Project: Creditworthiness

Step 1: Business and Data Understanding

Key Decisions:

Answer these questions

- What decisions needs to be made?
  There are 500 loan applications, and we need to decide who should receive credit and whom we should reject based on best model

- What data is needed to inform those decisions?
  1. Data on all past applications
  2. 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 use Binary model – Result should be YES or NO

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.

Fields	**Reason for removing or imputing**
Guarantors	low variability, heavily skew towards one type of data so **Removed**
Duration-in-Current-address	A lot of missing data so **Removed**
Age-years	2% missing data were found. **Imputed using median age** because this variable is important for our analysis and can affect other variables as well
Concurrent-Credits	low variability, data is entirely uniform so **Removed**
Occupation	low variability, data is entirely uniform so **Removed**
No-of-dependents	low variability, heavily skew towards one type of data so **Removed**
Telephone	irrelevant to creditworthiness so **Removed**
Foreign-Worker	low variability, heavily skew towards one type of data so **Removed**

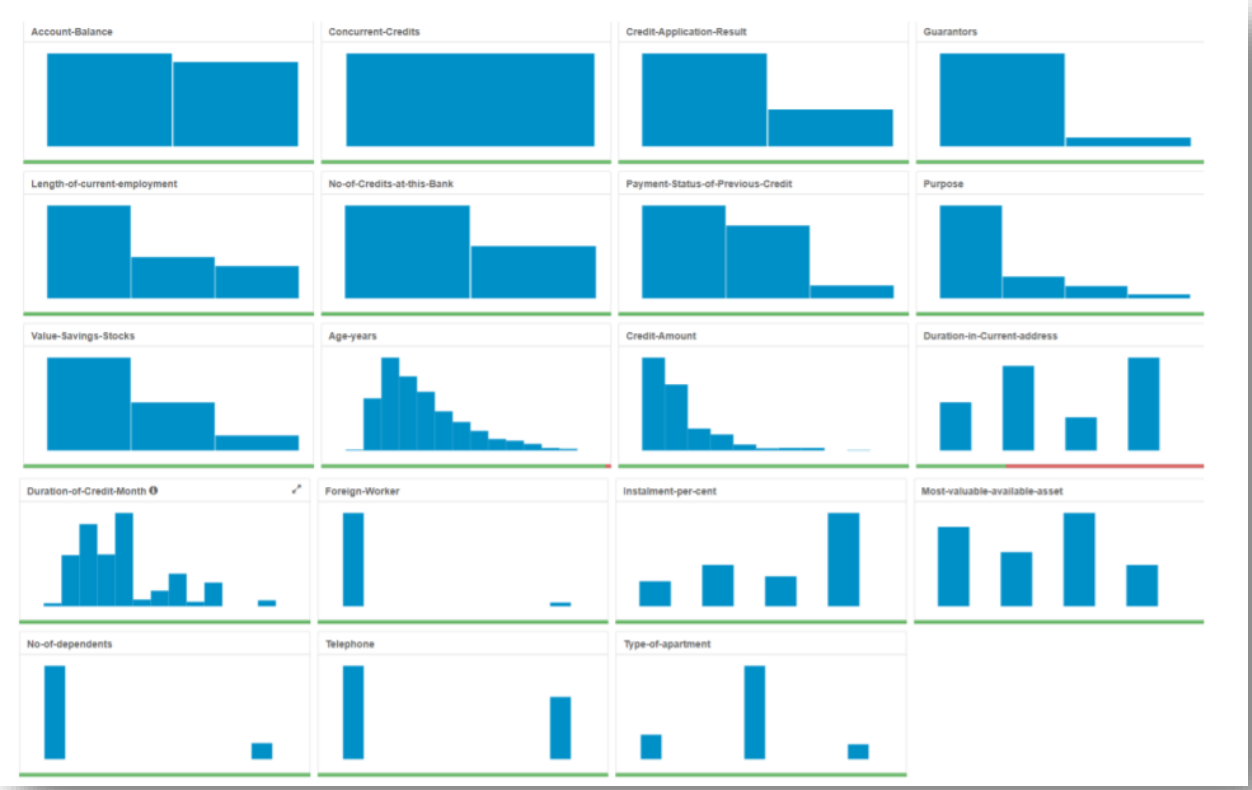


Fig : Summary of all fields

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.

