Project: Creditworthiness

Step 1: Business and Data Understanding

Provide an explanation of the key decisions that need to be made. (250 word limit)

Key Decisions:

Answer these questions

- **What decisions needs to be made?**

  The bank needs to classify the customers’ new loan applications can be approved for a loan or not. .

- **What data is needed to inform those decisions?**

The attributes of the existing customers who are already classified as creditworthy or not. The attributes such as

- ✧ Credit-Application-Result
- ✧ 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

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

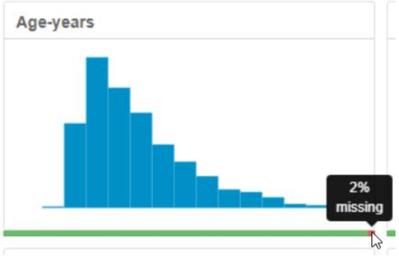
  Since the Target variable is to categorize the customer as Creditworthy or not, a Binary model should be used.

Step 2: Building the Training Set

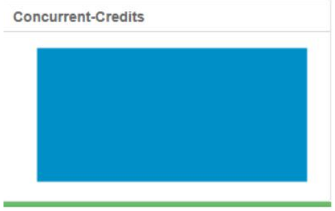
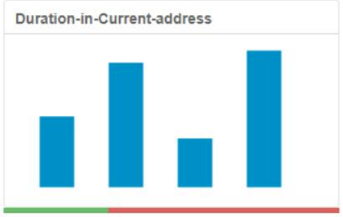
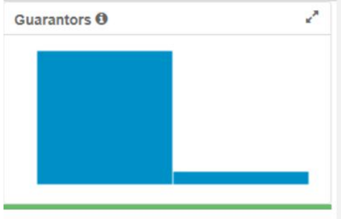
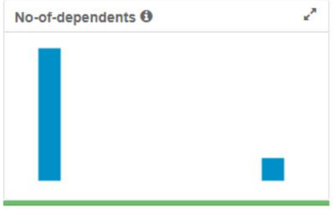
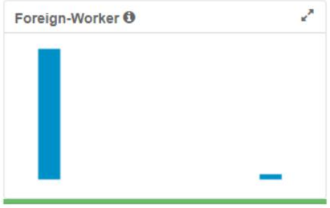
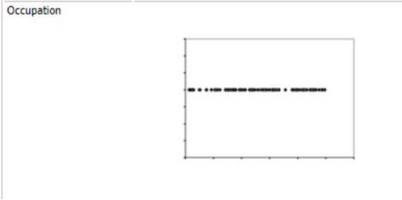
Answer this question:

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

Imputed fields:

Field Name	Reason	Visualization
Age-Years	Only 2% data missing-so the median of the Age-years column is imputed	

Removed fields:

Field Name	Reason	Visualization
Concurrent-Credits	Uniformity	
Duration in current address	Too many missing data	
Guarantors	Low variability	
Number of dependents	Low variability	
Foreign worker	Low variability	
Occupation	Low variability	
Telephone	No logical reason for including the variable	

