Dhuha Baqarish
PAND

Project: Creditworthiness

Step 1: Business and Data Understanding

Key Decisions:

Answer these questions

- What decisions needs to be made?

Identify whether customers who applied for loan are creditworthy to be extended one or not .

- What data is needed to inform those decisions?

- Account Balance
- Duration-of-Credit-Month
- Payment-Status-of-Previous-Credit
- Purpose
- Credit Amount
- Value-Savings-Stocks
- Length-of-current-employment
- Instalment-per-cent
- Most-valuable-available-asset
- Age-years
- Type-of-apartment
- No-of-Credits-at-this-Bank
- Occupation

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

Binary classification models such as Logistics Regression, Decision Tree, Forest Model and Boosted Model .

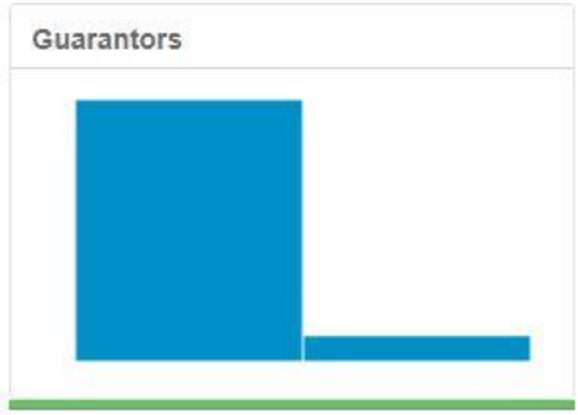
Step 2: Building the Training Set

Answer this question:

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

When we summarize the all fields of dataset called ("credit-data-training") , we take some steps depend on the appeared result

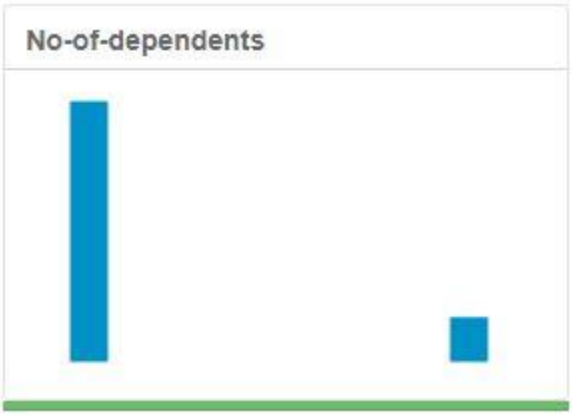
1. **Removed some fields , such :**
   - **Guarantors ,** it is low variability and the majority of the data is skewed toward "None" with 457 records , while the "Yes" with 43 records only .



   - **Duration-in-Current-address ,** because we have 69% from records is missing .

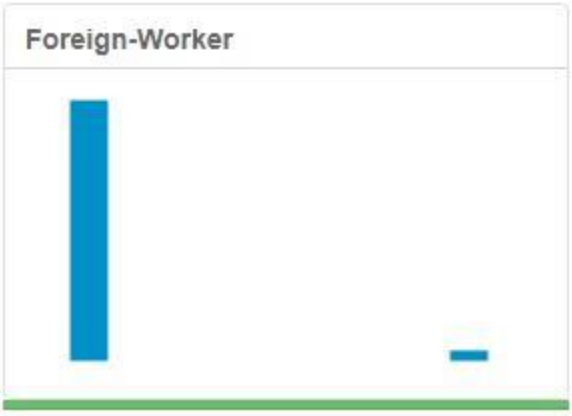
- **Concurrent-Credits ,** it is low variability and the data is entirely uniform .

- **No-of-dependents ,** it is low variability and the majority of the data is skewed toward "1.0 to 1.1" with 427 records , while the "2.0 to 2.1" with 73 records .



- **Telephone ,** because it is not useful column for predicting process .

- **Foreign-Worker ,** it is low variability and the majority of the data is skewed toward "1.0 to 1.1" with 300 records , while the "2.0 to 2.1" with 19 records only .