- 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?

1. **Logistic Regression:**

   Important variables : Account Balance, Credit Amount and Purpose

   The overall accuracy : 76%.

   The rate to predict Creditworthy correctly : 87.6%.

   The rate to predict Non-Creditworthy : 48.8%.

Report

Report for Logistic Regression Model Log_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 .

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

(Dispersion parameter for binomial taken to be 1)

Confusion matrix of Log_Stepwise

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

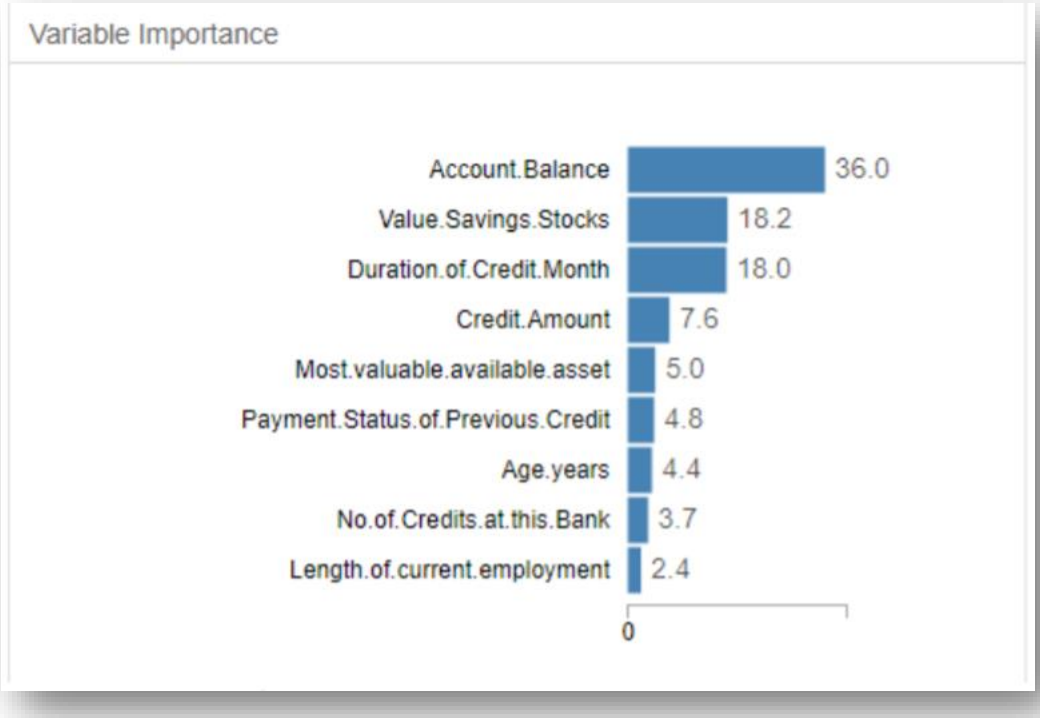
2. Decision Tree

Important variables : Account Balance, Value Saving Stocks and Duration of Credit Month.

The overall accuracy : 74.6%

The rate to predict Creditworthy correctly : 86.6%.

The rate to predict Non-Creditworthy : 46.6%.



Confusion matrix of Tree

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

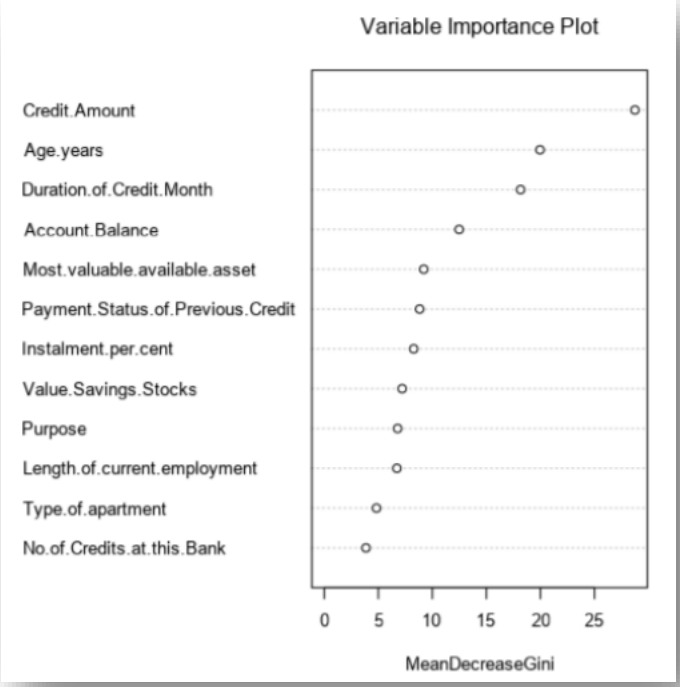
3. Forest Model

Important variables : Credit Amount, Age years and Duration of Credit Month.

