Step 1: Business and Data Understanding

Key Decisions:

- What decisions needs to be made?
  The decision needs to be made is whether the bank should approve or deny 500 new loan applications.

- What data is needed to inform those decisions?
  Historical data of all credit approvals the bank as ever completed. data should include fields such as credit application results, purpose of loan, credit amount, etc.

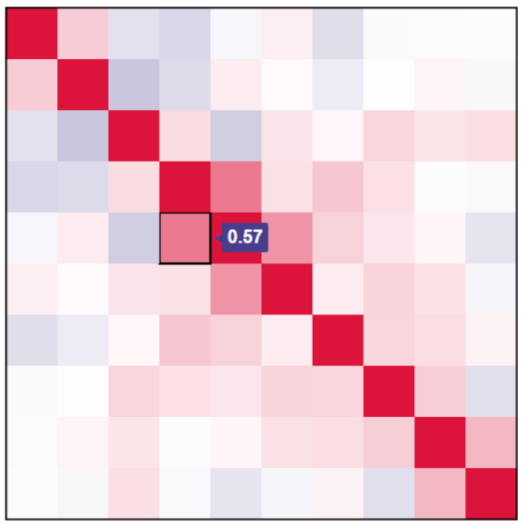
  The data of new loan applications to be determined to be approved. The data should include the same fields as the historical data.

- What kind of model (Continuous, Binary, Non-Binary, Time-Series) do we need to use to help make these decisions?
  The result of each loan application should be approved or denied, so Binary models should be used to make decisions.

Step 2: Building the Training Set

- For numerical data fields, are there any fields that highly-correlate with each other? The correlation should be at least .70 to be considered “high”.

  By looking at the Correlation Matrix generated by Association Analysis Tool, there wasn’t any fields that highly-correlated with each other.



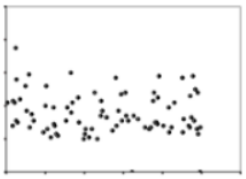
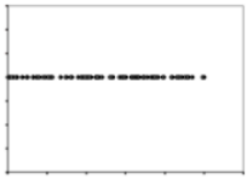
- Are there any missing data for each of the data fields? Fields with a lot of missing data should be removed
  Duration-in-Current-address field is missing 68.8% data.
  Age-years field is missing 2.4% data.

- Are there only a few values in a subset of your data field? Does the data field look very uniform (there is only one value for the entire field?). This is called “low variability” and you should remove fields that have low variability.
  Concurrent-Credits field has only 1 unique value for the entire field.
  Occupation field has only 1 unique value for the entire field.
  Guarantors field has majority of data skewed towards “None”.
  Foreign worker field has the majority of data skewed towards 1.
  No-of-dependents field has the majority of data skewed towards 1.

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

  Fields removed:
  Duration-in-Current-address field is missing 68.8% data.
  Concurrent-Credits field has only 1 unique value for the entire field.
  Occupation field has only 1 unique value for the entire field.
  Guarantors field has majority of data skewed towards “None”.
  Foreign worker field has the majority of data skewed towards 1.
  No-of-dependents field has the majority of data skewed towards 1.
  Telephone field was removed because it shouldn’t have too much effect on whether a loan should be approved or not.

  Field imputed:
  Age-years field is missing 2.4% data, and data was imputed using the median of entire data field.

Name	Plot	% Missing	Unique Values	Min	Mean	Median	Max	Std Dev
Age-years		2.4%	54	19.000	35.637	33.000	75.000	11.502
Occupation		0.0%	1	1.000	1.000	1.000	1.000	0.000

Step 3: Train your Classification Models

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.

- 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

These are the 4 most significant predictor variables for logistic regression model.

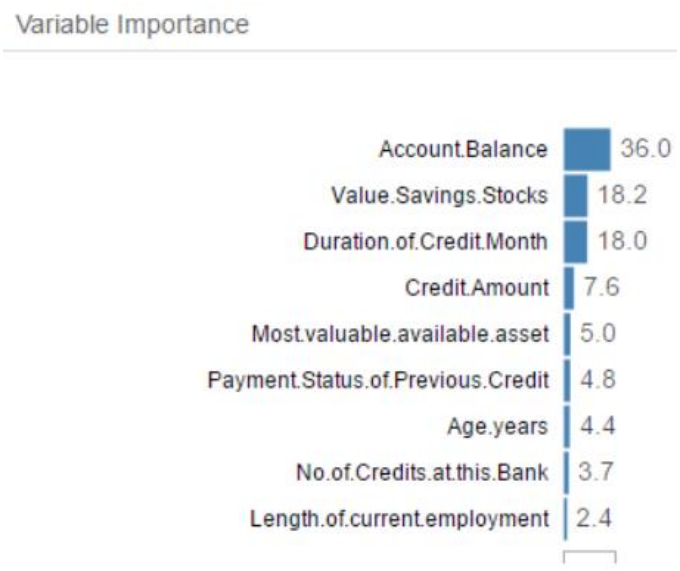
- Account-Balance
- Credit-Amount
- Purpose
- Instalment-per-cent

