Project4: Predicting Default Risk

PROJECT SUBMISSION WRITTEN BY CHIDIEBERE STEPHEN NWOSU

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 decision needs to be made?

The decision needed to be made is, I need to systematically evaluate an efficient solution to whether or not the new applicants are creditworthy for a loan.

What data is needed to inform those decisions?
Answer:
We have two datasets:
Credit-data-training.xlsx- This file contains all credit approvals from my past loan applicant the bank has ever approved
Customer-to-score.xlsx- This is the new set of customers that I need to score on the classification model I will create. These files contain the following:

- Account-Balance
- Duration-of-Credit-Month
- Payment-Status-of-Previous-Credit
- Purpose
- Credit-Amount
- Value-Saving-Stocks
- Length-of-Current-Employment
- Installment-per-cent
- Guarantors
- Duration-in-current-address
- Most-valuable-available-assets
- Age-years
- Current-credit
- Type-of-apartment
- No-of-Credit-at-this-Bank

- Occupation
- No-of-dependent
- Telephone
- Foreign-worker
- Credit-application-result

What kind of model (Continuous, Binary, Non-Binary, Time-Series) do we need to use to help make these decisions?
Answer:
Since we are interested to know whether the new customer is creditworthy or not creditworthy, Binary Classification model is used to help make this decision.

Step 2: Building the Training Set

1. For numerical data fields, are there any fields that highly correlate with each other? We need to verify if any possible group of the predictor variables are highly correlated with each other or not. From the 'Association Analysis' we can verify if they are highly correlated or not buy using the correlated plot matrix and the scatter plot side by side.

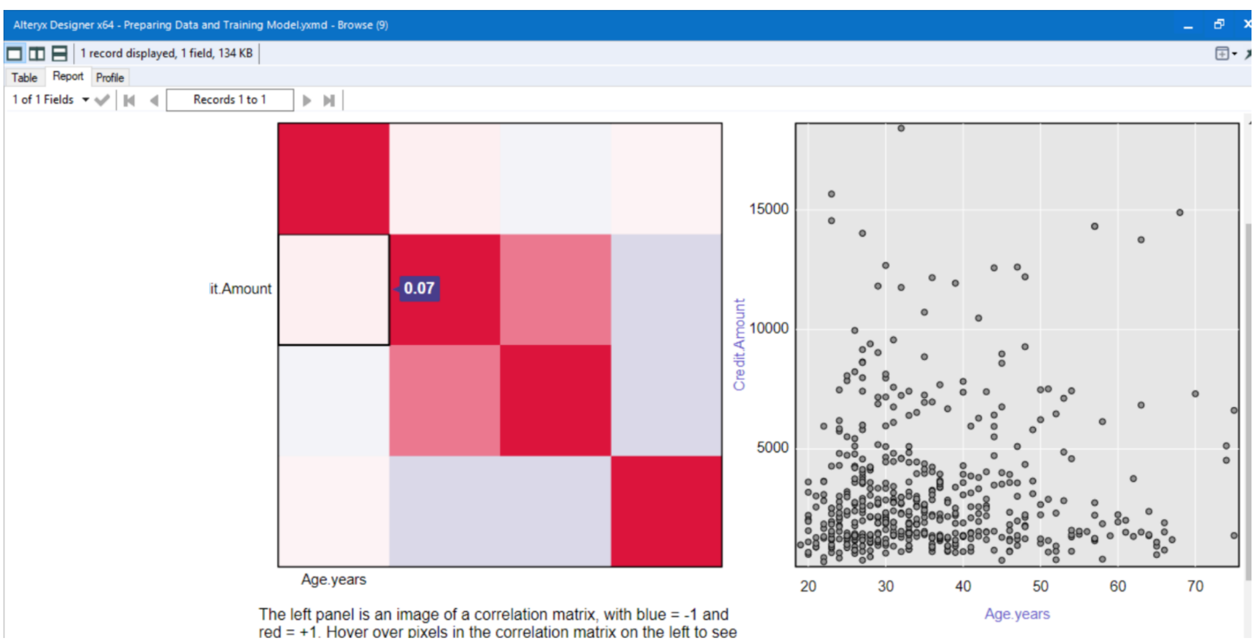


Fig1: Shows the multicollinearity Identification (Correlated plot matrix and Scatter plot)
Correlated plot matrix and scatter plot clearly shows that there is no highly correlated predictor variable. Correlated plot 0.07 value is valid and the scatter plot shows randomness.

