Project: Creditworthiness Classification

Business Needs

- The decision that needs to be made is: amongst the applications received, which clients are the most likely to be creditworthy for the bank loans.
- The data that will be needed: account-balance, duration of credit amount, payment status of previous credit, purpose of loan, credit amount, value savings, length of current employment, instalment percent, most valuable available asset, type of apartments, number of credits and age.
- The model that should be used for this project is based on the outcome that the business decision needs: are new applicants creditworthy or not creditworthy. Since there are only 2 categorical outcomes, the model that should be used to train the data will be binary models (Logistic regression and Decision Trees). However, since decision trees have an overfitting tendency, we will also run train the models using Forest Model and Boosted Model to check for accuracy of each algorithms.

Building the Training Set

- For numerical data fields, duration of credit and credit amount has correlation of 0.57
- There is a field with missing data that was removed and that was Duration of current address.
- I imputed the age of the applicants in the training set with its median value because only field was missing and it will not affect the overall pattern of the age column.
- There were fields with low variability and they are telephone, foreign-works, guarantors and concurrent-credits.
- I also removed occupation and no of dependents as occupation was only filled with one value and the no of dependents were very high at 2.

Training of classification models

Linear Regression:

Coefficients:

	Estimate	Std. Error	z value	Pr(>|z|)
(Intercept)	-3.0136120	1.013e+00	-2.8700	0.00292 **
Account.Balancesome Balance	-1.5433899	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	0.4054308	3.841e-01	1.0554	0.29124
Payment.Status.of.Previous.CreditSome Problems	1.2807175	5.335e-01	2.3632	0.01812 *
PurposeNew car	-1.7941034	6.276e-01	-2.7951	0.00519 **
PurposeOther	-0.3191177	8.341e-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	0.8074082	5.109e-01	1.1911	0.23391
Value.Savings.Stocks£100-£1000	0.1694413	5.649e-01	0.3000	0.7642
Length.of.current.employment4-7 yrs	0.5224158	4.930e-01	1.0598	0.28934
Length.of.current.employment< 1yr	0.7779402	3.959e-01	1.9644	0.04925 *
Instalment.per.cent	0.3109833	1.399e-01	2.2232	0.0262 *
Most.valuable.available.asset	0.3258705	1.556e-01	2.0945	0.03621 *
Type.of.apartment	-0.2803038	2.959e-01	-0.8805	0.3788
No.of.Credits.at.this.BankMore than 1	0.3819549	7.815e-01	0.9487	0.34275
Age.years	-0.0141206	1.335e-02	-0.9202	0.35747

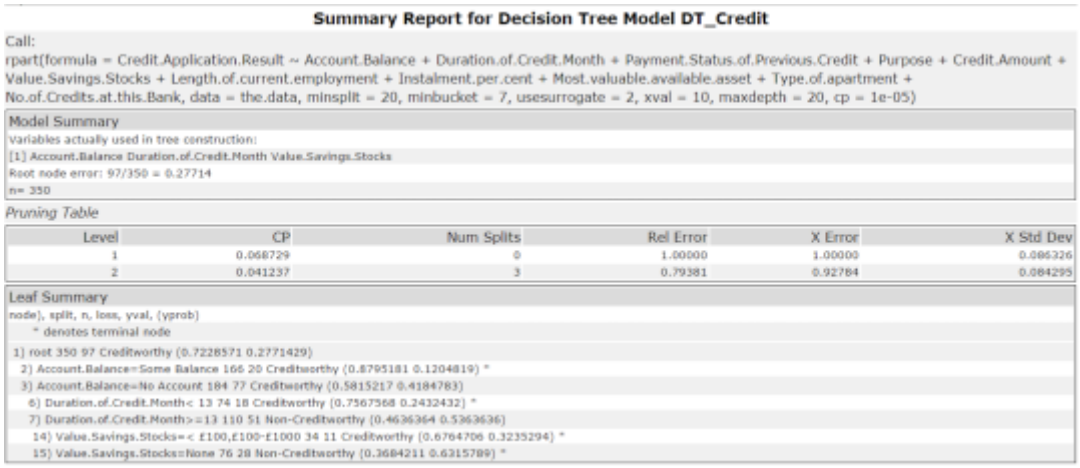
Significance codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(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, AIC: 358.3

Account balance, purpose, value savings, instalment and most valuable assets are significant

Decision Tree:



Summary Report for Decision Tree Model DT_Credit

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 + Type.of.apartment + No.of.Credits.at.this.Bank, data = the.data, minsplit = 20, minbucket = 7, usesurrogate = 2, xval = 10, maxdepth = 20, cp = 1e-05)

Model Summary

Variables actually used in tree construction:
[1] Account.Balance Duration.of.Credit.Month Value.Savings.Stocks
Root node error: 97/350 = 0.27714
n= 350

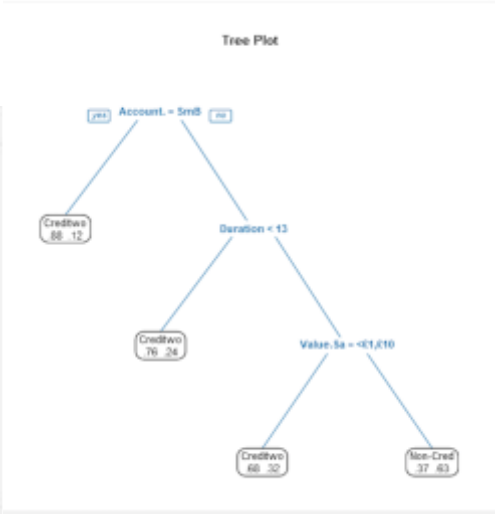
Pruning Table

Level	CP	Num Splits	Rel Error	X Error	X Std Dev
1	0.068729	0	1.00000	1.00000	0.086326
2	0.041237	3	0.79381	0.92784	0.084295

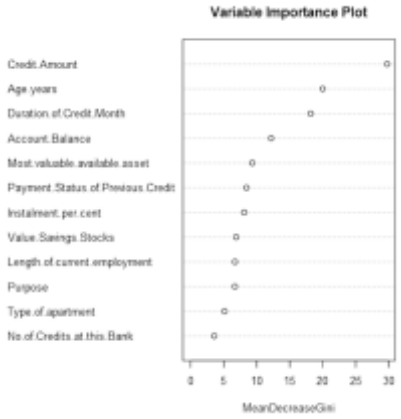
Leaf Summary

node), split, n, loss, yval, (yprob)
* denotes terminal node
1) root 350 97 Creditworthy (0.7228571 0.2771429)
2) Account.Balance=Some Balance 166 20 Creditworthy (0.8795181 0.1204819) *
3) Account.Balance=No Account 184 77 Creditworthy (0.5815217 0.4184783)
6) Duration.of.Credit.Month< 13 74 18 Creditworthy (0.7567568 0.2432432) *
7) Duration.of.Credit.Month>=13 110 51 Non-Creditworthy (0.4636364 0.5363636)
14) Value.Savings.Stocks=< £100,£100-£1000 34 11 Creditworthy (0.6764706 0.3235294) *
15) Value.Savings.Stocks=None 76 28 Non-Creditworthy (0.3684211 0.6315789) *

The variables used were account balance, duration of credit and values saving stocks with the following probability

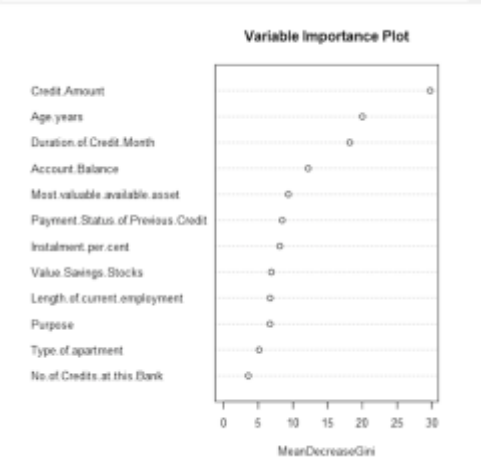


Forest Model:



The significant values were credit amount, account balance and duration of credit month.

Boosted Model:



The significant variables ere credit amount and account balance.

Overall Accuracy

These are the overall accuracy of the models:

Model Comparison Report

Fit and error measures

Model	Accuracy	F1	AUC	Accuracy_Creditworthy	Accuracy_Non-Creditworthy
Log_credit	0.7800	0.8520	0.7314	0.8051	0.6875
DT_Credit	0.7467	0.8273	0.7054	0.7913	0.6000
Forest_Credit	0.8000	0.8707	0.7419	0.7953	0.8261
Boosted_Credit	0.7933	0.8670	0.7509	0.7891	0.8182

From the model comparison, forest trees model offers higher overall accuracy at 0.80, accuracy for creditworthy is at 0.7953 and non creditworthy at 0.8261

Confusion Matrix Comparison

Confusion matrix of Boosted_Credit

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

Confusion matrix of DT_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	99	25
Predicted_Non-Creditworthy	6	20

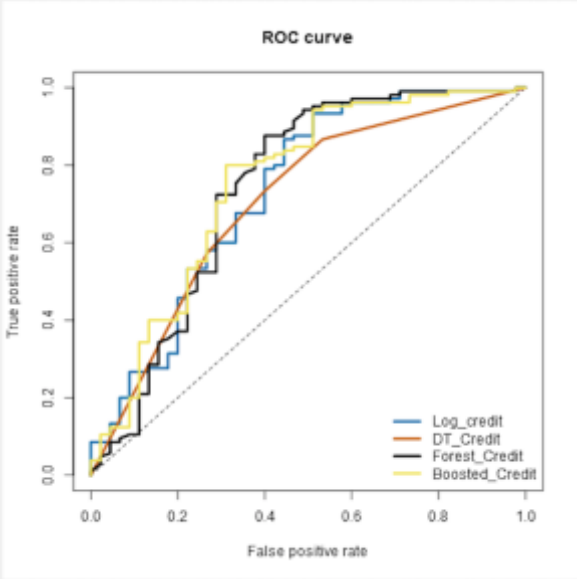
Confusion matrix of Log_credit

	Actual_Creditworthy	Actual_Non-Creditworthy
Predicted_Creditworthy	93	23
Predicted_Non-Creditworthy	12	22

Performance Diagnostic Plots

From the confusion matrix above, there seems to be a bias towards predicting for creditworthy. This may happen because most of the data given to train to the model were majorly creditworthy data.

ROC:



According the ROC curve, all 4 models have areas above 0.50. This shows that all 4 models are good models that can be used for this classification. But the Forest_Credit model has the highest curve amongst all 4. This shows that for every false positive rate, it will give a better number of true positive rate.

From this matrix comparison, accuracy measures and ROC curve, forest model was chosen to classify the new applications for their creditworthiness.

For the bank's solution, I used the formula tool to create a new column called "Potential Client".

Those with score_creditworthy > 0.50 were to be considered as potential clients. As a result of this, the bank would get a potential 415 clients out of the 500 applications that can be further assessed.