Project: Creditworthiness

The Business Problem

You work for a small bank and are responsible for determining if customers are creditworthy to give a loan to. Your team typically gets 200 loan applications per week and approves them by hand.

Due to a financial scandal that hit a competitive bank last week, you suddenly have an influx of new people applying for loans for your bank instead of the other bank in your city. All of a sudden you have nearly 500 loan applications to process this week!

Your manager sees this new influx as a great opportunity and wants you to figure out how to process all of these loan applications within one week.

Step 1: Business and Data Understanding

Key Decisions:

Answer these questions

- What decisions needs to be made?

Due to an influx of new people applying for loans there are now nearly 500 loan applications to process within a week.
It needs to be determined which of the new customers are creditworthy to be given a loan.

- What data is needed to inform those decisions?

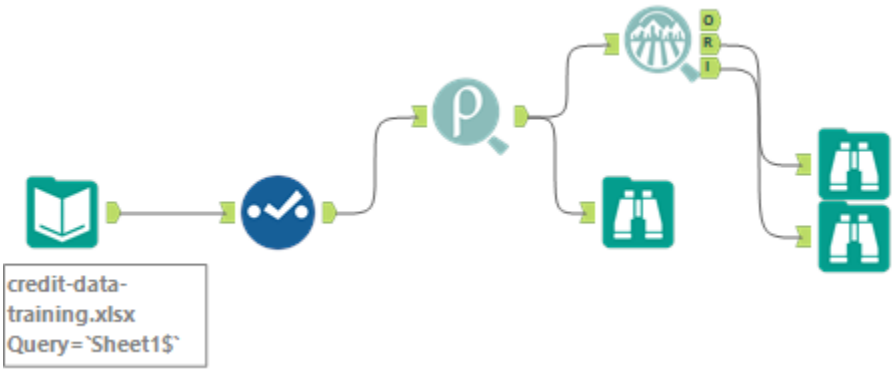
To inform these decisions data showing all credit approvals from past applicants will be used and also data provided by the new customers.

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

The outcome will be either creditworthy or not creditworthy, therefore a binary classification model will be used.

Step 2: Building the Training Set

I used the following workflow with the Pearson Correlation and the Field Summary to clean up the data and select the predictor variables to be removed.



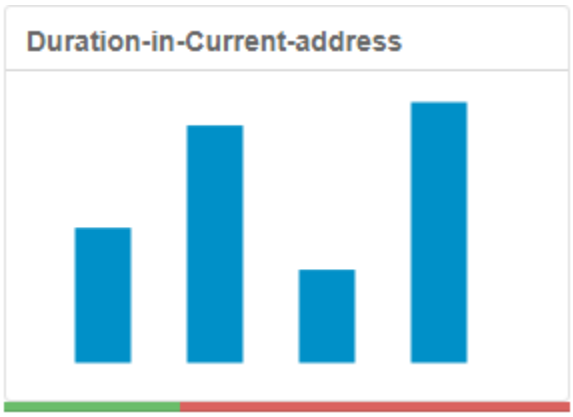
- For numerical data fields, are there any fields that highly-correlate with each other? The correlation should be at least .70 to be considered “high”.

None of the numerical data fields have a correlation of more that 0.70. There are no data fields which are highly correlated with each other.

FieldName	Duration-of-Credit-Month	Credit-Amount	Instalment-per-cent	Duration-in-Current-address	Most-valuable-available-asset	Age-years	Type-of-apartment	Occupation	No-of-dependents	Telephone	Foreign-Worker
Duration-of-Credit-Month	1	0.57	0.07		0.3		0.15		-0.07	0.14	-0.12
Credit-Amount	0.57	1	-0.29		0.33		0.17		0	0.29	0.03
Instalment-per-cent	0.07	-0.29	1		0.08		0.07		-0.13	0.03	-0.13
Duration-in-Current-address				1							
Most-valuable-available-asset	0.3	0.33	0.08		1		0.37		0.05	0.2	-0.15
Age-years						1					
Type-of-apartment	0.15	0.17	0.07		0.37		1		0.17	0.1	-0.09
Occupation								1			
No-of-dependents	-0.07	0	-0.13		0.05		0.17		1	-0.05	0.07
Telephone	0.14	0.29	0.03		0.2		0.1		-0.05	1	-0.06
Foreign-Worker	-0.12	0.03	-0.13		-0.15		-0.09		0.07	-0.06	1

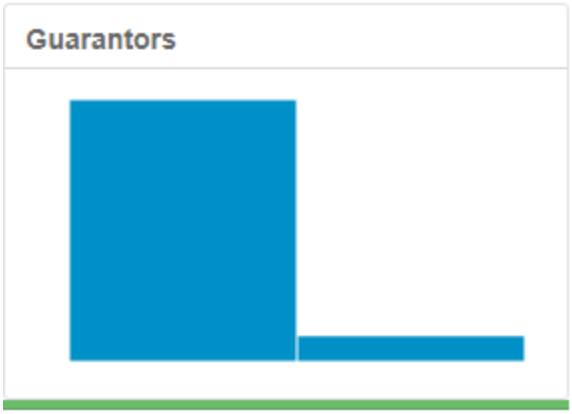
- Are there any missing data for each of the data fields? Fields with a lot of missing data should be removed

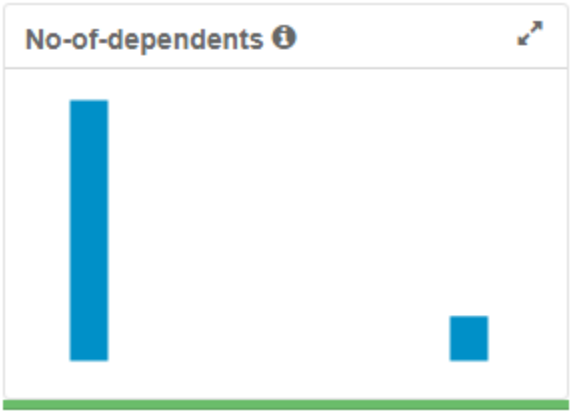
The Duration in Current Address field shows that it is missing 69% of the data, therefore this field was removed.

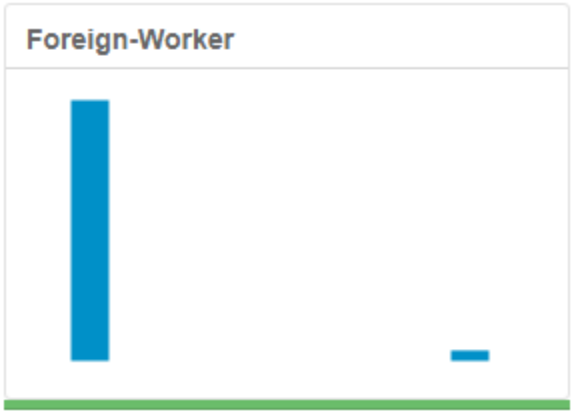


- Are there only a few values in a subset of your data field? Does the data field look very uniform (there is only one value for the entire field?). This is called “low variability” and you should remove fields that have low variability. Refer to the "Tips" section to find examples of data fields with low-variability.

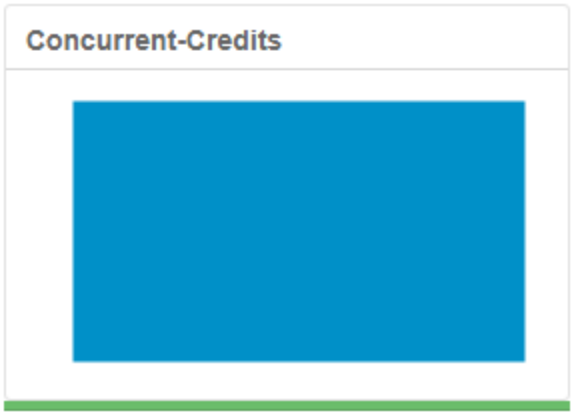
Low variability Guarantors (Yes = 43, None = 457), Number of Dependents (1 = 427, 2 = 73) and Foreign Worker (1 = 481, 2 = 19).