2. **Impute the null values of "Age-years" :**
   We did that by Median of Age-years , that equal 33 .

Step 3: Train your Classification Models

*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 Model :

As we see in the table below , the most important predictor variables for Logistic model are :

- Account Balance
- Credit Amount
- Purpose

Report

Report for Logistic Regression Model LM_Predicting

Basic Summary

Call:

glm(formula = Credit.Application.Result ~ Most.valuable.available.asset + Credit.Amount + Payment.Status.of.Previous.Credit + Account.Balance + Length.of.current.employment + Instalment.per.cent + Purpose, 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	***
Most.valuable.available.asset	0.2650267	1.425e-01	1.8599	0.06289	.
Credit.Amount	0.0001704	5.733e-05	2.9716	0.00296	**
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	*
Account.BalanceSome Balance	-1.6053228	3.067e-01	-5.2344	1.65e-07	***
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	*
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	.

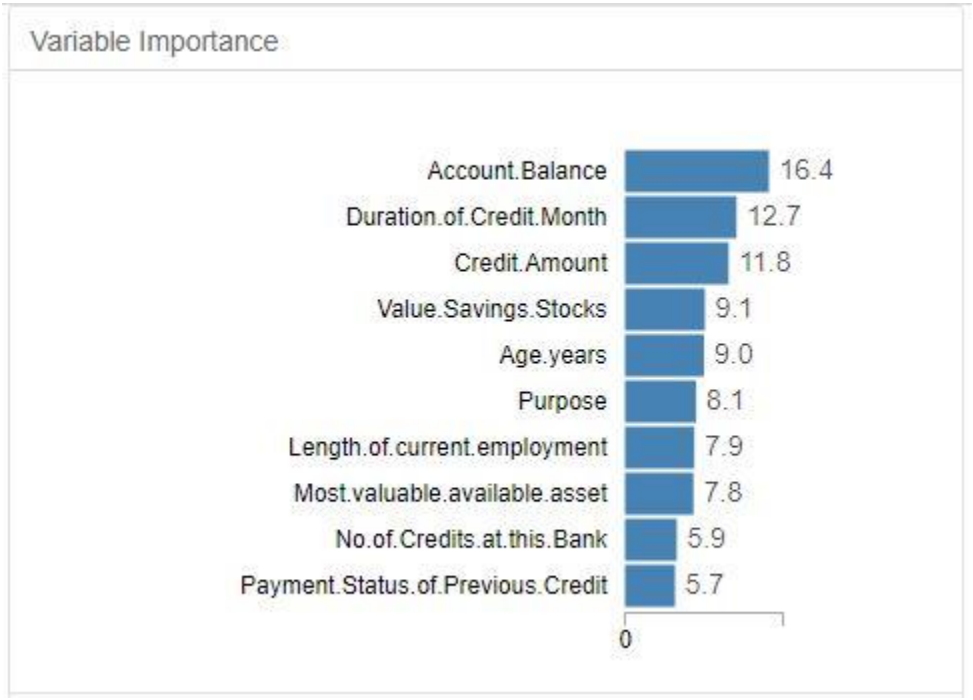
Significance codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial taken to be 1)

Decision Tree :

As we see in the image below , the most important predictor variables for Decision Tree are :