2. Missing Data

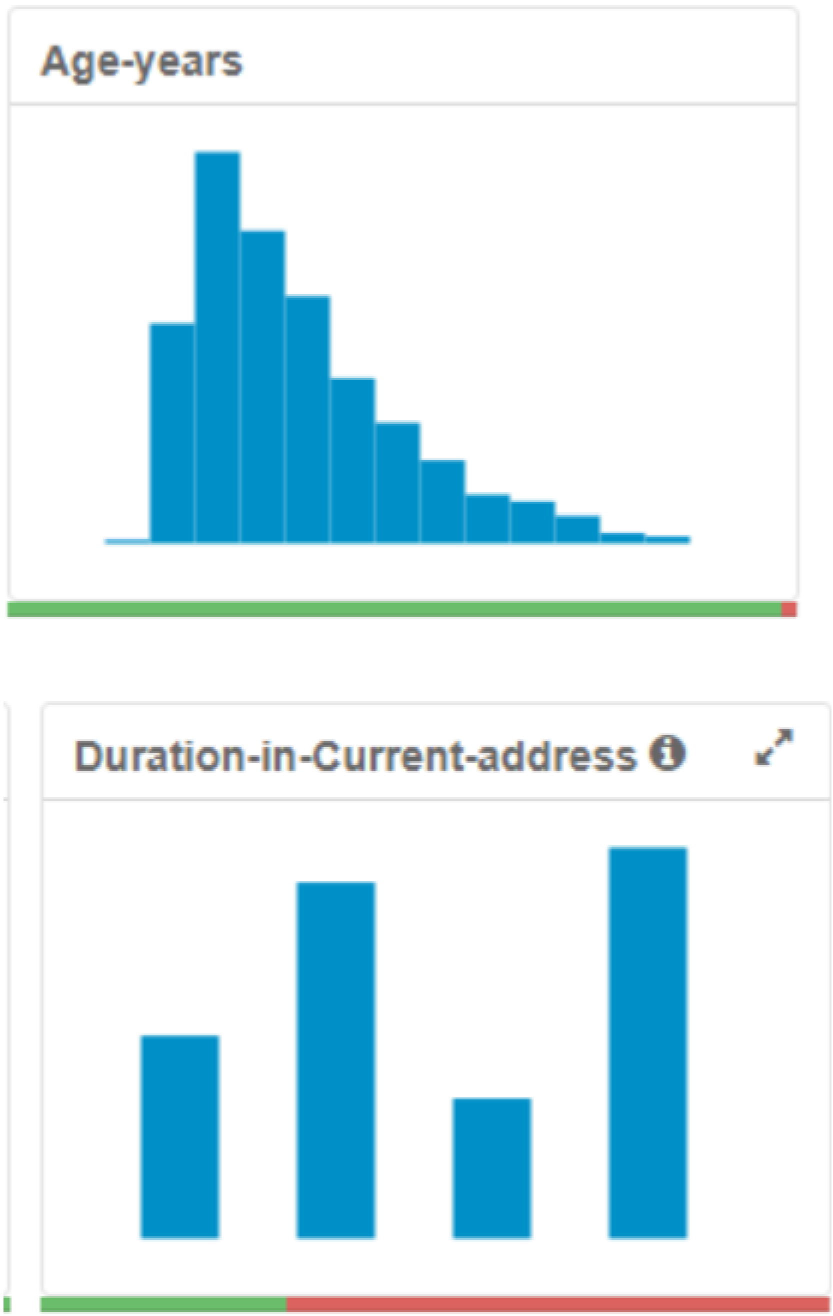


Figure2: Age-years and Duration-in-Current-address

The above visualization in Figure2, shows we have two fields with missing data.

