Project 3: Creditworthiness

Step 1: Business and Data Understanding

Key Decisions:

- What decisions needs to be made?

  Determine which loan applicants of the new list of 500 are creditworthy and deserve to be given a loan.

- What data is needed to inform those decisions?

  Data from the bank’s past loan applicants. This includes many factors such as credit score, purpose of loan, age, and several others.

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

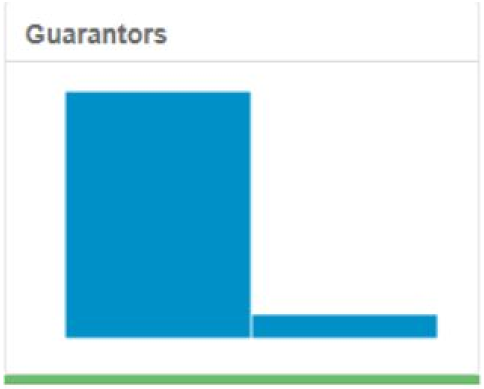
  We will need to use a Binary classification model to help make this decision as applicants will be classified as either Creditworthy or Non_Creditworthy.

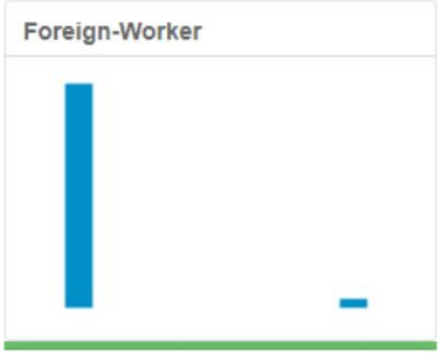
Step 2: Building the Training Set

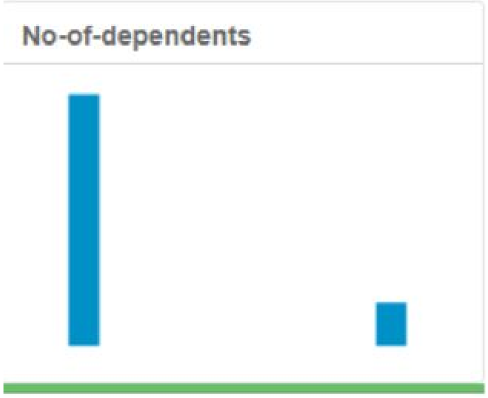
- In your cleanup process, which fields did you remove or impute? Please justify why you removed or imputed these fields. Visualizations are encouraged.

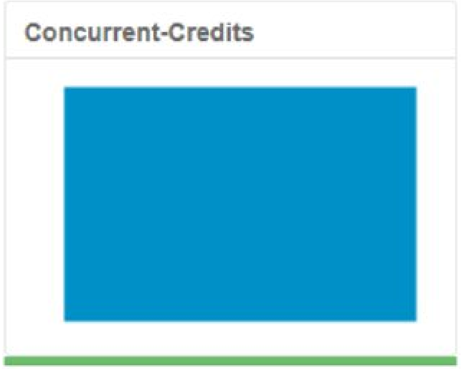
  During the cleanup process, I decided to remove 7 fields: Guarantors, Duration in Current Address, Concurrent Credits, Occupation, Number of Dependents, Telephone and Foreign Worker.

  The fields **Guarantors**, **Foreign Worker** and **Number of Dependents** all showed low variability with over 80% of the data being the same value. Additionally, the fields **Concurrent Credits** and **Occupation** only returned one value for all rows.

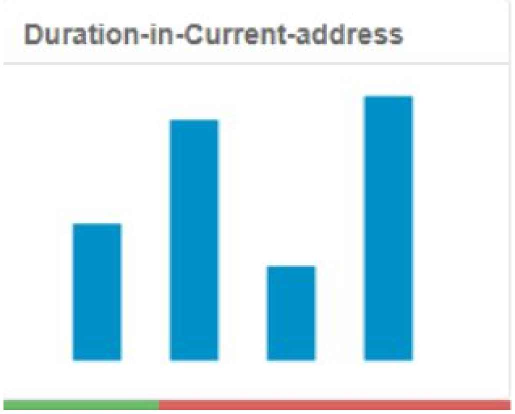








Telephone was removed due to the field being irrelevant to the necessary data while **Duration in Current Address** was removed as 69% of data in the field was missing. It is important to note that despite **Age** missing 2% of data, the missing values were imputed using the median age across the entire dataset.



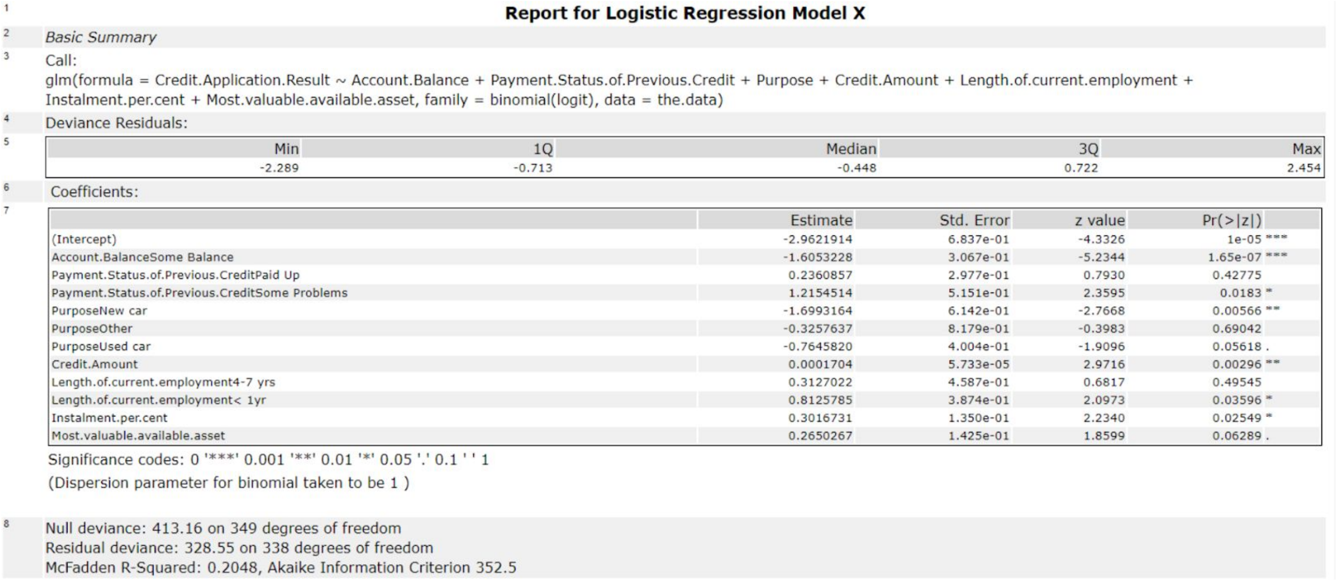
Step 3: Train your Classification Models

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

- 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?

Linear Regression

The predictor variables that showed the most significance in the Linear Regression were Account Balance, Purpose and Credit Amount.



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

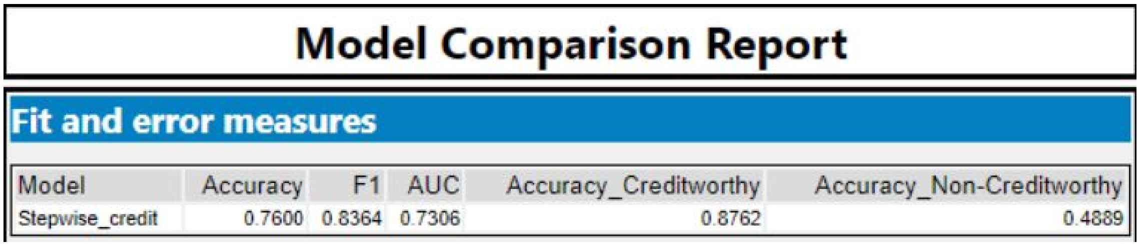
Coefficients:

	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, Akaike Information Criterion 352.5

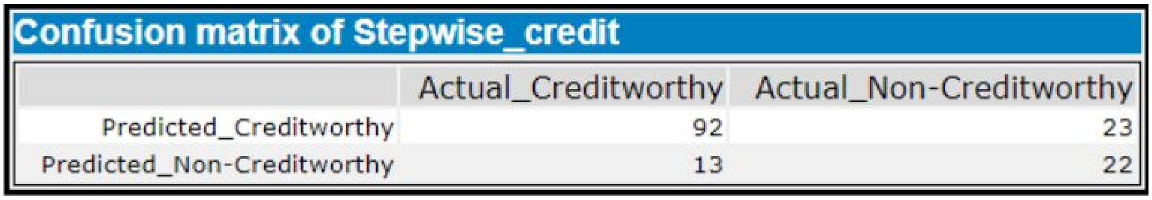
The overall accuracy of the linear regression was 76% and showed bias toward predicting customers to be creditworthy.



Model Comparison Report

Fit and error measures

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

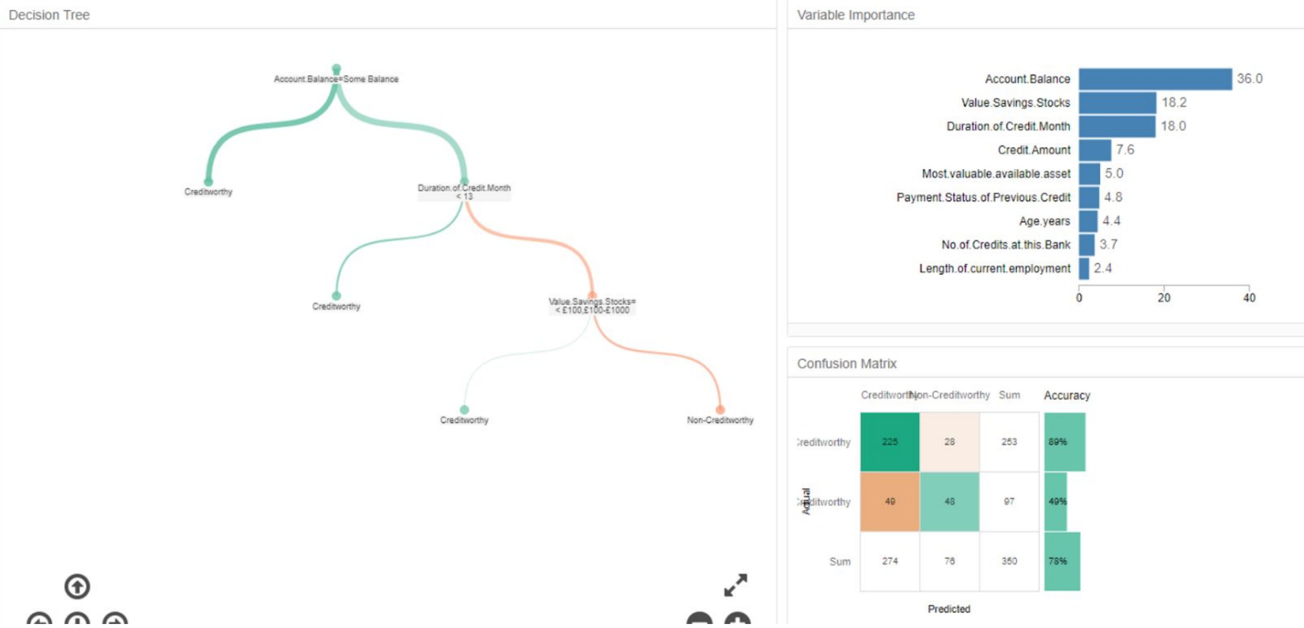


Confusion matrix of Stepwise_credit

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

