Project: Creditworthiness

Step 1: Business and Data Understanding

- What decisions needs to be made?

  We need to decide which customers are enough creditworthy to give them a loan.

- What data is needed to inform those decisions?

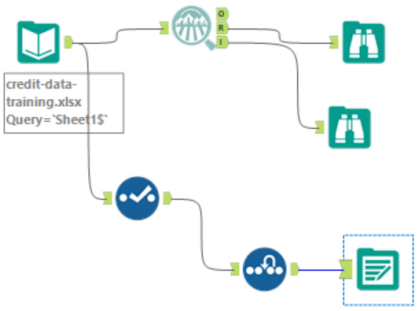
  To Make this decision first we need to see which are the most important variable to each model after that we will make decisions depend on those variables.

  Such as the age is important because the old customer have a time to earn more many than the young customer. Also the credit amount is one of the needed data to make our decision.

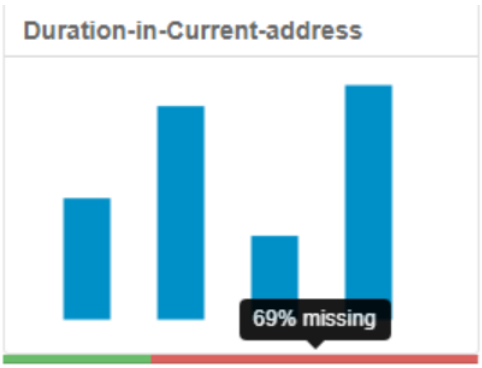
- What kind of model (Continuous, Binary, Non-Binary, Time-Series) do we need to use to help make these decisions?

  We will use Binary model due to the target variable Credit Application Result countians of two values (Creditworthy-Non Creditworthy).

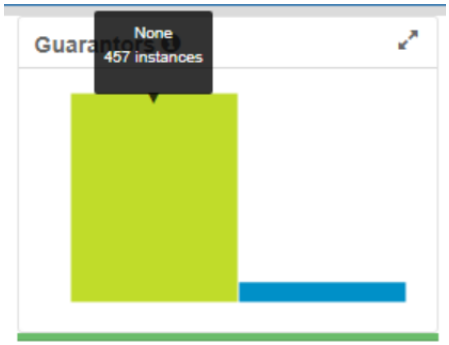
Step 2: Building the Training Set



1- I removed the telephone attribute because it's irrelevant to our target.
2- I removed Duration-in-Current-address attribute because the percentage of missing value is greater than 50% which is very high.

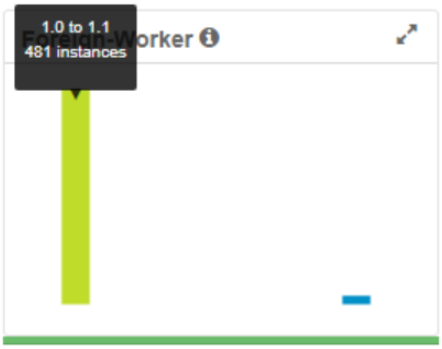


3- I removed Guarantors attribute due to (Low Variability)

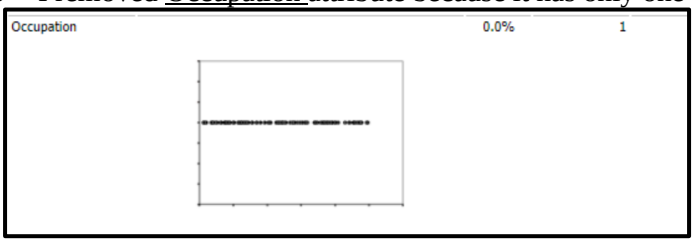


4- I removed No-of-dependents attribute due to (Low Variability)

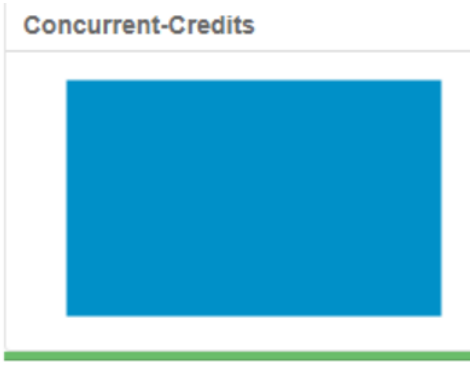
5- I removed Foreign-Worker attribute due to (Low Variability).



6- I removed Occupation attribute because it has only one value.

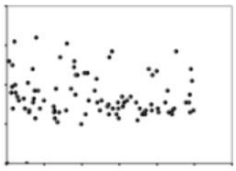


7- I removed Concurrent-Credits attribute because in has only one value



8- I impute Age-years attribute with median because it has 2.4% of missing value.