Step 3: Train your Classification Models

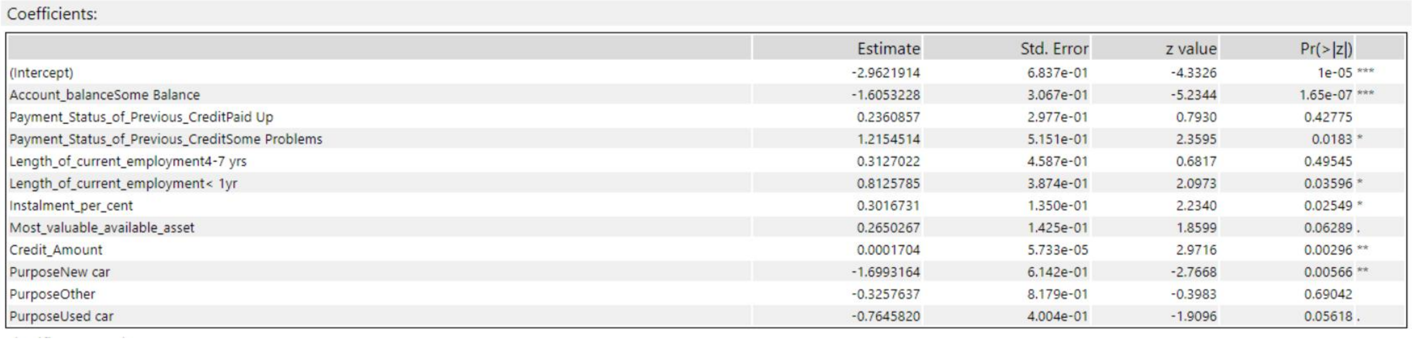
First, create your Estimation and Validation samples where 70% of your dataset should go to Estimation and 30% of your entire dataset should be reserved for Validation. Set the Random Seed to 1.

Create all of the following models: Logistic Regression, Decision Tree, Forest Model, Boosted Model

*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.

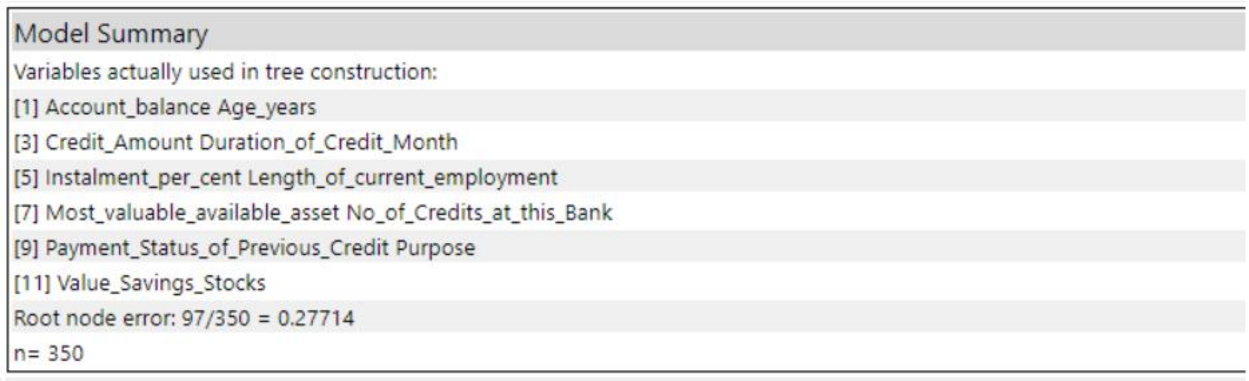
  Logistic Regression Model



Coefficients:

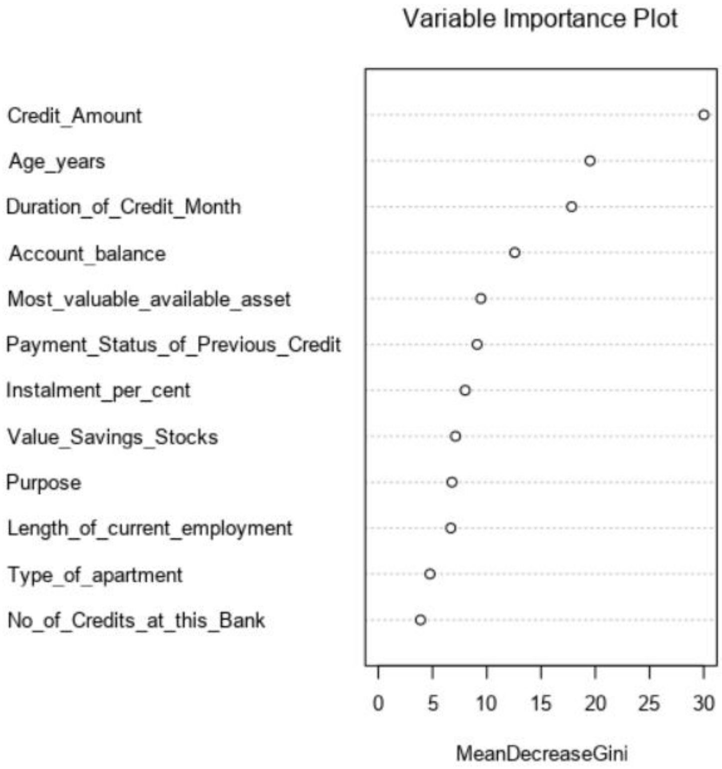
	Estimate	Std. Error	z value	Pr(>|z|)	
(Intercept)	-2.9621914	6.837e-01	-4.3326	1e-05	***
Account_balanceSome Balance	-1.6053228	3.067e-01	-5.2344	1.65e-07	***
Payment_Status_of_Previous_CreditPaid Up	0.2360857	2.977e-01	0.7930	0.42775	
Payment_Status_of_Previous_CreditSome Problems	1.2154514	5.151e-01	2.3595	0.0183	*
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	.
Credit_Amount	0.0001704	5.733e-05	2.9716	0.00296	**
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	.

  Decision Tree Model

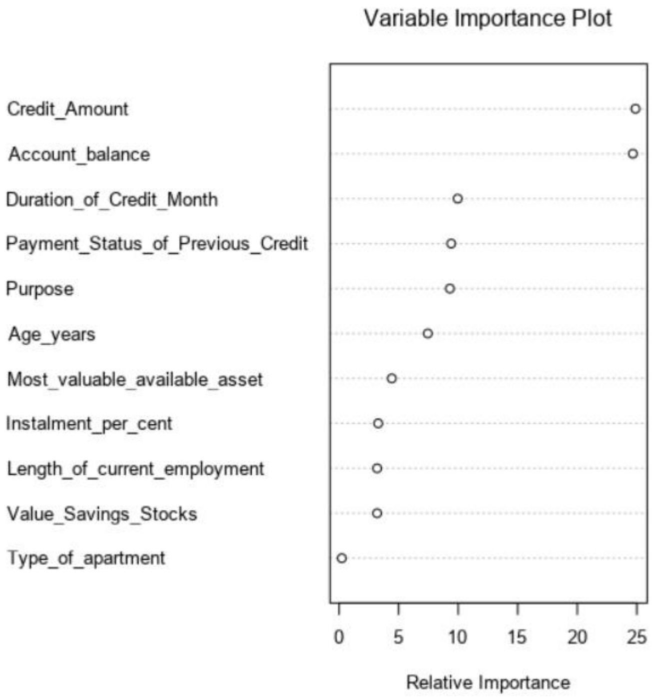


```
Model Summary
Variables actually used in tree construction:
[1] Account_balance Age_years
[3] Credit_Amount Duration_of_Credit_Month
[5] Instalment_per_cent Length_of_current_employment
[7] Most_valuable_available_asset No_of_Credits_at_this_Bank
[9] Payment_Status_of_Previous_Credit Purpose
[11] Value_Savings_Stocks
Root node error: 97/350 = 0.27714
n= 350
```