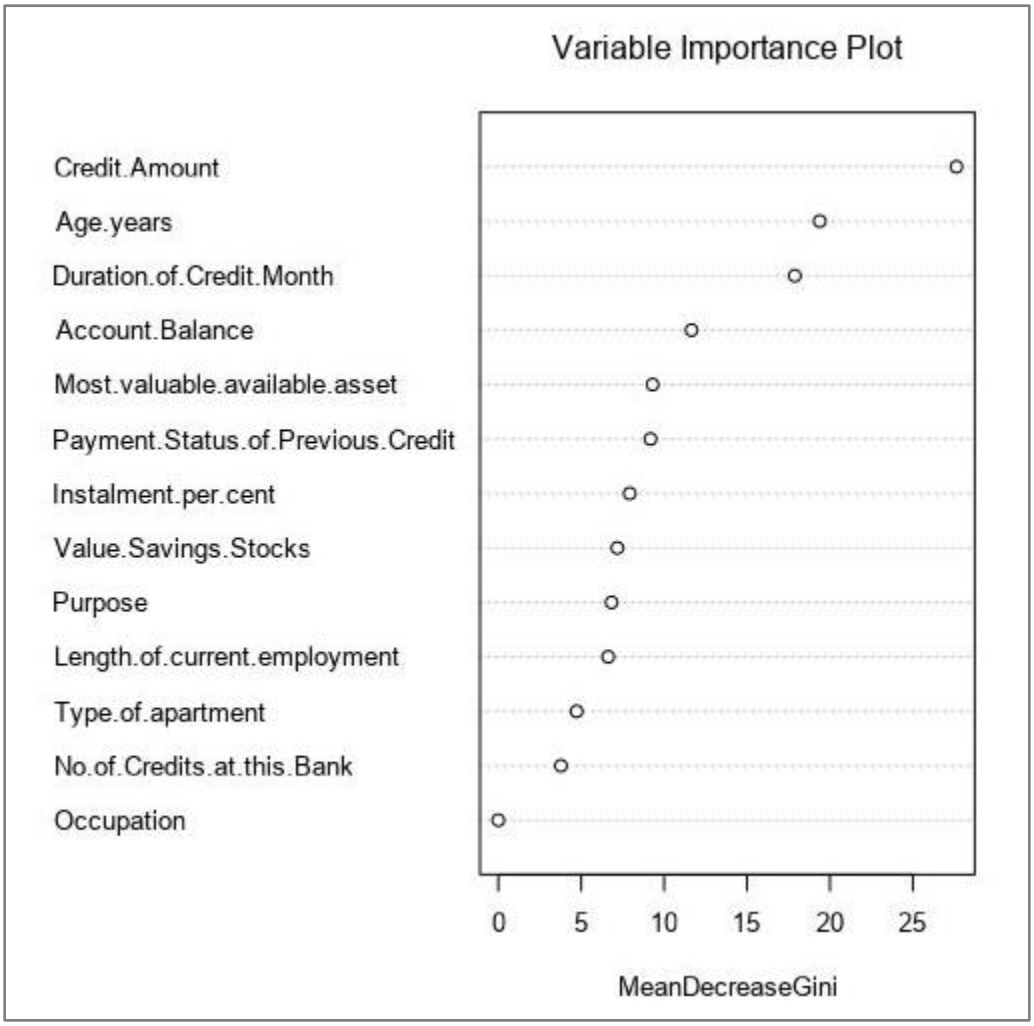
- Account Balance
- Duration of Credit Month
- Credit Amount



Forset Model :

As we see in the image below , the most important predictor variables for Forset Model are :

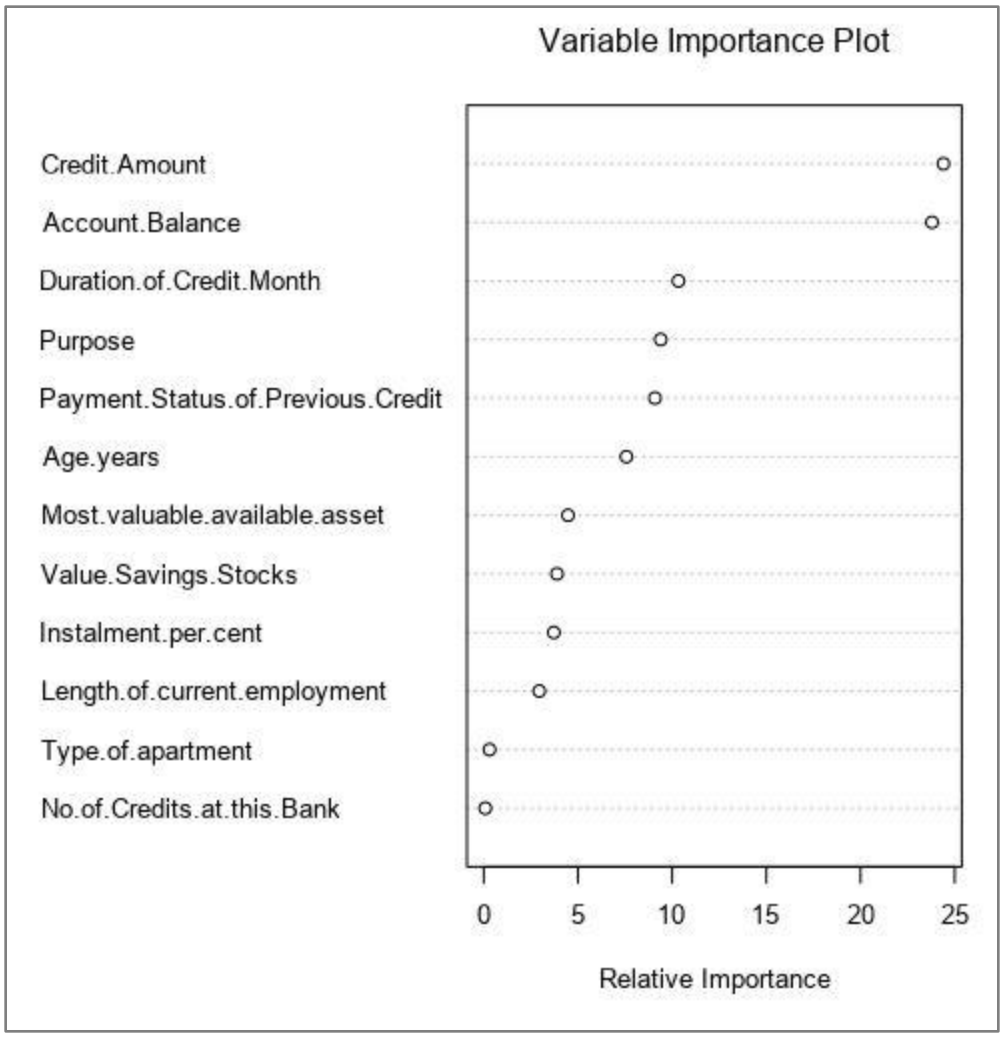
- Credit Amount
- Age years
- Duration of Credit Month



Boosted Model:

As we see in the image below , the most important predictor variables for Boosted Model are :

- Credit Amount
- Account Balance



- 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?

Model Comparison Report

Fit and error measures

Model	Accuracy	F1	AUC	Accuracy_Creditworthy	Accuracy_Non-Creditworthy
FM_Predicting	0.8133	0.8793	0.7498	0.9714	0.4444
BM_Predicting	0.7867	0.8632	0.7524	0.9619	0.3778
DT_Predicting	0.6667	0.7685	0.6272	0.7905	0.3778
LM_Predicting	0.7600	0.8364	0.7306	0.8762	0.4889

Model: model names in the current comparison.
Accuracy: overall accuracy, number of correct predictions of all classes divided by total sample number.
Accuracy_[class name]: accuracy of Class [class name] is defined as the number of cases that are **correctly** predicted to be Class [class name] divided by the total number of cases that actually belong to Class [class name], this measure is also known as *recall*.
AUC: area under the ROC curve, only available for two-class classification.
F1: F1 score, 2 * precision * recall / (precision + recall). The *precision* measure is the percentage of actual members of a class that were predicted to be in that class divided by the total number of cases predicted to be in that class. In situations where there are three or more classes, average precision and average recall values across classes are used to calculate the F1 score.

Confusion matrix of BM_Predicting

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

Confusion matrix of DT_Predicting

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

Confusion matrix of FM_Predicting

	Actual_Creditworthy	Actual_Non-Creditworthy
Predicted_Creditworthy	102	25
Predicted_Non-Creditworthy	3	20

Confusion matrix of LM_Predicting

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

As we see in the table above taken from Model Comparison report , the accuracy is :

- Foset Model 81%
- Boosted Model 79%
- Decision Tree 67%
- Logistic Model 76%

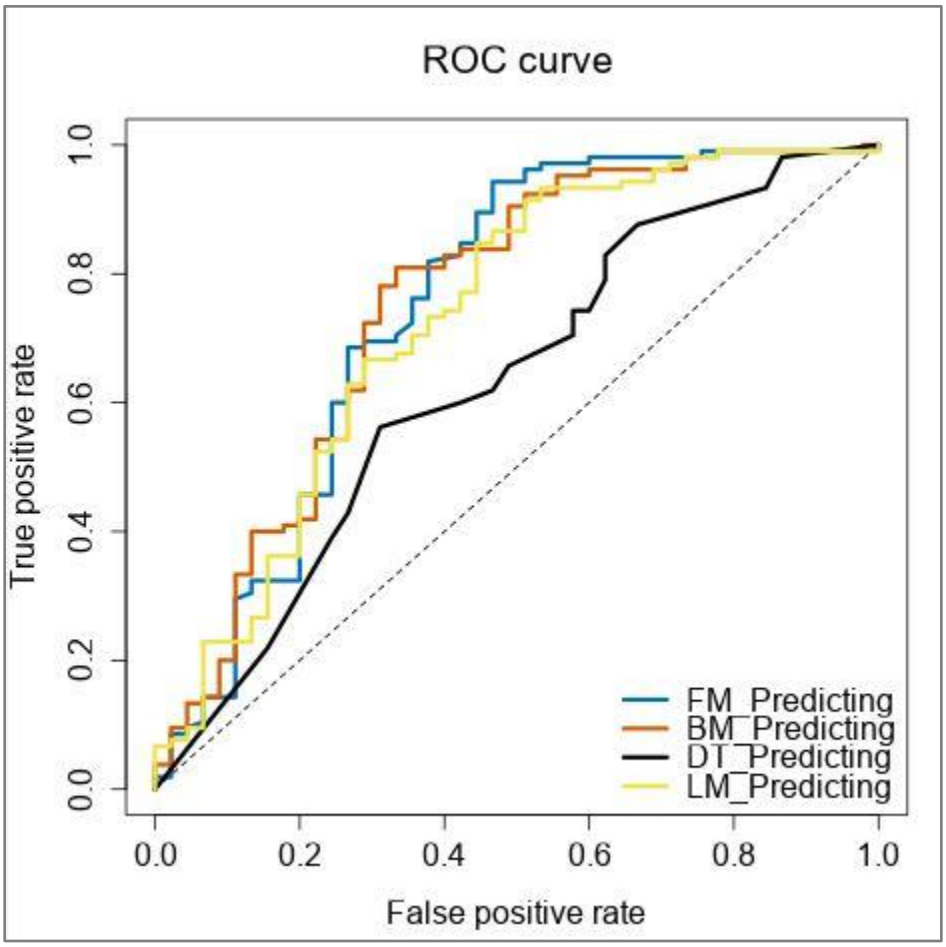
And as we show in Confusion matrix , there are bias in all models .

Step 4: Writeup

Answer these questions:

- 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:
    - Overall Accuracy against your Validation set
    - Accuracies within “Creditworthy” and “Non-Creditworthy” segments
    - ROC graph
    - Bias in the Confusion Matrices

- I choose **Forset Model** , because it have the highest accuracy from other models
- also the Creditworthy by 79% is greater than all Models and Non-Creditworthy segments by 44% is greater than all Models except **Logistic Model** , but overall I see the Forset Model is strong enough compared to other model
- In ROC graph , the Blue liner represent the **Forset Model** , and it is appear better than other models .



.

- The Confusion Matrices below represent the Bias in **Forest Model**

Confusion matrix of FM_Predicting

	Actual_Creditworthy	Actual_Non-Creditworthy
Predicted_Creditworthy	102	25
Predicted_Non-Creditworthy	3	20

- How many individuals are creditworthy?

the individuals are creditworthy is **413** .

The workflow below explain the steps of this decision , from cleaning data to the end in which if individual are creditworthy or not .