Project: Creditworthiness

Step 1: Business and Data Understanding

Key Decisions:

Answer these questions

- **What decisions needs to be made?**
  Predict that for customers who apply for the loan is credit worthy.

- **What data is needed to inform those decisions?**
  We need the data from customers who applied for loan before, no matter they are approved or not, along with other features of customers, like age, occupation, credit history, credit amount, account balance, etc.

- **What kind of model (Continuous, Binary, Non-Binary, Time-Series) do we need to use to help make these decisions?**
  This is simply the binary classification model, credit worthy yes or no.

Step 2: Building the Training Set

Build your training set given the data provided to you. The data has been cleaned up for you already so you shouldn't ***need to convert any data fields to the appropriate data types.***

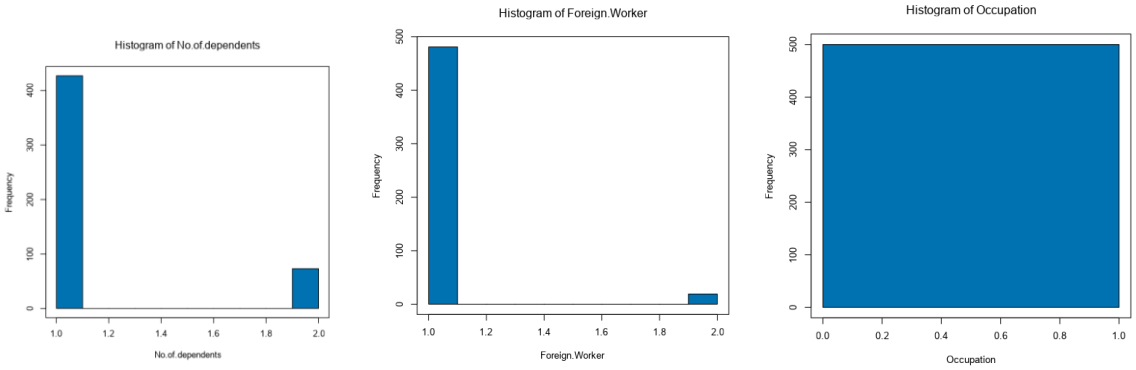
Here are some guidelines to help guide your data cleanup:

- **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".**
  Calculate the correlation between all numerical fields in the data with pearson correlation matrix, do not find any value larger than 0.7, which is highly correlated features.

- **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" has nearly 70% data missing; Thus, we will drop this column. The other column is "age-years", which has 2.4% data missing. We could fill in with media age of this column.

- **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.**

By check the histogram of each column, we can see that "*Concurrent credits*", "occupation" has just one value. "Guarantors", "No of dependens", "Foreign worker" these data are highly skewed with less variability. "Telephone" is also not important here. Thus, I will remove all these 7 columns.
Following figures show selected columns that is dropped from further analysis.



- **Your clean data set should have 13 columns where the Average of Age Years should be 36 (rounded up)**
  For age is Null, I replace the age with median 33 years, and confirm that the average of ages now is 35.57.

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.

Logistic Regression

Account balance, purpose, installment percentage and credit amount are most import predictors, since their p-value is much smaller than 0.05.

Overall accuracy for logistic regression model is 78%, accuracy for creditworthy is 90% and non-creditworthy is 49%. From confusion matrix, positive predictive value (PPV) = 95/(95+23) = 80%, negative predictive value (NPV) = 22/(22+10) = 69%, Thus, this model is bias when predict creditworthy customers.

