Project: Creditworthiness

Step 1: Business and Data Understanding

Key Decisions:

1. What decisions needs to be made?

We consider a small bank that has been in operations for two years. The company needs to implement an efficient solution to be able automated decisions to reject or accept new credit applications and achieve a high performance in discriminating good customers from bad customers. To help the company, we have to build a statistical model of default to able to predict how likely a customer is going to default on a loan. Classification models allow to make those predictions and understand what the risk is involved with lending.

2. What data is needed to inform those decisions?

The possible information about the individual taking the credit that can use to implement a statistical model of default:
- socioeconomic characteristics, such as age, gender, residential status, marital status, number of dependents, employment status, profession, income level;
- amount, types and temporal profile of actual indebtedness;
- credit activity and payment history to trace a debtor's profile and know whether we have paid on time or late, or missed payments;
- behavioral information such as preferences consumer, payment habits.

3. What kind of model (Continuous, Binary, Non-Binary, Time-Series) do we need to use to help make these decisions?

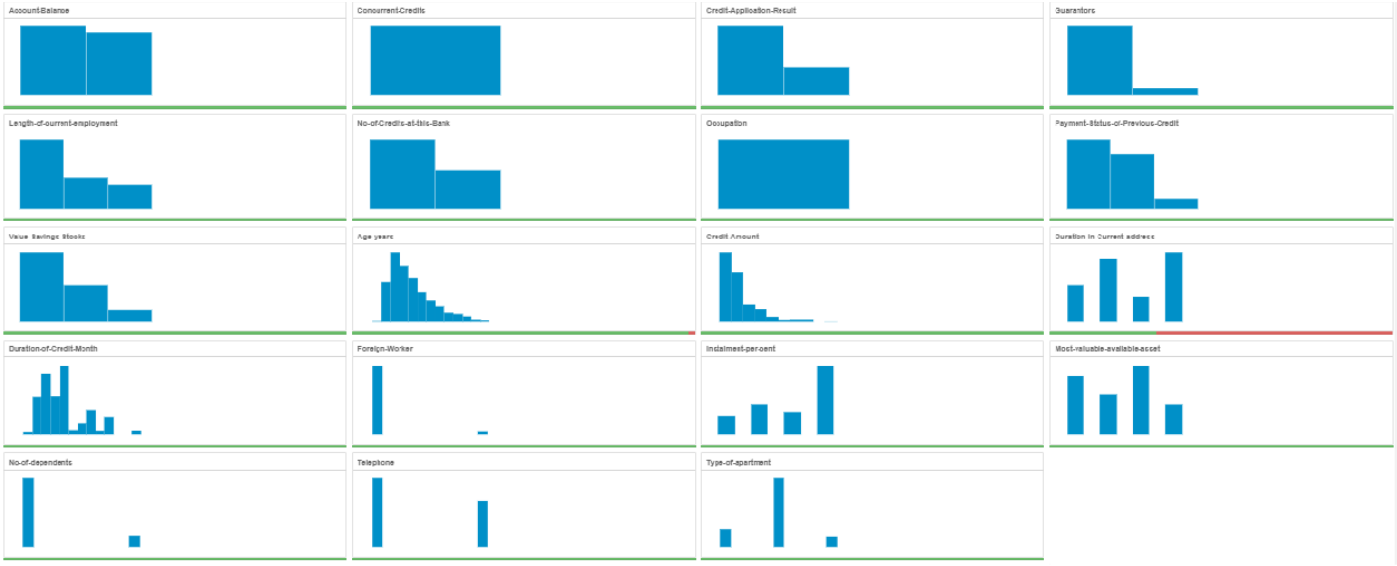
In the decision-making process of deciding whether or not to grant a loan to an applicant is typically used so-called credit scoring model. The aim of the credit-scoring model is to perform a classification to distinguish the good payers from bad payers. It is binary classification given that the task is to classify the object of a given set into two groups, predicting which group each one belongs to according to its characteristics.

Step 2: Building the Training Set

1. In your cleanup process, which fields did you remove or impute? Please justify why you removed or imputed these fields. Visualizations are encouraged.

To properly build the model, we need to explore and cleanup data. It requires understanding on the variables and a selection of the most significant ones. The available dataset includes 19 variables. To properly build the model, we need to explore and clean up data. It requires

understanding on the variables and a selection of the most significant ones. The available dataset includes 19 variables. Below, interactive dashboard, the output of “Field Summary Tool” of Alteryx.



The “Duration-in-Current-address” and “Age-years” have missing data, respectively 69% and 2%. Given the high number of missing data for “Duration-in-Current-address”, this variable is removed. In the case of “Age-years”, given the low number of missing data the best choice is to impute these values with the median age. It is chosen the median given that is more robust with respect to outliers than mean. The variables “Concurrent-Credits” and “Occupation”, “Guarantors”, ”No-of-dependents” and “Foreign-Worker” where the majority of data lies within one category are removed in order the analysis results are not skewed. We apply logic to get a list of potential predictor variables, then “Telephone” is removed due to its irrelevancy to predict default of applicants.
Below, Correlation Matrix between all numerical remaining variables.

Full Correlation Matrix

	Duration.of.Credit.Month	Credit.Amount	Instalment.per.cent	Most.valuable.available.asset	Age.years	Type.of.apartment
Duration.of.Credit.Month	1.000000	0.570441	0.079515	0.304734	-0.066319	0.153141
Credit.Amount	0.570441	1.000000	-0.285631	0.327762	0.068643	0.168683
Instalment.per.cent	0.079515	-0.285631	1.000000	0.078110	0.040540	0.082936
Most.valuable.available.asset	0.304734	0.327762	0.078110	1.000000	0.085437	0.379650
Age.years	-0.066319	0.068643	0.040540	0.085437	1.000000	0.333075
Type.of.apartment	0.153141	0.168683	0.082936	0.379650	0.333075	1.000000

There are not duplicate variables in fact none variable is highly correlate with each other.

Step 3: Train your Classification Models

1. Which predictor variables are significant or the most important? Please show the p-values or variable importance charts for all of your predictor variables.

2. Validate your model against the Validation set. What was the overall percent accuracy? Show the confusion matrix. Are there any bias seen in the model’s predictions?

We create all of the following models: Logistic Regression, Decision Tree, Forest Model, Boosted Model. The target variable is Credit-Application-Result.

- Logistic Regression - Stepwise

Below the coefficients in the logistic regression and correspondent p-value. The most important variables are “Account-Balance”, “Purpose” and “Credit-Amount” that are significance level of 0.05.

	Estimate	Std. Error	z value	Pr(>|z|)
(Intercept)	-2.9621914	6.837e-01	-4.3326	1e-05***
Account.BalanceSome Balance	-1.6053228	3.067e-01	-5.2344	1.65e-07***
Payment.Status.of.Previous.CreditPaid Up	0.2360857	2.977e-01	0.7930	0.42775
Payment.Status.of.Previous.CreditSome Problems	1.2154514	5.151e-01	2.3595	0.0183*
PurposeNew car	-1.6993164	6.142e-01	-2.7668	0.00566**
PurposeOther	-0.3257637	8.179e-01	-0.3983	0.69042
PurposeUsed car	-0.7645820	4.004e-01	-1.9096	0.05618.
Credit.Amount	0.0001704	5.733e-05	2.9716	0.00296**
Length.of.current.employment4-7 yrs	0.3127022	4.587e-01	0.6817	0.49545
Length.of.current.employment< 1yr	0.8125785	3.874e-01	2.0973	0.03596*
Instalment.per.cent	0.3016731	1.350e-01	2.2340	0.02549*
Most.valuable.available.asset	0.2650267	1.425e-01	1.8599	0.06289.

Below fit and error measures for logistic regression and confusion matrix. The overall accuracy of the model is 76%. Accuracies for creditworthy and non-creditworthy are 80% and 62.8% then the model is biased towards predicting customers as non-creditworthy.

Fit and error measures

Model	Accuracy	F1	AUC	Accuracy_Creditworthy	Accuracy_Non-Creditworthy
Stepwise_Log	0,7600	0,8364	0,7306	0,8000	0,6286

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], number of samples that are correctly predicted to be Class [class name] divided by number of samples predited to be Class [class name]

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

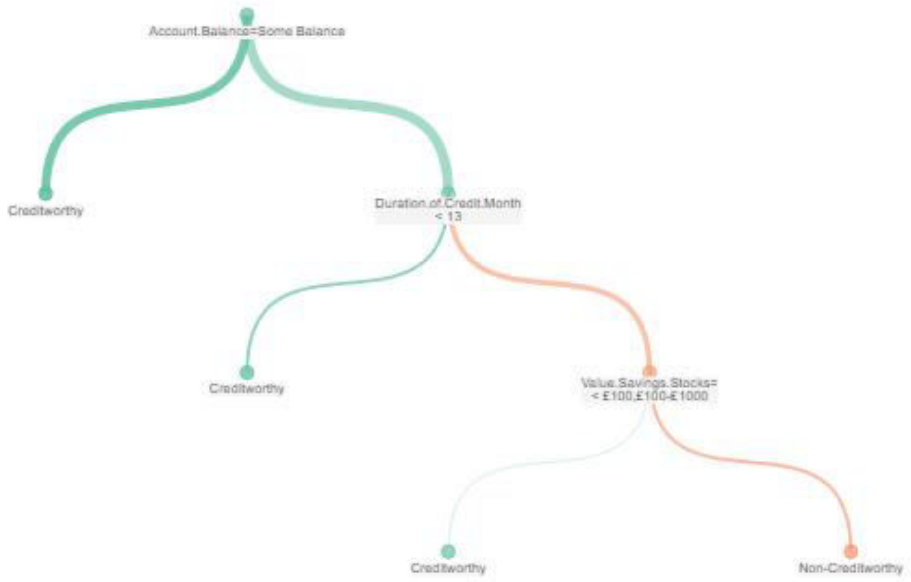
F1: F1 score, precision * recall / (precision + recall)

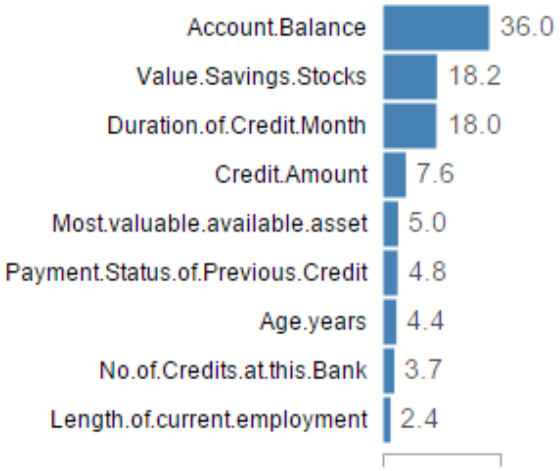
Confusion matrix of Stepwise_Log

	Actual_Creditworthy	Actual_Non-Creditworthy
Predicted_Creditworthy	92	23
Predicted_Non-Creditworthy	13	22

- Decision Tree

Below the decision tree and the variable importance plot. The most important variables are “Account-Balance”, “Value-Savings-Stocks” and “Duration-of-Credit-Month”.





Below fit and error measures for decision tree and confusion matrix. The overall accuracy of the model is 74.6%. The model is biased towards predicting customers as non-creditworthy given that the accuracies for creditworthy and non-creditworthy are 79.1% and 60% respectively.

Fit and error measures

Model	Accuracy	F1	AUC	Accuracy_Creditworthy	Accuracy_Non-Creditworthy
DT_Default	0,7467	0,8273	0,7054	0,7913	0,6000

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], number of samples that are correctly predicted to be Class [class name] divided by number of samples predited to be Class [class name]

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

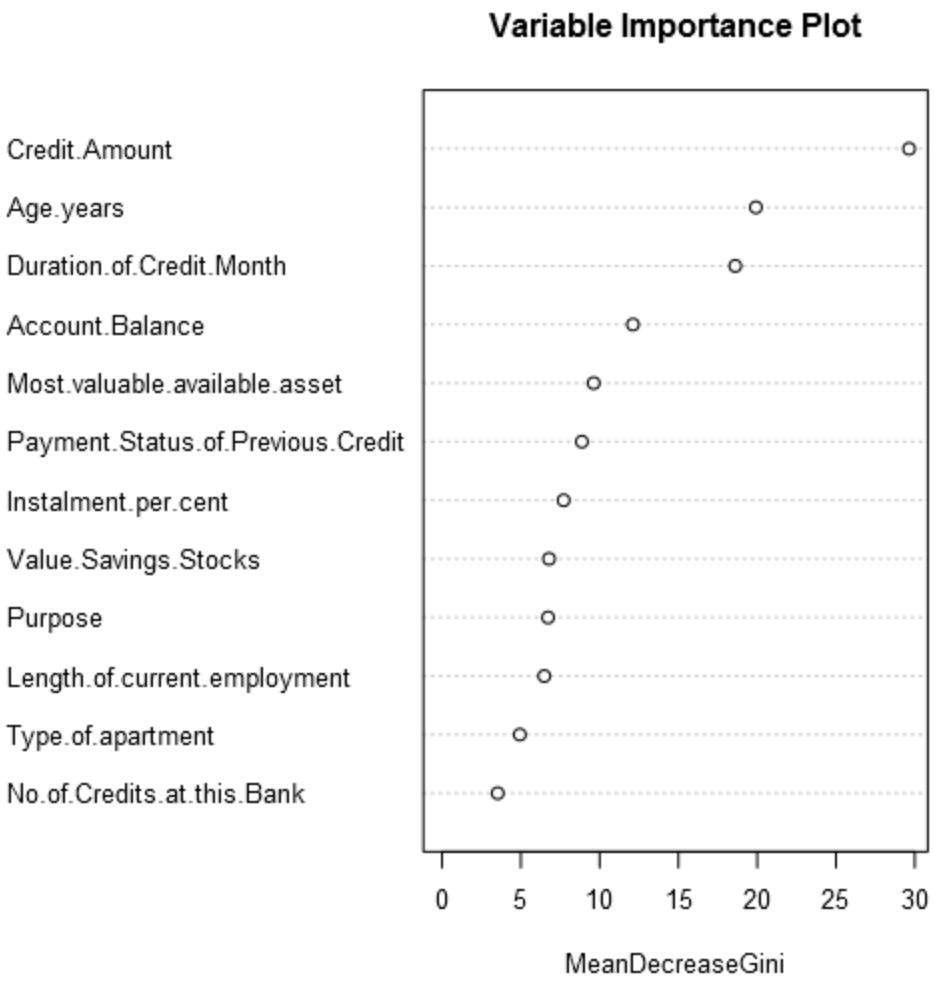
F1: F1 score, precision * recall / (precision + recall)

Confusion matrix of DT_Default

	Actual_Creditworthy	Actual_Non-Creditworthy
Predicted_Creditworthy	91	24
Predicted_Non-Creditworthy	14	21

- Random Forest

Below the variable importance plot. The most important variables are “Credit-Amount”, “Age-years” and “Duration-of-Credit-Month”.



Below fit and error measures for random forest and confusion matrix. The overall accuracy of the model is 80.6%. Accuracies for creditworthy and non-creditworthy are 79.6% and 86.3% respectively, the model is not biased.

Fit and error measures

Model	Accuracy	F1	AUC	Accuracy_Creditworthy	Accuracy_Non-Creditworthy
RF_Default	0,8067	0,8755	0,7392	0,7969	0,8636

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], number of samples that are correctly predicted to be Class [class name] divided by number of samples predited to be Class [class name]

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

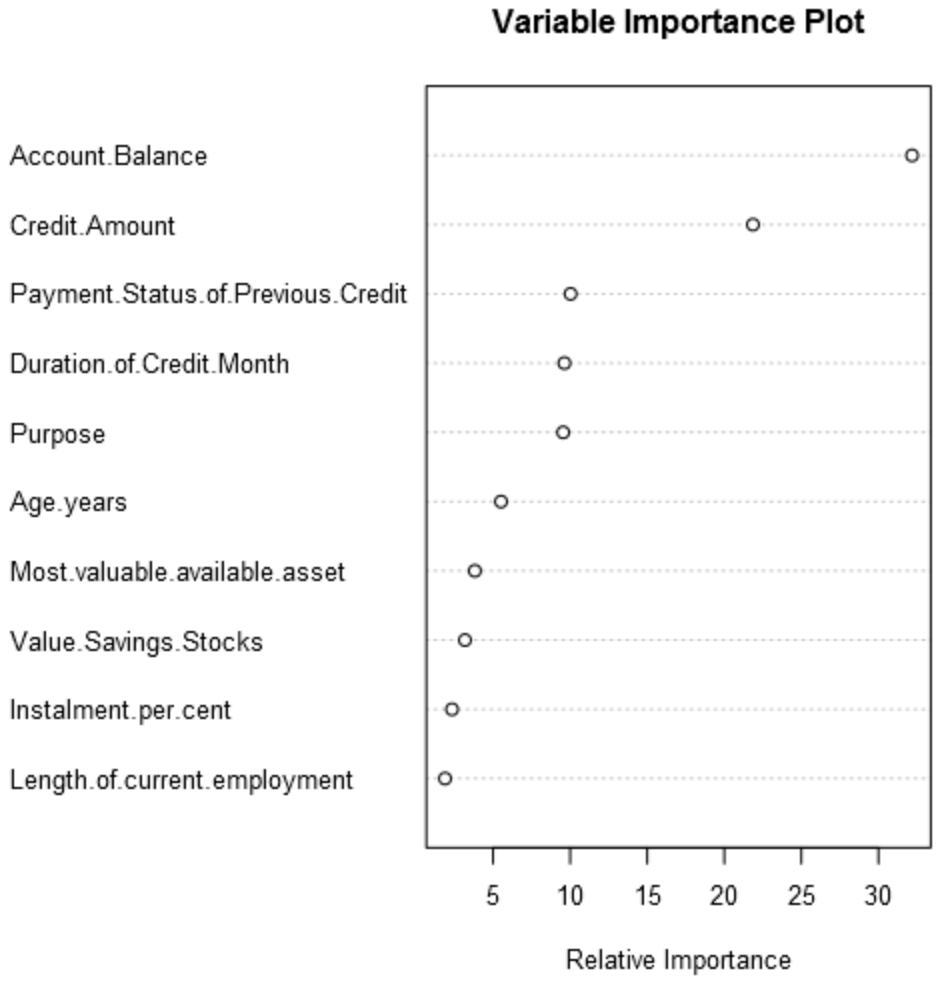
F1: F1 score, precision * recall / (precision + recall)

Confusion matrix of RF_Default

	Actual_Creditworthy	Actual_Non-Creditworthy
Predicted_Creditworthy	102	26
Predicted_Non-Creditworthy	3	19

- Boosted

Below the variable importance plot. The most important variables are “Account-Balance”, “Credit-Amount” and “Payment-Status-of-Previous-Credit”.



Below fit and error measures for boosted and confusion matrix. The overall accuracy of the model is 78.6%. The model is not biased given that the accuracies for creditworthy and non-creditworthy are 78.2% and 80.9% respectively.

Fit and error measures

Model	Accuracy	F1	AUC	Accuracy_Creditworthy	Accuracy_Non-Creditworthy
Boosted_Default	0,7867	0,8632	0,7524	0,7829	0,8095

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], number of samples that are correctly predicted to be Class [class name] divided by number of samples predited to be Class [class name]

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

F1: F1 score, precision * recall / (precision + recall)

Confusion matrix of Boosted_Default

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

Step 4: Writeup

1. Which model did you choose to use? Please justify your decision using only the following techniques:
    a. Overall Accuracy against your Validation set
    b. Accuracies within “Creditworthy” and “Non-Creditworthy” segments
    c. ROC graph
    d. Bias in the Confusion Matrices

Below, the model comparison report. The model chosen is Random Forest. The overall accuracy of this model is highest than other models. Furthermore, the accuracies for creditworthy and non-creditworthy are high and given that their difference is low, the model is not biased.

Model Comparison Report

Fit and error measures

Model	Accuracy	F1	AUC	Accuracy_Creditworthy	Accuracy_Non-Creditworthy
DT_Default	0,7467	0,8273	0,7054	0,7913	0,6000
RF_Default	0,8067	0,8755	0,7392	0,7969	0,8636
Boosted_Default	0,7867	0,8632	0,7524	0,7829	0,8095
Stepwise_Log	0,7600	0,8364	0,7306	0,8000	0,6286

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], number of samples that are correctly predicted to be Class [class name] divided by number of samples predited to be Class [class name]

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

F1: F1 score, precision * recall / (precision + recall)

Confusion matrix of Boosted_Default

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

Confusion matrix of DT_Default

	Actual_Creditworthy	Actual_Non-Creditworthy
Predicted_Creditworthy	91	24
Predicted_Non-Creditworthy	14	21

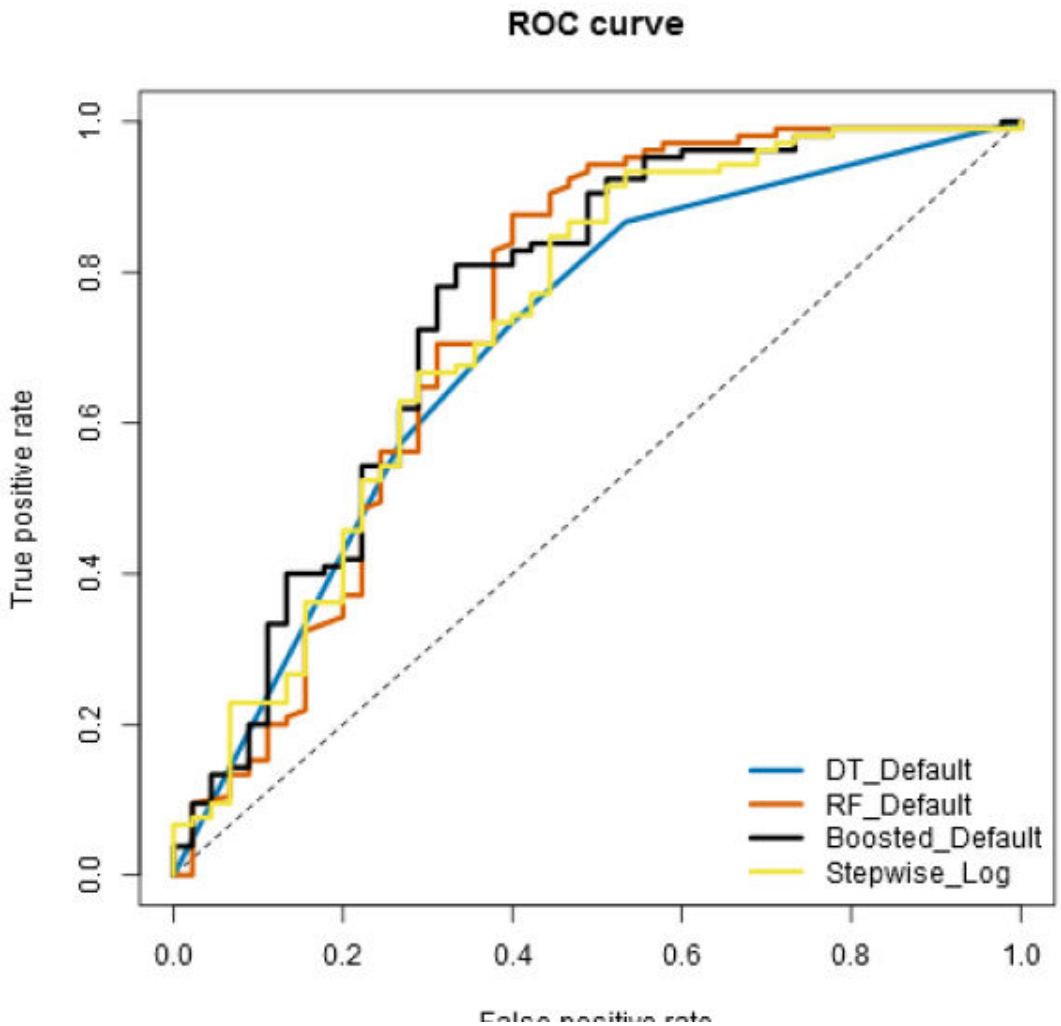
Confusion matrix of RF_Default

	Actual_Creditworthy	Actual_Non-Creditworthy
Predicted_Creditworthy	102	26
Predicted_Non-Creditworthy	3	19

Confusion matrix of Stepwise_Log

	Actual_Creditworthy	Actual_Non-Creditworthy
Predicted_Creditworthy	92	23
Predicted_Non-Creditworthy	13	22

Below the plot of ROC curve. It is a plot of the true positive rate against the false positive rate for the different possible cut off that separating two categories. The accuracy of separates two group is measured by the area under the ROC curve, AUC. The model has an AUC of 73.5 %, which is a good value.



2. How many individuals are creditworthy?

The persons are labeled as “Creditworthy” if “Score_Creditworthy” is greater than “Score_NonCreditworthy”. We use forest models to score new customers, we have 408 creditworthy customers.