Random Forest Tree Model



Boosted Model

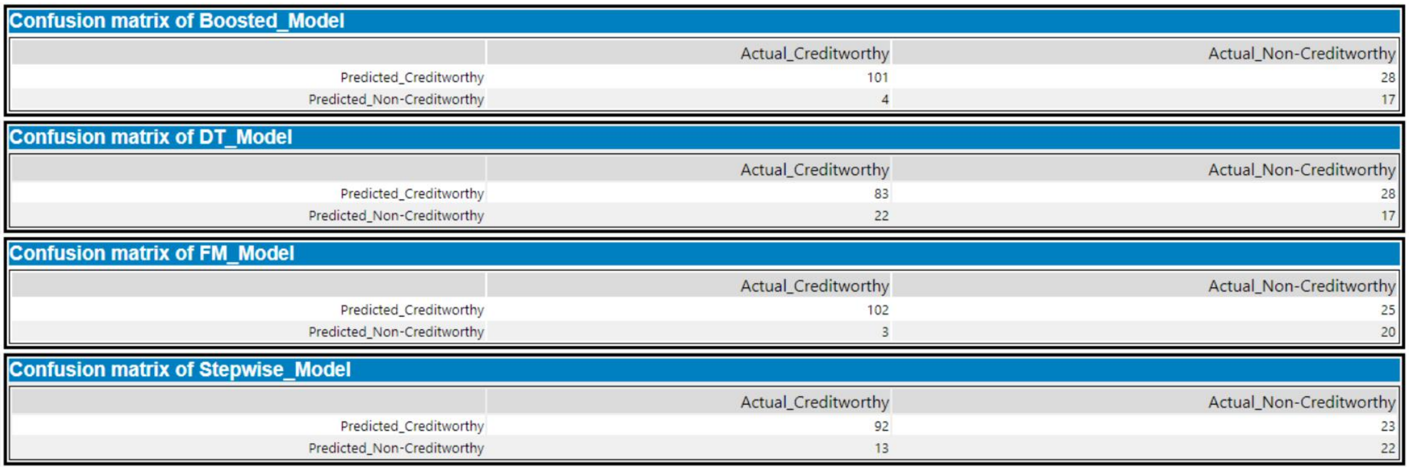


- Validate your model against the Validation set. What was the overall percent accuracy? Show the confusion matrix. Are there any bias seen in the model’s predictions?

Overall accuracy of the models

Model	Overall Accuracy
Logistic Regression-Stepwise Model	0.7600
Decision Tree Model	0.6667
Random Forest Tree Model	0.8133
Boosted Model	0.7867

Confusion Matrix



Confusion matrix of Boosted_Model

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

Confusion matrix of DT_Model

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

Confusion matrix of FM_Model

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

Confusion matrix of Stepwise_Model

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

Step 4: Writeup

Answer these questions:

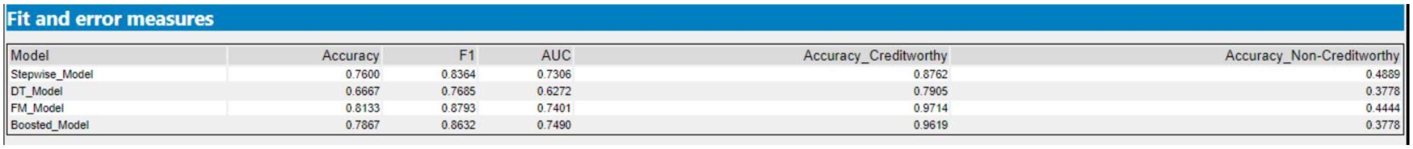
- Which model did you choose to use? Please justify your decision using **all** of the following techniques. Please only use these techniques to justify your decision:
  - Overall Accuracy against your Validation set
  - Accuracies within “Creditworthy” and “Non-Creditworthy” segments
  - ROC graph
  - Bias in the Confusion Matrices

I choose Random Forest Model.

Reasons:

The Random Forest Model has the highest overall accuracy (0.7571) with compared to other models.

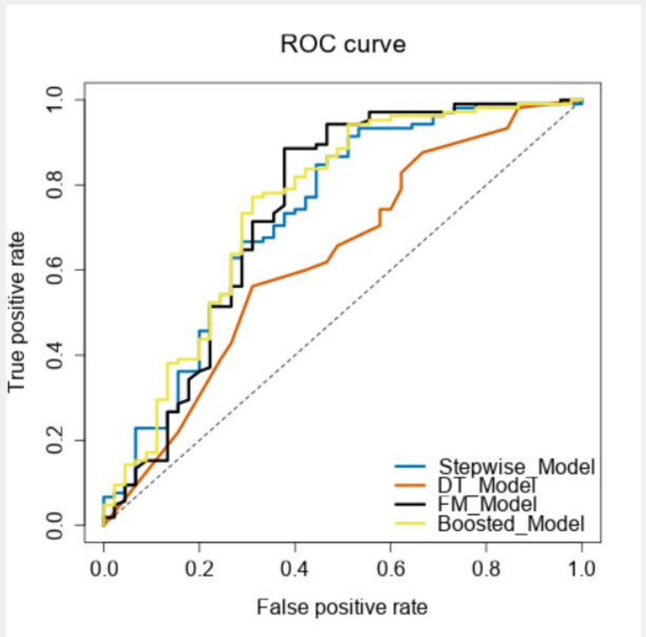
And also the Accuracies within “Creditworthy” and “Non-Creditworthy” segments are higher than other models. Please refer to the below screen capture for more details.



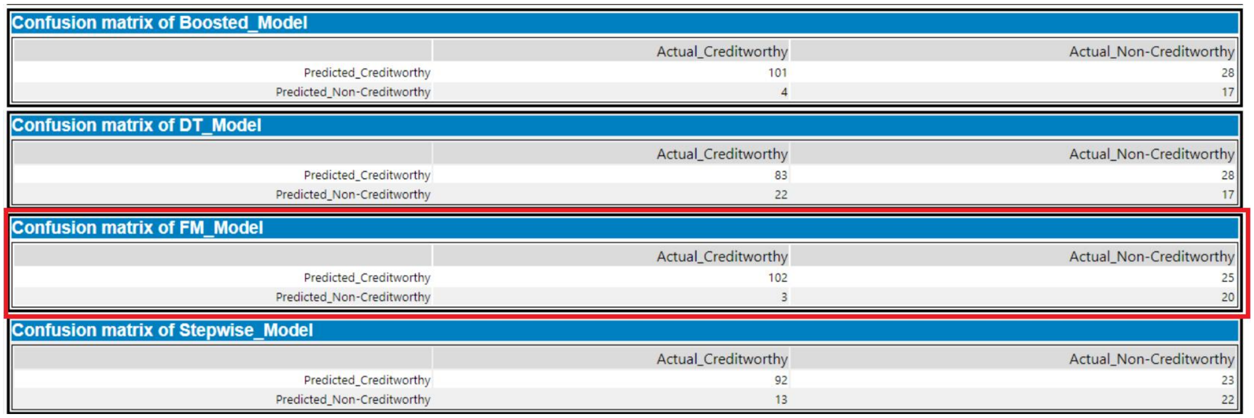
Fit and error measures

Model	Accuracy	F1	AUC	Accuracy_Creditworthy	Accuracy_Non-Creditworthy
Stepwise_Model	0.7600	0.8364	0.7306	0.8762	0.4889
DT_Model	0.6667	0.7685	0.6272	0.7905	0.3778
FM_Model	0.8133	0.8793	0.7401	0.9714	0.4444
Boosted_Model	0.7867	0.8632	0.7490	0.9619	0.3778

The ROC graph



With compared to the other models Random Forest model, the confusion matrix perform better.



Confusion matrix of Boosted_Model

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

Confusion matrix of DT_Model

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

Confusion matrix of FM_Model

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

Confusion matrix of Stepwise_Model

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

PPV and NPV comparison to check the bias of the model.

Model	PPV	NPV
Boosted Model	0.7829	0.8095
Decision Tree Model	0.7477	0.4359
Random Forest Model	0.8031	0.8696
Logistic Regression Model	0.8000	0.6286

Based on the PPV and NPV values there is least bias in the forest model.

- **How many individuals are creditworthy?**

  The number of individuals that are creditworthy is 407