Report for Logistic Regression Model Logistic_regression

Basic Summary

Call:
glm(formula = Credit.Application.Result ~ Account.Balance + Credit.Amount + Duration.of.Credit.Month + Instalment.per.cent + Length.of.current.employment + Most.valuable.available.asset + No.of.Credits.at.this.Bank + Payment.Status.of.Previous.Credit + Purpose + Type.of.apartment + Value.Savings.Stocks + Age_years, family = binomial(logit), data = the.data)

Deviance Residuals:

Min	1Q	Median	3Q	Max
-2.088	-0.719	-0.430	0.686	2.542

Coefficients:

	Estimate	Std. Error	z value	Pr(>|z|)	
(Intercept)	-3.0136120	1.013e+00	-2.9760	0.00292	**
Account.BalanceSome Balance	-1.5433699	3.232e-01	-4.7752	1.79e-06	***
Credit.Amount	0.0001764	6.838e-05	2.5798	0.00989	**
Duration.of.Credit.Month	0.0064973	1.371e-02	0.4738	0.63565	
Instalment.per.cent	0.3109833	1.399e-01	2.2232	0.0262	*
Length.of.current.employment4-7 yrs	0.5224158	4.930e-01	1.0596	0.28934	
Length.of.current.employment< 1yr	0.7779492	3.956e-01	1.9664	0.04925	*
Most.valuable.available.asset	0.3258706	1.556e-01	2.0945	0.03621	*
No.of.Credits.at.this.BankMore than 1	0.3619545	3.815e-01	0.9487	0.34275	
Payment.Status.of.Previous.CreditPaid Up	0.4054309	3.841e-01	1.0554	0.29124	
Payment.Status.of.Previous.CreditSome Problems	1.2607175	5.335e-01	2.3632	0.01812	*
PurposeNew car	-1.7541034	6.276e-01	-2.7951	0.00519	**
PurposeOther	-0.3191177	8.342e-01	-0.3825	0.70206	
PurposeUsed car	-0.7839554	4.124e-01	-1.9008	0.05733	.
Type.of.apartment	-0.2603038	2.956e-01	-0.8805	0.3786	
Value.Savings.StocksNone	0.6074082	5.100e-01	1.1911	0.23361	
Value.Savings.Stocks£100-£1000	0.1694433	5.649e-01	0.3000	0.7642	
Age_years	-0.0141206	1.535e-02	-0.9202	0.35747	

Model Comparison Report

Fit and error measures

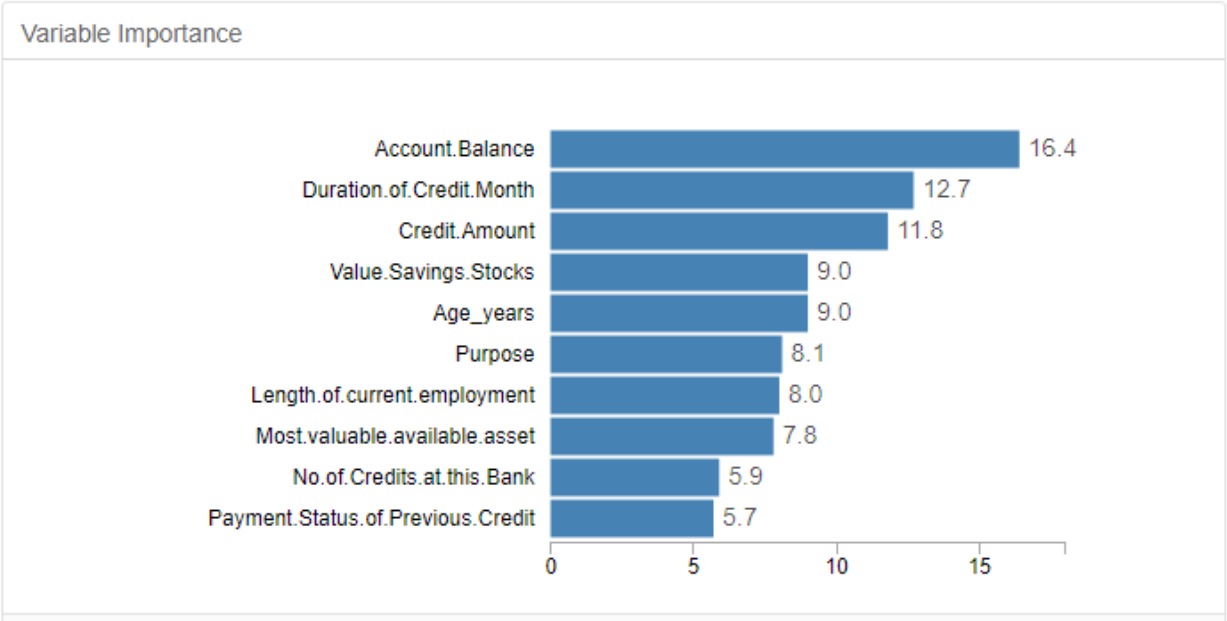
Model	Accuracy	F1	AUC	Accuracy_Creditworthy	Accuracy_Non-Creditworthy
Logistic_regression	0.7800	0.8520	0.7314	0.9048	0.4889