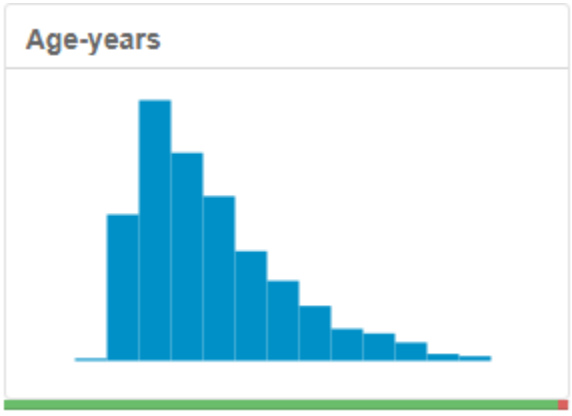


Two data fields which show only one answer are the Concurrent Credits and Occupation fields, therefore these two fields have been removed from the model.



The Telephone field has been excluded as there is no logical reason for this variable to be included.

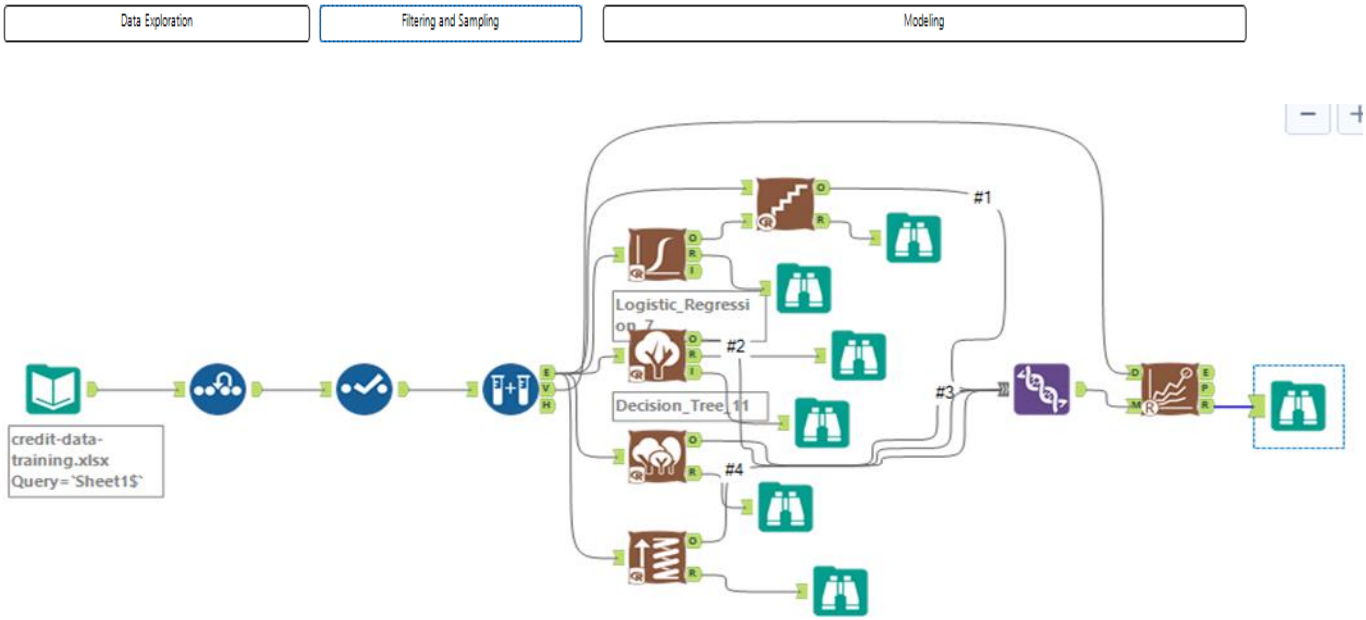
The Age-Years field showed some missing data (12 missing values). It is suggested that the median value of the field should be imputed for the sake of consistency.
Once the median values has been imputed the Age-Years column has an average of 36 (rounded up) as recommended.



Step 3: Train your Classification Models

The model is created with the Estimation and Validation samples, 70% of the dataset goes to Estimation and 30% of the entire dataset is reserved for Validation. The Random Seed is set to 1.

