Project: Creditworthiness

Step 1: Business and Data Understanding

Key Decisions:

- What decisions needs to be made?

  We must predict if a loan applicant is creditworthy or not, to decide if the bank loans money to the applicant.

- What data is needed to inform those decisions?

  First, we need a dataset to train our model containing the attribute creditworthy or not creditworthy with several other prediction variables.
  After we trained the model we need a validation set to measure the accuracy of our model. If we have decided on a model to use, we use it on the dataset of new loan applicants to get the probability of the classification. With this new prediction we can decide on who gets a loan and who doesn't.

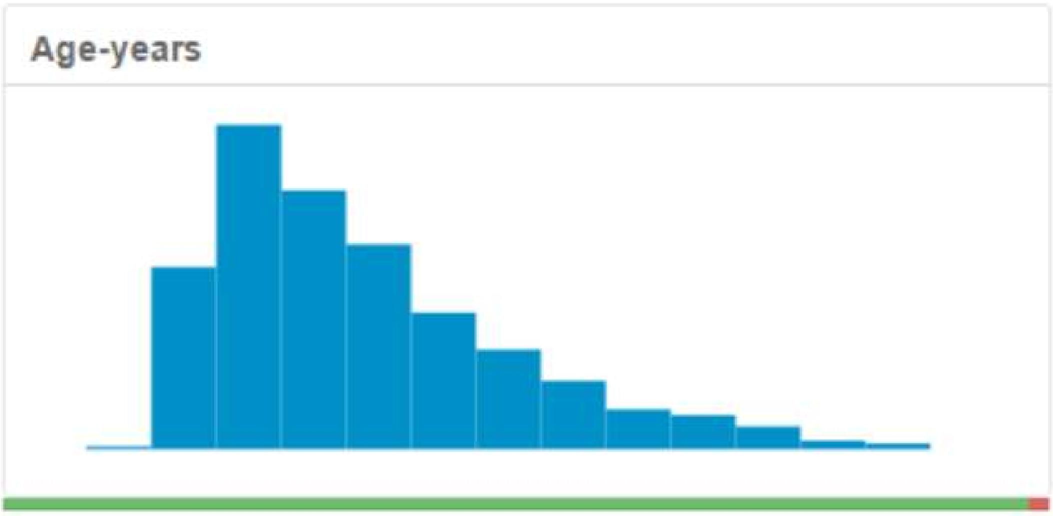
- What kind of model (Continuous, Binary, Non-Binary, Time-Series) do we need to use to help make these decisions?

  We must classify the applicants into binary attributes of creditworthy and not creditworthy. Therefore, we must use a binary classification model.

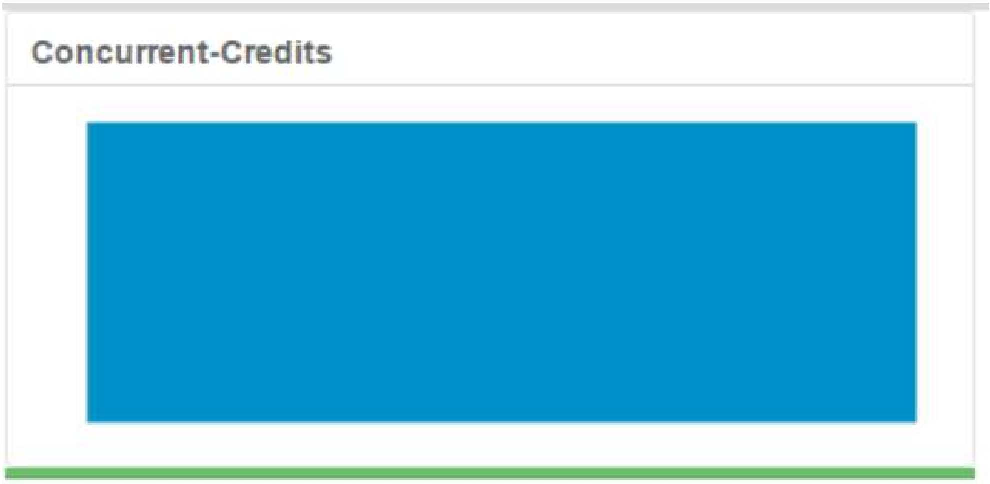
Step 2: Building the Training Set

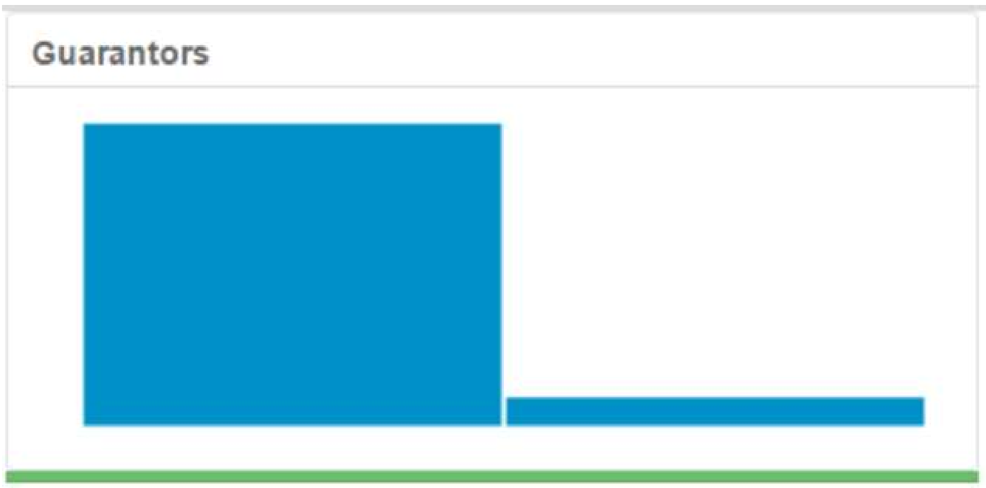
The field Duration-in-current-adress has too many missing values, so it is removed from the dataset.

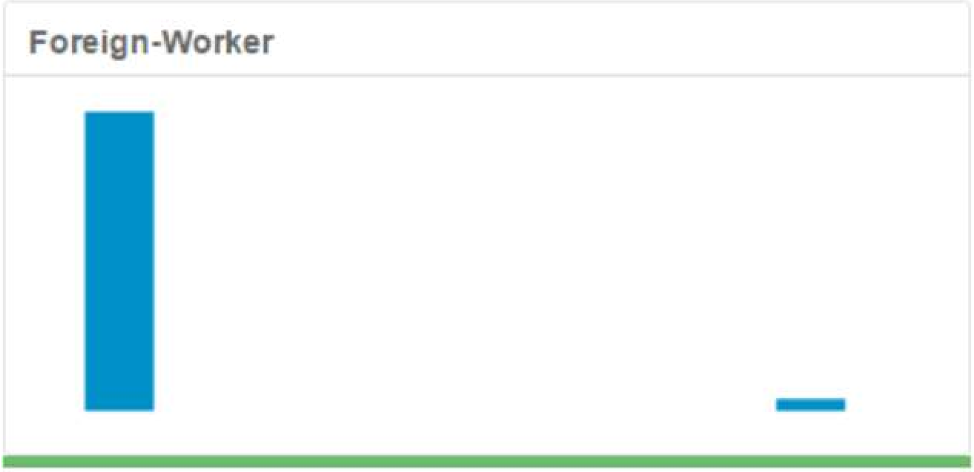
The field age-years is imputed. It is skewed so its better to use the median than the average.

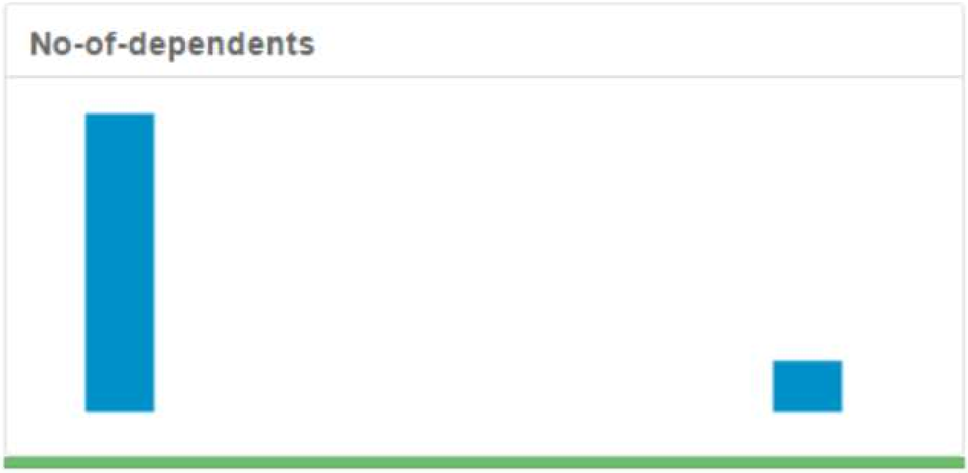


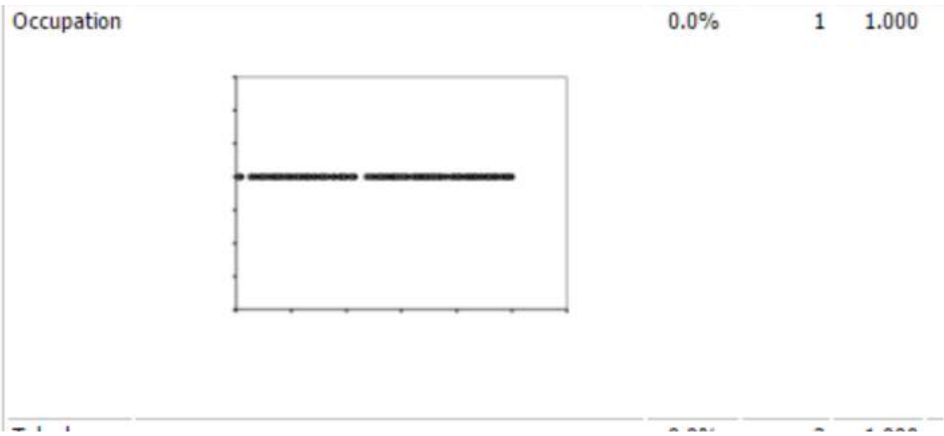
The fields concurrent-credits, guarantors, occupation, no-of-dependents and foreign-worker are removed due to low variability.





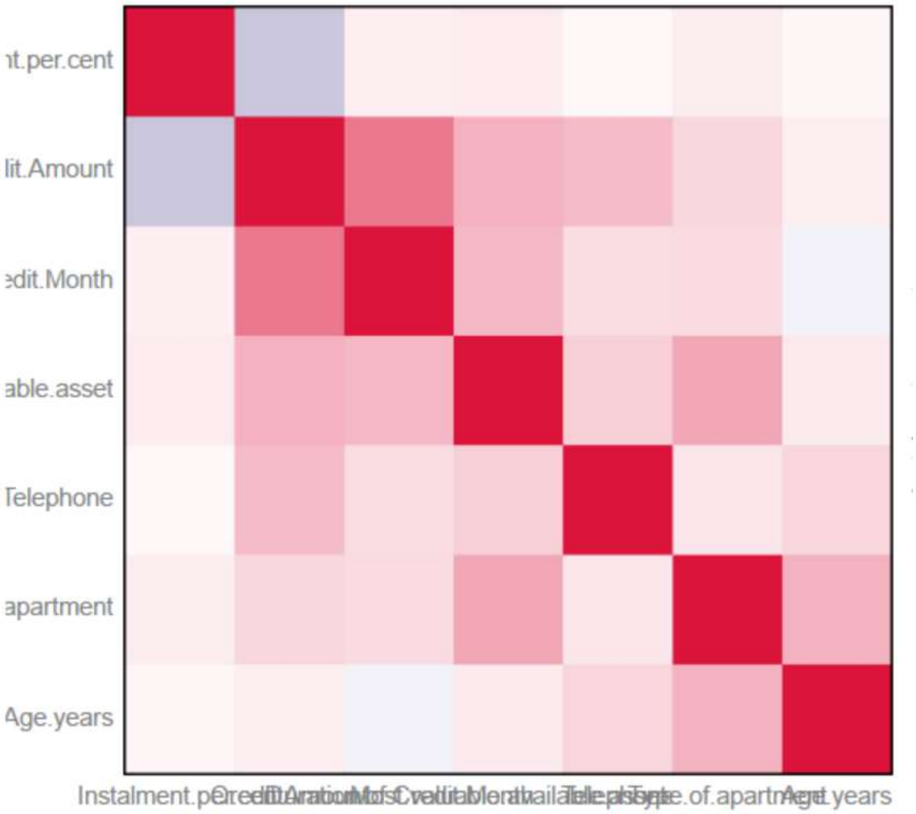






The field telephone should be removed due to its irrelevancy to the applicants creditworthy.

The correlation matrix shows that no variable is highly correlated with any other variable (a correlation higher than 0.7). Therefore, all other variables are kept to train the model.



Step 3: Train your Classification Models

a.) Logistic Regression (stepwise)

Using Credit Application Result as the target variable, the most significant prediction variables are account balance, purpose and credit amount. (p-value of less than 0.05)

The accuracy of the model is around 0.76. the accuracy for creditworthy (0.88) is higher than the accuracy for not creditworthy (0.49). It seems the model is biased to predict the applicants as non-creditworthy.

Report

Report for Logistic Regression Model X

Basic Summary

Call:
glm(formula = Credit.Application.Result ~ Account.Balance + Payment.Status.of.Previous.Credit + Purpose + Credit.Amount + Length.of.current.employment + Instalment.per.cent + Most.valuable.available.asset, family = binomial(logit), data = the.data)

Deviance Residuals:

Min	1Q	Median	3Q	Max
-2.289	-0.713	-0.448	0.722	2.454

Coefficients:

	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 .

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: 328.55 on 338 degrees of freedom
McFadden R-Squared: 0.2048, AIC: 352.5

Number of Fisher Scoring iterations: 5

Type II Analysis of Deviance Tests

Model Comparison Report

Fit and error measures

Model	Accuracy	F1	AUC	Accuracy_Creditworthy	Accuracy_Non-Creditworthy
StepReg_credit	0.7600	0.8364	0.7306	0.8762	0.4889

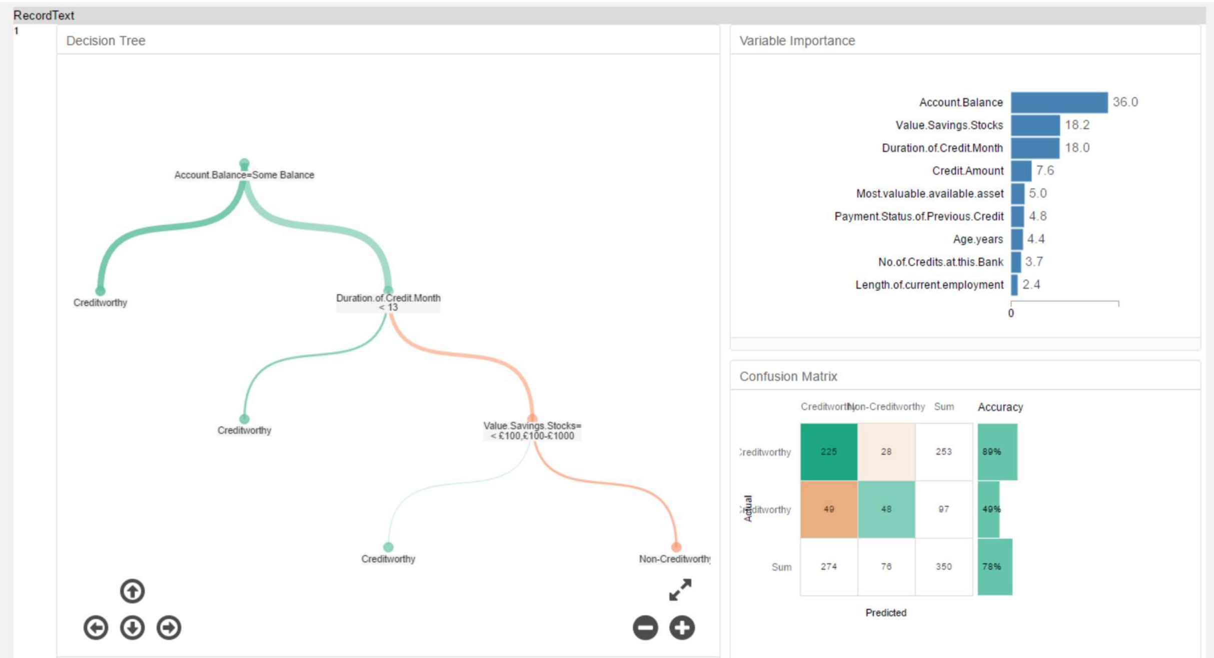
Confusion matrix of StepReg_credit

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

b.) Decision Tree

Using Credit Application Result as the target variable, shows that the most significant prediction variables of the model are account balance, savings stocks, duration of credit.

The overall accuracy is around 0.75. the accuracy for creditworthy is higher (0.87) than the accuracy for the predictions of non-creditworthy (0.47). This model also seems to be biased towards predicting applicants as non-creditworthy.



Model Comparison Report

Fit and error measures

Model	Accuracy	F1	AUC	Accuracy_Creditworthy	Accuracy_Non-Creditworthy
Dtree_credit	0.7467	0.8273	0.7054	0.8667	0.4667

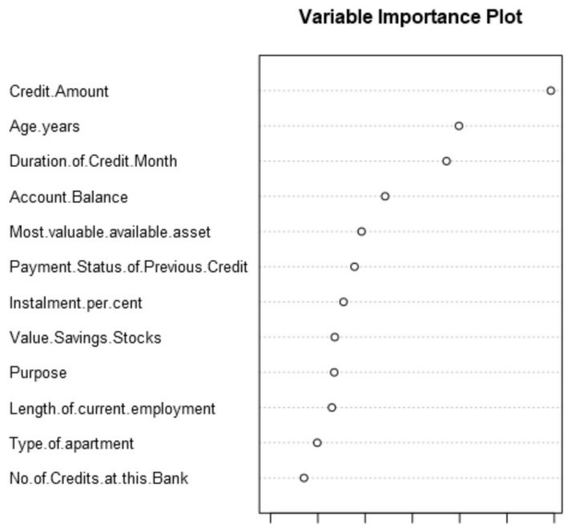
Confusion matrix of Dtree_credit

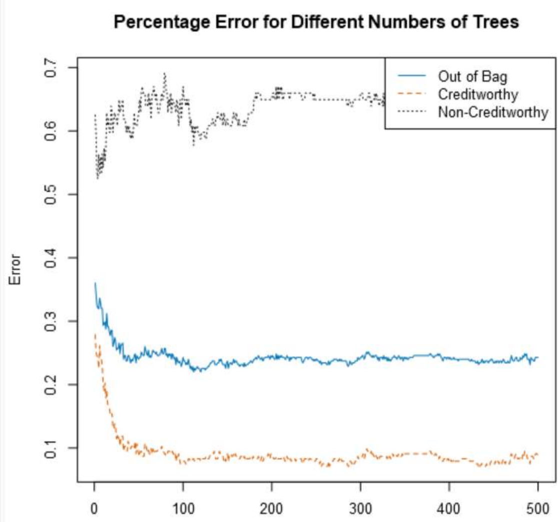
	Actual_Creditworthy	Actual_Non-Creditworthy
Predicted_Creditworthy	91	24
Predicted_Non-Creditworthy	14	21

c.) Forest Model

Using the target variable Credit Application Result again, we see that the Top 3 prediction variables with the most significance are credit amount, age years and duration of credit.

We see that the percentage of error flatlines around 100 to 120 trees. The overall accuracy is 0.81. With the accuracy for creditworthy (0.97) much higher than the accuracy for non-creditworthy (0.42). The confusion matrix suggests that the model isn't biased too much. The accuracy for non-creditworthy may be low but the precision (0.797) and the negative predictive value (NPV) are comparable (0.864). This means that we can't conclude if the model is biased to predict more towards creditworthy or non-creditworthy compared to the real distribution of both attributes in the actual dataset.





Model Comparison Report

Fit and error measures

Model	Accuracy	F1	AUC	Accuracy_Creditworthy	Accuracy_Non-Creditworthy
Forest_credit	0.8067	0.8755	0.7392	0.9714	0.4222

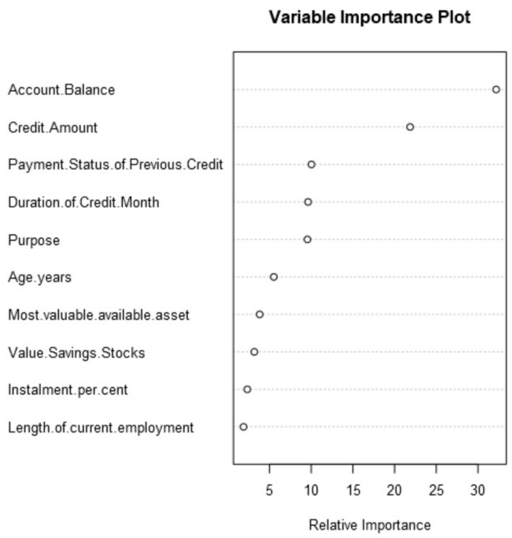
Confusion matrix of Forest_credit

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

d.) Boosted Model

Using Credit Application Result as the target variable, the most significant prediction variables are account balance, credit amount and payment status of previous credit.

The overall accuracy is 0.79. The accuracy for creditworthy (0.96) is higher than the accuracy for non-creditworthy (0.38). The precision is around 0.783 while the NPV is 0.81. These derivations of the confusion matrix suggest that the model is not biased towards any of the two attributes.



Model Comparison Report

Fit and error measures

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

Confusion matrix of Boosted_credit

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

Step 4: Writeup

The best model according to the overall accuracy against the validation dataset is the forest model with 80% accuracy. In addition, it is one of the models which isn't biased and has a high precision as well as a high NPV.

The forest model has the highest accuracy in the creditworthy segment but doesn't perform as well as other models in the non-creditworthy segment.

Model Comparison Report

Fit and error measures

Model	Accuracy	F1	AUC	Accuracy_Creditworthy	Accuracy_Non-Creditworthy
Dtree_credit	0.7467	0.8273	0.7054	0.8667	0.4667
Forest_credit	0.8067	0.8755	0.7392	0.9714	0.4222
Boosted_credit	0.2067	0.0630	0.2491	0.0381	0.6000
StepReg_credit	0.7600	0.8364	0.7306	0.8762	0.4889

Confusion matrix of Boosted_credit

	Actual_Creditworthy	Actual_Non-Creditworthy
Predicted_Creditworthy	4	18
Predicted_Non-Creditworthy	101	27

Confusion matrix of Dtree_credit

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

Confusion matrix of Forest_credit

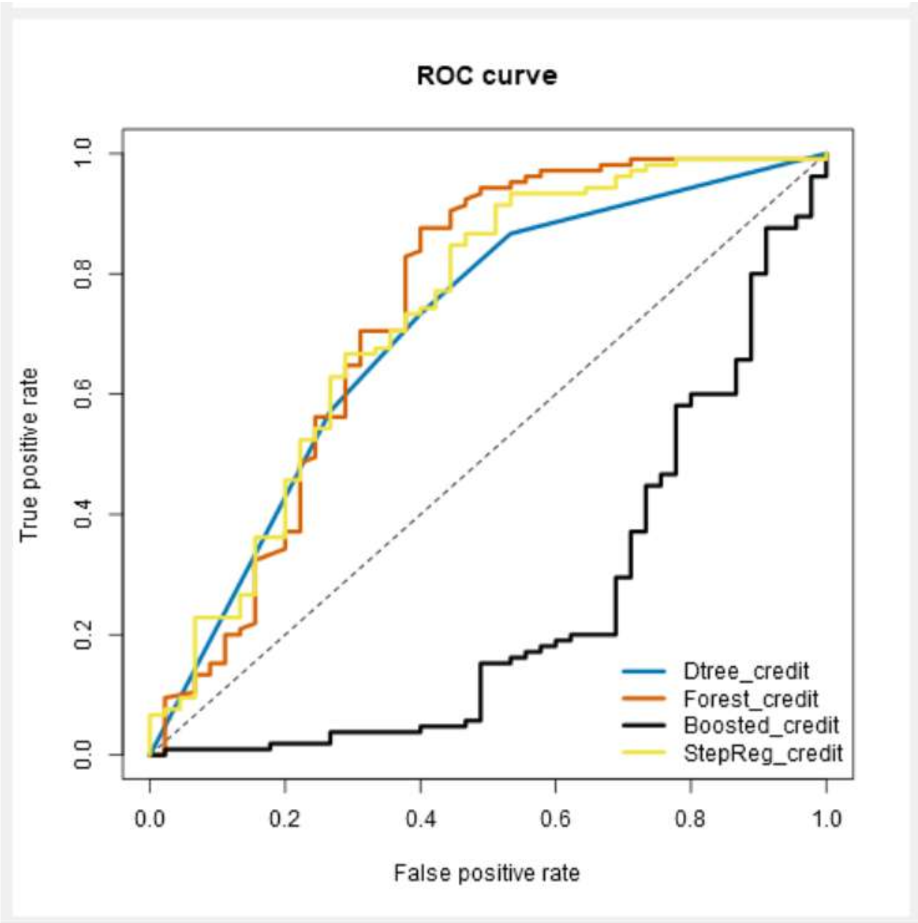
	Actual_Creditworthy	Actual_Non-Creditworthy
Predicted_Creditworthy	102	26
Predicted_Non-Creditworthy	3	19

Confusion matrix of StepReg_credit

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

The ROC curve shows that the forest model reaches the true positive rate the fastest and is therefore the preferable model according to this measure.

The low bias of the model is very important as we must avoid lending money to applicants with a high probability of defaulting while ensuring opportunities are not overlooked by not lending money to creditworthy applicants.



Summing up all the above, the forest model should be our choice. After scoring the model with the dataset of new loan applicants 408 creditworthy applicants and 92 non-creditworthy applicants were predicted.