Project: Creditworthiness

Step 1: Business and Data Understanding

Key Decisions:

- What decisions needs to be made?

To classify the new customers into two categories based on their creditworthiness in order to help the bank manager approving their loan applications.

- What data is needed to inform those decisions?

Two datasets are given for this case and they will be utilized as follow:

- ***credit-data-training.xlsx***: this file provides all the historical records of the credit approvals that the bank has completed before.
- ***customers-to-score.xlsx***: this file contains the list of the new customers who submitted a loan application and need to be evaluated in terms of their creditworthiness.

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

Since the purpose of our model in this case is to predict if the customer is creditworthy or not -two possible outcomes only- then a binary classification model will be built and used to get a reliable estimation.

Step 2: Building the Training Set

- In your cleanup process, which fields did you remove or impute?

Starting by running the association analysis & field summary tool, I found the following;

- No numerical values are highly correlated with each other -no correlation higher that .70-

Pearson Correlation Analysis

Full Correlation Matrix

	Duration.of.Credit.Month	Credit.Amount	Instalment.per.cent	Duration.in.Current.address	Most.valuable.available.asset	Age.years
Duration.of.Credit.Month	1.000000	0.565054	0.145637	-0.032494	0.128814	-0.018171
Credit.Amount	0.565054	1.000000	-0.253286	-0.136621	0.457147	0.040486
Instalment.per.cent	0.145637	-0.253286	1.000000	0.131231	0.115114	0.111456
Duration.in.Current.address	-0.032494	-0.136621	0.131231	1.000000	-0.047386	0.301966
Most.valuable.available.asset	0.128814	0.457147	0.115114	-0.047386	1.000000	0.123579
Age.years	-0.018171	0.040486	0.111456	0.301966	0.123579	1.000000
Type.of.apartment	0.126967	0.100413	0.178926	-0.163386	0.182744	0.208552
No.of.dependents	-0.185180	0.082721	-0.293380	-0.036814	0.019435	0.046996
Telephone	0.238437	0.192532	0.038515	0.055112	0.083395	0.141103
Foreign.Worker	-0.207298	-0.045994	-0.155458	-0.015787	0.071932	-0.020939

	Type.of.apartment	No.of.dependents	Telephone	Foreign.Worker
Duration.of.Credit.Month	0.126967	-0.185180	0.238437	-0.207298
Credit.Amount	0.100413	0.082721	0.192532	-0.045994
Instalment.per.cent	0.178926	-0.293380	0.038515	-0.155458
Duration.in.Current.address	-0.163386	-0.036814	0.055112	-0.015787
Most.valuable.available.asset	0.182744	0.019435	0.083395	0.071932
Age.years	0.208552	0.046996	0.141103	-0.020939
Type.of.apartment	1.000000	-0.010189	0.179688	-0.026742
No.of.dependents	-0.010189	1.000000	-0.097632	0.218454
Telephone	0.179688	-0.097632	1.000000	-0.168472
Foreign.Worker	-0.026742	0.218454	-0.168472	1.000000

Figure 1

- **Duration in Current Address** has 69% missing data while **Age Years** has only 2% missing data, therefore **Duration in Current Address** will be removed and the missing values for the **Age Years** will be imputed with the median value "which equals 33" since the distribution of this variable is asymmetric -skewed to the left-.
- **Concurrent Credits** & **Occupation** has one value only "uniform " while **Guarantors**, **Foreign Worker** and **No. of Dependents** show low variability as the majority of the data are skewed towards one value, they all will be removed.
- **Telephone** should be excluded as well since it doesn't play any role in predicting the customers' creditworthiness -irrelevant variable-.

So, my final dataset has 13 columns and the average of the Age Years is around 36. The graph below "Field Summary Report" illustrates the previous points clearly.

Figure 2

Due to the **Occupation** field uniform values, it couldn't be captured in the previous report. The graph below shows that the **Occupation** has a very small number of unique values and hence it's should be removed.

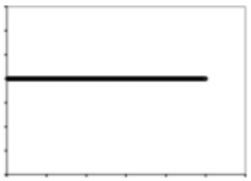
Name	Plot	% Missing	Unique Values	Min	Mean	Median	Max	Std Dev	Remarks
Occupation		0.0%	1	1.000	1.000	1.000	1.000	0.000	This field has a small number of unique values, and appears to be a categorical field. Consider changing the field data type to "string".

Figure 3

Step 3: Train your Classification Models

- Which predictor variables are significant or the most important?
- 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?

  - **Logistic Regression (Stepwise)**

By sitting the **Credit Application Result** as the target variable, I found that the highlighted variables are the most significant ones with p value less than .05.

Report for Logistic Regression Model Stepwise_Creditworthniess

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

Figure 4

- **Logistic Regression (Stepwise)- Continued**

The overall accuracy is around 76% while the accuracy of the creditworthy is higher than the non-creditworthy -88% & 49% respectively-. Also, PPV= 80% and NPV= 63%, this gap between the two segment accuracies can be considered as a bias towards predicting the creditworthy customers.

Model Comparison Report

Fit and error measures

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

Confusion matrix of Stepwise_Creditworthniess

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

Figure 5

- **Decision Tree**

As shown in Figure 6 below; **Account Balance**, **Value Savings Stocks** and **Duration of Credit Month** are the most important variables. Figure 7 shows that the overall accuracy is around 75% while 87% accuracy for the creditworthy and 47% accuracy for the non-creditworthy. However, based on the confusion matrix the PPV= 79% and NPV= 60% which indicates that the model is biased towards predicting the customer creditworthiness.

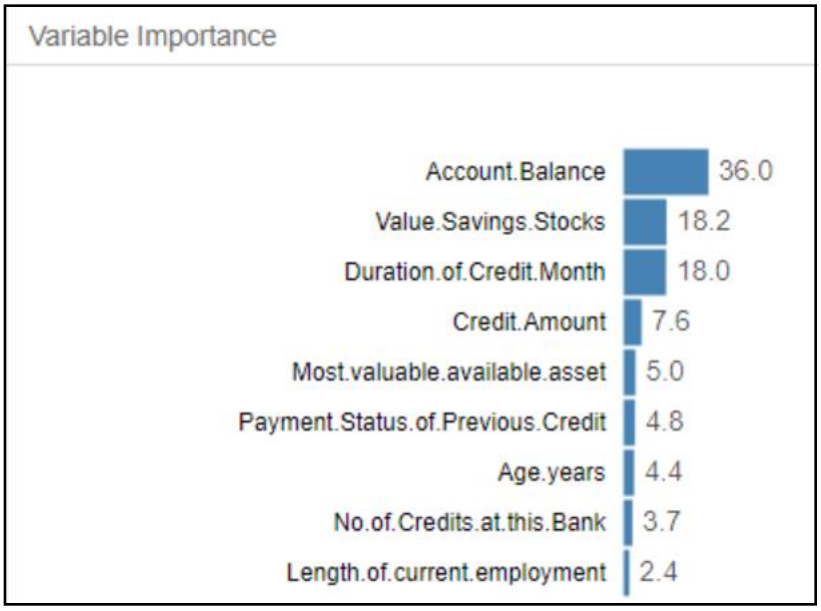


Figure 6

Model Comparison Report

Fit and error measures

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

Confusion matrix of DT_Creditworthniess

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

Figure 7

- **Forest Model**

According to graph 8 below, **Credit Amount**, **Age Years** & **Duration of credit Month** are the most significate variables. Total accuracy of this model is 79% and the calculated PPV= 78% and NPV= 85%, since there's a small difference between the accuracies in this case I can consider the forest model as an unbiased model.

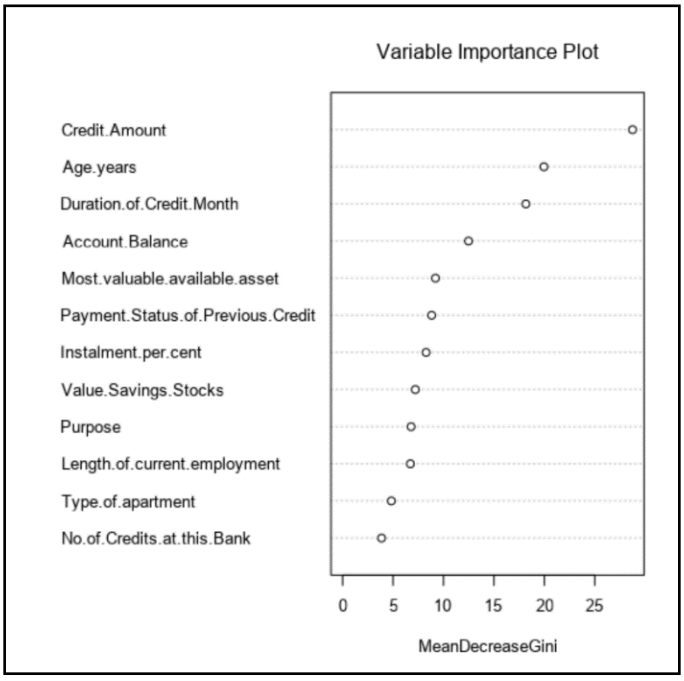


Figure 8

Model Comparison Report

Fit and error measures

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

Confusion matrix of FM_Creditworthiness

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

Figure 9

- **Boosted Model**

Referring to Graph 10 in the next page, the most important variables are; **Credit Amount** & **Account Balance.** For this model, the total accuracy as shown in the graph next page is around 79% and the calculated PPV= 78% while the NPV= 81% and it's considered as an unbiased model since the difference between the two segments' accuracies is small.

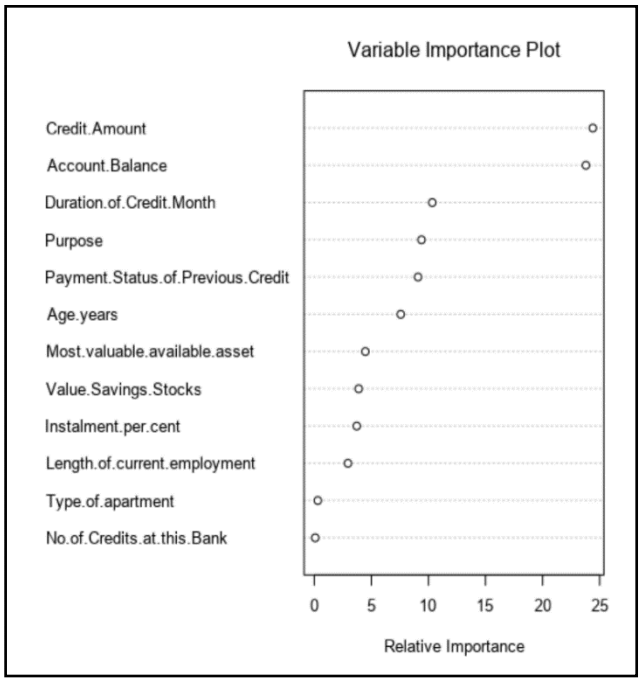


Figure 10

Model Comparison Report

Fit and error measures

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

Confusion matrix of BM_Creditworthniess

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

Figure 11

Step 4: Writeup

- Which model did you choose to use?

Forest Model has a high percentage of the overall accuracy among other models and it provides a better accuracy for both segments creditworthy & non-creditworthy. Also, the difference between these two segments are mostly comparable and that could be used as an indicator that this model has the least bias. Choosing unbiased model has a major impact in facilitating the process of validating the loan applications, as it will enable the bank to fairly ensure the creditworthiness of their customers, minimize the credit risk and seize the opportunity of having new customers.

Model Comparison Report

Fit and error measures

Model	Accuracy	F1	AUC	Accuracy_Creditworthy	Accuracy_Non-Creditworthy
DT_Creditworthniess	0.7467	0.8273	0.7054	0.8667	0.4667
FM_Creditworthiness	0.7933	0.8681	0.7368	0.9714	0.3778
BM_Creditworthniess	0.7867	0.8632	0.7524	0.9619	0.3778
Stepwise_Creditworthniess	0.7600	0.8364	0.7306	0.8762	0.4889

Figure 12

Moreover, the ROC curve below shows that the Forest Models reached the top faster than other models and has the highest curve. This's important because it indicates the model ability to give the best number of true positive prediction for a given set of wrongly predicted values.

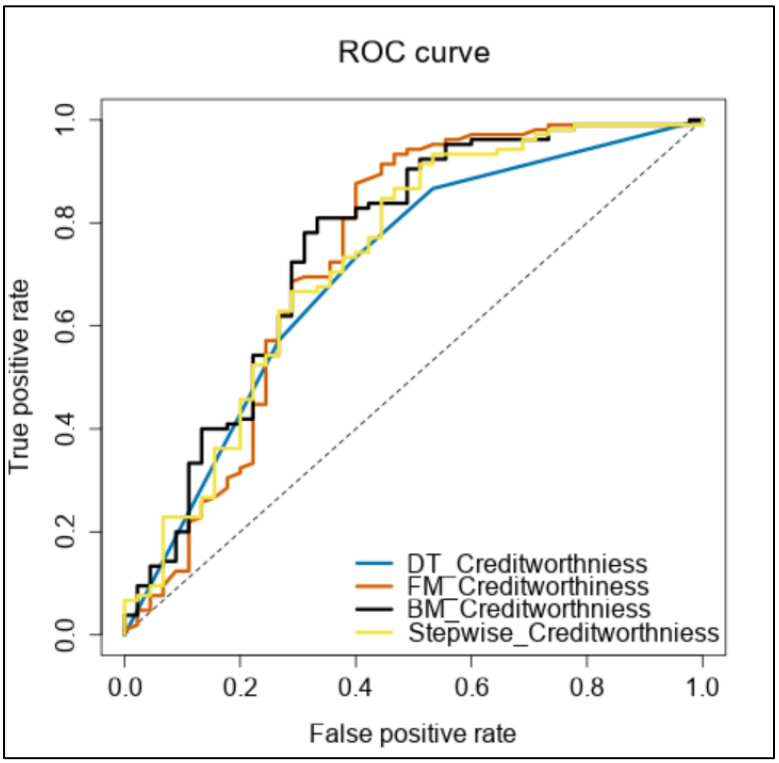


Figure 13

- How many individuals are creditworthy?

Out of the 500 applicants, 410 customers seem to be creditworthy "creditworthiness score ≥ .50"

Alteryx Workflow

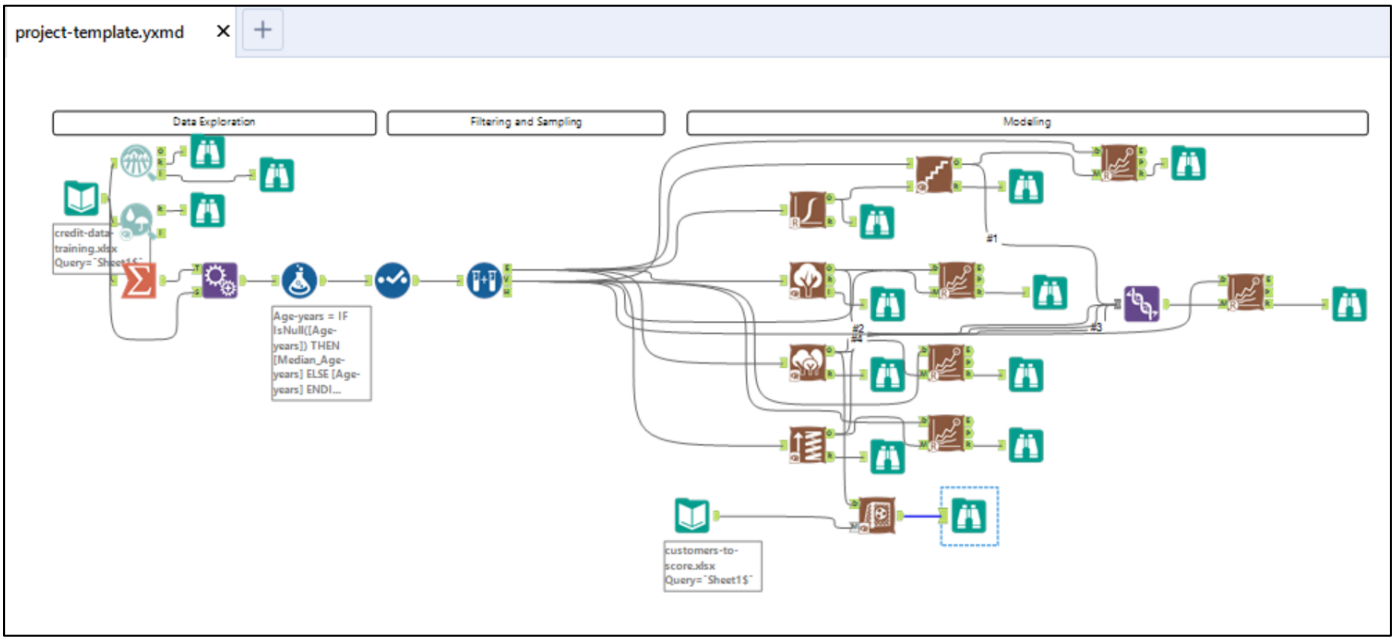


Figure 14