Project: Creditworthiness

I work for a small bank responsible for determining if customers are creditworthy to give a loan to. My team typically gets 200 loan applications per week and approves them by hand. Due to a financial scandal that hit a competitive bank last week, we have nearly 500 loan applications to process this week.

Step 1: Business and Data Understanding

- What decisions needs to be made?
  Determining if customers applying for loan are creditworthy to give a loan to within one week.

- What data is needed to inform those decisions?
  - Data on all past applications
  - The list of customers that need to be processed in the next few days.

  To determine wither the customers creditworthy or not, there are some factors that affects this decision. This is such as customers credit balance, their incomes, their ages, the length of their current employer, and the purpose of applying to loan.
- What kind of model (Continuous, Binary, Non-Binary, Time-Series) do we need to use to help make these decisions?
  Binary classification models. This is including Logistic regression model, Decision tree model, Forest model and Boosted tree model.

Step 2: Building the Training Set

- In your cleanup process, which fields did you remove or impute? Please justify why you removed or imputed these fields. Visualizations are encouraged.
  There are some fields need to be removed from the data set for some reasons. Concurrent credit and Occupation fields need to be removed since they have Low variability where they have only one type of data. In addition, Guarantors, No of dependents, and Foreign worker fields need to be removed as well where they have Low variability. Also, Duration in current address field need to be removed because it has 68.8% of missing data. Lastly, Telephone field need to be removed since it is not relevant to the decision of creditworthy of the customer for the loan.

  On the other side, although Age years field has 2% of missing data, it should be imputed to avoid having effect on other attributes. For that, missing data replaced with “median”.

Step 3: Train your Classification Models

a. Which predictor variables are significant or the most important? Please show the p-values or variable importance charts for all of your predictor variables.

b. 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 model

a. The predictor variables that are most important to the target variable "Credit Application Result" are Account.BalanceSome Balance, PurposeNew car, and Credit.Amount. The p-value of all these variables are below 0.01.

Report for Logistic Regression Model Logistic_Regression

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)

b. This model has an overall 76% of accuracy. The accuracy of predicts the creditworthy is 80% and 62% for non- creditworthy. The model biases of predicting some creditworthy customers as a non- creditworthy.

Model Comparison Report

Fit and error measures

Model	Accuracy	F1	AUC	Accuracy_Creditworthy	Accuracy_Non-Creditworthy
Stepwise	0.7600	0.8364	0.7306	0.8000	0.6286
Decision_Tree	0.7467	0.8273	0.7054	0.7913	0.6000
Forest_Model	0.8000	0.8718	0.7243	0.7907	0.8571
Boosted_Model	0.7867	0.8632	0.7524	0.7829	0.8095

Confusion matrix of Stepwise

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

2. Decision tree model

a. The predictor variables that are most important to the target variable “Credit Application Result” are Account Balance, Value Savings Stocks and Duration of Credit Month.

b. This model has an overall 74.67% of accuracy. The accuracy of predicts the creditworthy is 79% and 60% for non- creditworthy. The model biases of predicting some creditworthy customers as a non- creditworthy.

Model Comparison Report

Fit and error measures

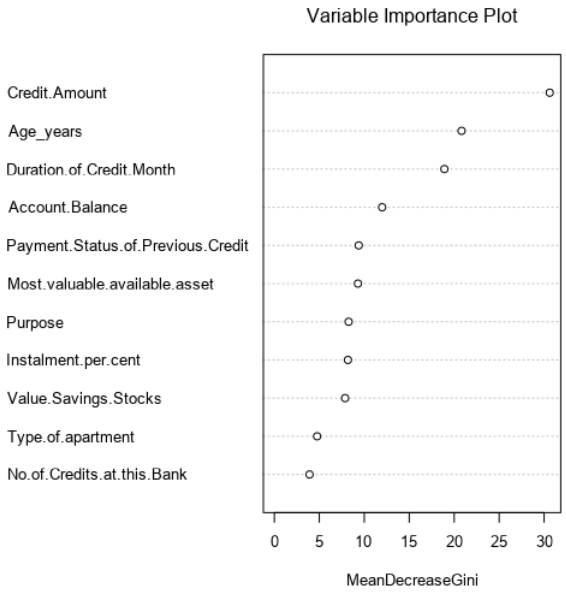
Model	Accuracy	F1	AUC	Accuracy_Creditworthy	Accuracy_Non-Creditworthy
Stepwise	0.7600	0.8364	0.7306	0.8000	0.6286
Decision_Tree	0.7467	0.8273	0.7054	0.7913	0.6000
Forest_Model	0.8000	0.8718	0.7243	0.7907	0.8571
Boosted_Model	0.7867	0.8632	0.7524	0.7829	0.8095

Confusion matrix of Decision_Tree

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

3. Forest model

a. The predictor variables that are most important to the target variable “Credit Application Result” are Credit Amount, Age years and Duration of Credit Month.



b. This model has an overall 80% of accuracy. The accuracy of predicts the creditworthy is 79% and 85.71% for non- creditworthy. The model has no biases in predicting the creditworthy customers and non- creditworthy customers.

Model Comparison Report

Fit and error measures

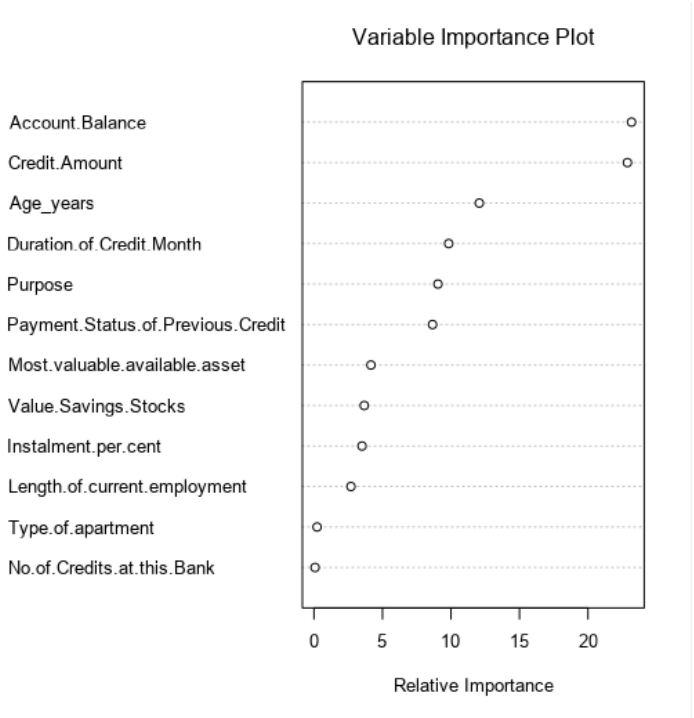
Model	Accuracy	F1	AUC	Accuracy_Creditworthy	Accuracy_Non-Creditworthy
Stepwise	0.7600	0.8364	0.7306	0.8000	0.6286
Decision_Tree	0.7467	0.8273	0.7054	0.7913	0.6000
Forest_Model	0.8000	0.8718	0.7243	0.7907	0.8571
Boosted_Model	0.7867	0.8632	0.7524	0.7829	0.8095

Confusion matrix of Forest_Model

	Actual_Creditworthy	Actual_Non-Creditworthy
Predicted_Creditworthy	102	27
Predicted_Non-Creditworthy	3	18

4. Boosted tree model

a. The predictor variables that are most important to the target variable “Credit Application Result” are Credit Amount and Account Balance.



b. This model has an overall 78.67% of accuracy. The accuracy of predicts the creditworthy is 78.29% and 80.95% for non- creditworthy. The model lack biases of predicting creditworthy customers and non- creditworthy customers.

Model Comparison Report

Fit and error measures

Model	Accuracy	F1	AUC	Accuracy_Creditworthy	Accuracy_Non-Creditworthy
Stepwise	0.7600	0.8364	0.7306	0.8000	0.6286
Decision_Tree	0.7467	0.8273	0.7054	0.7913	0.6000
Forest_Model	0.8000	0.8718	0.7243	0.7907	0.8571
Boosted_Model	0.7867	0.8632	0.7524	0.7829	0.8095

Confusion matrix of Boosted_Model

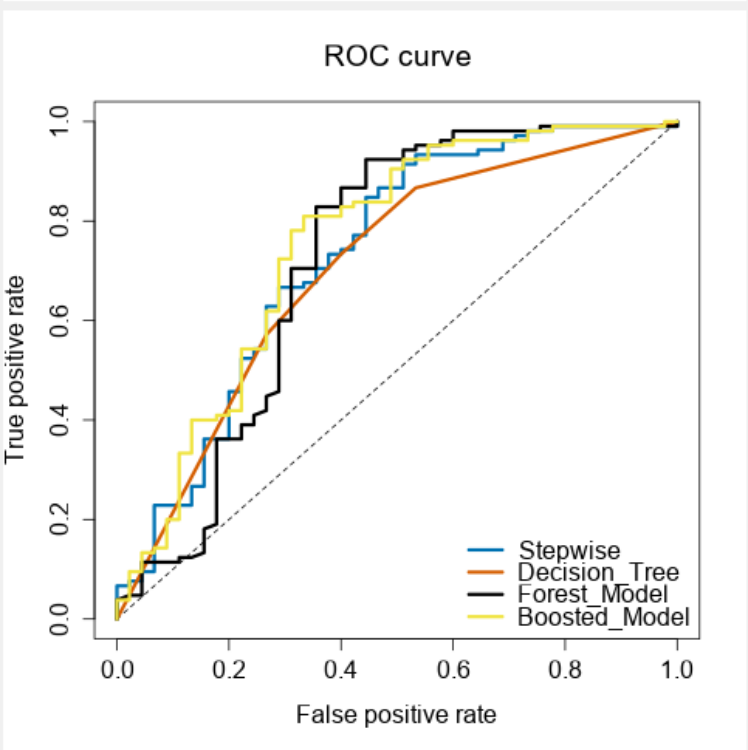
	Actual_Creditworthy	Actual_Non-Creditworthy
Predicted_Creditworthy	101	28
Predicted_Non-Creditworthy	4	17

Step 4: Writeup

- 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:

  I choose the Forest model for many reasons. It has the highest overall accuracy against your validation set with 80% of accuracy. Accuracies within Creditworthy and Non-Creditworthy segments is the highest as well with 79% and 85.71% respectively. In addition, this model has no biases in predicting the creditworthy customers and non-creditworthy customers. This is very important because it gives the idea of the customer probability of defaulting to avoid loaning the customers with high level of defaulting probability at the same time taking the opportunities to loan the creditworthy customers.

  - ROC graph



- How many individuals are creditworthy?
  Using Forest model, there are 408 creditworthy customers.