Project: Creditworthiness

Step 1: Business and Data Understanding

- What decisions needs to be made?

Among the new 500 loan applications, how many individuals are creditworthy.

- What data is needed to inform those decisions?

We need the past applications data including an attributes(fields) that can help in building a predicting model, we also need the new applications data to feed them into our generated model in order to predict the worthiness for individual.

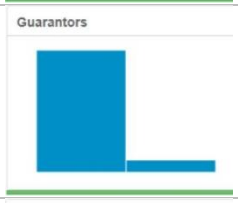
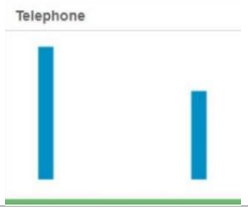
- What kind of model (Continuous, Binary, Non-Binary, Time-Series) do we need to use to help make these decisions?

Since we need to decide whether the applicants should be approved or not, we need to use a Binary 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.

I removed 7 fields, and imputed one field as shown below :

Cleanup	Field	Visualization	Reason
Remove	**Duration-in-Current-address**		Missing Data *This field has over 65% missing values, so removing the field is a reasonable.*
	Occupation		Completely Uniformed
	Concurrent-Credits		
	Foreign-Worker		Low Variability
	Guarantors		
	No-of-dependents		
	Telephone		
Impute	**Age-years**		I imputed with a median of 33, because Age is likely to be one of main factors in which application get approved or not.

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.

I passed the cleaned data to four models (Logistic, Decision Tree, Forest, Boosted) with one target variable which is credit application result. The table below shows the most important predictors for each model:

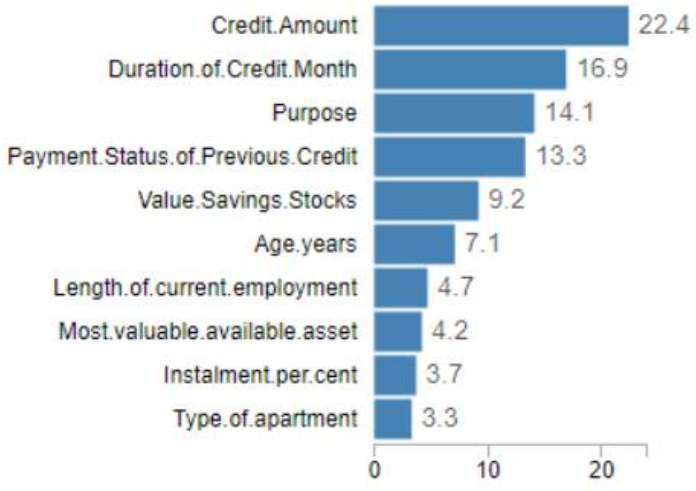
Logistic Model

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

Most important Variables:

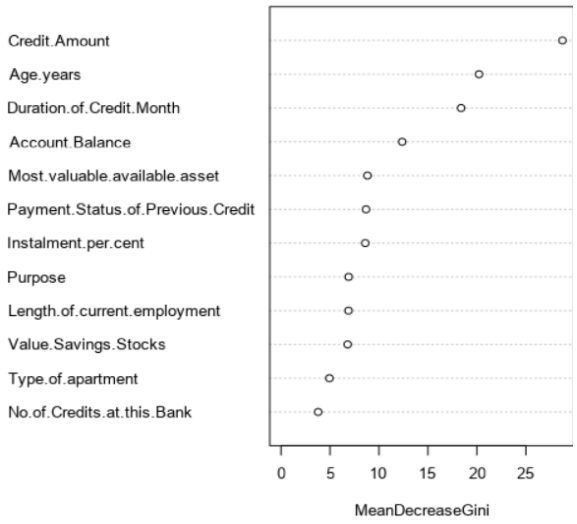
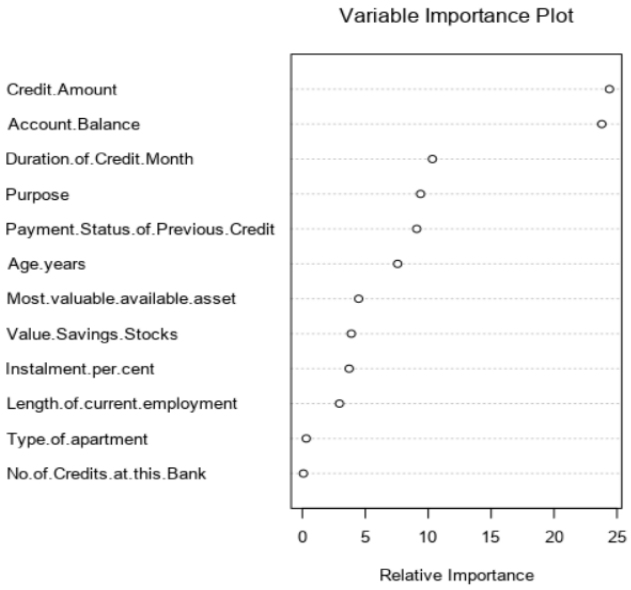
We can see the lowest P value is Account Balance categorical variable

Decision Tree Model



Most important Variables:

- Credit Amount
- Duration of Credit Month
- Purpose

Forest Model	
	**Most important Variables:** • Credit Amount • Age_years • Duration_of_Credit_ Month
Boosted Model	
	**Most important Variables:** • Credit Amount • Account Balance

- 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?

I validated all four models using Model Comparison Tool, the table below shows the Accuracy for each model:

Model	Accuracy	F1	AUC	Accuracy_Creditworthy	Accuracy_Non-Creditworthy
modForest	0.8067	0.8745	0.7359	0.9619	0.4444
modDec	0.7400	0.8235	0.6751	0.8667	0.4444
modLogistic	0.7800	0.8520	0.7314	0.9048	0.4889
modBoost	0.7867	0.8632	0.7524	0.9619	0.3778

Here is the confusion matrix:

Confusion matrix of modBoost

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

Confusion matrix of modDec

	Actual_Creditworthy	Actual_Non-Creditworthy
Predicted_Creditworthy	91	25
Predicted_Non-Creditworthy	14	20

Confusion matrix of modForest

	Actual_Creditworthy	Actual_Non-Creditworthy
Predicted_Creditworthy	101	25
Predicted_Non-Creditworthy	4	20

Confusion matrix of modLogistic

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

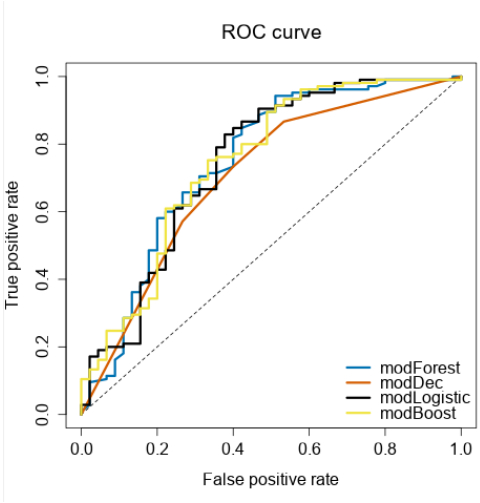
For bias calculation, we should look at the confusion matrix and calculate the accuracy of predicting Creditworthy represented by true positive rate (TPR) and the accuracy of predicting Non-Creditworthy represented by true negative rate (TNR). If the true positive value (TPR) and true negative rate (TNR) are close to each other, then we say that the model is unbiased, and if these values are not close then the model is biased.

Model	**TNV**	**TPV**	**Biased ?**
	TN/Actual Negatives	**TP/Actual Positives**	
Logistic	22/45= 48%	95/105=90%	Yes
Decision Tree	20/45=44%	91/105=86%	Yes
Forest Tree	20/45=44%	101/105=96%	Yes
Boosted	17/45=37%	101/105=96%	Yes

Step 4: Writeup

- Which model did you choose to use?

After comparing models (Logistic, Decision Tree, Forest Tree, Boosted), I've concluded the forest tree model is the best because it has a heights overall accuracy at 0.79. it has the heights Creditworthy accuracy at 0.96, and 0.40 Non-Creditworthy segments.
Looking at the ROC curve below, we can see that the Forest Tree model is best classifying predictor model:



- 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?
  After I applied Score tool for scoring a new 500 applicants, 410 among them could be approved.