Model: model names in the current comparison.
Accuracy: overall accuracy, number of correct predictions of all classes divided by total sample number.
Accuracy_[class name]: accuracy of Class [class name] is defined as the number of cases that are **correctly** predicted to be Class [class name] divided by the total number of cases that actually belong to Class [class name], this measure is also known as *recall*.
AUC: area under the ROC curve, only available for two-class classification.
F1: F1 score, 2 * precision * recall / (precision + recall). The *precision* measure is the percentage of actual members of a class that were predicted to be in that class divided by the total number of cases predicted to be in that class. In the situations where there are three or more classes, average precision and average recall values across classes are used to calculate the F1 score.

Confusion matrix of Logistic_regression

	Actual_Creditworthy	Actual_Non-Creditworthy
Predicted_Creditworthy	95	23
Predicted_Non-Creditworthy	10	22

Decision Tree



Model Comparison Report

Fit and error measures

Model	Accuracy	F1	AUC	Accuracy_Creditworthy	Accuracy_Non-Creditworthy
Decision_tree	0.6733	0.7721	0.6296	0.7905	0.4000

Model: model names in the current comparison.
Accuracy: overall accuracy, number of correct predictions of all classes divided by total sample number.
Accuracy_[class name]: accuracy of Class [class name] is defined as the number of cases that are **correctly** predicted to be Class [class name] divided by the total number of cases that actually belong to Class [class name], this measure is also known as *recall*.
AUC: area under the ROC curve, only available for two-class classification.
F1: F1 score, 2 * precision * recall / (precision + recall). The *precision* measure is the percentage of actual members of a class that were predicted to be in that class divided by the total number of cases predicted to be in that class. In the situations where there are three or more classes, average precision and average recall values across classes are used to calculate the F1 score.

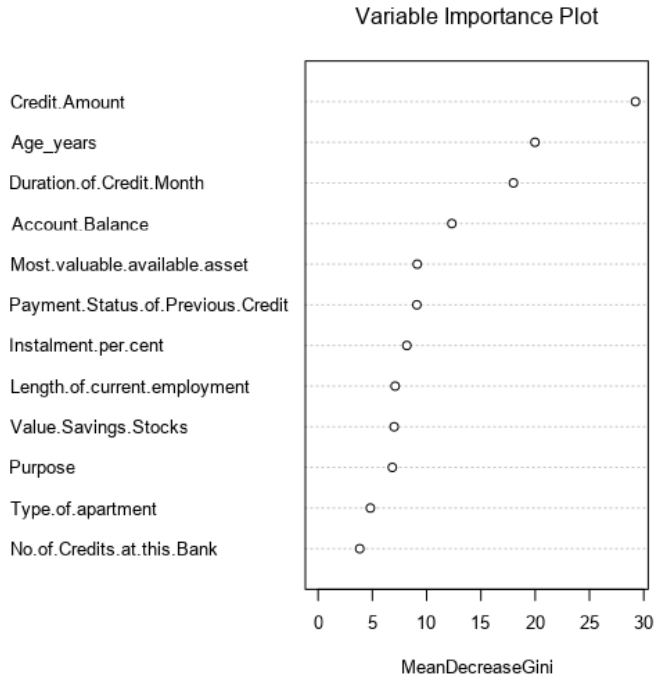
Confusion matrix of Decision_tree

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

Account balance, duration of credit month and credit amount are most import predictors, since their variable importance is the highest

Overall accuracy for decision tree model is 67%, accuracy for creditworthy is 79% and non-creditworthy is 40%. From confusion matrix, positive predictive value (PPV) = 83/(83+27) = 75%, negative predictive value (NPV) = 18/(22+18) = 45%, Thus, this model is bias when predict creditworthy customers.

Forest Model



Model Comparison Report

Fit and error measures

Model	Accuracy	F1	AUC	Accuracy_Creditworthy	Accuracy_Non-Creditworthy
Forest_model	0.8133	0.8793	0.7412	0.9714	0.4444

Model: model names in the current comparison.
Accuracy: overall accuracy, number of correct predictions of all classes divided by total sample number.
Accuracy_[class name]: accuracy of Class [class name] is defined as the number of cases that are **correctly** predicted to be Class [class name] divided by the total number of cases that actually belong to Class [class name], this measure is also known as *recall*.
AUC: area under the ROC curve, only available for two-class classification.
F1: F1 score, 2 * precision * recall / (precision + recall). The *precision* measure is the percentage of actual members of a class that were predicted to be in that class divided by the total number of cases predicted to be in that class. In the situations where there are three or more classes, average precision and average recall values across classes are used to calculate the F1 score.

Confusion matrix of Forest_model

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

Duration of credit month, age years and credit amount are most import predictors, since their variable importance is the highest
Overall accuracy for decision tree model is 81%, accuracy for creditworthy is 97% and non-creditworthy is 44%. From confusion matrix, positive predictive value (PPV) = 102/127 = 80%, negative predictive value (NPV) = 20/23 = 87%, Thus, this model is less bias when predict customers.

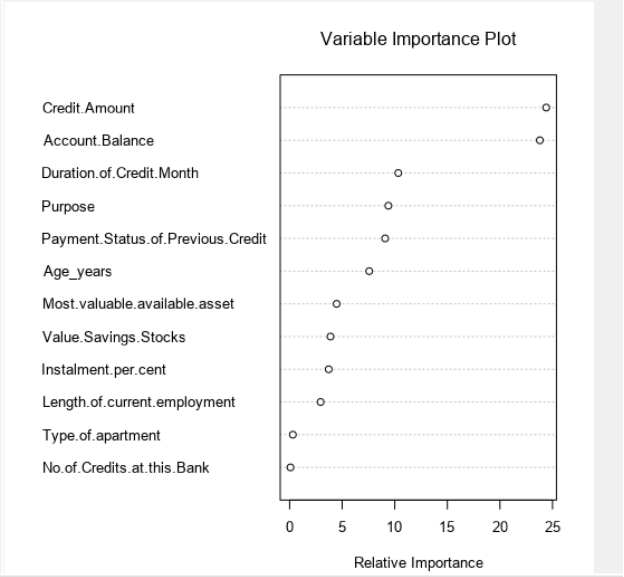
Boost method:

Report for Boosted Model Boosted_model

Basic Summary:

Loss function distribution: Bernoulli
Total number of trees used: 4000
Best number of trees based on 5-fold cross validation: 2036

Plots:



Model Comparison Report

Fit and error measures

Model	Accuracy	F1	AUC	Accuracy_Creditworthy	Accuracy_Non-Creditworthy
Boosted_model	0.7867	0.8632	0.7524	0.9619	0.3778

Model: model names in the current comparison.
Accuracy: overall accuracy, number of correct predictions of all classes divided by total sample number.
Accuracy_[class name]: accuracy of Class [class name] is defined as the number of cases that are **correctly** predicted to be Class [class name] divided by the total number of cases that actually belong to Class [class name], this measure is also known as *recall*.
AUC: area under the ROC curve, only available for two-class classification.
F1: F1 score, 2 * precision * recall / (precision + recall). The *precision* measure is the percentage of actual members of a class that were predicted to be in that class divided by the total number of cases predicted to be in that class. In the situations where there are three or more classes, average precision and average recall values across classes are used to calculate the F1 score.

Confusion matrix of Boosted_model

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

Account balance, duration of credit month and credit amount are most import predictors, since their variable importance is the highest

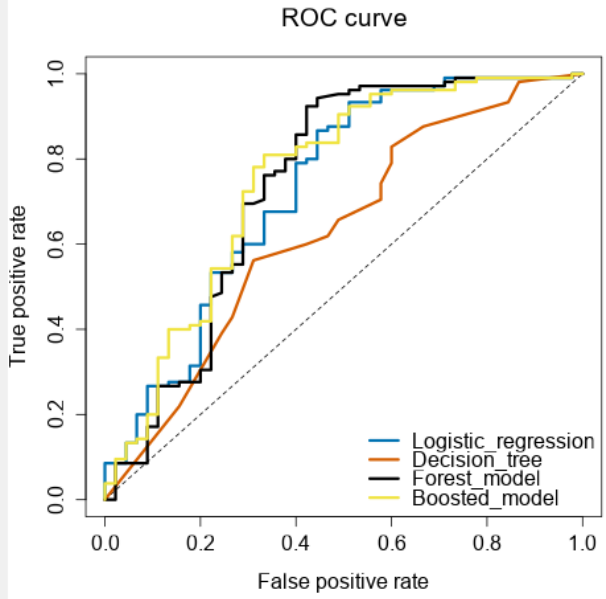
Overall accuracy for decision tree model is 77%, accuracy for creditworthy is 96% and non-creditworthy is 38%. From confusion matrix, positive predictive value (PPV) = 101/129 = 78%, negative predictive value (NPV) = 17/21 = 81%. Thus, this model is less bias when predict customers.

Step 4: Writeup

Model Comparison Report

Fit and error measures

Model	Accuracy	F1	AUC	Accuracy_Creditworthy	Accuracy_Non-Creditworthy
Logistic_regression	0.7800	0.8520	0.7314	0.9048	0.4889
Decision_tree	0.6733	0.7721	0.6296	0.7905	0.4000
Forest_model	0.8133	0.8793	0.7412	0.9714	0.4444
Boosted_model	0.7867	0.8632	0.7524	0.9619	0.3778



Confusion matrix of Boosted_model

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

Confusion matrix of Decision_tree

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

Confusion matrix of Forest_model

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

Confusion matrix of Logistic_regression

	Actual_Creditworthy	Actual_Non-Creditworthy
Predicted_Creditworthy	95	23
Predicted_Non-Creditworthy	10	22

- **Which model did you choose to use? Please justify your decision using all the following techniques. Please only use these techniques to justify your decision:**

  Overall Accuracy against your Validation set

  The final model that I choose is the forest model, because it has the highest accuracy of 81% against all other models, when I use 30% of validation dataset to evaluate the model accuracy.

Accuracies within "Creditworthy" and "Non-Creditworthy" segments
It also has the highest accuracies within "Creditworthy" and 2nd highest accuracy in "Non-Creditworthy" segments.

ROC Graph
Although its ROC curve is a little smaller than boosted model, but forest model reaches top quickest than all the other models. This means that for a given amount of false positive predictions (wrongly predicted creditworthy people), forest model will give the best number of true positive predictions (correctly predicted creditworthy people).

Bias in the Confusion Matrices
As discussed earlier, the bias of the models is lower for forest model with PPV = 0.80, NPV= 0.87. Less bias in model prediction. Therefore, we will choose random forest model.

- **How many individuals are creditworthy?**
  Scoring the remaining 500 new users, **408 of them are creditworthy**, since their score is higher than 0.5.

Workflow: