Creditworthiness of Loan Applicants.

1 - Business and Data Understanding.

What business decisions need to be made?

Due to a financial scandal that hit a competitive bank last week, there has been a sudden influx of new people applying for loans at the bank. All of a sudden there are nearly 500 loan applications to process this week as opposed to typical 200 loan applications per week which are approved by hand.

This new influx is a great opportunity and the bank wants to figure out how to systematically evaluate the creditworthiness of these new loan applicants.

What data is needed to inform these decisions?

The data needed will come from “credit-data-training.xlsx”. The data has already been cleaned, however it will still need to be checked for missing data and later used to train four different models. The models will be compared to find the most suitable.

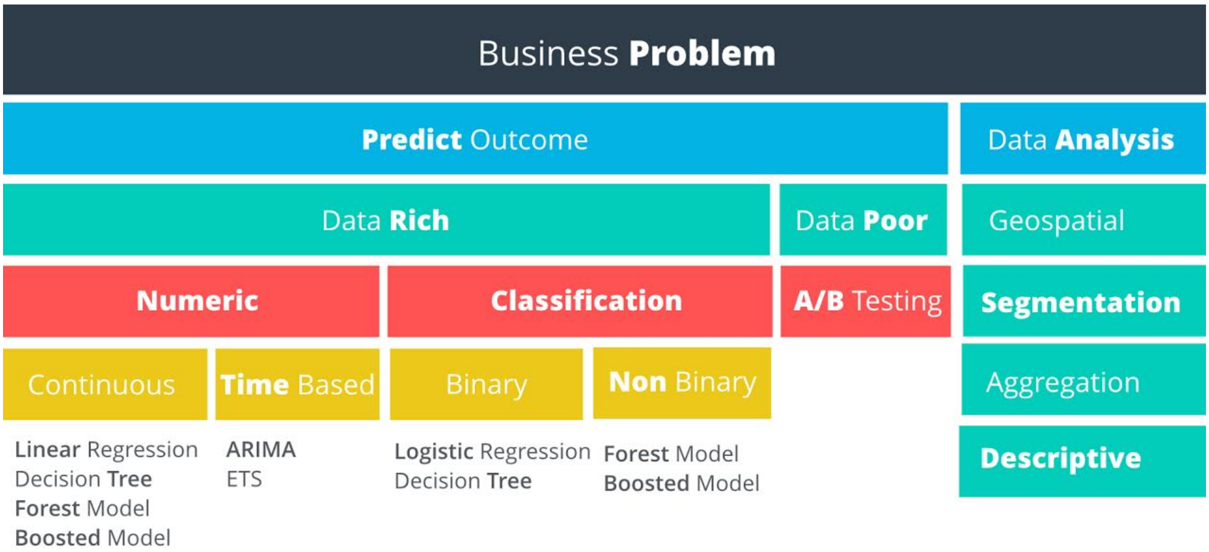
The chosen model will be used to predict the loan applicants worthy of a loan from “customers-to-score.xlsx”.

The columns used from “credit-data-training.xlsx” are:

Columns Names
Credit-Application-Result
Account-Balance
Duration-of-Credit-Month
Payment-Status-of-Previous-Credit
Purpose
Credit-Amount
Value-Savings-Stocks
Length-of-current-employment
Most-valuable-available-asset
No-of-Credits-at-this-Bank
Type-of-Apartment
Instalment-per-cent
Age-years

What kind of model do we need to use to help make these decisions?

Using the methodology map below to aid my decision making:



I can see that the business problem requires me to:

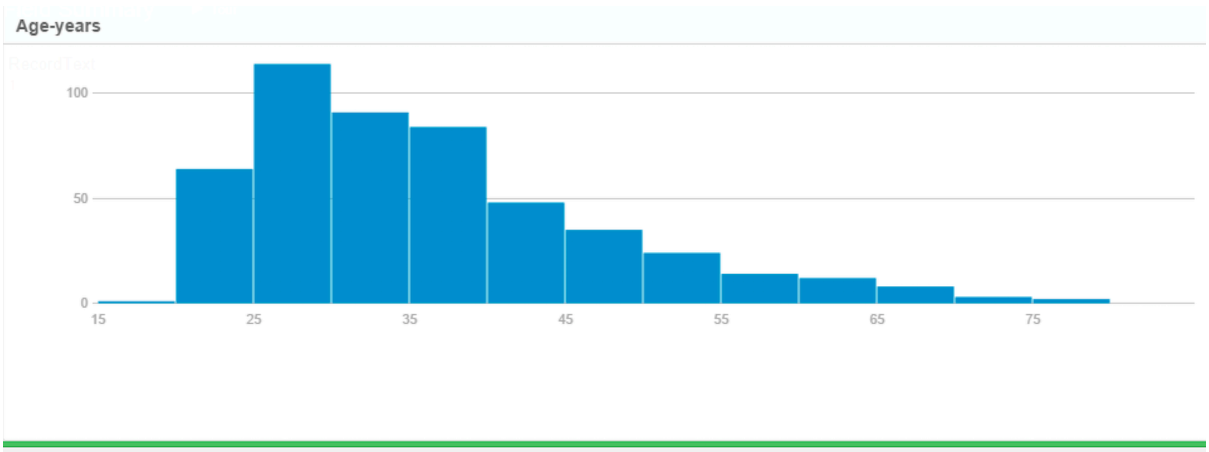
a) Predict an Outcome
b) Use Rich Data
c) Classify the available data
d) Obtain a binary outcome ie, give the applicant a loan or not.

The model used will likely be a binary classification model.

2 - Building the training set.

The data used to train the model will come from "credit-data-training.xlsx". Predictor variables will need to be chosen based on their relationship with the target variable which is whether the applicant will be creditworthy.

We will be looking at variability of data, an example of which is the Age-years variable which is demonstrated in the below chart.



Correlation of the data is another factor used to help find suitable predictor variables. The full correlation matrix is summarised further in the document.

For numerical data fields, are there any fields that are highly correlated?

Below is a summary of the data.

Name	Field Category	Min	Max	Median	Std. Dev.	Percent Missing	Unique Values	Mean	Shortest_Value	Longest_Value	MinValueCount	MaxValueCount
Age-years	Numeric	19	75	33	11.50152219	2.4	54	35.63729508				
Credit-Amount	Numeric	276	18424	2236.5	2831.386861	0	464	3199.98				
Duration-in-Current-address	Numeric	1	4	2	1.150017082	68.8	5	2.66025641				
Duration-of-Credit-Month	Numeric	4	60	18	12.30742009	0	30	21.434				
Foreign-Worker	Numeric	1	2	1	0.191387718	0	2	1.038				
Instalment-per-cent	Numeric	1	4	3	1.113723826	0	4	3.01				
Most-valuable-available-asset	Numeric	1	4	3	1.064267509	0	4	2.36				
No-of-dependents	Numeric	1	2	1	0.353459853	0	2	1.146				
Occupation	Numeric	1	1	1	0	0	1	1				
Telephone	Numeric	1	2	1	0.490388583	0	2	1.4				
Type-of-apartment	Numeric	1	3	2	0.539813669	0	3	1.928				
Account-Balance	String					0	2		No Account	Some Balance	238	262
Concurrent-Credits	String					0	1		Other Banks/Depts	Other Banks/Depts	500	500
Credit-Application-Result	String					0	2		Creditworthy	Non-Creditworthy	142	358
Guarantors	String					0	2		Yes	None	43	457
Length-of-current-employment	String					0	3		< 1yr	1-4 yrs	97	279
No-of-Credits-at-this-Bank	String					0	2		1	More than 1	180	320
Payment-Status-of-Previous-Credit	String					0	3		Paid Up	No Problems (in this bank)	36	260
Purpose	String					0	4		Other	Home Related	15	355
Value-Savings-Stocks	String					0	3		None	£100-£1000	48	298

Below is the correlation matrix of all the variables with Credit_Applicant_Result as the target variable.

Full Correlation Matrix

	Credit.Applicati	Duration.of.	Credit.	Instalmen	Duration.in.C	Most.valuable.
Credit.Applicati	1.0000000	-0.1900741	-0.07921	-0.1165998	0.0792585	-0.0525198
Duration.of.Cre	-0.1900741	1.0000000	0.59061	0.1040048	-0.0506493	0.1195555
Credit.Amount	-0.0792182	0.5906171	1.00000	-0.2653537	-0.1580690	0.3012233
Instalment.per	-0.1165998	0.1040048	-0.26535	1.0000000	0.1733930	0.1341344
Duration.in.Cur	0.0792585	-0.0506493	-0.15806	0.1733930	1.0000000	0.1092968
Most.valuable.	-0.0525198	0.1195555	0.30122	0.1341344	0.1092968	1.0000000
Type.of.apartm	-0.0423327	0.1201070	0.10696	0.1369001	-0.1575495	0.0938777
No.of.depende	0.0294867	-0.1959091	0.06386	-0.3127847	-0.0566456	-0.0479319
Telephone	0.0322363	0.2103393	0.17151	0.0526591	0.0849249	0.1788326
Foreign.Worker	0.0714765	-0.2184723	-0.05635	-0.1898275	-0.0365874	-0.0013900
Age_years	0.1205908	-0.0172588	0.03854	0.1072625	0.2866444	0.0638176

	Type.of.apartm	No.of.depen	Teleph	Foreign.W	Age_years
Credit.Applicati	-0.0423327	0.0294867	0.03223	0.0714765	0.1205908
Duration.of.Cre	0.1201070	-0.1959091	0.21033	-0.2184723	-0.0172588
Credit.Amount	0.1069607	0.0638629	0.17151	-0.0563574	0.0385492
Instalment.per	0.1369001	-0.3127847	0.05265	-0.1898275	0.1072625
Duration.in.Cur	-0.1575495	-0.0566456	0.08492	-0.0365874	0.2866444
Most.valuable.	0.0938777	-0.0479319	0.17883	-0.0013900	0.0638176
Type.of.apartm	1.0000000	0.0039290	0.19053	-0.0087732	0.1919314
No.of.depende	0.0039290	1.0000000	-0.10550	0.2699279	0.0461411
Telephone	0.1905344	-0.1055013	1.00000	-0.1718538	0.1350691
Foreign.Worker	-0.0087732	0.2699279	-0.17185	1.0000000	-0.0200493
Age_years	0.1919314	0.0461411	0.13506	-0.0200493	1.0000000

Looking through the correlation matrix and using 0.7 as the benchmark for high correlation, there seems to be nothing of high correlation with the numerical data fields.

Are there any missing data fields?

Looking at the summary, there are missing fields in Age-years and Duration-in-apartment.

There are too many missing fields in Duration-in-apartment and therefore this field will be excluded from the analysis.

Age-in-years is missing only 2.4% of its data and in turn I will substitute missing data here with the average age of the dataset.

Are there any fields with low variability?

Below are columns that potentially show low variability due to the majority of its data being one sided:

Foreign-worker
Guarantors
Concurrent-Credits
Telephone
Occupation
No-of-dependents

Data with low variability will be excluded from the model.

The below is my final dataset for modelling.

Columns Names
Credit-Application-Result
Account-Balance
Duration-of-Credit-Month
Payment-Status-of-Previous-Credit
Purpose
Credit-Amount
Value-Savings-Stocks
Length-of-current-employment
Most-valuable-available-asset
No-of-Credits-at-this-Bank
Type-of-Apartment
Instalment-per-cent
Age-years

A sanity check of the 13 final columns and gives an **average age of 35.637 or 36 yrs** (rounded up to the nearest year)

Creating the model.

In order to create the models, a 70/30 split was done to create an estimation and validation dataset.

The models were run and each of the model summaries are below.

Report for Logistic Regression Model LM_Result

Basic Summary

Call:

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

Deviance Residuals:

Min	1Q	Median	3Q	Max
-2.084	-0.719	-0.429	0.691	2.543

Coefficients:

	Estimate	Std. Error	z value	Pr(>|z|)
(Intercept)	-2.990817	1.013e+00	-2.9527	0.00315**
Payment.Status.of.Previous.CreditPaid Up	0.402974	3.843e-01	1.0487	0.2943
Payment.Status.of.Previous.CreditSome Problems	1.259683	5.334e-01	2.3616	0.0182*
PurposeNew car	-1.755074	6.278e-01	-2.7954	0.00518**
PurposeOther	-0.290165	8.359e-01	-0.3471	0.72848
PurposeUsed car	-0.785627	4.124e-01	-1.9049	0.05679.
Type.of.apartment	-0.254565	2.958e-01	-0.8605	0.38949
Value.Savings.StocksNone	0.609298	5.099e-01	1.1949	0.23213
Value.Savings.Stocks£100-£1000	0.172241	5.649e-01	0.3049	0.76046
No.of.Credits.at.this.BankMore than 1	0.362688	3.816e-01	0.9505	0.34184
Credit.Amount	0.000177	6.841e-05	2.5879	0.00966**
Account.BalanceSome Balance	-1.543669	3.233e-01	-4.7745	1.80e-06***
Age.years	-0.015092	1.539e-02	-0.9809	0.32666
Length.of.current.employment4-7 yrs	0.530959	4.932e-01	1.0767	0.28163
Length.of.current.employment< 1yr	0.777372	3.957e-01	1.9646	0.04946*
Most.valuable.available.asset	0.325606	1.557e-01	2.0918	0.03645*
Duration.of.Credit.Month	0.006391	1.371e-02	0.4660	0.6412
Instalment.per.cent	0.310524	1.399e-01	2.2197	0.02644*

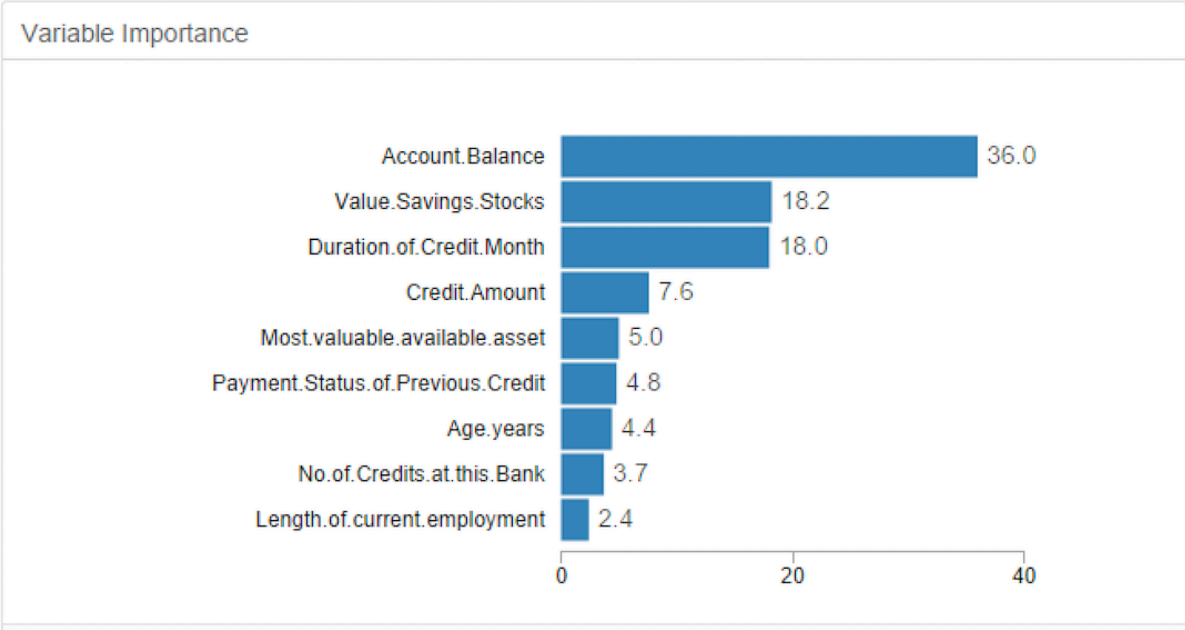
Significance codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial taken to be 1)

For the Logistic Model, the most significant predictor variables are:

Variable Name	p-value
Payment.Status.of.Previous.CreditSome.Problems	0.0182
PurposeNew car	0.00518
	0.00966
Account.BalanmceSome balance	1.80e-06
Length.of.current.em,plyment<1yr	0.04946
Most.valuable.available.asset	0.03645
Instalment.per.cent	0.02644

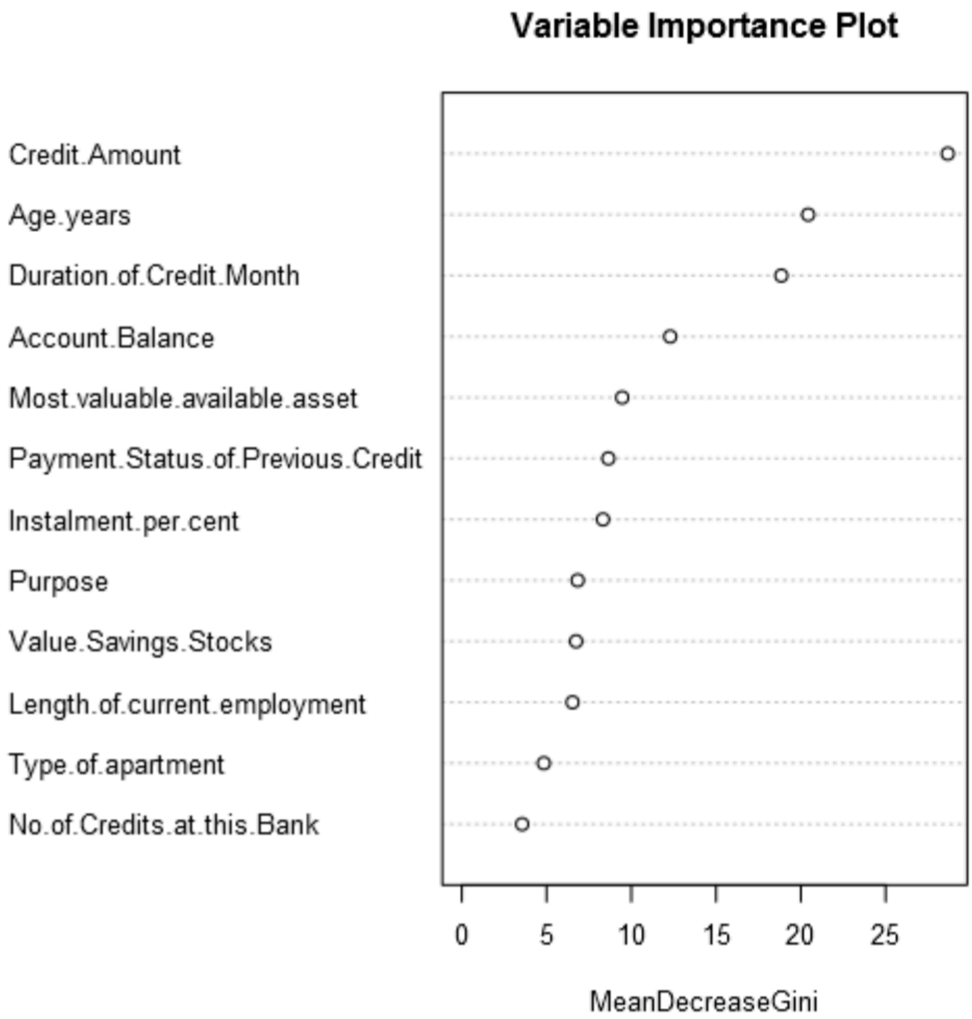
Variable Importance Chart for the decision tree model.



The most important variables in the decision tree model are:

Account Balance
Value.Savings.Stocks
Duration.of.Credit.Month

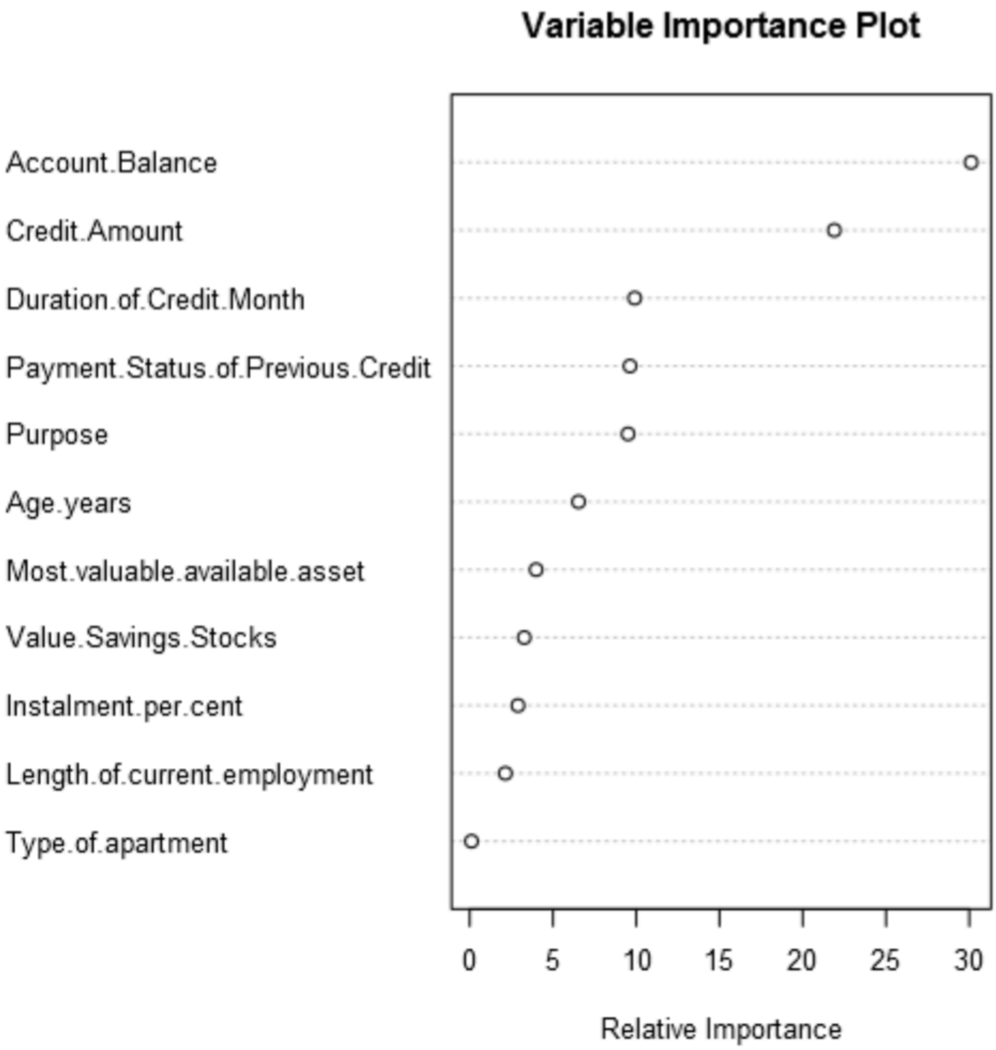
Below is the variable importance chart for the random forest model.



The chart indicates the most important predictor variables for the random forest model are:

Credit.Amount
Age.years
Duration.of.Credit.Month

Below is the variable importance plot of the boosted model.



The most important variables for the boosted model are:

Credit.Amout
Account.Balance

Below is the accuracy and confusion matrix to each model having been validated through the validation dataset.

Model Comparison Report

Fit and error measures

Model	Accuracy	F1	AUC	Accuracy_Creditworthy	Accuracy_Non-Creditworthy
LM_Result	0.7800	0.8520	0.7310	0.8051	0.6875
DT_Results	0.7467	0.8273	0.7054	0.7913	0.6000
RM_Results	0.8133	0.8793	0.7380	0.8031	0.8696
Bosted_Results	0.7867	0.8621	0.7526	0.7874	0.7826

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 Bosted_Results

	Actual_Creditworthy	Actual_Non-Creditworthy
Predicted_Creditworthy	100	27
Predicted_Non-Creditworthy	5	18

Confusion matrix of DT_Results

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

Confusion matrix of LM_Result

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

Confusion matrix of RM_Results

	Actual_Creditworthy	Actual_Non-Creditworthy
Predicted_Creditworthy	102	25
Predicted_Non-Creditworthy	3	20

Performance Diagnostic Plots

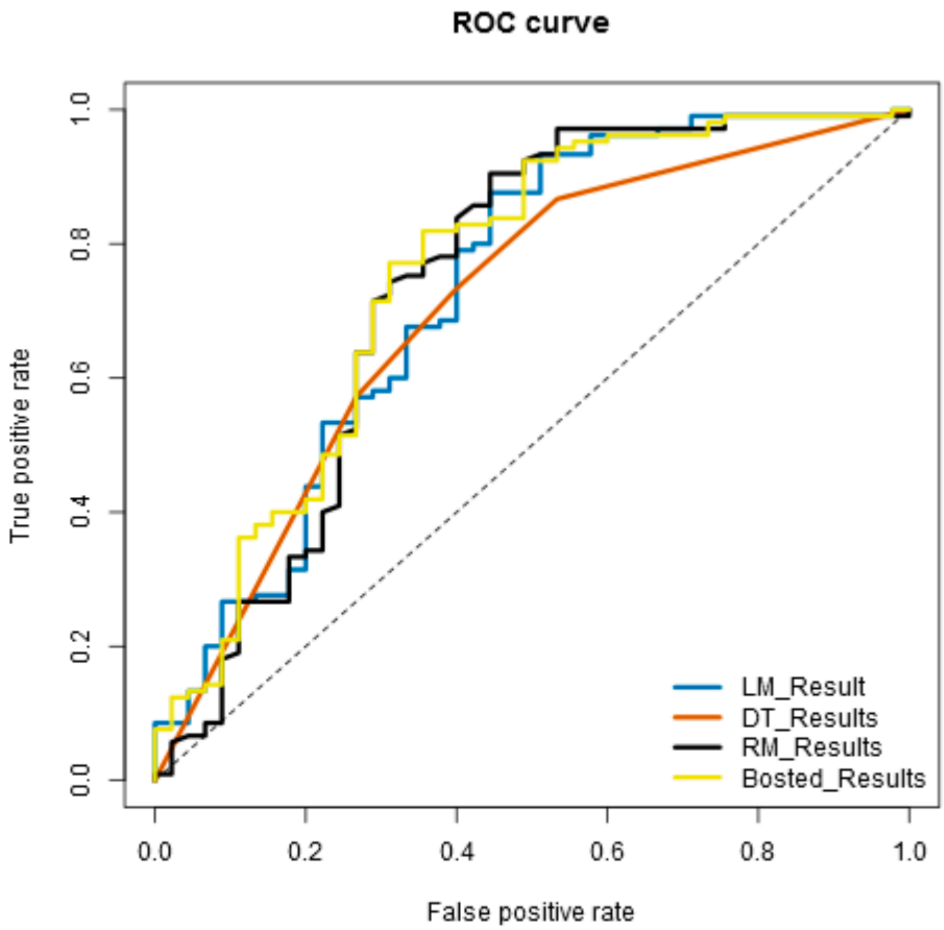
The model with the highest accuracy score is the Random Forest Model at 0.8133.

The models appear to predict Creditworthy more accurately than Non-Creditworthy. It also looks like there are more applicants that are creditworthy and not.

Final Model.

The final model used for prediction will be the Random Forest model due to its highest overall accuracy at 0.8133. It has a high accuracy, 0.803, score for predicting Creditworthy applicants and also the highest accuracy score, 0.8696, for predicting non-creditworthy applicants

Below is the ROC chart for the models.



The ROC plots shows the Random Forest model to be the second best with an AUC of 0.7380.

Applying the model to the new dataset, customers-to-score.xls and taking any applicant that has a greater Creditworthy accuracy score than non-creditworthy to mean the applicant should be granted a loan, the final count of **individuals whom are creditworthy are 411**.