3. Fields with low variability

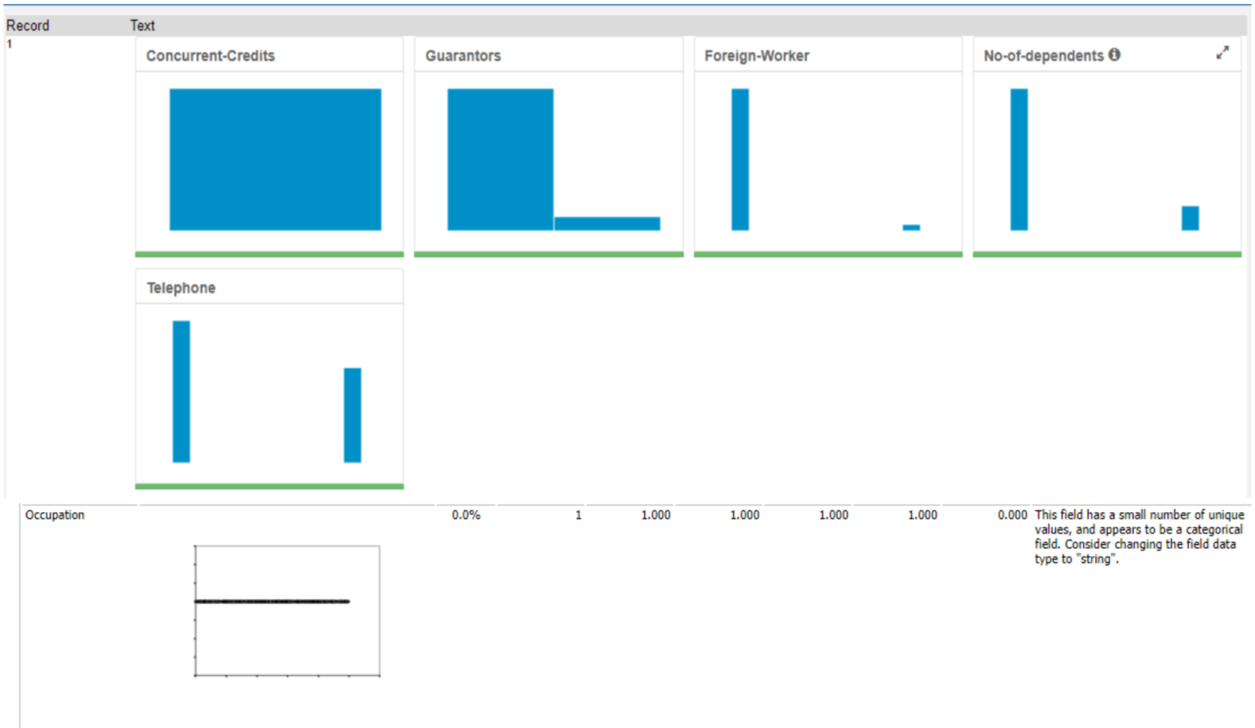


Figure3: Shows low variability with 6 fields

Answer this question

In your clean up process, which fields did you remove or input? Please justify why you remove or imputed these fields. Visualization are encouraged.

Answer:

In Figure2 above, Duration-in-Current-address with 69% shouldn't filter this out since it will reduce our dataset way too much, as well if we input for such a large number of missing records, it will most likely cause a bias in our dataset. Duration-in-Current-address should be removed completely.

Age-years, indicate that there are about 2% missing data. Considering the fact that age is one of the important predictor variables, the missing ages should be replaced by impute age median.

Fields with low variability, their 6 fields with low variability, and they are Concurrent-credit, Guarantors, Foreign-Workers, No-of-dependents, Telephone and Occupation, they should all be removed because they are highly skewed towards one direction and the Concurrent-credits, shows no variation in data.

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.

Answer:

Figure4: Logistic Regression Model and P-Value

The most important Significant predictor variables are; Account-Balance, Purpose and Credit-Amount

Report for Logistic Regression Model LR_Loan_Evaluate

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

Figure4: Logistic Regression Model

Figure5: Decision Tree

From the Decision Tree below, the most important significant variables are; Account Balance, Value Saving Stock and Duration of Credit Month.

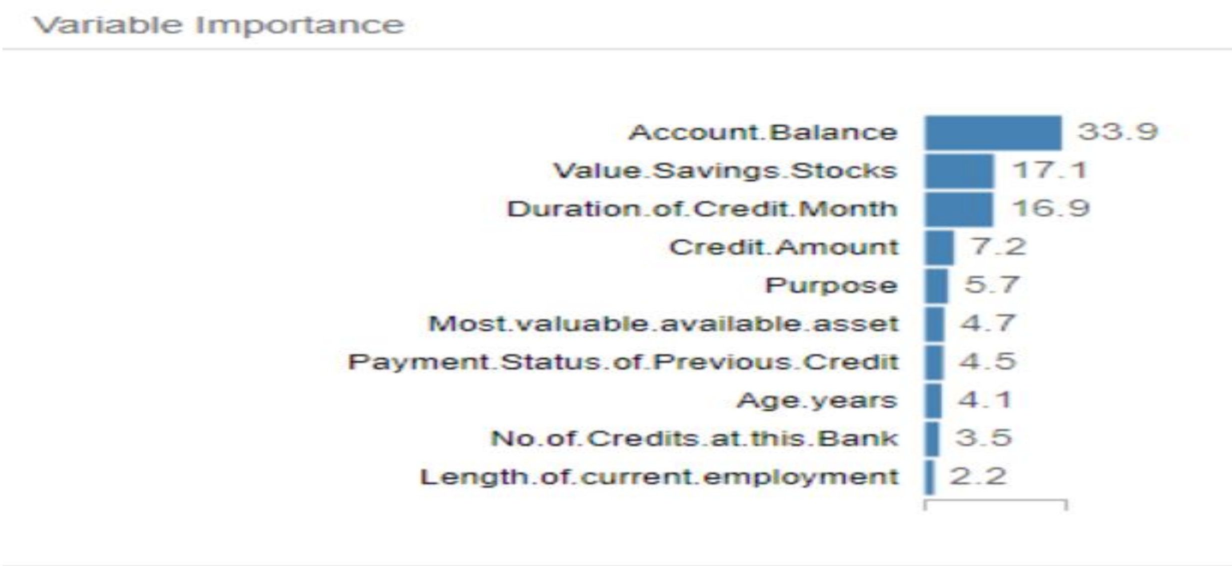


Figure5: Decision Tree

Figure6: Forrest Model

From the Forest model plot below, the most important significant variables are; Credit Amount, Age years and Duration of Credit Month.

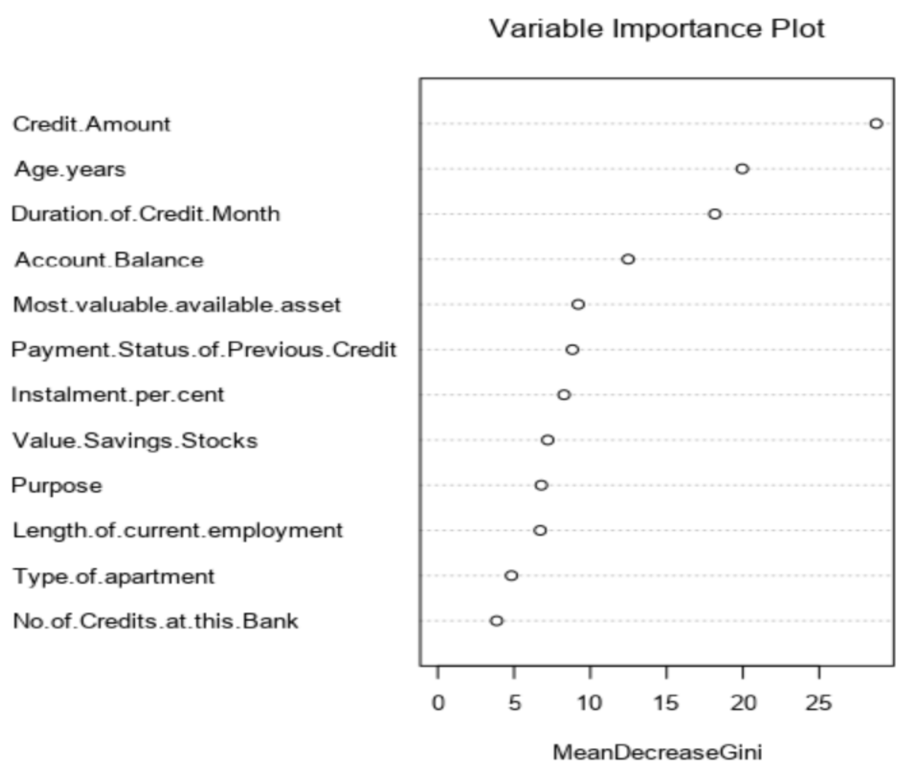


Figure6: Forest Model

Figure7: Boosted Model

From the Boosted Model plot below, the most important significant variables are; Account Balance, Credit Amount and Duration of Credit Month

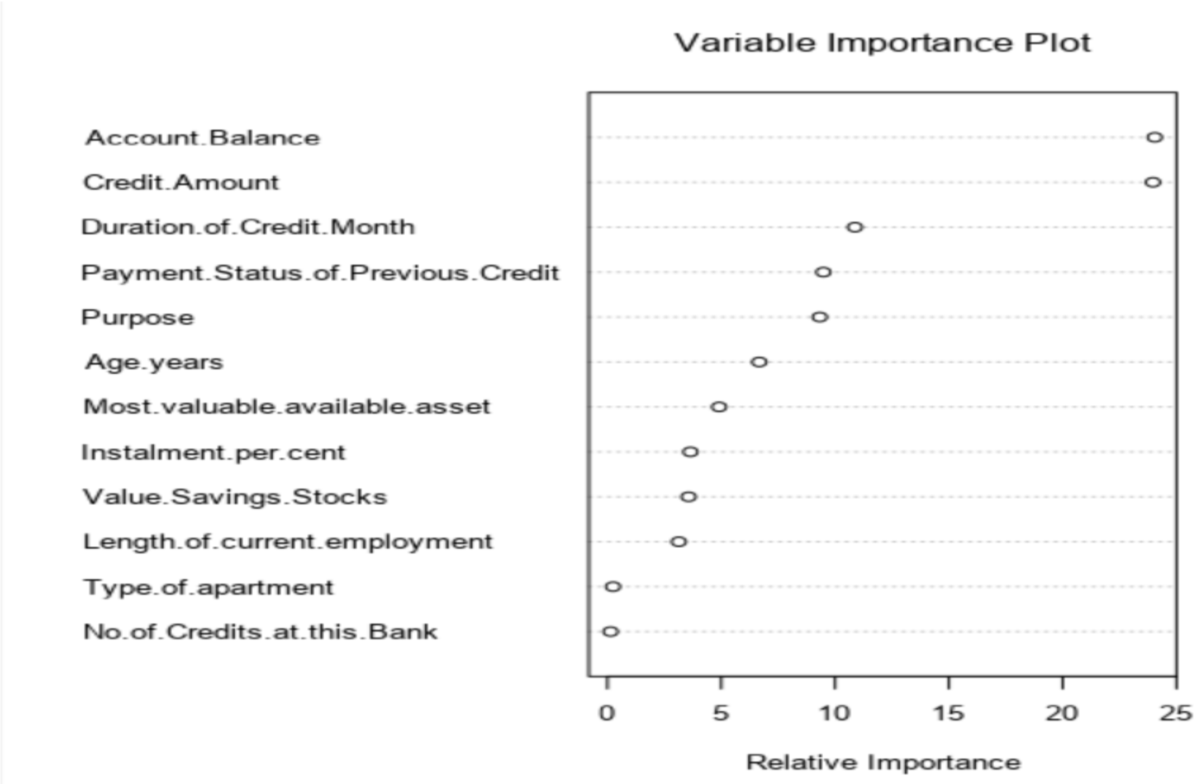


Figure7: Boosted Model

- 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?

Answer:

Logistic Regression Model:

Fit and error measures

Model	Accuracy	F1	AUC	Accuracy_Creditworthy	Accuracy_Non-Creditworthy
DT_Loan_Evaluate	0.7467	0.8304	0.7035	0.8857	0.4222
FM_Loan_Evaluate	0.7933	0.8681	0.7368	0.9714	0.3778
BM_Loan_Evaluate	0.7933	0.8670	0.7505	0.9619	0.4000
LR_Loan_Evaluate	0.7600	0.8364	0.7306	0.8762	0.4889

Confusion matrix of LR_Loan_Evaluate

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

From the Logistic Regression Model (LR_Loan_Evaluate) above, we can see that the overall accuracy result is 76%, the accuracy for predicting Creditworthy is approximately 88%, while the accuracy for predicting non-Creditworthy is about 49% approx. In this case, the model is biased towards Creditworthy than non-Creditworthy.

Decision Tree

From the Decision Tree (DT_Loan_Evaluate) below, we can see that the overall accuracy is approximately 75%, the accuracy for predicting Creditworthy is approximately 89%, while the accuracy for predicting non-Creditworthy is about 42%. In this case, the Decision tree model is biased towards Creditworthy than non-Creditworthy.

Fit and error measures

Model	Accuracy	F1	AUC	Accuracy_Creditworthy	Accuracy_Non-Creditworthy
DT_Loan_Evaluate	0.7467	0.8304	0.7035	0.8857	0.4222
FM_Loan_Evaluate	0.7933	0.8681	0.7368	0.9714	0.3778
BM_Loan_Evaluate	0.7933	0.8670	0.7505	0.9619	0.4000
LR_Loan_Evaluate	0.7600	0.8364	0.7306	0.8762	0.4889

Confusion matrix of DT_Loan_Evaluate

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

Forest Model

From the Forest Model (FM_Loan_Evaluate) below, we can see that the overall accuracy is 79%, the accuracy for predicting Creditworthy is about 97%, and the accuracy for predicting non-Creditworthy is approximately 38%. In this case, the Forest model is biased or skew towards the Creditworthy than non-Creditworthy.

