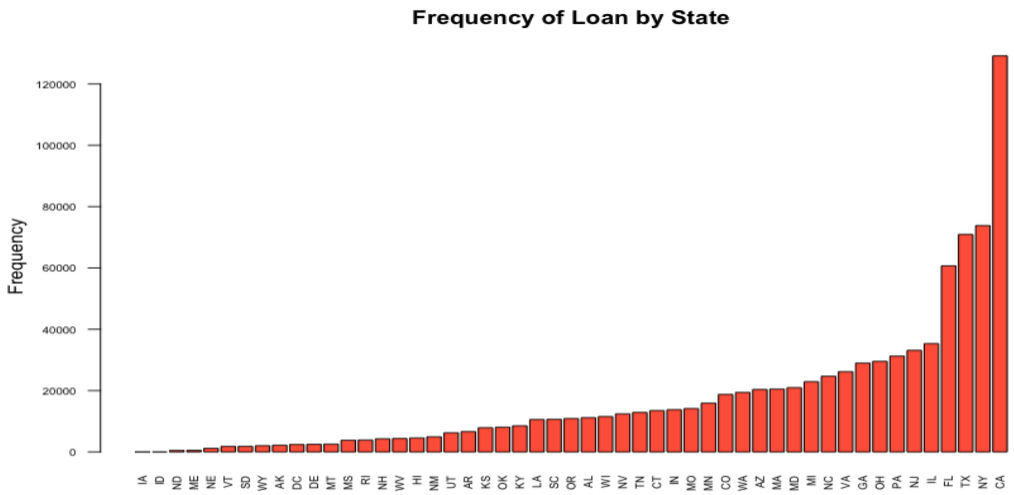


As you can see in the diagram labelled "Frequency, distribution is slight skewed to the right. 70 percent of borrowers which is majority are labelled as current status. 30 percent of borrowers are labelled as fully paid under loan status. The second top category which means loan has been fully repaid, either at the expiration of the 3 or 5-year term as a result of prepayment. 10 percent of borrowers are labelled as charged off which implies that the borrower who are labelled as status, means that no longer a reasonable expectation of further payments according to Lending club website.

State

We wanted to visualize from what states the borrowers apply from, we could conclude that California is twice as much compare to the rest of the states where the applicants are from. The graph is gradually

increasing in ascending order. California has 120,000 applicants from that state compare to New York which is half of the size compare to California. We can draw a conclusion that for the top four states California, New York, Texas, and Florida are metropolitan's states with more jobs to payback the loans they borrowed from Lending Club.



Research Question

We aim to address the following question that is very critical to the lending industry – Given the borrower's risk, should we lend him/her? This is answered indirectly by trying to predict if the borrower will repay the loan by its mature date or not.

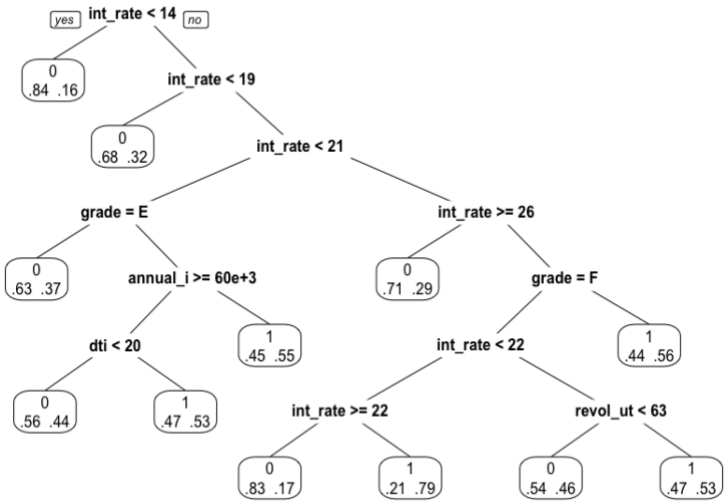
Modelling

Under sample the Good loans

The data available to us is highly unbalanced with approximately 1% bad vs 99% good loans. In such scenario, our model may not give us required results. One of the methods to solve this problem is to under sample the good loans. After under sampling, we will have approx. 75%-25% good vs bad loans, which will improve our model prediction.

Decision Tree

Tree model is used to predict either a qualitative or quantitative result based on what region the observation falls into. For regression tree, we built the model and use the best split to minimize RSS each step, while for classification tree, we use misclassification error rate instead. Once tree model has been developed, we can just go through each node and the associated branch for predicted value.



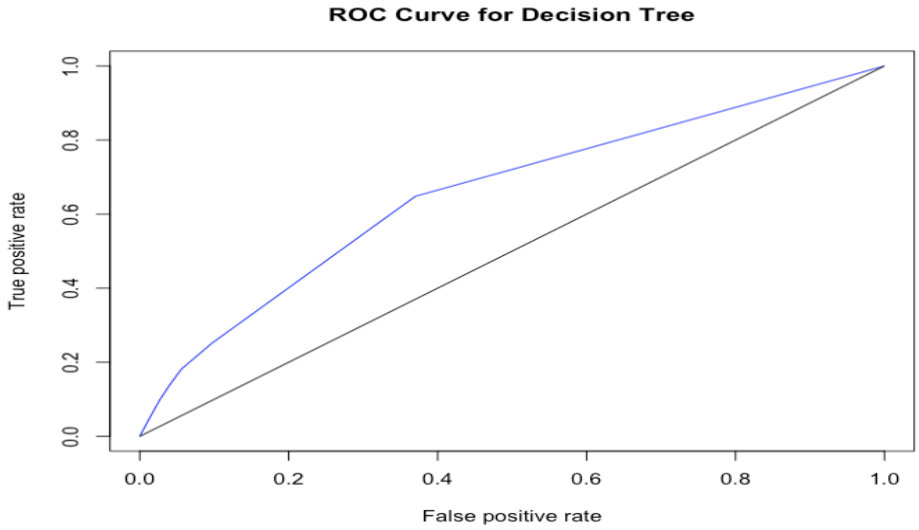

```
## Confusion Matrix and
Statistics (cut-off = 20%)
##
##           Actual
## Prediction     0     1
##          0 37338  6953
##          1 21965 12814
##
##       Accuracy : 0.6343
##    Sensitivity : 0.6483
##    Specificity : 0.6296
```

```
Confusion Matrix and Statisti
cs (cut-off = 25%)
##
##           Actual
## Prediction     0     1
##          0 37338  6953
##          1 21965 12814
##
##       Accuracy : 0.6343
```

```
##    Sensitivity : 0.6483
##    Specificity : 0.6296
```

```
Confusion Matrix and
Statistics (cut-off = 30%)
##
##            Actual
## Prediction     0     1
##          0 37416  6984
##          1 21887 12783
##
##    Accuracy : 0.6349
   Sensitivity : 0.6467
## Specificity : 0.6309
```

```
Confusion Matrix and
Statistics (cut-off = 50%)
##
##            Actual
## Prediction     0     1
##          0 56834  16949
##          1 2469    2818
##
##    Accuracy : 0.7544
   Sensitivity : 0.1425
## Specificity : 0.9583
```




```
## [1] "Area Under Curve for decision tree is :  0.65631"
```

Bagging

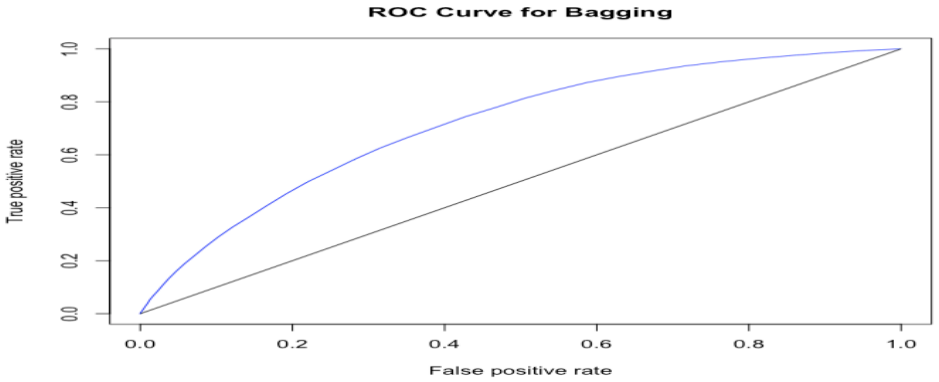
Bagging is the method of taking repeated sample from a single training dataset to reduce the variance and hence increase the prediction accuracy. To start with, we need to create a finite number of datasets using sample with replacement. Then, we create one classifier for each dataset. Lastly, we take average for numeric prediction but use majority vote for categorical outcome prediction.

```
## Confusion Matrix and
Statistics (cut-off = 20%)
##
##            Actual
## Prediction     0     1
##          0 29313  3678
##          1 29990 16089
##
##       Accuracy : 0.5742
##    Sensitivity : 0.8139
##    Specificity : 0.4943
```

```
## Confusion Matrix and Sta
tistics (cut-off = 25%)
##
##            Actual
## Prediction     0     1
##          0 33963  5062
##          1 25340 14705
##
##       Accuracy : 0.6155
##    Sensitivity : 0.7493
##    Specificity : 0.5727
```

```
Confusion Matrix and
Statistics (cut-off = 30%)
##
##            Actual
## Prediction     0     1
##          0 40606  7401
##          1 18697 12366
##
##    Accuracy : 0.6699
   Sensitivity : 0.6256
## Specificity : 0.6847
```

```
Confusion Matrix and
Statistics (cut-off = 50%)
##
##            Actual
## Prediction     0     1
##          0 55102  15455
##          1  4201   4312
##
##    Accuracy : 0.7514
   Sensitivity : 0.2181
## Specificity : 0.9291
```




```
## [1] "Area Under Curve for decision tree is :  0.71656"
```

Random Forest

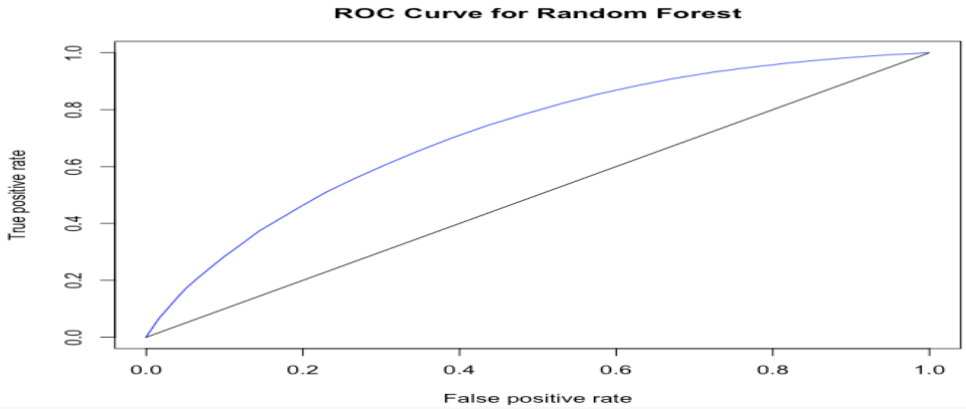
It is a method involving averaging multiple deep decision trees, trained on different parts of the same training set, with the goal of overcoming over-fitting problem of individual decision tree. It is an ensemble learning method that operates by constructing a lot of decision trees at training time and outputting the class that is the mode of the classes output by individual trees For many data sets, it produces a highly accurate classifier. It has an effective method for estimating missing data and maintains accuracy when a large proportion of the data are missing.

```
## Confusion Matrix and
Statistics (cut-off = 20%)
##
##            Actual
## Prediction     0     1
##          0 27899  3537
##          1 31404 16230
##
##       Accuracy : 0.5581
##    Sensitivity : 0.8211
##    Specificity : 0.4704
```

```
Confusion Matrix and Statist
ics (cut-off = 25%)
##
##            Actual
## Prediction     0     1
##          0 33500  5061
##          1 25803 14706
##
##       Accuracy : 0.6097
##    Sensitivity : 0.7440
##    Specificity : 0.5649
```

```
Confusion Matrix and
Statistics (cut-off = 30%)
##
##            Actual
## Prediction     0     1
##          0 41230  7797
##          1 18073 11970
##
##    Accuracy : 0.6728
   Sensitivity : 0.6056
## Specificity : 0.6952
```

```
Confusion Matrix and
Statistics (cut-off = 50%)
##
##            Actual
## Prediction     0     1
##          0 56257  16332
##          1  3046   3435
##
##    Accuracy : 0.7549
   Sensitivity : 0.1737
## Specificity : 0.9486
```




```
## [1] "Area Under Curve for decision tree is :  0.71229"
```

Boosting

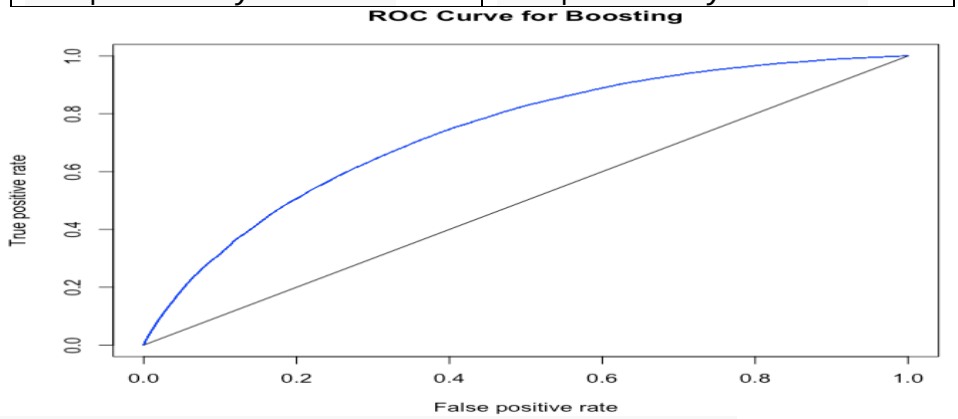
It is a machine learning ensemble meta-algorithm for primarily reducing bias, and also variance in supervised learning, and a family of machine learning algorithms which convert weak learners to strong ones. The main advantage of boosting is the speed. As opposed to random forests, in boosting, the growth of a particular tree takes into account the other trees that have already been grown and thus smaller trees are sufficient, which aids in interpretability.

```
## Confusion Matrix and
Statistics (cut-off = 20%)
##
##            Actual
## Prediction     0     1
##          0 30708  3663
##          1 28595 16104
##
##       Accuracy : 0.5920
##    Sensitivity : 0.8147
##    Specificity : 0.5178
```

```
Confusion Matrix and Statis
tics (cut-off = 25%)
##
##            Actual
## Prediction     0     1
##          0 37965  5794
##          1 21338 13973
##
##       Accuracy : 0.6569
##    Sensitivity : 0.7069
##    Specificity : 0.6402
```

```
Confusion Matrix and
Statistics (cut-off = 30%)
##
##            Actual
## Prediction     0     1
##          0 44123  8166
##          1 15180 11601
##
##    Accuracy : 0.7047
   Sensitivity : 0.5869
## Specificity : 0.7440
```

```
Confusion Matrix and
Statistics (cut-off = 50%)
##
##            Actual
## Prediction     0     1
##          0 56711  16382
##          1  2592   3385
##
##    Accuracy : 0.7600
   Sensitivity : 0.1712
## Specificity : 0.9562
```




```
## [1] "Area Under Curve for decision tree is :  0.73566"
```

Logistic Regression

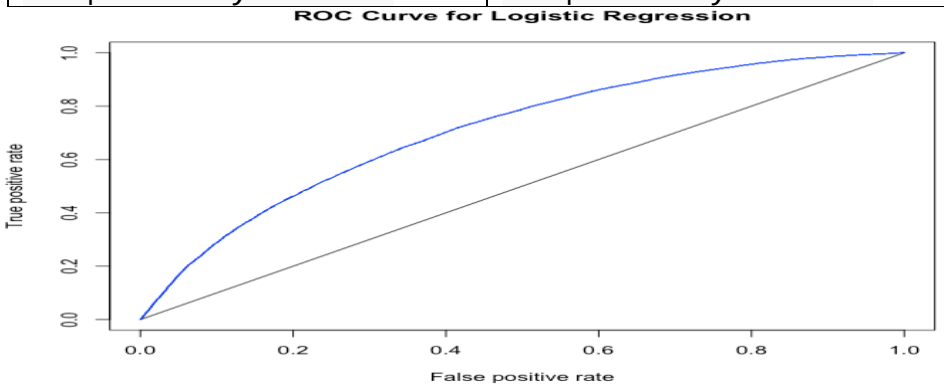
Logistic Regression is a form of regression that allows the prediction of discrete variables by a mix of continuous and discrete predicators. Logistic regression is often used because the relationship between discrete variable(DV) and predicator is non-linear. Instead of using Y (or p) as the dependent variable, we use a function of it, which is called logit. The key advantage of logits maps any value of the dependent variables into a probability [0,1]. The logit, it turns out, can be modelled as a linear function of predicators. Once the logit has been predicted, it can be mapped back to a probability p.

```
## Confusion Matrix and
Statistics (cut-off = 20%)
##
##            Actual
## Prediction     0     1
##          0 29369  4085
##          1 29934 15682
##
##       Accuracy : 0.5698
##    Sensitivity : 0.7933
##    Specificity : 0.4952
```

```
Confusion Matrix and Statis
tics (cut-off = 25%)
##
##            Actual
## Prediction     0     1
##          0 37762  6655
##          1 21541 13112
##
##       Accuracy : 0.6434
##    Sensitivity : 0.6633
##    Specificity : 0.6368
```

```
Confusion Matrix and
Statistics (cut-off = 30%)
##
##            Actual
## Prediction     0     1
##          0 44528  9294
##          1 14775 10473
##
##    Accuracy : 0.6956
   Sensitivity : 0.5298
## Specificity : 0.7509
```

```
Confusion Matrix and
Statistics (cut-off = 50%)
##
##            Actual
## Prediction     0     1
##          0 56853  17051
##          1  2450   2716
##
##    Accuracy : 0.7534
   Sensitivity : 0.1374
## Specificity : 0.9586
```




```
## [1] "Area Under Curve for decision tree is :  0.7083"
```

LASSO

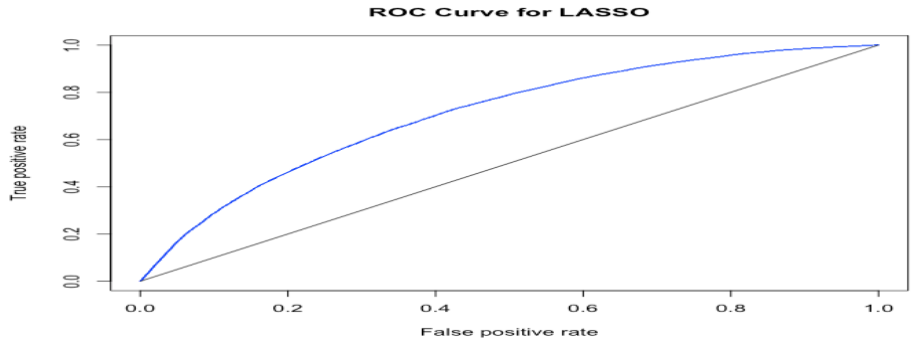
The LASSO (Least Absolute Shrinkage and Selection Operator) is a regression method that involves penalizing the absolute size of the regression coefficients. By penalizing (or equivalently constraining the sum of the absolute values of the estimates) you end up in a situation where some of the parameter estimates may be exactly zero. This is convenient when we want some automatic feature/variable selection, or when dealing with highly correlated predictors, where standard regression will usually have regression coefficients that are 'too large'.

```
## Confusion Matrix and
Statistics (cut-off = 20%)
##
##           Actual
## Prediction     0     1
##          0 29070  3996
##          1 30233 15771
##
##       Accuracy : 0.5671
##    Sensitivity : 0.7978
##    Specificity : 0.3428
```

```
Confusion Matrix and Statis
tics (cut-off = 25%)
##
##           Actual
## Prediction     0     1
##          0 37565  6578
##          1 21738 13189
##
##       Accuracy : 0.6419
##    Sensitivity : 0.6672
##    Specificity : 0.6334
```

```
Confusion Matrix and
Statistics (cut-off = 30%)
##
##           Actual
## Prediction     0     1
##          0 44505  9294
##          1 14798 10473
##
##    Accuracy : 0.6953
   Sensitivity : 0.5298
## Specificity : 0.7505
```

```
Confusion Matrix and
Statistics (cut-off = 50%)
##
##           Actual
## Prediction     0     1
##          0 56951  17157
##          1  2352   2610
##
##    Accuracy : 0.7533
   Sensitivity : 0.1320
## Specificity : 0.9560
```




```
## [1] "Area Under Curve for decision tree is :  0.70818"
```

Naïve Bayes

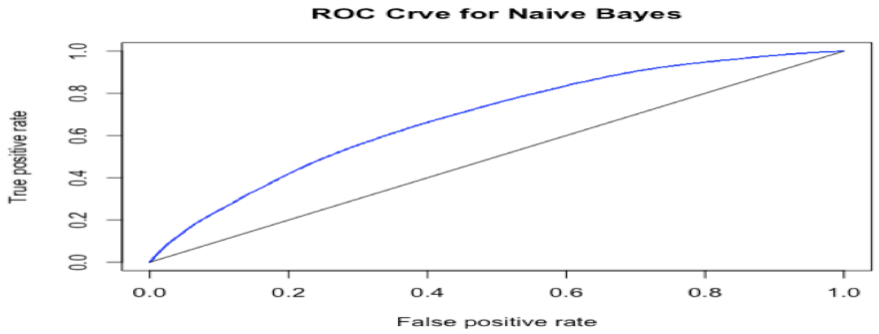
The Naïve Bayes classifier technique is based on the Bayes theorem and assumes the predicators to be independent, which means knowing the value of one attribute does influence the value of any other attribute. The independence assumption is what makes Naïve Bayes naïve. Naïve Bayes classifiers are easy to build, do not involve any iterative process, and work very well with large datasets. Despite its simplicity, Naïve Bayes is known to have often outperformed other classification methods.

```
## Confusion Matrix and
Statistics (cut-off = 20%)
##
##            Actual
## Prediction     0     1
##          0 31939  5563
##          1 27364 14204
##
##       Accuracy : 0.5863
##    Sensitivity : 0.7186
##    Specificity : 0.5286
```

```
Confusion Matrix and Statis
tics (cut-off = 25%)
##
##            Actual
## Prediction     0     1
##          0 36429  6950
##          1 22874 12817
##
##       Accuracy : 0.6228
##    Sensitivity : 0.6484
##    Specificity : 0.6143
```

```
Confusion Matrix and
Statistics (cut-off = 30%)
##
##            Actual
## Prediction     0     1
##          0 39773  8137
##          1 19530 11630
##
##    Accuracy : 0.6501
   Sensitivity : 0.5884
## Specificity : 0.6707
```

```
Confusion Matrix and
Statistics (cut-off = 50%)
##
##            Actual
## Prediction     0     1
##          0 49202  12484
##          1 10101   7283
##
##    Accuracy : 0.7144
   Sensitivity : 0.3684
## Specificity : 0.8296
```




```
## [1] "Area Under Curve for decision tree is :  0.68321"
```

Conclusion

From the performance measures of all the models, we can see that boosting has the largest area under curve of 0.73566. Also in credit risk analysis, accuracy does not play a major role in analysing performance. Predicting a good loan as bad will have less impact on the business than predicting a bad loan as good. In our case sensitivity i.e. the ability of a classifier to predict important class (class = 1) plays a major role. It means how good is our classifier in predicting a default loan. We can see that in most of the cases, ensemble methods outperforms all the other methods. The business implication for

this performance summary table is great. A client can see and understand how their model performs at various cut-off values.

Example Scenario: 25% cut-off (loan with default probability more than .25 is termed as bad loan)

In this scenario, looking at the sensitivity, we can say that ensemble method bagging with sensitivity of 0.7493 has the best results followed by random forest with sensitivity of 0.744

Cur-offs	Models -->	Decision Tree	Bagging	Random Forest	Boosting	Logistic Regression	LASSO	Naïve Bayes
20% cut-off	Accuracy	0.6343	0.5742	0.5581	0.592	0.5698	0.5671	0.5863
	Sensitivity	0.6483	0.8139	0.8211	0.8147	0.7933	0.7978	0.7186
	Specificity	0.6296	0.4943	0.4704	0.5178	0.4952	0.3428	0.5286
25% cut-off	Accuracy	0.6343	0.6155	0.6097	0.6569	0.6434	0.6419	0.6228
	Sensitivity	0.6483	0.7493	0.744	0.7069	0.6633	0.6672	0.6484
	Specificity	0.6296	0.5727	0.5649	0.6402	0.6368	0.6334	0.6143
30% cut-off	Accuracy	0.6349	0.6699	0.6728	0.7047	0.6956	0.6953	0.6501
	Sensitivity	0.6467	0.6256	0.6056	0.5869	0.5298	0.5298	0.5884
	Specificity	0.6309	0.6847	0.6952	0.744	0.7509	0.7505	0.6707
50% cut-off	Accuracy	0.7544	0.7514	0.7549	0.76	0.7534	0.7533	0.7144
	Sensitivity	0.1425	0.2181	0.1737	0.1712	0.1374	0.132	0.3684
	Specificity	0.9583	0.9291	0.9486	0.9562	0.9586	0.956	0.8296
Area under Curve		0.65631	0.71656	0.71229	0.73566	0.7083	0.70818	0.68321

Business Recommendations:

1. Avoid loan with debt-to-income ratio more than 40
2. Loan with greater interest rate gives better return but their probability to go default is even more as the interest rate is highly influenced by the grades provided by Lending Club.
3. Charge more interest rate for borrower with verified source as they are less likely to default.
4. Borrower with Mortgage or rented loan are more likely to default.
5. Better model management that spans the entire modelling life cycle.

6. Data visualization capabilities and business intelligence tools that get important information into the hands of those who need it, when they need it.
7. Implementation of ensemble methods in production.