The overall accuracy : 79.3%

The rate to predict Creditworthy correctly : 97.7%.

The rate to predict Non-Creditworthy : 37.7%



Confusion matrix of Forest

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

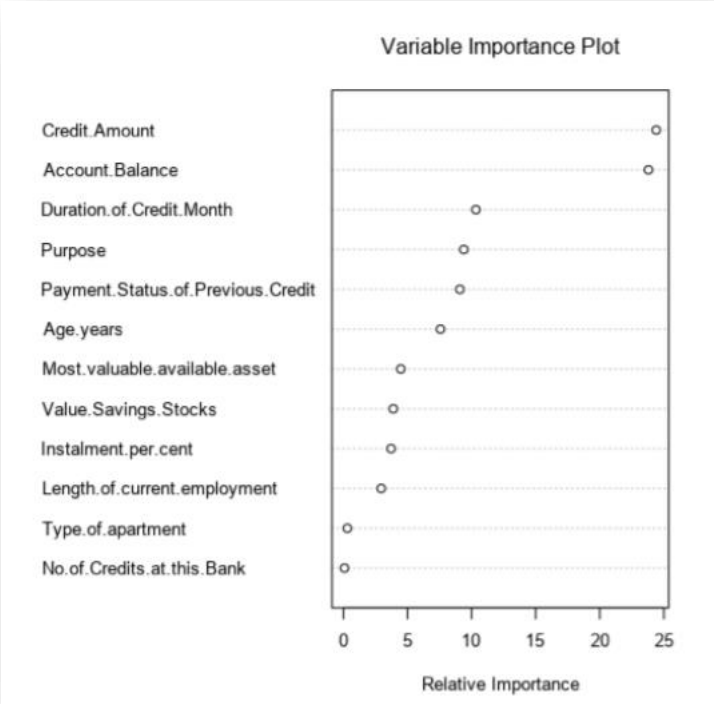
4. Boosted Model

Important variables : Credit Amount, Account Balance.

The overall accuracy : 78.6%

The rate to predict Creditworthy correctly : 96.6%.

The rate to predict Non-Creditworthy : 37.7%.



Confusion matrix of Boosted

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

Step 4: Writeup

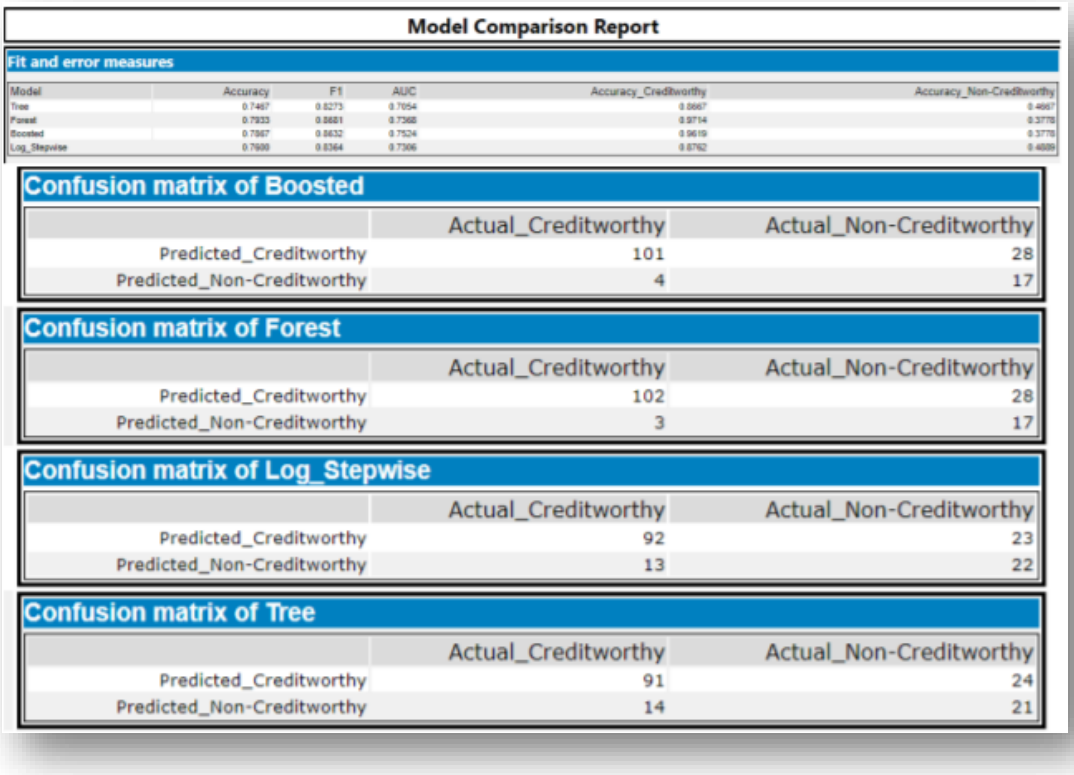
Answer these questions:

- 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

Chosen : **The Forest Model**

- It has the highest overall accuracy among all models with rate of **79.3**%.
- It has the highest accuracy for predicting "Creditworthy" with rate of **97.1%** which is extremely good in order to not overlook the potential opportunities.
- It has a rate **of 37.7%** to predict "Non-Creditworthy". It's a low rate which could lead us giving loans to Non-Creditworthy. However, such thing can be avoided with some extra procedures. Our priority is to find out the "Creditworthy" because if we ignore them, we will lose some opportunities.

ROC graph shows the Forest Model reaching high True Positive rate and taking area under the curve above other models. This is a good indicator.



Model Comparison Report

Fit and error measures

Model	Accuracy	F1	AUC	Accuracy_Creditworthy	Accuracy_Non-Creditworthy
Tree	0.7467	0.8273	0.7054	0.8667	0.4667
Forest	0.7933	0.8681	0.7368	0.9714	0.3778
Boosted	0.7867	0.8632	0.7524	0.9619	0.3778
Log_Stepwise	0.7600	0.8364	0.7306	0.8762	0.4889

Confusion matrix of Boosted

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

Confusion matrix of Forest

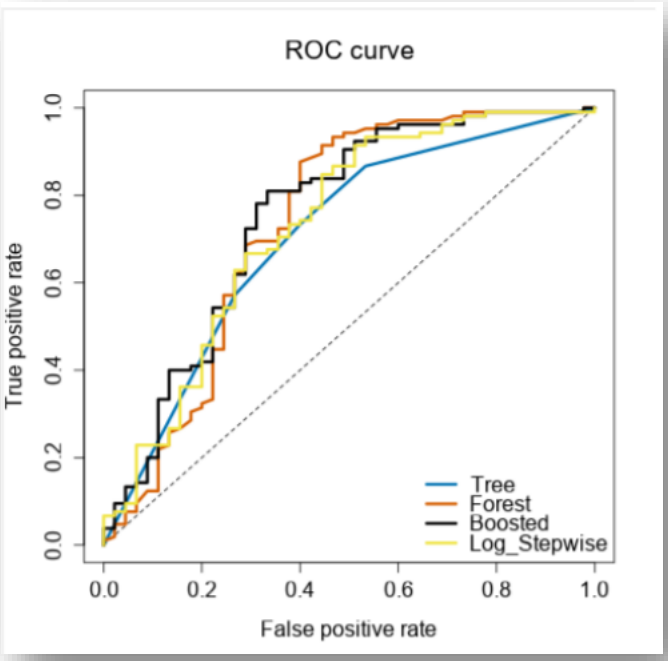
	Actual_Creditworthy	Actual_Non-Creditworthy
Predicted_Creditworthy	102	28
Predicted_Non-Creditworthy	3	17

Confusion matrix of Log_Stepwise

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

Confusion matrix of Tree

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



- How many individuals are creditworthy?

Record	Credit_Worthiness	Count
1	No	92
2	Yes	408

There are 408 people who are creditworthy.

Alteryx flow: