Project: Creditworthiness

Step 1: Business and Data Understanding

Key Decisions:

- What decisions needs to be made?
    - Decide whether the new clients are creditworthy to give a loan or not.
- What data is needed to inform those decisions?
    - Data on all past applications (Salaried or not, Income, Age, Purpose, Credit Amount, Account Balance, Duration of Credit…).
    - The list of customers that need to be processed in the next few days.
- What kind of model (Continuous, Binary, Non-Binary, Time-Series) do we need to use to help make these decisions?
    - We need to classify client into two groups Creditworthy or Non- Creditworthy, to do that the Binary classification model is what we need.

Step 2: Building the Training Set

The correlation should be at least 0.70 to be considered “high”, none of the numerical data fields in our dataset surpassed that value. As we can see in the figure below the highest value is 0.57.

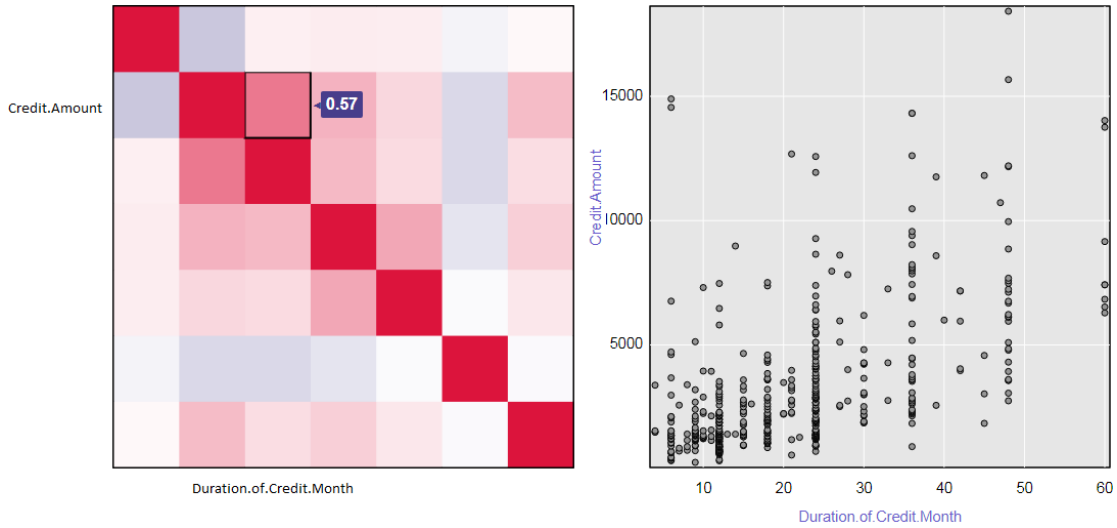

Figure 1: Correlation Matrix with ScatterPlot

Field that should be removed:

- Fields with a lot of missing data:
  - **Duration-in-Current-address:** 68.8% of the data are messing.

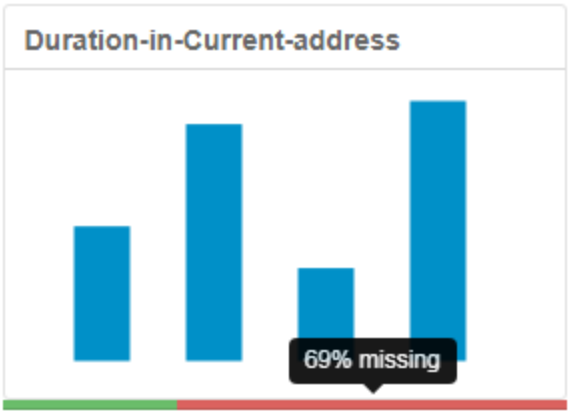

Figure 2: Histogram of Duration-in-Current-address

- Fields that have low variability:
  - **Concurrent-Credits, Occupation**: the data in these field is entirely uniform (there is only one value for the entire field).

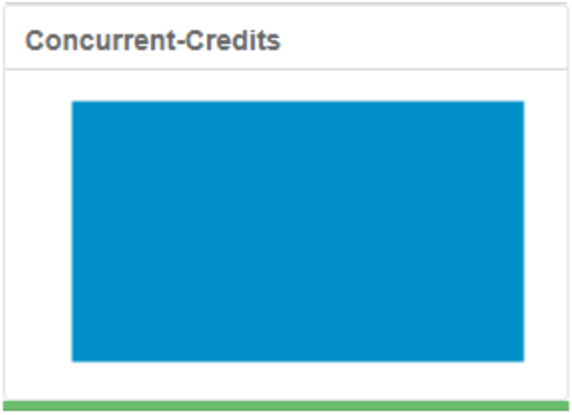

Figure 3: Bar chart of Concurrent-Credits

Figure 4: Bar chart of Occupation

- **Foreign-Worker, Guarantors, No-of-dependents:** The majority of data in these fields are skewed towards “1”, “None” and “1” respectively.