	Estimate	Std. Error	z value	Pr(>|z|)	
(Intercept)	-2.5783608	6.414e-01	-4.0202	6e-05	***
Account.BalanceSome Balance	-1.5715598	3.037e-01	-5.1742	2.28e-07	***
Payment.Status.of.Previous.CreditPaid Up	0.2117362	2.952e-01	0.7174	0.47316	
Payment.Status.of.Previous.CreditSome Problems	1.3053044	5.089e-01	2.5648	0.01032	*
PurposeNew car	-1.6344313	6.137e-01	-2.6633	0.00774	**
PurposeOther	-0.4435055	8.242e-01	-0.5381	0.59049	
PurposeUsed car	-0.7315961	3.976e-01	-1.8400	0.06577	.
Credit.Amount	0.0002076	5.453e-05	3.8070	0.00014	***
Length.of.current.employment4-7 yrs	0.3678284	4.537e-01	0.8107	0.41752	
Length.of.current.employment< 1yr	0.7564408	3.833e-01	1.9733	0.04846	*
Instalment.per.cent	0.3426933	1.325e-01	2.5873	0.00967	**

Decision Tree Model

These are the 4 most important predictor variables for decision tree model.

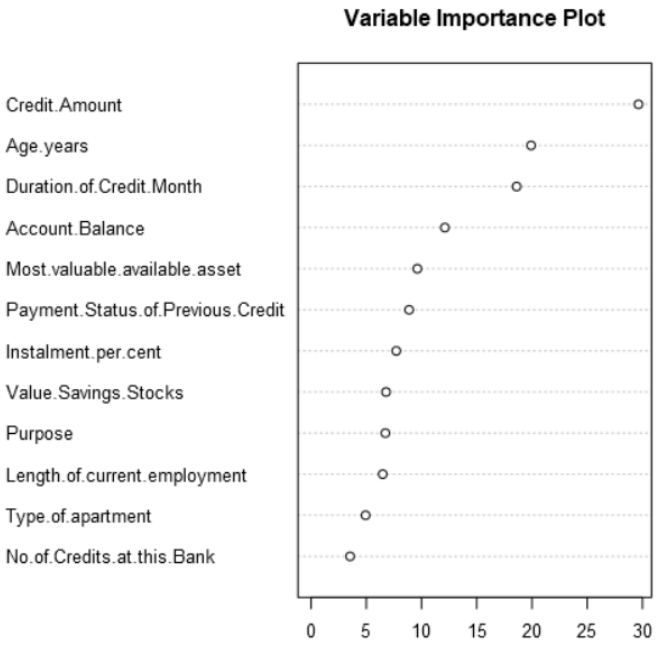
- Account-Balance
- Value-Saving-Stocks
- Duration-of-Credit-Month
- Credit-Amount



Forest Model

These are the 4 most important predictor variables for Forest model.

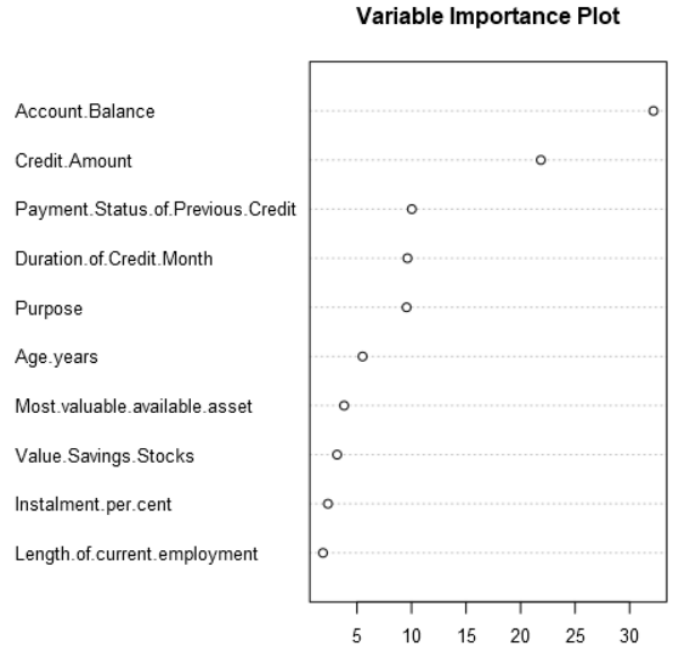
- Credit-Amount
- Age-years
- Duration-of-Credit-Month
- Account-Balance



Boosted Model

These are the 4 most important predictor variables for Boosted model.

- Account-Balance
- Credit-Amount
- Payment-Status-of-Previous-Credit
- Duration-of-Credit-Month



- 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 percent accuracy for all 4 models are followings. The best model for determining whether a loan should be approved or not is Forest Modal which has 80.67% accuracy.

Logistic Model and Boosted Model have 78% and 78.67% accuracy, and Decision Tree model has the worst percent accuracy of 74.67%

Model	Accuracy	F1	AUC	Accuracy_Creditworthy	Accuracy_Non-Creditworthy
LR_LoanApproval	0.7800	0.8507	0.7352	0.8103	0.6765
DT_LoanApproval	0.7467	0.8273	0.7054	0.7913	0.6000
FM_LoanApproval	0.8067	0.8755	0.7392	0.7969	0.8636
BM_LoanApproval	0.7867	0.8632	0.7524	0.7829	0.8095

While Boosted Model and Forest Model performed really well with 96% and 97% of predicting creditworthy correctly, Decision Tree Model has the worst percent accuracy of 87% for creditworthy. Logistic Model were able to predict 90% accuracy of predicting creditworthy correctly.

All 4 models only predicted 50% or less correctly for non-creditworthy. Logistic Model has 51% accuracy and is slightly better than the other models. Decision Tree Model and Forest Model predicted 47% and 42% of non-creditworthy correctly. Boosted Model has the worst accuracy of 38% for predicting non-creditworthy. All models have bias towards predicting non-creditworthy as creditworthy.

Confusion matrix of BM_LoanApproval

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

Confusion matrix of DT_LoanApproval

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

Confusion matrix of FM_LoanApproval

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

Confusion matrix of LR_LoanApproval

	Actual_Creditworthy	Actual_Non-Creditworthy
Predicted_Creditworthy	94	22
Predicted_Non-Creditworthy	11	23

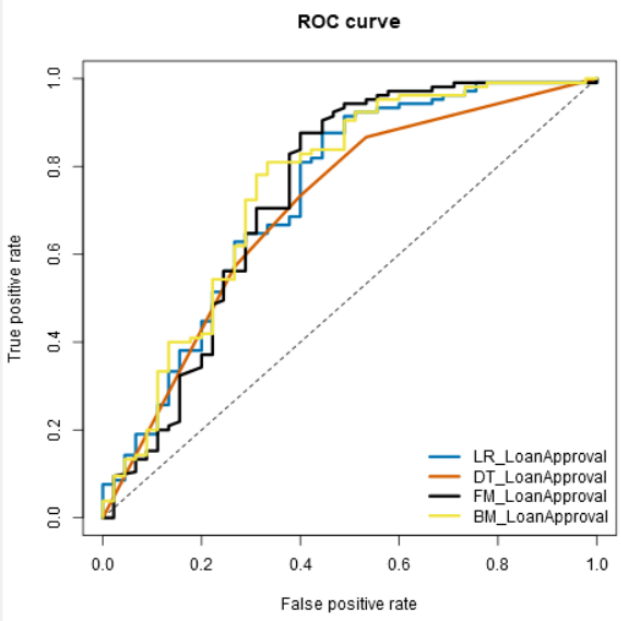
Step 4: Writeup

- How many individuals are creditworthy?
- 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 have predicted 408 new customers would qualify for a loan, and I chose Forest Model to predict it.

Although it doesn’t have high accuracy and only 42% for predicting non-creditworthy, Forest Model has the highest overall percent accuracy of 80% against validation set and the highest percent accuracy of 97% out of all 4 models for predicting creditworthy.

The accuracy of test is measured by the area under the ROC curve. The closer the curve follows the left-hand border and then the top border of the ROC space, the more accurate the test. Forest Model does has the worst curve starting, but as the curves pass the middle and getting to the end, Forest model curve is the closest to the top border.



Since all 4 models have bias towards predicting non-creditworthy as creditworthy, there was no one model stood out for bias in the confusion metrices.