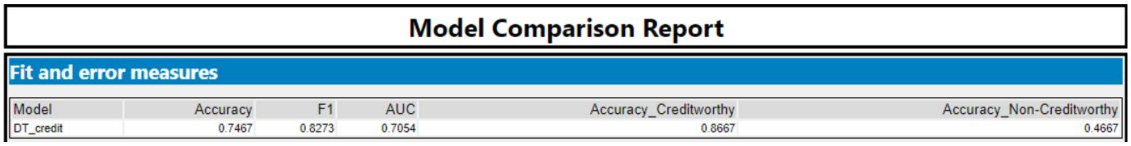
Decision Tree

The variables that showed the most significance in the Decision Tree were Account Balance, Value of Saving Stocks and Duration of Credit Month.



	Creditworthy	Non-Creditworthy	Sum	Accuracy
:reditworthy	225	28	253	89%
:editworthy	49	48	97	49%
Sum	274	76	350	78%

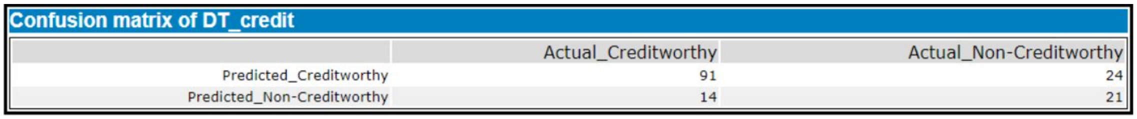
The overall accuracy of the Decision Tree was nearly 75% and showed bias toward predicting customers to be creditworthy.



Model Comparison Report

Fit and error measures

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

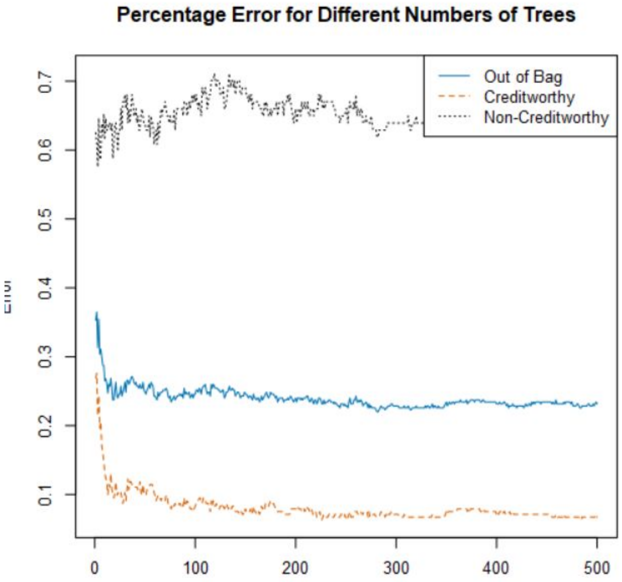


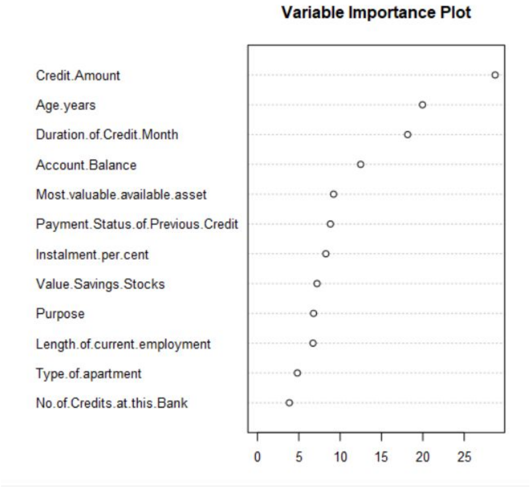
Confusion matrix of DT_credit

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

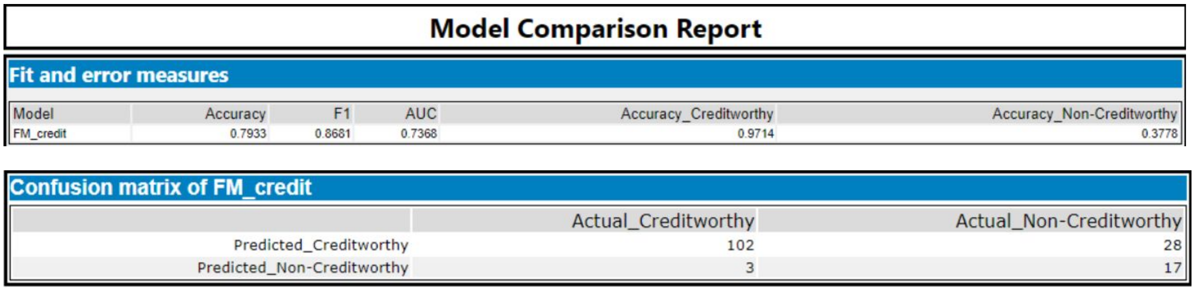
Forest Model

The variables that showed the most significance in the Forest Model were Credit Amount, Age, Duration of Credit Month and Account Balance.





The overall accuracy of the Forest Model clocked in at 79% and showed bias toward predicting customers as being creditworthy.



Model Comparison Report

Fit and error measures

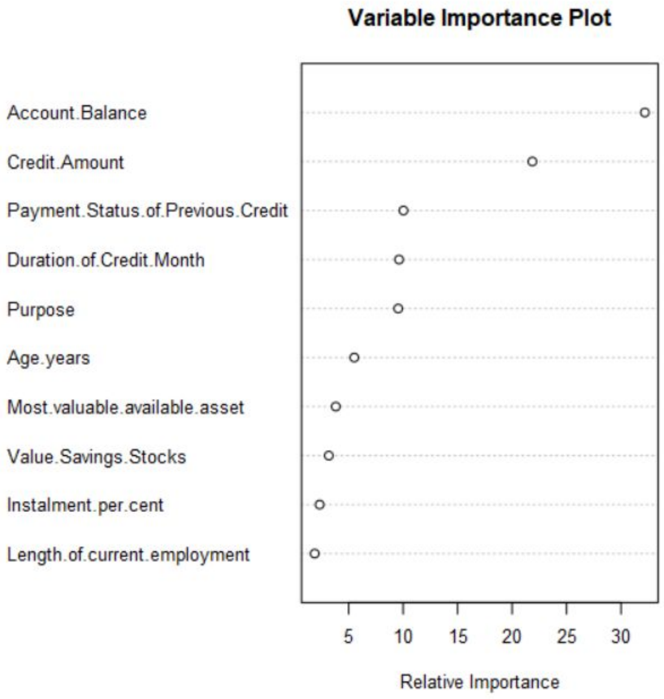
Model	Accuracy	F1	AUC	Accuracy_Creditworthy	Accuracy_Non-Creditworthy
FM_credit	0.7933	0.8681	0.7368	0.9714	0.3778

Confusion matrix of FM_credit

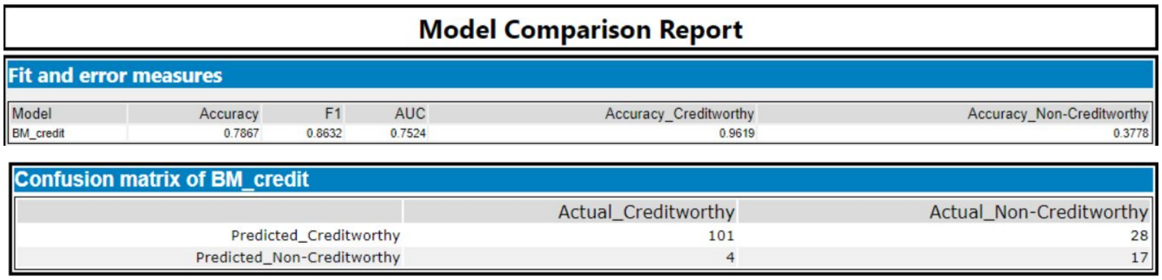
	Actual_Creditworthy	Actual_Non-Creditworthy
Predicted_Creditworthy	102	28
Predicted_Non-Creditworthy	3	17

Boosted Model

The predictor variables that showed the most significance in the Boosted Model were Account Balance, Credit Amount and Payment Status of Previous Credit.



The overall accuracy of the Boosted Model was almost 79% and showed bias toward predicting customers as being creditworthy.



Model Comparison Report

Fit and error measures

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

Confusion matrix of BM_credit

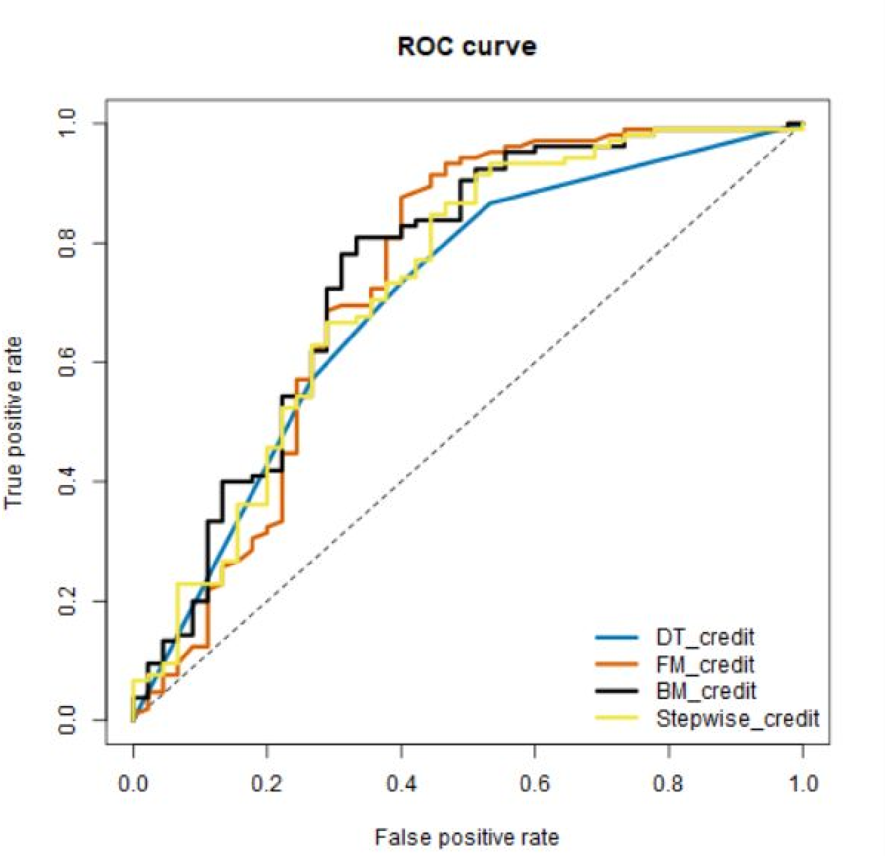
	Actual_Creditworthy	Actual_Non-Creditworthy
Predicted_Creditworthy	101	28
Predicted_Non-Creditworthy	4	17

Step 4: Writeup

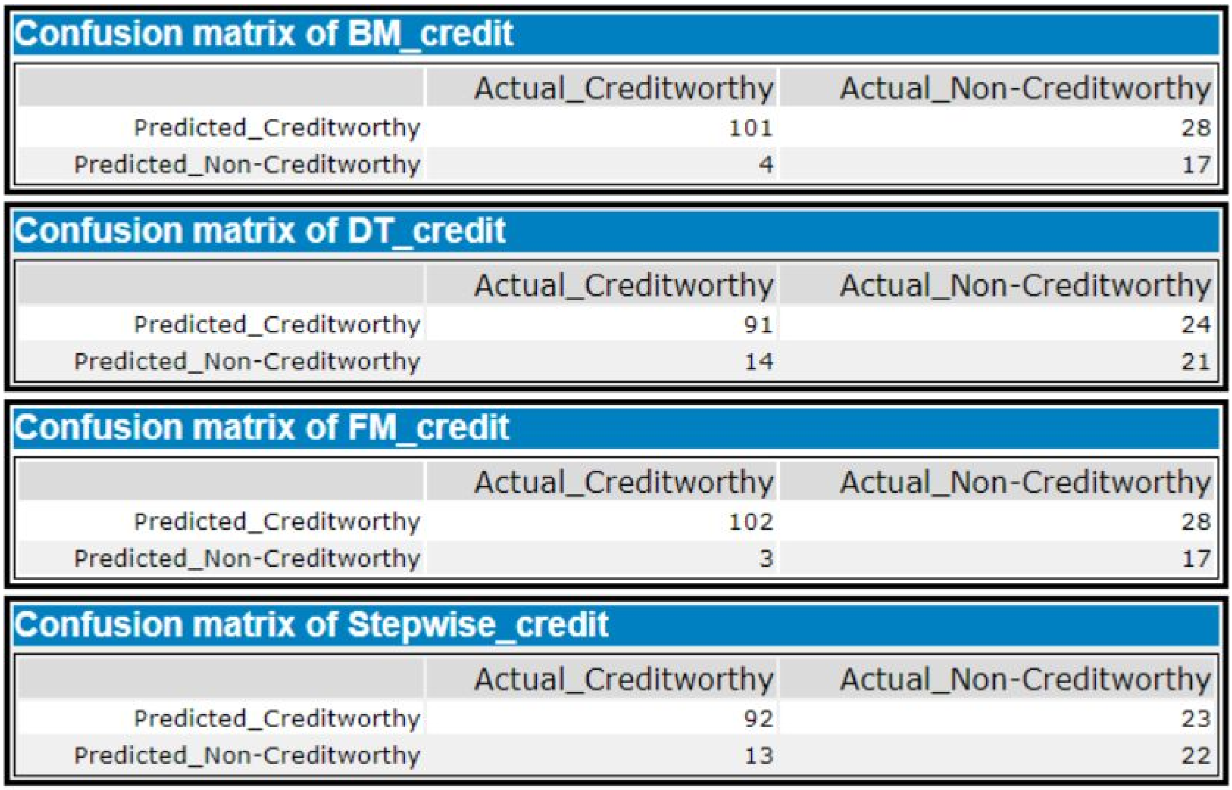
- Which model did you choose to use? Please justify your decision using **all** of the following techniques. Please only use these techniques to justify your decision:

I chose to use the Forest Model as it had the highest accuracy rate at 79.33%. Although the model had low accuracy in determining Non-Creditworthy loan applicants, its incredibly high accuracy of 97% in determining Creditworthy loan applicants made this model the best choice.

Additionally, all the models showed considerable bias towards predicting customers as being creditworthy so using the model with the highest accuracy was important.



Although it is hard to tell in the ROC graph, the Forest Model does have a higher curve than the other models. Further proof can be seen in the comparison of confusion matrices as the orest Model had the most true-positives.



Confusion matrix of BM_credit

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

Confusion matrix of DT_credit

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

Confusion matrix of FM_credit

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

Confusion matrix of Stepwise_credit

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

- How many individuals are creditworthy?

408 individuals are creditworthy.