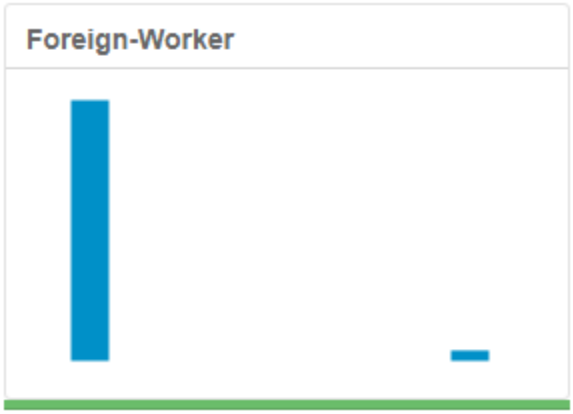

Figure 5: Histogram of Foreign-Worker

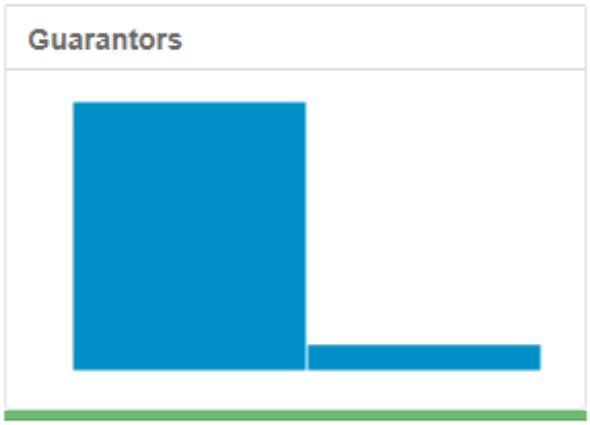

Figure 6: Bar chart of Guarantors

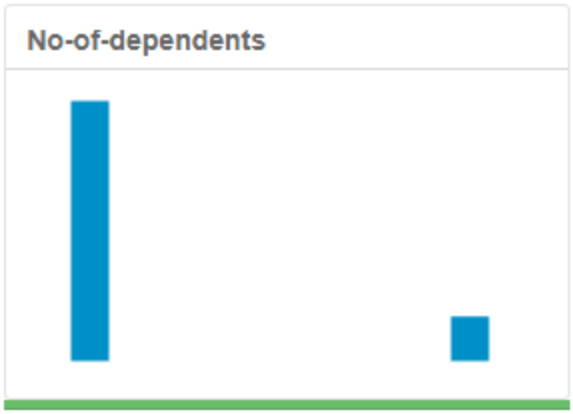

Figure 7: Histogram of No-of-dependents

- Fields that logically does not have a relationship with the target variable:
  - **Telephone:** There is no relationship with Credit-Application-Result.

Field to impute:

- Fields with missing data:
  - **Age-years:** 2.4% of the data are missing and since the variable is skewed to the right, the mean is biased by the values at the far end of the distribution. Therefore, the median which equals 33 is a better representation of the majority of the values in the variable.

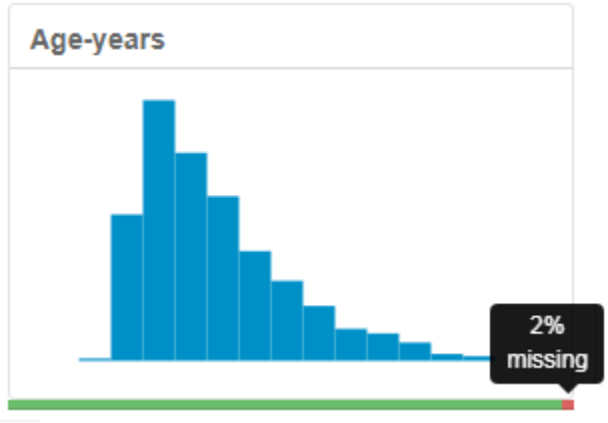

Figure 8: Histogram of Age-years

Step 3: Train your Classification Models

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

	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 9: Predictor variables coefficients of the Linear Regression Model