Name	Plot	% Missing	Unique Values
Age-years		2.4%	54



Step 3: Train your Classification Models

- Logistic Regression Model

Coefficients:

	Estimate	Std. Error	z value	Pr(>|z|)	
(Intercept)	-2.9621914	6.837e-01	-4.3326	1e-05	***
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	.
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	*
Most.valuable.available.asset	0.2650267	1.425e-01	1.8599	0.06289	.
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	*
Credit.Amount	0.0001704	5.733e-05	2.9716	0.00296	**
Account.BalanceSome Balance	-1.6053228	3.067e-01	-5.2344	1.65e-07	***

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

Depend on the table above the most three important variables are Account-Balance, Credit_Amount, and Purpose.

Fit and error measures

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

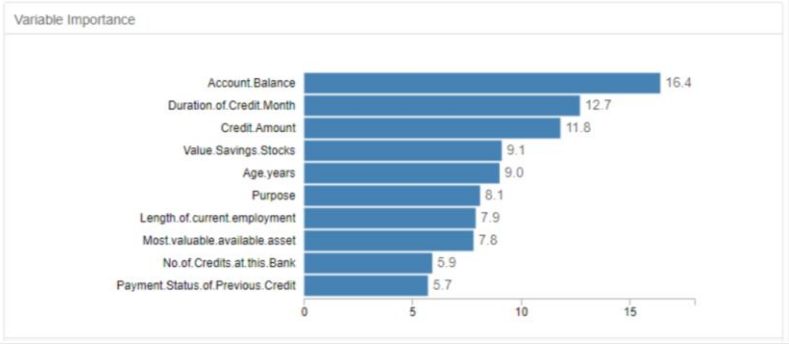
As shown in the above table the accuracy of this model was 0.760

Confusion matrix of Stepwise_Log

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

This confusion matrix showed us there is a bias in this model due to the actual predicted creditworthy is more than the number of actual predicted Non- creditworthy.

- Decision Tree



Depend on the variable importance plot above the most three important variables in this model are Account_ Balance, Duration_Of_Credit_Month, and Credit_Amount.

Fit and error measures

Model	Accuracy	F1	AUC	Accuracy_Creditworthy	Accuracy_Non-Creditworthy
DT_Model	0.6667	0.7685	0.6272	0.7905	0.3778

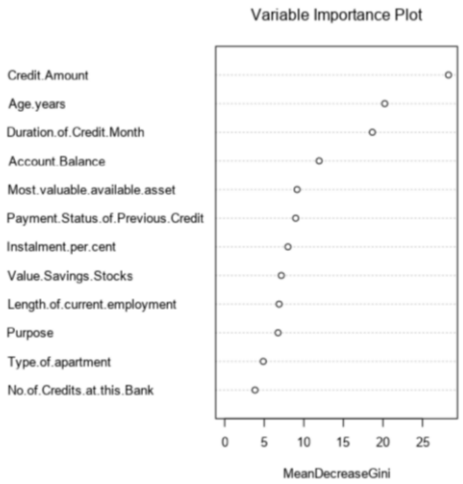
As shown in the above table the accuracy of this model was 0.66

Confusion matrix of DT_Model

	Actual_Creditworthy	Actual_Non-Creditworthy
Predicted_Creditworthy	83	28
Predicted_Non-Creditworthy	22	17

This confusion matrix showed us there is a bias in this model due to the actual predicted creditworthy is more than the number of actual predicted Non- creditworthy.

- Forest Model



Depend on the variable importance plot above the three most important variables in this model are Credit_Amount, Age_Years, and Duration_Of_Credit_Month.

Fit and error measures

Model	Accuracy	F1	AUC	Accuracy_Creditworthy	Accuracy_Non-Creditworthy
Forest_Model	0.8067	0.8755	0.7531	0.9714	0.4222

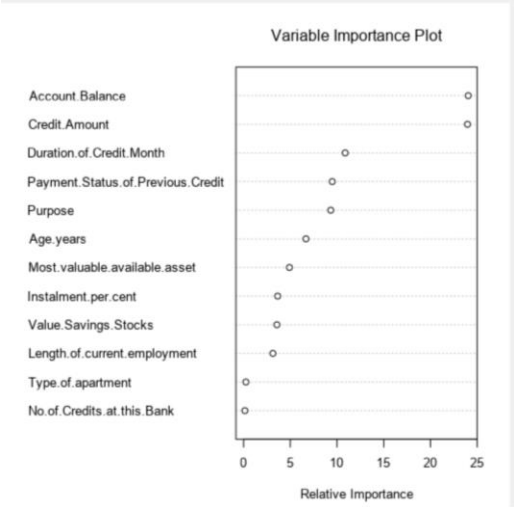
As shown in the above table the accuracy of this model was 0.80

Confusion matrix of Forest_Model

	Actual_Creditworthy	Actual_Non-Creditworthy
Predicted_Creditworthy	102	26
Predicted_Non-Creditworthy	3	19

This confusion matrix showed us there is a bias in this model due to the actual predicted creditworthy is more than the number of actual predicted Non- creditworthy.

- Boosted Model



Depend on the variable importance plot above the most three important variables in this model are Account_Balance, Credit_Amount, and Duration_Of_Credit_Month.

Fit and error measures

Model	Accuracy	F1	AUC	Accuracy_Creditworthy	Accuracy_Non-Creditworthy
Boosted_Model	0.7933	0.8670	0.7505	0.9619	0.4000

As shown in the above table the accuracy of this model was 0.79

Confusion matrix of Boosted_Model

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

This confusion matrix showed us there is a bias in this model due to the actual predicted creditworthy is more than the number of actual predicted Non- creditworthy.

Step 4: Writeup

After training all the model I decided to use “Forest Model” and to make this decision I did relied on different things:

1- I compared the accuracy for all models with each other

Fit and error measures

Model	Accuracy	F1	AUC	Accuracy_Creditworthy	Accuracy_Non-Creditworthy
DT_Model	0.6733	0.7721	0.6296	0.7905	0.4000
Forest_model	0.8067	0.8755	0.7531	0.9714	0.4222
Boosted_Model	0.7933	0.8670	0.7505	0.9619	0.4000
Stepwise_Log	0.7600	0.8364	0.7306	0.8762	0.4889

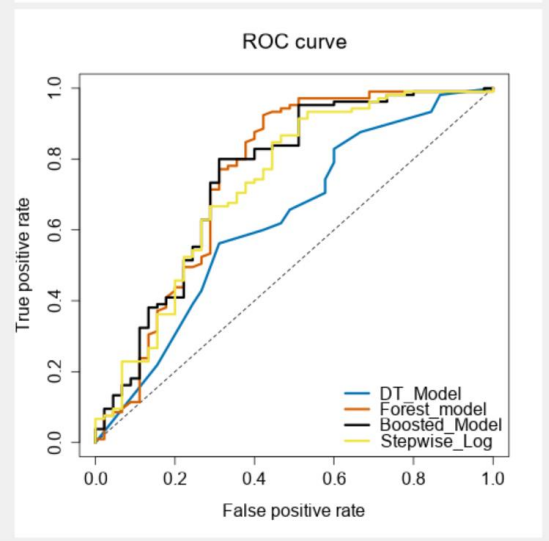
As shown in above table the Forest model has highest accuracy, F1, and AUC.

2- I compared the Accuracy of creditworthy between all the model and the Forest model has the highest creditworthy accuracy.

Fit and error measures

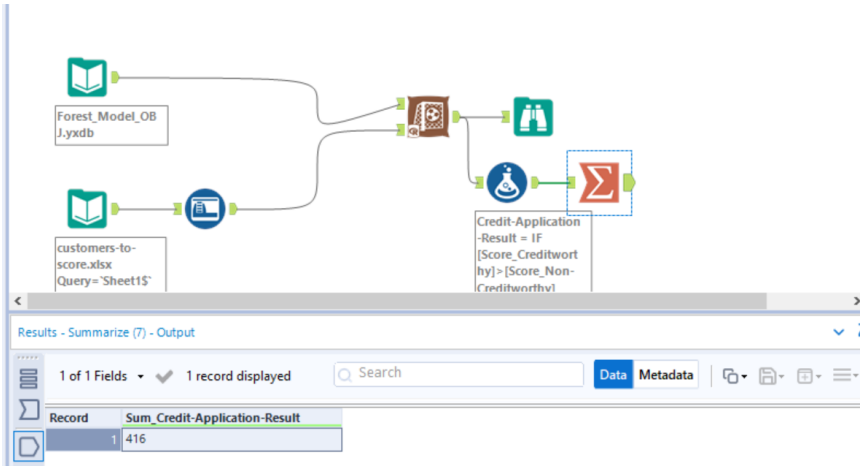
Model	Accuracy	F1	AUC	Accuracy_Creditworthy	Accuracy_Non-Creditworthy
DT_Model	0.6733	0.7721	0.6296	0.7905	0.4000
Forest_model	0.8067	0.8755	0.7531	0.9714	0.4222
Boosted_Model	0.7933	0.8670	0.7505	0.9619	0.4000
Stepwise_Log	0.7600	0.8364	0.7306	0.8762	0.4889

3- I compared the models in ROC graph



The forest model is very close to 1 that’s mean it’s the best model.

The bias in all the models was explain in the previse section.



Finally, the number of individuals is creditworthy is 416.