Project: Creditworthiness

By: Anthony Nguyen

Step 1: Business and Data Understanding

Key Decisions:

- What decisions needs to be made?
  A local bank just acquired a mortgage company. Prior to the acquisition, roughly 200 loan applications each week are processed. With the recent acquisition, the number of loan applicants has increased to 500 this week. The decision that needs to be made is determining the creditworthy status of the new bank loan customers by utilizing a classification model to speed up the approval process.

- What data is needed to inform those decisions?
  The data that is needed to inform the creditworthy status decisions are:
  - Credit Application Result
  - Age Years
  - Most Valuable Available Asset
  - Installment Per-Cent
  - Value Savings Stocks
  - Duration of Credit Month
  - No. of Dependents
  - Credit Amount
  - Account Balance
  - Payment Status of Previous Credit
  - Purpose
  - Length of Current Employment
  - 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?
  In order to determine whether the new bank customers are creditworthy or non-creditworthy of a loan, the kind of model that is needed to help make these decisions are binary classification models.

Step 2: Building the Training Set

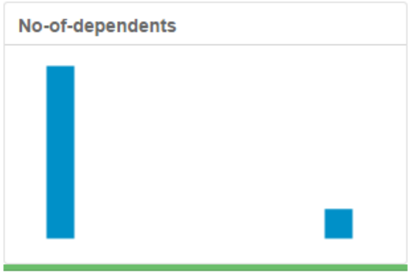
- For numerical data fields, are there any fields that highly-correlate with each other? The correlation should be at least .70 to be considered "high".
  No, there are not any numerical data fields that highly-correlate with each other.
- Are there any missing data for each of the data fields? Fields with a lot of missing data should be removed
  Yes, there are 2 data fields that are missing data: Duration in Current Address + Age Years.
- 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

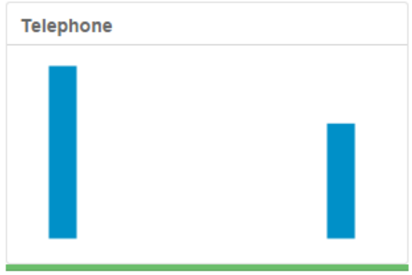
you should remove fields that have low variability. Refer to the "Tips" section to find examples of data fields with low-variability.
Yes, there are 5 data fields that reflect low-variability: Concurrent Credits, Gurantors, No. of Credits at this Bank, Telephone, & Occupation

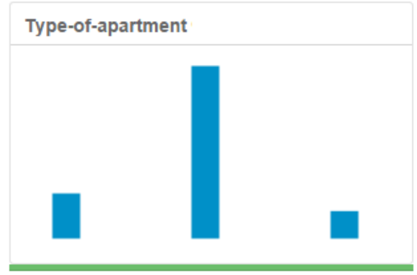
- In your cleanup process, which fields did you remove or impute? Please justify why you removed or imputed these fields. Visualizations are encouraged.
  The following fields were removed:

Field Removed	Reason Field Removed
1. Concurrent Credits	Entire dataset has only one observation.
2. Occupation	Entire dataset has only one observation.
3. Telephone	Doesn't contribute to classification of customer.
4. Guarantors	Low variability data.
5. Foreign Workers	Low variability data.
6. Number of Dependents	Low numbers of observation.
7. Duration in Current Address	Low numbers of observation.

Imputed Field	Reason for Imputation
1. Age-Years	The missing values were imputed with the median value.







Step 3: Train your Classification Models

- 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 most important variables = Credit Application Result, Account Balance, Payment Status of Previous Credit – Some Problems, Purpose – New Car, Credit Amount, Installment Per.Cent, Most Valuable Available Asset (listed directly below)

	Estimate	Std. Error	z value	Pr(>|z|)
(Intercept)	-3.1969303	9.849e-01	-3.2460	0.00117 **
Account.Balance1	-1.5837266	3.201e-01	-4.9473	7.52e-07 ***
Payment.Status.of.Previous.CreditPaid Up	0.4272881	3.848e-01	1.1104	0.26682
Payment.Status.of.Previous.CreditSome Problems	1.2853969	5.338e-01	2.4079	0.01605 *
PurposeNew car	-1.7484975	6.274e-01	-2.7868	0.00532 **
PurposeOther	-0.2368860	8.322e-01	-0.2846	0.77593
PurposeUsed car	-0.7675243	4.108e-01	-1.8682	0.06174 .
Credit.Amount	0.0001742	6.838e-05	2.5479	0.01084 *
Value.Savings.StocksNone	0.6021148	5.065e-01	1.1887	0.23456
Value.Savings.Stocks£100-£1000	0.1849187	5.622e-01	0.3289	0.74223
Length.of.current.employment4-7 yrs	0.5356571	4.935e-01	1.0854	0.27776
Length.of.current.employment< 1yr	0.7769992	3.952e-01	1.9660	0.0493 *
Instalment.per.cent	0.2967259	1.384e-01	2.1435	0.03208 *
Most.valuable.available.asset	0.2884695	1.489e-01	1.9370	0.05275 .
Age.years	-0.0191174	1.479e-02	-1.2925	0.1962
No.of.Credits.at.this.BankMore than 1	0.3918336	3.812e-01	1.0279	0.30402
Duration.of.Credit.Month	0.0057243	1.365e-02	0.4193	0.67499

  Decision Tree most important variables = Account Balance, Duration of Credit Month, Credit Amount, Value Savings Stock, Age Years, Purpose, Length of Current Employment, Most Valuable Available Asset, No. of Credits at this Bank, Payment Status of Previous Credit (listed directly below)

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

Forest Model most important variables = Account Balance, Payment Status of Previous Credit, Purpose, Credit Amount, Value Savings Stocks, Length of Current Employment, Installment Per.Cent, Most Valuable Available Asset, Age Years, No. of Credits at this Bank, Duration of Credit Month (listed directly below)

Basic Summary

Call:
randomForest(formula = Credit.Application.Result ~ Account.Balance + Payment.Status.of.Previous.Credit + Purpose + Credit.Amount + Value.Savings.Stocks + Length.of.current.employment + Instalment.per.cent + Most.valuable.available.asset + Age.years + No.of.Credits.at.this.Bank + Duration.of.Credit.Month, data = the.data, ntree = 500)

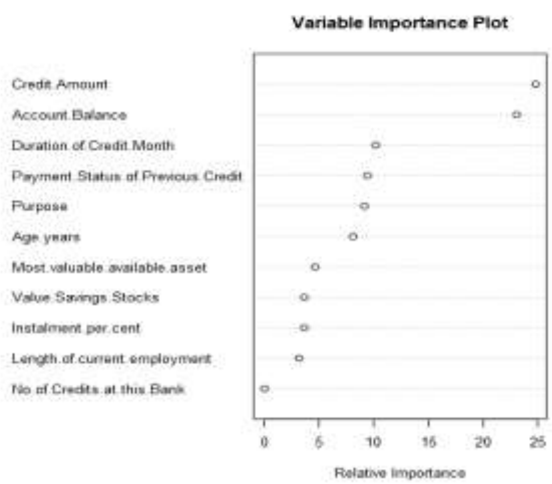
Type of forest: classification
Number of trees: 500
Number of variables tried at each split: 3

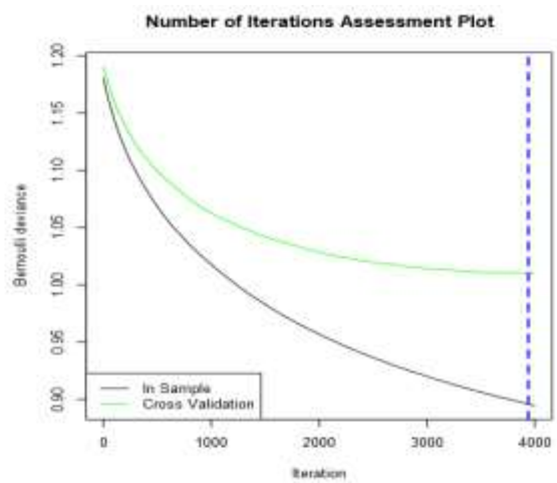
OOB estimate of the error rate: 36.1%

Confusion Matrix:

	Classification Error	Creditworthy	Non-Creditworthy
Creditworthy	0.103	227	26
Non-Creditworthy	0.619	60	37

Boosted Model most important variables = Credit Amount, Account Balance, Duration of Credit Month, Payment Status of Previous Credit, Purpose, Age Years, Most Valuable Available Asset, Value Savings Stock, Installment Per.Cent, Length of Current Employment (listed directly below)





- 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?
  The following are the overall percent model accuracy against the validation set:
  - Logistic Regression = 80%
  - Decision Tree = 67.33%
  - Forest Model = 81.33%
  - Boosted Model = 79.33%

Yes there is a bias in the model’s predictions. All 4 model’s confusion actual non-creditworthy matrix’s predicted roughly the same amount for both predicted creditworthy and predicted non-creditworthy. This holds true because it is slighly harder to predict for non-creditworthy because there are far more creditworthy values. The accuracy of creditworthy predictions is greater than non-creditworthy predictions. So the bias is towards creditworthy predictions.

Fit and error measures

Model	Accuracy	F1	AUC	Accuracy_Creditworthy	Accuracy_Non-Creditworthy
Logistic_Reg_PredictDefaultRisk	0.8000	0.8661	0.7380	0.8151	0.7419
DecisionTree_PredictDefaultRisk	0.6733	0.7721	0.6296	0.7545	0.4500
Forest_PredictDefaultRisk	0.8133	0.8783	0.7419	0.8080	0.8400
Boosted_PredictDefaultRisk	0.7933	0.8670	0.7537	0.7891	0.8182

Confusion matrix of Boosted_PredictDefaultRisk

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

Confusion matrix of DecisionTree_PredictDefaultRisk

	Actual_Creditworthy	Actual_Non-Creditworthy
Predicted_Creditworthy	83	27
Predicted_Non-Creditworthy	22	18

Confusion matrix of Forest_PredictDefaultRisk

	Actual_Creditworthy	Actual_Non-Creditworthy
Predicted_Creditworthy	101	24
Predicted_Non-Creditworthy	4	21

Confusion matrix of Logistic_Reg_PredictDefaultRisk

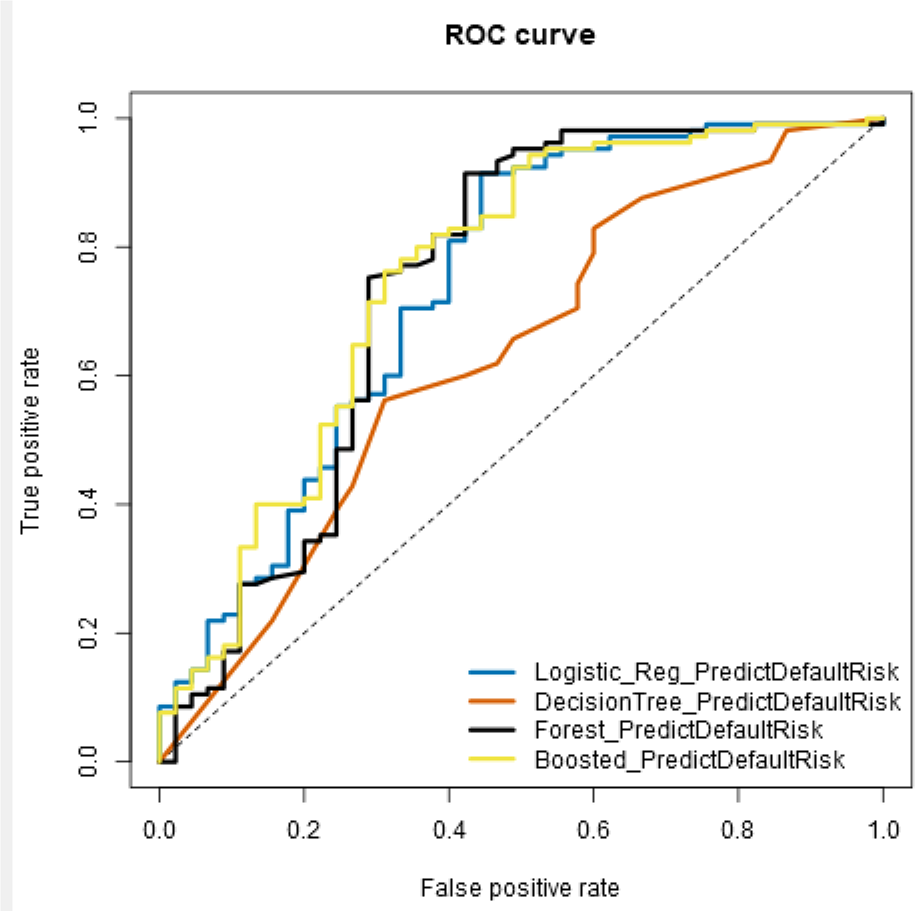
	Actual_Creditworthy	Actual_Non-Creditworthy
Predicted_Creditworthy	97	22
Predicted_Non-Creditworthy	8	23

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:
    - Overall Accuracy against your Validation set
    - Accuracies within “Creditworthy” and “Non-Creditworthy” segments
    - ROC graph
    - Bias in the Confusion Matrices

- How many individuals are creditworthy? **406**

The classification model that was chosen to score new bank loan customers was the Forest Model. The overall accuracy of the Forest Model proved to be highest amongst all classification models as verified against the validation set. The accuracy for the Creditworthy segment is 80.80% and Non-Creditworthy is 84%. The ROC curve for the Forest Model does a fair job at separating the creditworthy class and non-creditworthy class. This is determined by reviewing where the ROC curve is in relation to the dotted line. The current bias for the Forest Model in the confusion matrix is that it is heavily on the creditworthy predictions. Reason being is that there are far more creditworthy values compared to non-creditworthy values. Ultimately, with the prediction model that I have created using the Forest Model, a total of 406 new customers are creditworthy and would qualify for a loan.


ROC curve
True positive rate
False positive rate
0.0
0.2
0.4
0.6
0.8
1.0
Logistic_Reg_PredictDefaultRisk
DecisionTree_PredictDefaultRisk
Forest_PredictDefaultRisk
Boosted_PredictDefaultRisk