From the above figure we can see that the most three important variables for the model are:

- Account.Balance
- Purpose
- Credit.Amount

    - Decision Tree

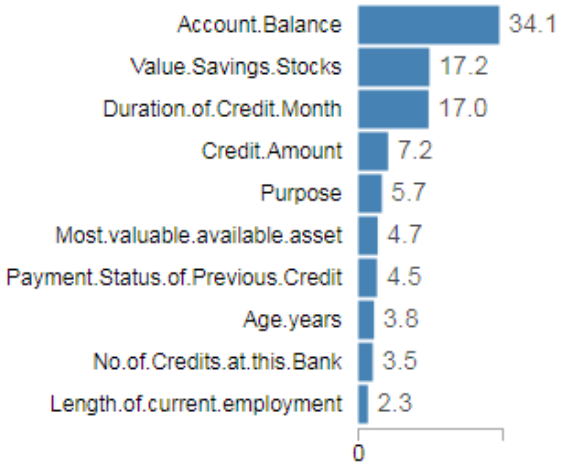

Figure 10: Variable importance of the Decision Tree Model

From the above figure we can see that the most three important variables for the model are:

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

- Forest Model

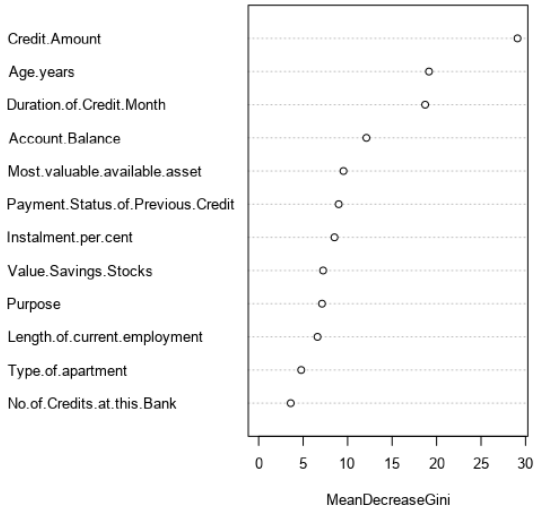

Figure 11: Variable importance plot of the Forest Model

From the above figure we can see that the most three important variables for the model are:

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

- Boosted Tree

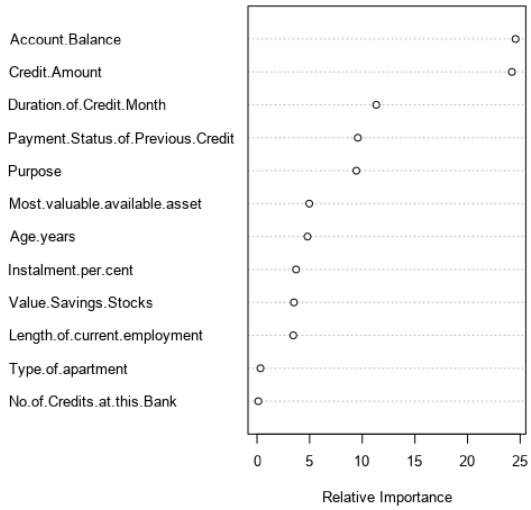

Figure 12: Variable importance plot of the Boosted Model

From the above figure we can see that the most three important variables for the model are:

- Account.Balance
- Credit.Amount
- Duration.of.Credit.Month

- Validate your model against the Validation set. What was the overall percent accuracy?

Model	Accuracy
Decision_Tree	0.7467
Forest_Model	0.7933
Boosted_Model	0.7867
Logistic_Regression	0.7600

Figure 13: Accuracy of the models

- Show the confusion matrix. Are there any bias seen in the model's predictions?

Confusion matrix of Boosted_Model

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

Confusion matrix of Decision_Tree

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

Confusion matrix of Forest_Model

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

Confusion matrix of Logistic_Regression

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

Figure 14: Confusion matrices of the models

We can see from the confusion matrices that 4, 12, 4 and 13 records that were predicted Non-Creditworthy by Boosted Model, Decision Tree, Forest Model and Logistic regression respectively that were actually Creditworthy. Yet there are 28, 26, 27 and 23 records that were predicted Creditworthy that were actually Non-Creditworthy by Boosted Model, Decision Tree, Forest Model and Logistic regression respectively.

This result shows us that the models are biased to predict Creditworthy, we have way more Non-Creditworthy records that were predicted Creditworthy. This typically happens when we have one category that is much dominant than the other in the raining data.

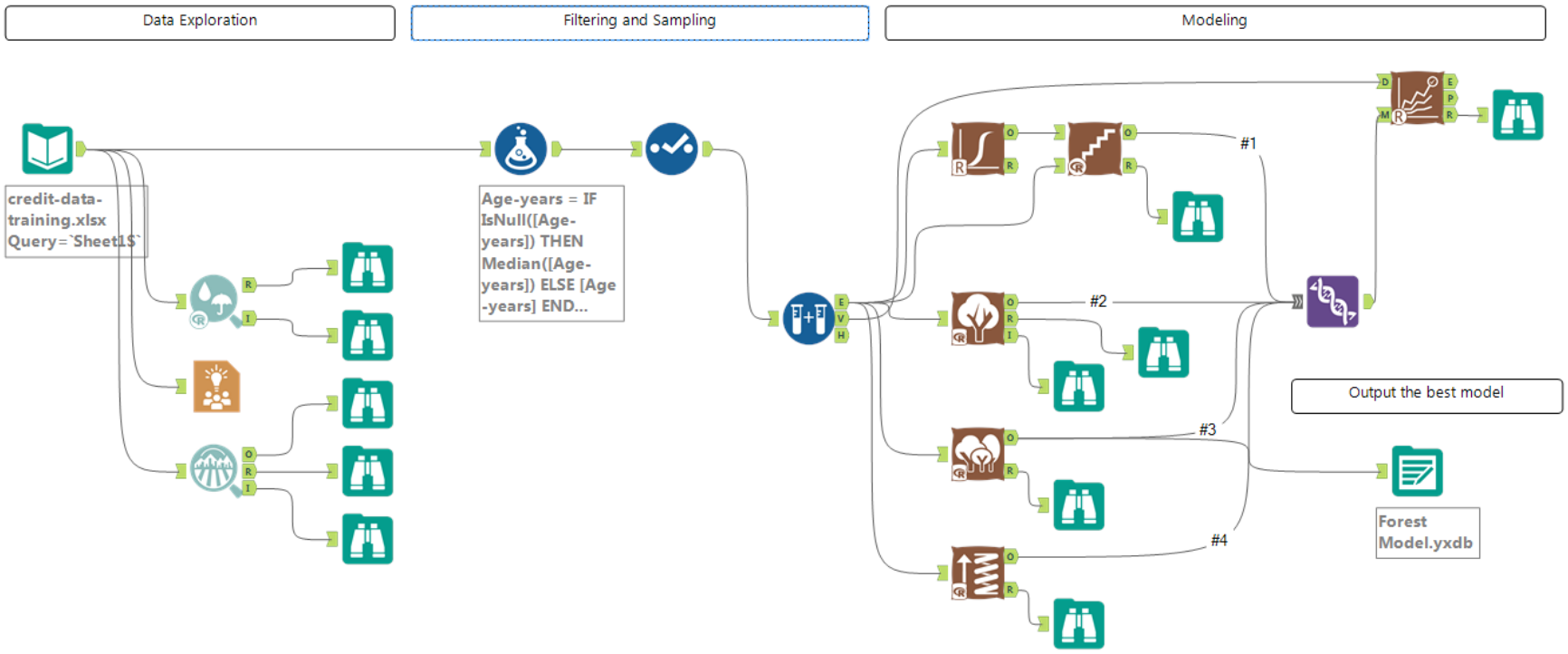

Figure 15: Training workflow

Step 4: Writeup

From the below figure we can see that the validation dataset was predicted quite well by the four models, the overall accuracy and the Creditworthy accuracy were the highest for the Forest Model at 0.9619 and 0.7933 respectively, while the Non-Creditworthy records were a little bit tougher to predict, at only 0.4000 accuracy which is the second best value.

Model	Accuracy	F1	AUC	Accuracy_Creditworthy	Accuracy_Non-Creditworthy
Decision_Tree	0.7467	0.8304	0.7035	0.8857	0.4222
Forest_Model	0.7933	0.8670	0.7403	0.9619	0.4000
Boosted_Model	0.7867	0.8632	0.7467	0.9619	0.3778
Logistic_Regression	0.7600	0.8364	0.7306	0.8762	0.4889

Figure 16: Fit and error measures

The ROC curve for both Boosted and Forest Models are the best and it's quite hard to compare between the two (Figure 17), but from the above figure we can see that the AUC for Boosted Model is the best at 0.7505.

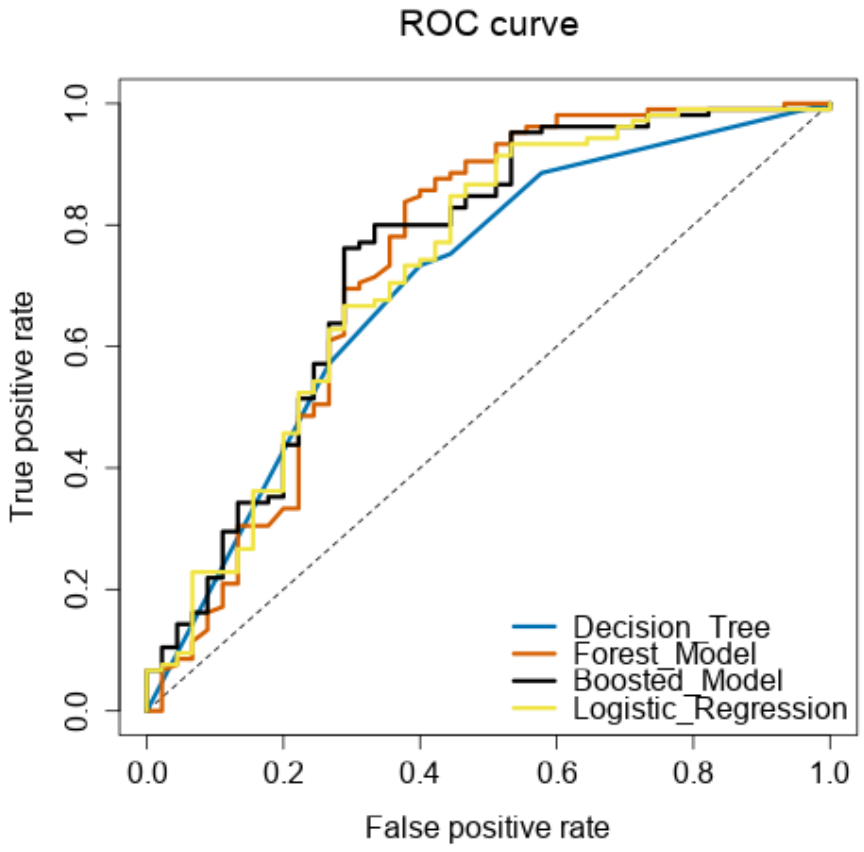

Figure 17: ROC curves of the models

Forest Model has less bias in its predictions compared to the other model (Figure 14) we can see in the confusion matrix of the Forest Model that 4 records were predicted Non-Creditworthy that were actually Creditworthy and 27 records were predicted Creditworthy that were actually Non-Creditworthy.

In conclusion and by taking into consideration that the boss only cares about prediction accuracy for Creditworthy and Non-Creditworthy segments the "**Forest Model"** will be the best choice.

- How many individuals are creditworthy?
    - 410

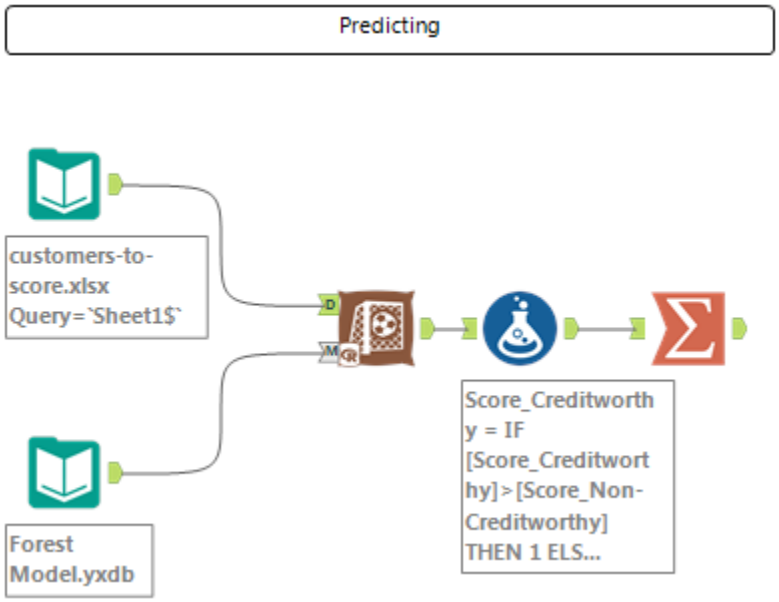

Figure 18: Predicting workflow