Fit and error measures

Model	Accuracy	F1	AUC	Accuracy_Creditworthy	Accuracy_Non-Creditworthy
DT_Loan_Evaluate	0.7467	0.8304	0.7035	0.8857	0.4222
FM_Loan_Evaluate	0.7933	0.8681	0.7368	0.9714	0.3778
BM_Loan_Evaluate	0.7933	0.8670	0.7505	0.9619	0.4000
LR_Loan_Evaluate	0.7600	0.8364	0.7306	0.8762	0.4889

Confusion matrix of FM_Loan_Evaluate

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

Boosted Model

From the Boosted Model (BM_Loan_Evaluate) below, accuracy is about 79%, the accuracy for predicting Creditworthy is 96% and the accuracy for predicting non-Creditworthy is 40%. In this case the Boosted model is biased towards the Creditworthy than non-Creditworthy.

Fit and error measures

Model	Accuracy	F1	AUC	Accuracy_Creditworthy	Accuracy_Non-Creditworthy
DT_Loan_Evaluate	0.7467	0.8304	0.7035	0.8857	0.4222
FM_Loan_Evaluate	0.7933	0.8681	0.7368	0.9714	0.3778
BM_Loan_Evaluate	0.7933	0.8670	0.7505	0.9619	0.4000
LR_Loan_Evaluate	0.7600	0.8364	0.7306	0.8762	0.4889

Confusion matrix of BM_Loan_Evaluate

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

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

Answer:

The 4 models; Logistic Regression, Decision Tree, Forest, and Boosted model, can be compared side by side, by looking at the following.

Model Comparison Report

Fit and error measures

Model	Accuracy	F1	AUC	Accuracy_Creditworthy	Accuracy_Non-Creditworthy
DT_Loan_Evaluate	0.7467	0.8304	0.7035	0.8857	0.4222
FM_Loan_Evaluate	0.7933	0.8681	0.7368	0.9714	0.3778
BM_Loan_Evaluate	0.7933	0.8670	0.7505	0.9619	0.4000
LR_Loan_Evaluate	0.7600	0.8364	0.7306	0.8762	0.4889

Confusion matrix of BM_Loan_Evaluate

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

Confusion matrix of DT_Loan_Evaluate

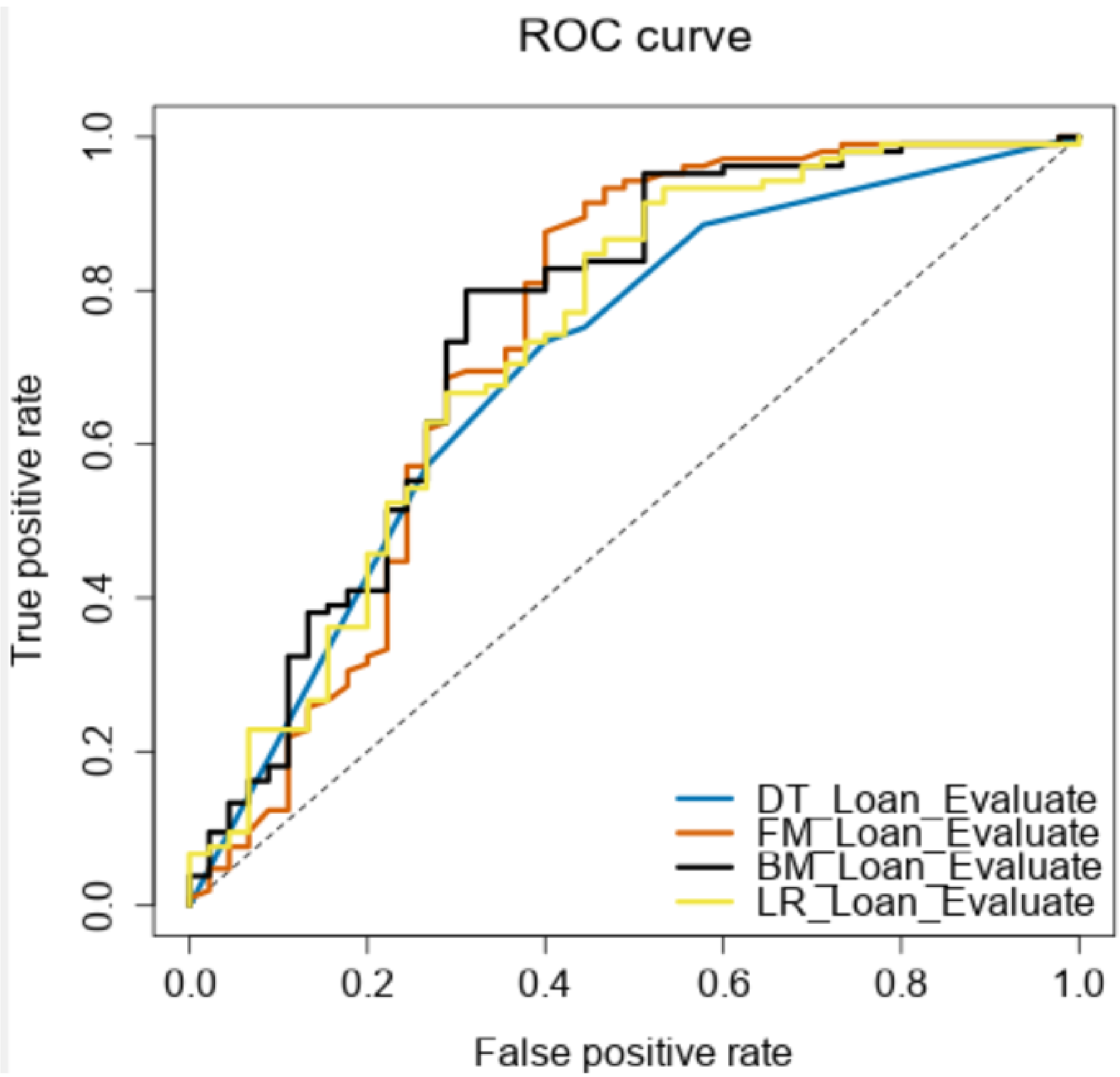
	Actual_Creditworthy	Actual_Non-Creditworthy
Predicted_Creditworthy	93	26
Predicted_Non-Creditworthy	12	19

Confusion matrix of FM_Loan_Evaluate

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

Confusion matrix of LR_Loan_Evaluate

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

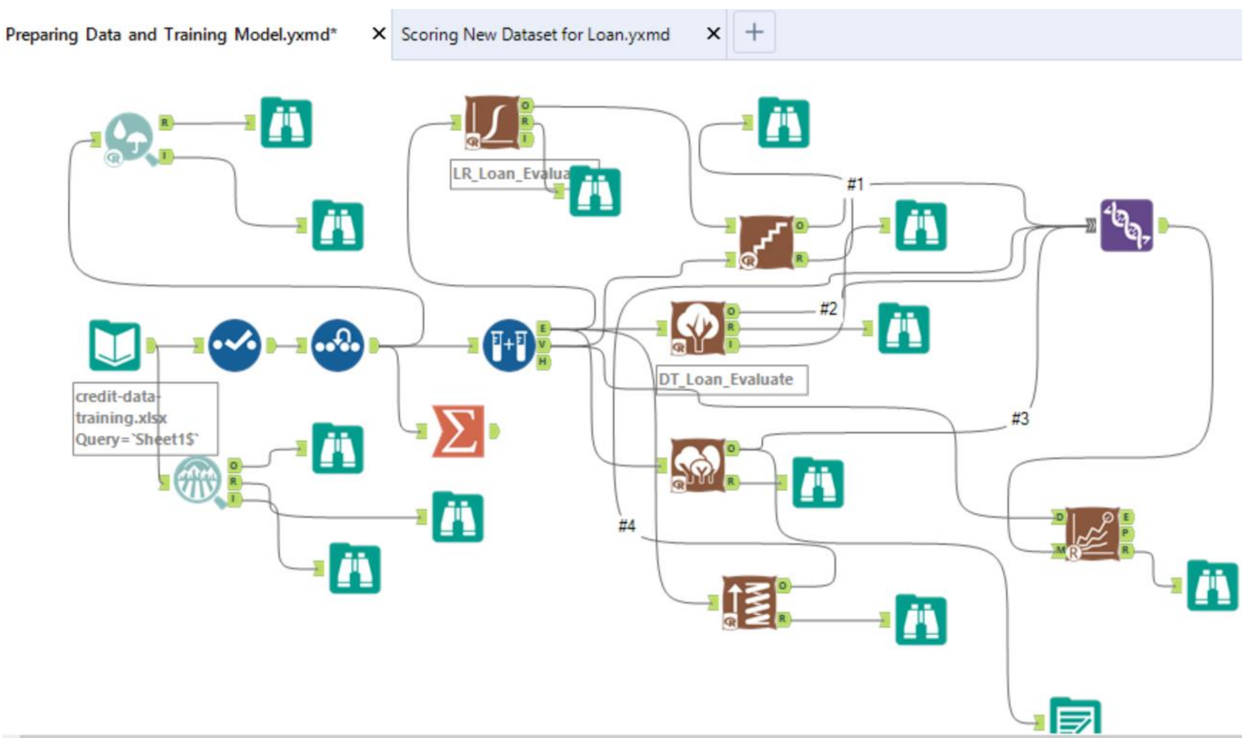


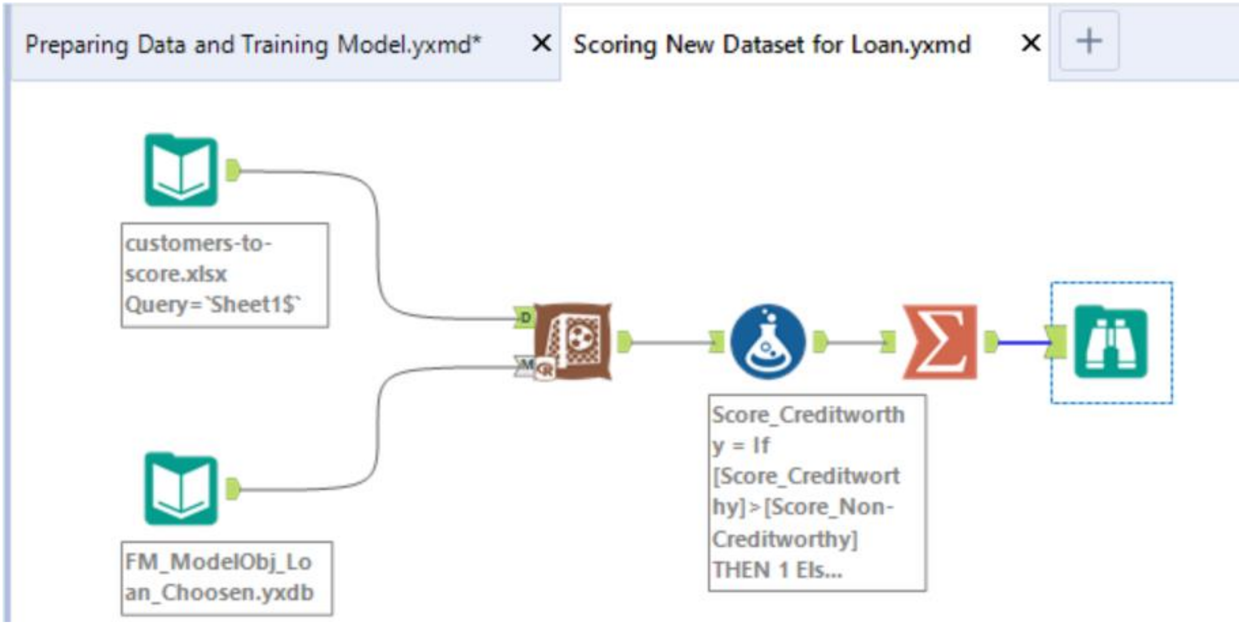
Carefully looking at the 4 models above, we can prove that the overall accuracy, for both Forest model and Boosted model are highest at 79%. For predicting Creditworthy, we can also see that the Forest model has the highest percentage Accuracy_Creditworthy at 97.14%, followed by Boosted model at 96.19%.

The ROC Curve plot, shows that the Forest model has the highest true positive rate value. Since we are interested in predicting creditworthiness the new applicant, Forest model should be chosen as the best fit model.

- How many individuals are creditworthy?

Since Forest model is chosen as the best fit model, then it should be applied to the new dataset(customers-to-score). The individual that are creditworthy is seen with following result below.





	Sum_Score_Creditworthy	Sum_Score_Non-Creditworthy
1	408	92

408 individuals are creditworthy.