# Creditworthiness

## Business and Data Understanding

I work for a small bank and are responsible for determining if customers are creditworthy to give a loan to. We received 500 new loan applications this week and we need to select which customers are creditworthy by performing analysis on previous applications and applying the model to these 500 new applications.

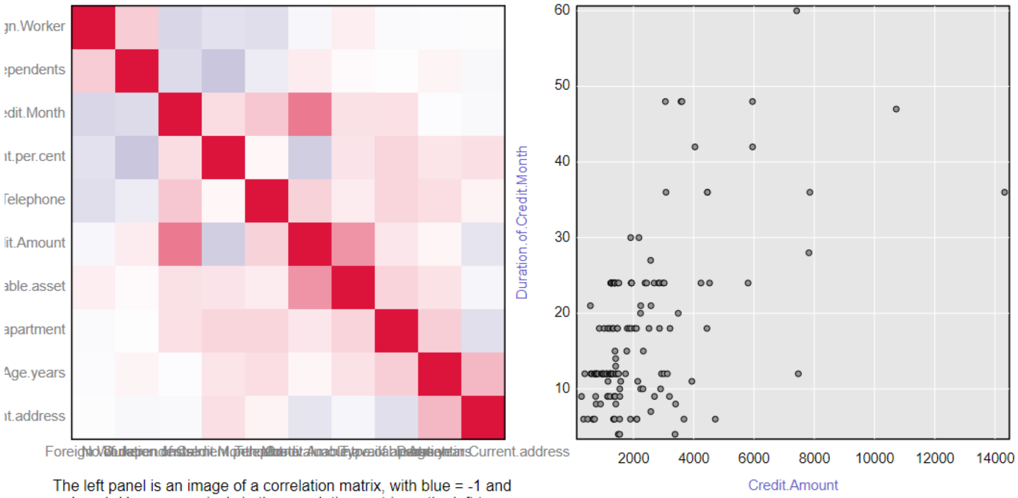
After we navigate the previous dataset, we picked below variables as predicted variables: Account Balance, Duration of Credit Month, Payment status of previous credit, Purpose, Credit-Amount, Value Savings Stocks, Length of current employment, Installment per cent, Most valuable available asset, Age years, Type of apartment and No of credits at this bank.

Since we need to categorize the outcome as creditworthy or noncreditworthy, I rule out the non-binary model. I picked up Binary model and tested the data on 4 different models: Logistic Regression, Decision Tree, Forest Model and Boosted Tree.

## Building the Training Set

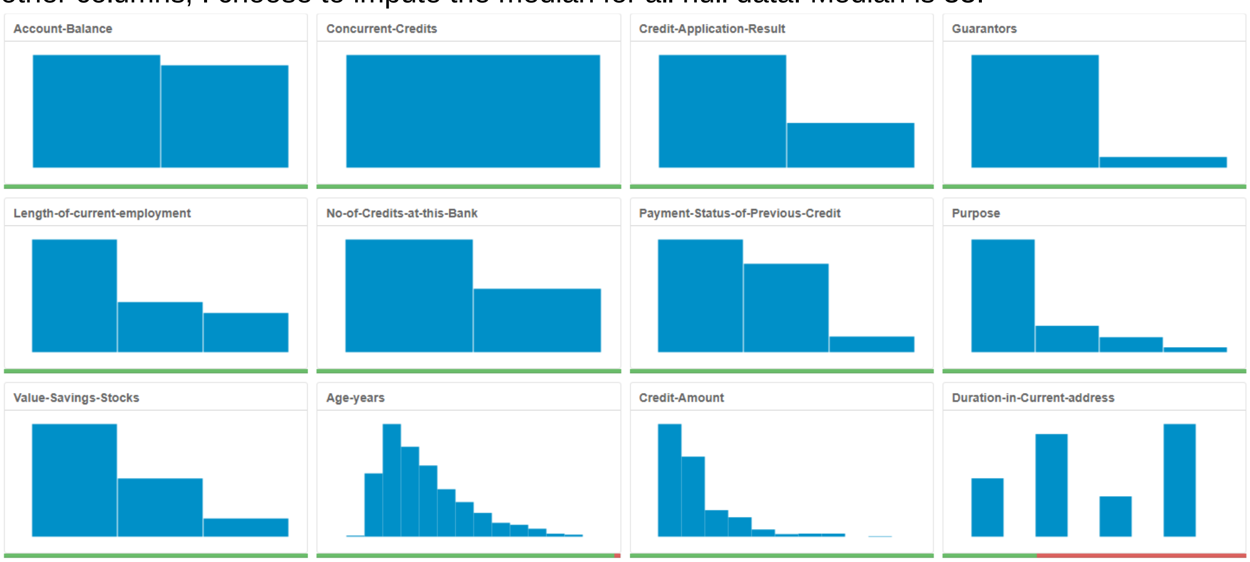
Before we build out model, we first need to clean the data set, remove unrelated values and select predicted variables.

- For numerical data fields, there are no field that highly-correlate with each other. All correlation is below 0.7. Please report from association analysis tool below.

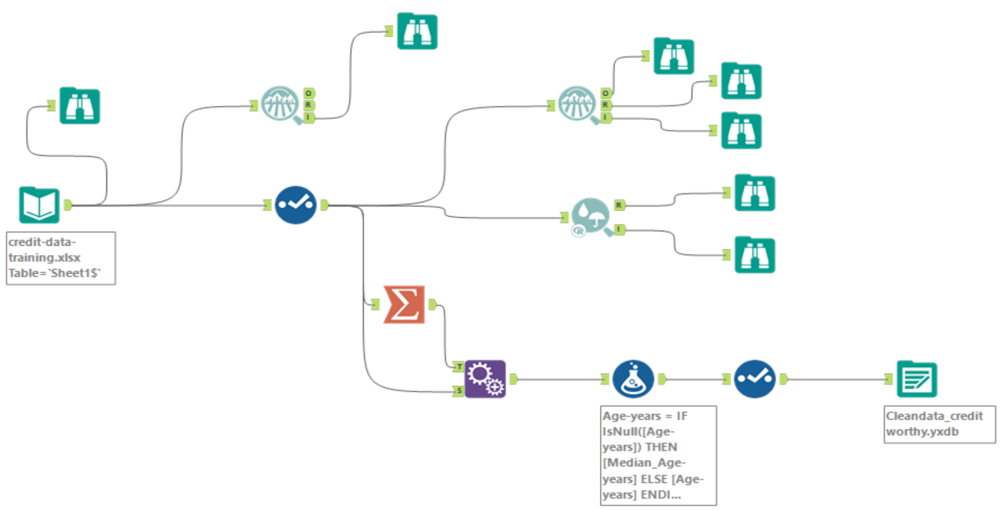



The left panel is an image of a correlation matrix, with blue = -1 and red = +1. Hover over pixels in the correlation matrix on the left to see the values; click to see the corresponding scatterplot on the right. The variables have been clustered based on degree of correlation, so that highly correlated variables appear adjacent to each other.

- There are some missing data in the dataset. Please see report from field summary. From below report we could see column Duration_to _current_address have large missing data. Thus, I choose to remove this field when we build our model. Another field Age_years also have 2% missing data. Since it's a small amount but it might impact other columns, I choose to impute the median for all null data. Median is 33.




- From report above, we also have some fields are low variability. We decided to remove these fields: **Occupation, Concurrent-credits, Guarantors, Foreign-worker, No-of-dependents and Telephone.**
- **Alteryx screenshot and file below:**




data wrangling.yxmd

# Train your Classification Models

First, I created my Estimation and Validation samples where 70% of your dataset should go to Estimation and 30% of your entire dataset should be reserved for Validation. Set the Random Seed to 1.

Create all of the following models: Logistic Regression, Decision Tree, Forest Model, Boosted Model.

Logistic Regression:
From this model, Duration of credit month are most important.
The overall percent accuracy is 78%.
Bias calculation:
PPV = true positives/ (true positives+ false positives) = 95 / (95+23) = .8
NPV= true negatives \ (true negatives + false negatives) = 22/ (22+10) = .68
There's bias towards correctly predicting creditworthy.

**Report for Logistic Regression Model LR_creditworthy**

*Basic Summary*

Call:
glm(formula = Credit.Application.Result ~ Account.Balance + Duration.of.Credit.Month + Payment.Status.of.Previous.Credit + Purpose + Credit.Amount + Value.Savings.Stocks + Length.of.current.employment + Instalment.per.cent + Most.valuable.available.asset + Age.years + Type.of.apartment + No.of.Credits.at.this.Bank, family = binomial(logit), data = the.data)

Deviance Residuals:

| Min | 1Q | Median | 3Q | Max |
|---|---|---|---|---|
| -2.088 | -0.719 | -0.430 | 0.686 | 2.542 |

Coefficients:

| | Estimate | Std. Error | z value | Pr(>\|z\|) | |
|---|---|---|---|---|---|
| (Intercept) | -3.0136120 | 1.013e+00 | -2.9760 | 0.00292 | ** |
| Account.BalanceSome Balance | -1.5433699 | 3.232e-01 | -4.7752 | 1.79e-06 | *** |
| Duration.of.Credit.Month | 0.0064973 | 1.371e-02 | 0.4738 | 0.63565 | |
| Payment.Status.of.Previous.CreditPaid Up | 0.4054309 | 3.841e-01 | 1.0554 | 0.29124 | |
| Payment.Status.of.Previous.CreditSome Problems | 1.2607175 | 5.335e-01 | 2.3632 | 0.01812 | * |
| PurposeNew car | -1.7541034 | 6.276e-01 | -2.7951 | 0.00519 | ** |
| PurposeOther | -0.3191177 | 8.342e-01 | -0.3825 | 0.70206 | |
| PurposeUsed car | -0.7839554 | 4.124e-01 | -1.9008 | 0.05733 | . |
| Credit.Amount | 0.0001764 | 6.838e-05 | 2.5798 | 0.00989 | ** |
| Value.Savings.StocksNone | 0.6074082 | 5.100e-01 | 1.1911 | 0.23361 | |
| Value.Savings.Stocks£100-£1000 | 0.1694433 | 5.649e-01 | 0.3000 | 0.7642 | |
| Length.of.current.employment4-7 yrs | 0.5224158 | 4.930e-01 | 1.0596 | 0.28934 | |
| Length.of.current.employment< 1yr | 0.7779492 | 3.956e-01 | 1.9664 | 0.04925 | * |
| Instalment.per.cent | 0.3109833 | 1.399e-01 | 2.2232 | 0.0262 | * |
| Most.valuable.available.asset | 0.3258706 | 1.556e-01 | 2.0945 | 0.03621 | * |
| Age.years | -0.0141206 | 1.535e-02 | -0.9202 | 0.35747 | |
| Type.of.apartment | -0.2603038 | 2.956e-01 | -0.8805 | 0.3786 | |
| No.of.Credits.at.this.BankMore than 1 | 0.3619545 | 3.815e-01 | 0.9487 | 0.34275 | |

Significance codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
(Dispersion parameter for binomial taken to be 1 )

Null deviance: 413.16 on 349 degrees of freedom
Residual deviance: 322.31 on 332 degrees of freedom
McFadden R-Squared: 0.2199, Akaike Information Criterion 358.3
Number of Fisher Scoring iterations: 5

**Model Comparison Report**

**Fit and error measures**

| Model | Accuracy | F1 | AUC | Accuracy_Creditworthy | Accuracy_Non-Creditworthy |
|---|---|---|---|---|---|
| LR_creditworthy | 0.7800 | 0.8520 | 0.7314 | 0.9048 | 0.4889 |

Model: model names in the current comparison.
Accuracy: overall accuracy, number of correct predictions of all classes divided by total sample number.
Accuracy_[class name]: accuracy of Class [class name] is defined as the number of cases that are **correctly** predicted to be Class [class name] divided by the total number of cases that actually belong to Class [class name], this measure is also known as *recall*.
AUC: area under the ROC curve, only available for two-class classification.
F1: F1 score, 2 * precision * recall / (precision + recall). The *precision* measure is the percentage of actual members of a class that were predicted to be in that class divided by the total number of cases predicted to be in that class. In situations where there are three or more classes, average precision and average recall values across classes are used to calculate the F1 score.

**Confusion matrix of LR_creditworthy**

| | Actual_Creditworthy | Actual_Non-Creditworthy |
|---|---|---|
| Predicted_Creditworthy | 95 | 23 |
| Predicted_Non-Creditworthy | 10 | 22 |

Decision Tree:
From this model, we can see account balance and age years are most important.
The overall percent accuracy is 67%.
Bias calculation:
PPV = true positives/ (true positives+ false positives) = 83 / (83+28) = .75
NPV= true negatives \ (true negatives + false negatives) = 17/ (17+23) = .43
In this model, there's also bias towards correctly predicting creditworthy.

**Summary Report for Decision Tree Model DT_creditworhty**

Call:
rpart(formula = Credit.Application.Result ~ Account.Balance + Duration.of.Credit.Month + Payment.Status.of.Previous.Credit + Purpose + Credit.Amount + Value.Savings.Stocks + Length.of.current.employment + Instalment.per.cent + Most.valuable.available.asset + Age.years + Type.of.apartment + No.of.Credits.at.this.Bank, data = the.data, minsplit = 20, minbucket = 7, usesurrogate = 0, xval = 10, maxdepth = 20, cp = 0)

Model Summary
Variables actually used in tree construction:
[1] Account.Balance Age.years
[3] Credit.Amount Duration.of.Credit.Month
[5] Instalment.per.cent Length.of.current.employment
[7] Most.valuable.available.asset No.of.Credits.at.this.Bank
[9] Payment.Status.of.Previous.Credit Purpose
[11] Value.Savings.Stocks
Root node error: 97/350 = 0.27714
n= 350

*Pruning Table*

| Level | CP | Num Splits | Rel Error | X Error | X Std Dev |
|---|---|---|---|---|---|
| 1 | 0.0687285 | 0 | 1.00000 | 1.00000 | 0.086326 |
| 2 | 0.0412371 | 3 | 0.79381 | 0.92784 | 0.084295 |
| 3 | 0.0257732 | 4 | 0.75258 | 0.91753 | 0.083987 |
| 4 | 0.0206186 | 8 | 0.64948 | 0.92784 | 0.084295 |
| 5 | 0.0103093 | 9 | 0.62887 | 1.00000 | 0.086326 |
| 6 | 0.0017182 | 12 | 0.59794 | 1.06186 | 0.087894 |
| 7 | 0.0000000 | 18 | 0.58763 | 1.05155 | 0.087644 |

## Model Comparison Report

### Fit and error measures

| Model | Accuracy | F1 | AUC | Accuracy_Creditworthy | Accuracy_Non-Creditworthy |
|---|---|---|---|---|---|
| DT_creditworhty | 0.6667 | 0.7685 | 0.6272 | 0.7905 | 0.3778 |

Model: model names in the current comparison.
Accuracy: overall accuracy, number of correct predictions of all classes divided by total sample number.
Accuracy_[class name]: accuracy of Class [class name] is defined as the number of cases that are **correctly** predicted to be Class [class name] divided by the total number of cases that actually belong to Class [class name], this measure is also known as *recall*.
AUC: area under the ROC curve, only available for two-class classification.
F1: F1 score, 2 * precision * recall / (precision + recall). The *precision* measure is the percentage of actual members of a class that were predicted to be in that class divided by the total number of cases predicted to be in that class. In situations where there are three or more classes, average precision and average recall values across classes are used to calculate the F1 score.

### Confusion matrix of DT_creditworhty

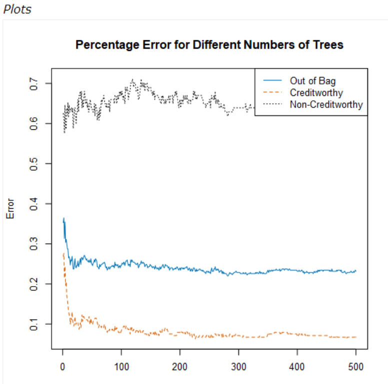
| | Actual_Creditworthy | Actual_Non-Creditworthy |
|---|---|---|
| Predicted_Creditworthy | 83 | 28 |
| Predicted_Non-Creditworthy | 22 | 17 |

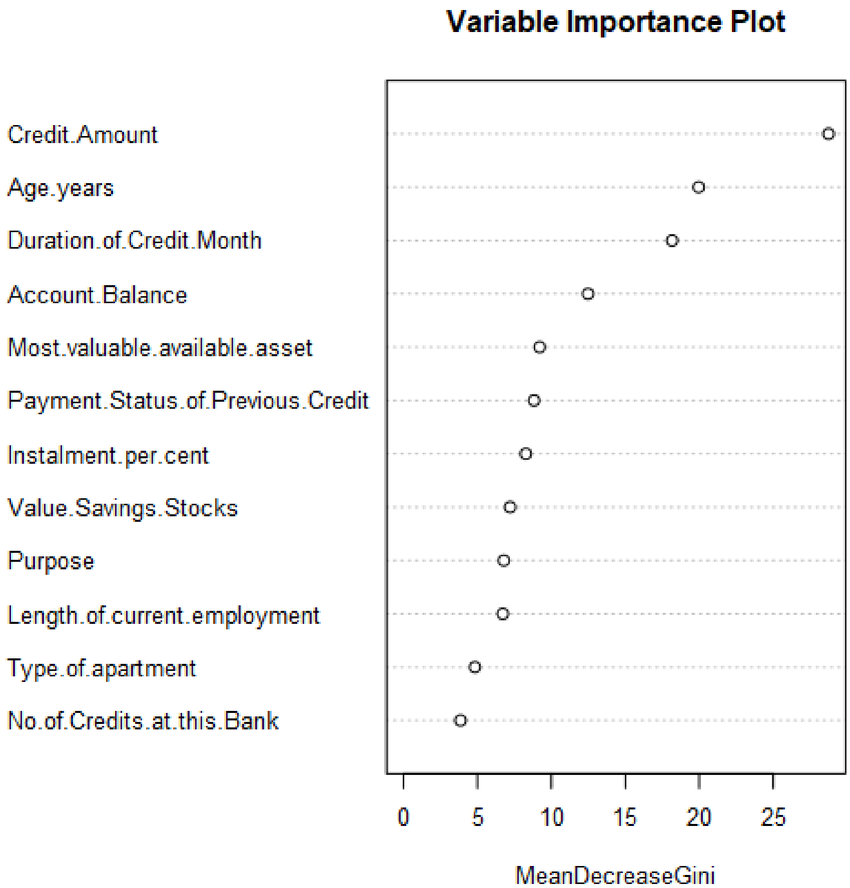
Forest Model:
From this model, we can see credit amount is most important.
The overall percent accuracy is 79%.
Bias calculation:
PPV = true positives/ (true positives+ false positives) = 102 / (102+28) = .78
NPV= true negatives \ (true negatives + false negatives) = 17/ (17+3) = .85
There's almost no in this model.

*Basic Summary*
Call:
randomForest(formula = Credit.Application.Result ~ Account.Balance + Duration.of.Credit.Month + Payment.Status.of.Previous.Credit + Purpose + Credit.Amount + Value.Savings.Stocks + Length.of.current.employment + Instalment.per.cent + Most.valuable.available.asset + Age.years + Type.of.apartment + No.of.Credits.at.this.Bank, data = the.data, ntree = 500, replace = TRUE)
Type of forest: classification
Number of trees: 500
Number of variables tried at each split: 3
OOB estimate of the error rate: 23.1%
Confusion Matrix:

| | Classification Error | Creditworthy | Non-Creditworthy |
|---|---|---|---|
| Creditworthy | 0.067 | 236 | 17 |
| Non-Creditworthy | 0.66 | 64 | 33 |

*Plots*






**Model Comparison Report**

**Fit and error measures**

| Model | Accuracy | F1 | AUC | Accuracy_Creditworthy | Accuracy_Non-Creditworthy |
|---|---|---|---|---|---|
| RF_creditworthy | 0.7933 | 0.8681 | 0.7368 | 0.9714 | 0.3778 |

Model: model names in the current comparison.
Accuracy: overall accuracy, number of correct predictions of all classes divided by total sample number.
Accuracy_[class name]: accuracy of Class [class name] is defined as the number of cases that are **correctly** predicted to be Class [class name] divided by the total number of cases that actually belong to Class [class name], this measure is also known as *recall*.
AUC: area under the ROC curve, only available for two-class classification.
F1: F1 score, 2 * precision * recall / (precision + recall). The *precision* measure is the percentage of actual members of a class that were predicted to be in that class divided by the total number of cases predicted to be in that class. In situations where there are three or more classes, average precision and average recall values across classes are used to calculate the F1 score.

**Confusion matrix of RF_creditworthy**

| | Actual_Creditworthy | Actual_Non-Creditworthy |
|---|---|---|
| Predicted_Creditworthy | 102 | 28 |
| Predicted_Non-Creditworthy | 3 | 17 |

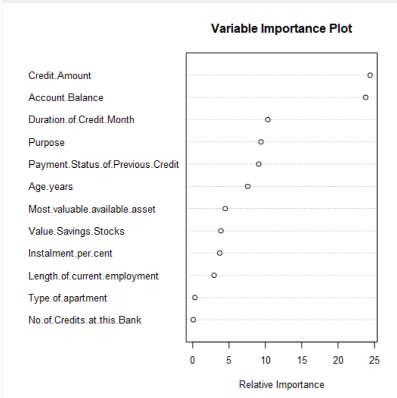
Boosted model:
From this model, we can see credit amount is most important.
The overall percent accuracy is 79%.
Bias calculation:
PPV = true positives/ (true positives+ false positives) = 101 / (101+28) = .78
NPV= true negatives \ (true negatives + false negatives) = 17/ (17+4) = .81
There’s almost no bias in this model.

**Report for Boosted Model Boosted_creditworthy**

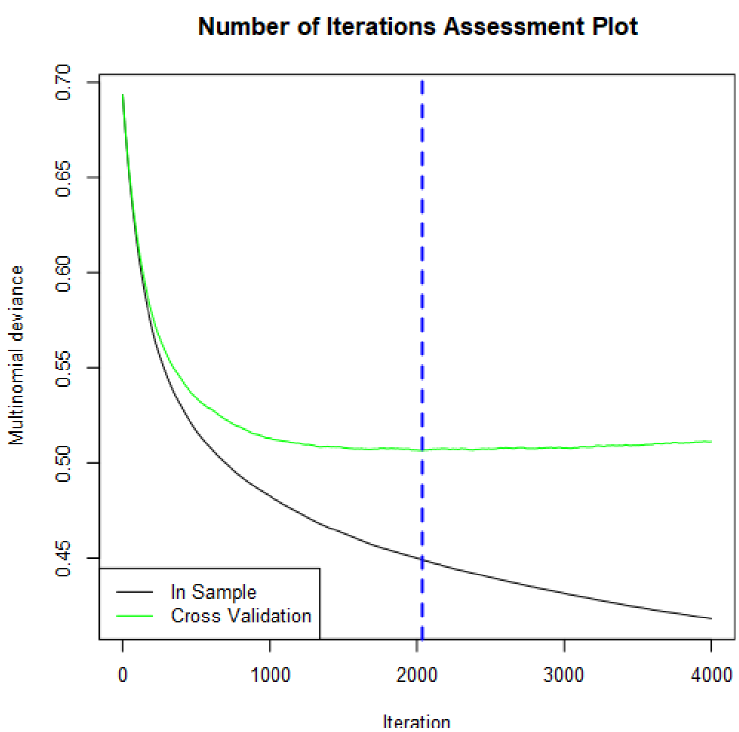
Basic Summary:

Loss function distribution: Bernoulli
Total number of trees used: 4000
Best number of trees based on 5-fold cross validation: 2036

Plots:




The Variable Importance Plot provides information about the relative importance of each predictor field. The measures are normalized to sum to 100, and the value for each field gives the relative percentage importance of that field to the overall model.




## Model Comparison Report

### Fit and error measures

| Model | Accuracy | F1 | AUC | Accuracy_Creditworthy | Accuracy_Non-Creditworthy |
|---|---|---|---|---|---|
| Boosted_creditworthy | 0.7867 | 0.8632 | 0.7524 | 0.9619 | 0.3778 |

Model: model names in the current comparison.
Accuracy: overall accuracy, number of correct predictions of all classes divided by total sample number.
Accuracy_[class name]: accuracy of Class [class name] is defined as the number of cases that are **correctly** predicted to be Class [class name] divided by the total number of cases that actually belong to Class [class name], this measure is also known as *recall*.
AUC: area under the ROC curve, only available for two-class classification.
F1: F1 score, 2 * precision * recall / (precision + recall). The *precision* measure is the percentage of actual members of a class that were predicted to be in that class divided by the total number of cases predicted to be in that class. In situations where there are three or more classes, average precision and average recall values across classes are used to calculate the F1 score.

### Confusion matrix of Boosted_creditworthy

| | Actual_Creditworthy | Actual_Non-Creditworthy |
|---|---|---|
| Predicted_Creditworthy | 101 | 28 |
| Predicted_Non-Creditworthy | 4 | 17 |

We can see there are some bias in the model’s predictions since the accuracy is different between Estimate report and validation report for each model.

# Writeup

**Model Comparison Report**

**Fit and error measures**

| Model | Accuracy | F1 | AUC | Accuracy_Creditworthy | Accuracy_Non-Creditworthy |
|---|---|---|---|---|---|
| LR_creditworthy | 0.7800 | 0.8520 | 0.7314 | 0.9048 | 0.4889 |
| DT_creditworthy | 0.6667 | 0.7685 | 0.6272 | 0.7905 | 0.3778 |
| RF_creditworthy | 0.7933 | 0.8681 | 0.7368 | 0.9714 | 0.3778 |
| Boosted_creditworthy | 0.7867 | 0.8632 | 0.7524 | 0.9619 | 0.3778 |

Model: model names in the current comparison.

Accuracy: overall accuracy, number of correct predictions of all classes divided by total sample number.

Accuracy_[class name]: accuracy of Class [class name] is defined as the number of cases that are **correctly** predicted to be Class [class name] divided by the total number of cases that actually belong to Class [class name], this measure is also known as *recall*.

AUC: area under the ROC curve, only available for two-class classification.

F1: F1 score, 2 * precision * recall / (precision + recall). The *precision* measure is the percentage of actual members of a class that were predicted to be in that class divided by the total number of cases predicted to be in that class. In situations where there are three or more classes, average precision and average recall values across classes are used to calculate the F1 score.

**Confusion matrix of Boosted_creditworthy**

| | Actual_Creditworthy | Actual_Non-Creditworthy |
|---|---|---|
| Predicted_Creditworthy | 101 | 28 |
| Predicted_Non-Creditworthy | 4 | 17 |

**Confusion matrix of DT_creditworhty**

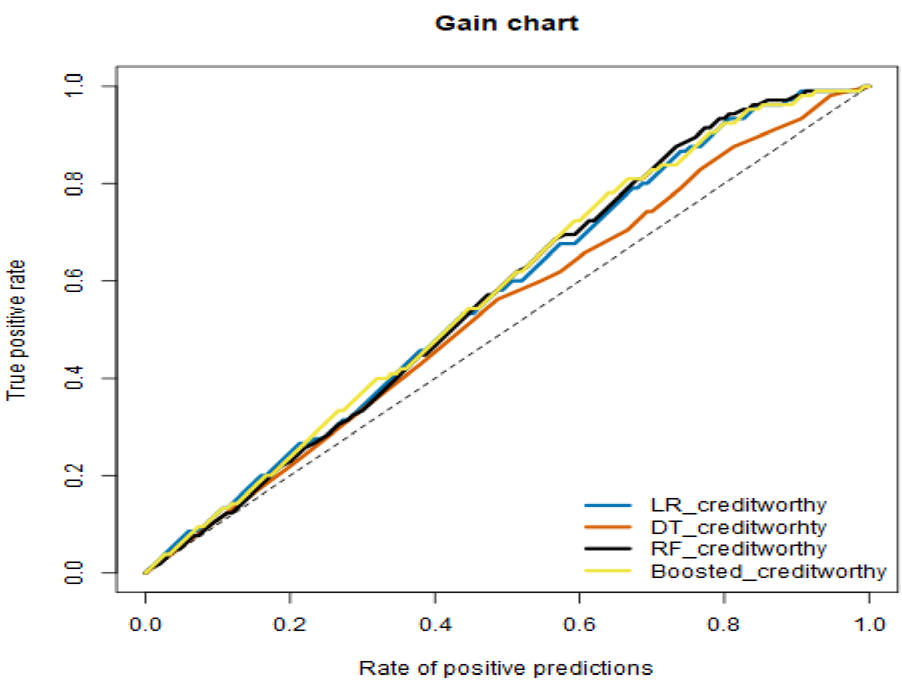
| | Actual_Creditworthy | Actual_Non-Creditworthy |
|---|---|---|
| Predicted_Creditworthy | 83 | 28 |
| Predicted_Non-Creditworthy | 22 | 17 |

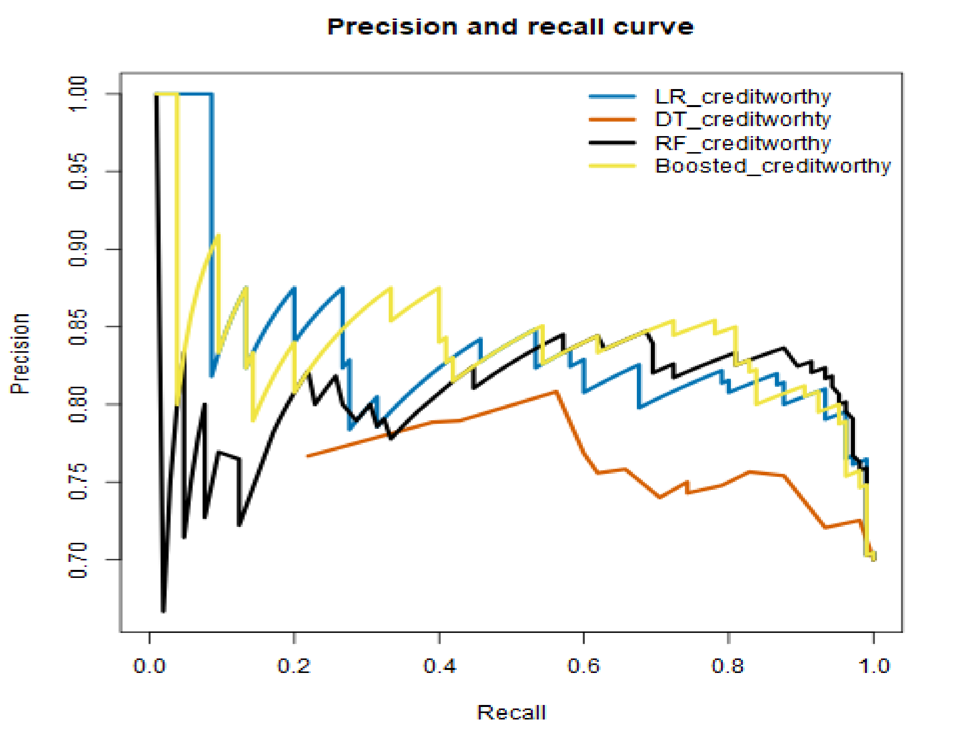
**Confusion matrix of LR_creditworthy**

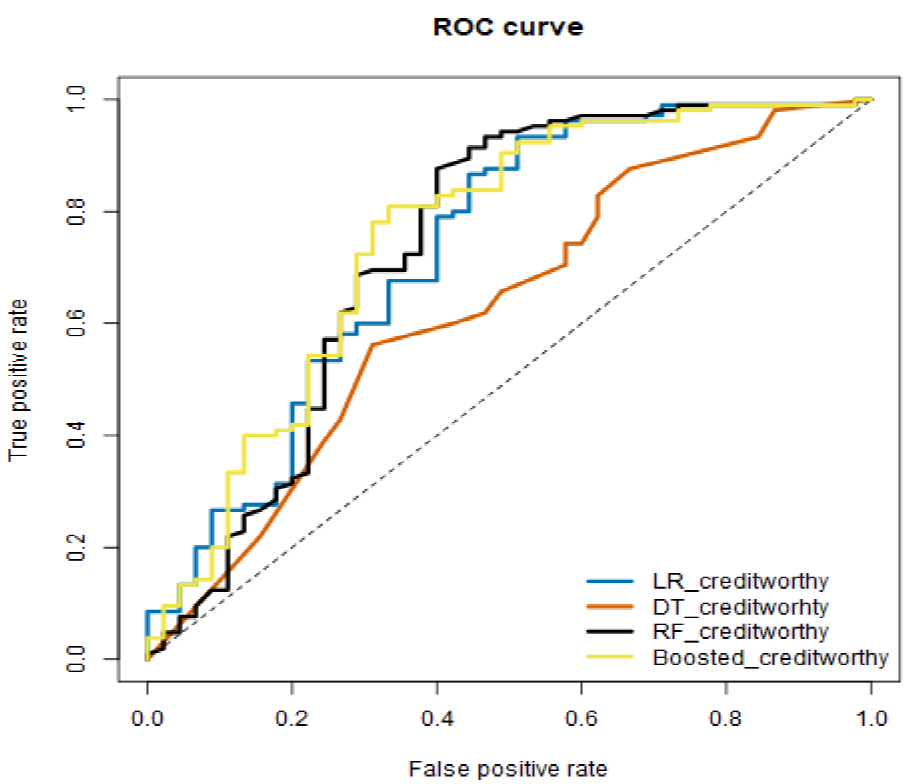
| | Actual_Creditworthy | Actual_Non-Creditworthy |
|---|---|---|
| Predicted_Creditworthy | 95 | 23 |
| Predicted_Non-Creditworthy | 10 | 22 |

**Confusion matrix of RF_creditworthy**

| | Actual_Creditworthy | Actual_Non-Creditworthy |
|---|---|---|
| Predicted_Creditworthy | 102 | 28 |
| Predicted_Non-Creditworthy | 3 | 17 |






After testing on 4 different models, I chose Random Forest model for below 4 reason:

1. It provided highest overall validation accuracy 79% which is highest among 4 different models.
2. This model has almost no bias. From below calculation, we see that 0.78 and 0.85 are quite close. Thus, forest model also works well on confusion metrix.
   Bias calculation:
   PPV = true positives/ (true positives+ false positives) = 102 / (102+28) = .78
   NPV= true negatives \ (true negatives + false negatives) = 17/ (17+3) = .85
3. Forest model also has highest accuracies with creditworthy and non-creditworthy prediction, as we can see this model correctly predicted 102 for actual creditworthy and 17 for non-actual creditworthy.
4. From the ROC graph, Forest model has the highest curve in these models and boosted model rises the fastest. After compare 2 models in chart, we still choose Forest model as it reached top left corner which means that for a given amount of false positive predictions, this model will give the best number of true positive predictions.

For final calculation, after we applied the Forest model to 500 new applications, we had 407 individuals are creditworthy.