The following models are created: Logistic Regression, Decision Tree, Forest Model, Boosted Model



*Answer these questions for **each model** you created:*

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

Logistic Regression significant variables

- Account Balance
- Payment Status
- Purpose
- Credit Amount
- Length of Current Employment
- Instalment Percent

Report for Logistic Regression Model PDR_Stepwise

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:

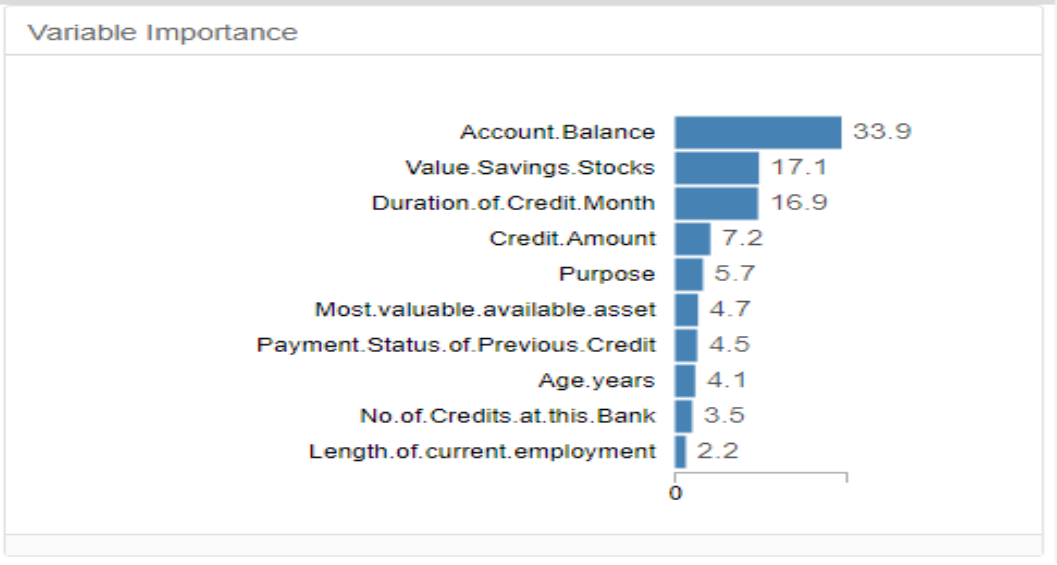
Min	1Q	Median	3Q	Max
-2.289	-0.713	-0.448	0.722	2.454

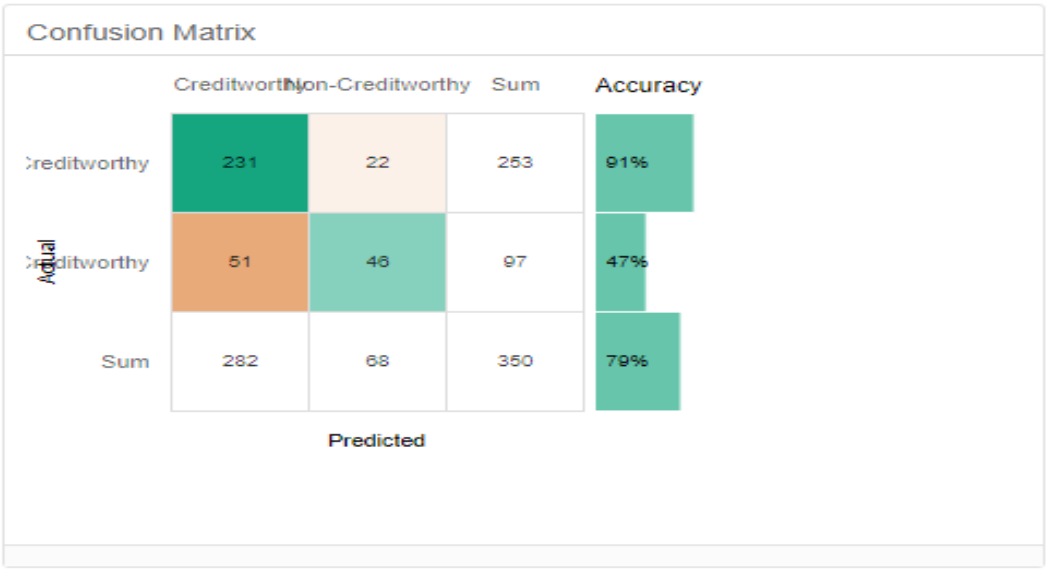
Coefficients:

	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	.

Decision Tree significant variables

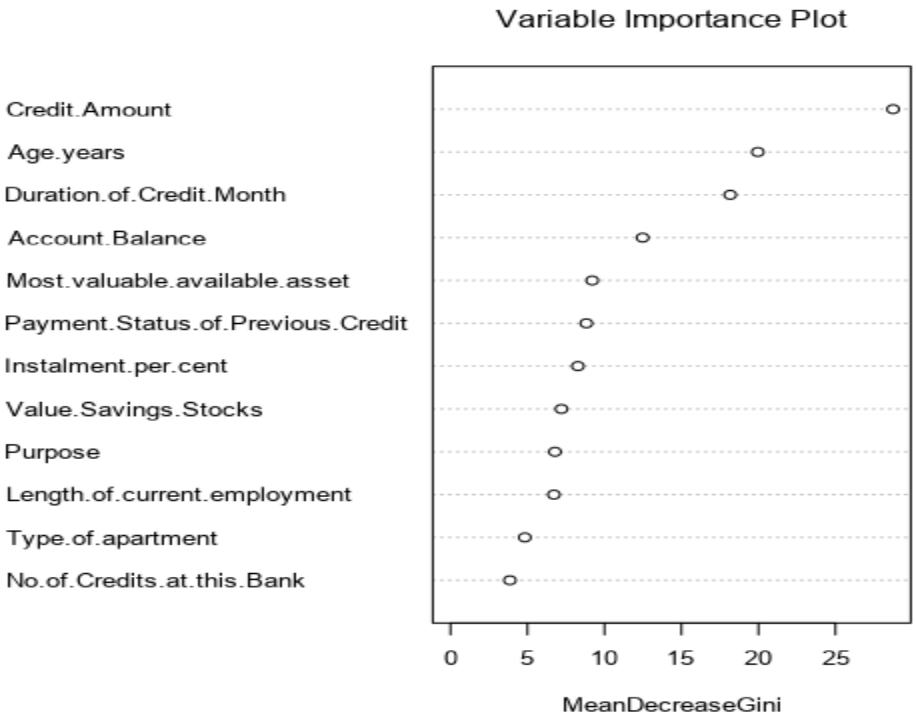
- Account Balance
- Value-Savings-Stocks
- Duration of Credit in Months





Forest Model significant variables

- Credit Amount
- Age in Years
- Duration of Credit in Months



Boosted Model significant variables

- Credit Amount
- Account Balance
- Duration of Credit in Months

Report for Boosted Model Boosted_PDR

Basic Summary:

Loss function distribution: Bernoulli
Total number of trees used: 4000
Best number of trees based on 5-fold cross validation: 1808

Model Comparison Report

Fit and error measures

Model	Accuracy	F1	AUC	Accuracy_Creditworthy	Accuracy_Non-Creditworthy
PDR_Stepwise	0.7600	0.8364	0.7306	0.8762	0.4889
Decision_Tree_PDR	0.7467	0.8304	0.7035	0.8857	0.4222
FM_PDR	0.7933	0.8681	0.7368	0.9714	0.3778
Boosted_PDR	0.7867	0.8632	0.7490	0.9619	0.3778

Confusion matrix of Boosted_PDR

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

Confusion matrix of Decision_Tree_PDR

	Actual_Creditworthy	Actual_Non-Creditworthy
Predicted_Creditworthy	93	26
Predicted_Non-Creditworthy	12	19

Confusion matrix of FM_PDR

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

Confusion matrix of PDR_Stepwise

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

Performance Diagnostic Plots

The Logistic Regression model showed a 76% accuracy a good Creditworthy Accuracy of 88% and a low Non-Creditworthy Accuracy of 49%.

The Decision Tree model showed the lowest level of accuracy at 75% a good Creditworthy Accuracy of 89% and a low Non-Creditworthy Accuracy of 42%.

The Forest Model showed the highest level of accuracy at 79% a good Accuracy Creditworthy of 97% and a low Non-Creditworthy Accuracy of 38%.

The Boosted model showed a hight level of accuracy at 79%, a good Creditworthy Accuracy of 96% and a low Non-Creditworthy Accuracy of 38%.

There seems to be a bias towards Actual Creditworthy in the Confusion Matrix however there is a larger sample of Creditworthy individuals in the data file.

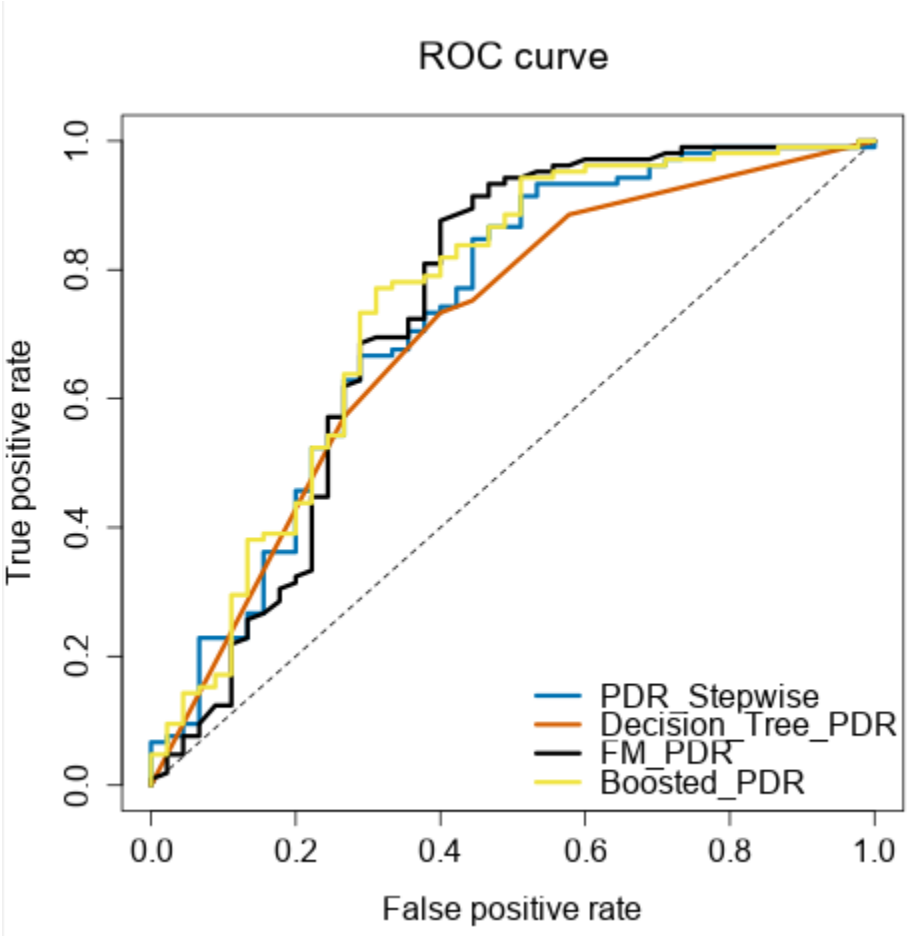
Step 4: Writeup

Model Comparison Report

Fit and error measures

Model	Accuracy	F1	AUC	Accuracy_Creditworthy	Accuracy_Non-Creditworthy
PDR_Stepwise	0.7600	0.8364	0.7306	0.8762	0.4889
Decision_Tree_PDR	0.7467	0.8304	0.7035	0.8857	0.4222
FM_PDR	0.7933	0.8681	0.7368	0.9714	0.3778
Boosted_PDR	0.7867	0.8632	0.7490	0.9619	0.3778

The greatest level of accuracy was displayed by the Forest Model (79%) against the validation set.
Within the Creditworthy and Non-Creditworthy segments the Forest Model shows the highest accuracy to predict Creditworthy.
Comparing the models in the ROC curve graph below, the Forest model produces the best result.
There also seems to be a bias towards Creditworthy individuals.



The most accurate model in the validation set is the Forest Model.
Within the Creditworthy and Non-Creditworthy segments the Forest Model shows the highest accuracy to predict Creditworthy.
In the ROC curve graph, the Forest model produces the best result.

The Forest Model is used in the following workflow to predict how many of the new customers will be Creditworthy.
The model predicts that 408 of the new customers will be Creditworthy.


credit-data-training.xlsx Query=`Sheet1$`
customers-to-score.xlsx Query=`Sheet1$`
Score_Creditworthy = IF [Score_Creditworthy] > [Score_Non-Creditworthy] THEN 1 E...