Project: Creditworthiness

Step 1: Business and Data Understanding

Provide an explanation of the key decisions that need to be made. (250 word limit)

Key Decisions:

Answer these questions

- What decisions needs to be made?

Need to evaluate the creditworthiness of new loan applications based on customer data.

- What data is needed to inform those decisions?

We will decide the final creditworthiness based on Account-Balance, Duration-of-Credit-Month, Payment-Status-of-Previous-Credit, and so on.

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

We will evaluate creditworthiness using Binary Classification Models, including Logistic Model, Decision Tree, Forest Model, and Boosted Model.

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".
- Are there any missing data for each of the data fields? Fields with a lot of missing data should be removed
- 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.
- Your clean data set should have 13 columns where the Average of **Age Years** should be 36 (rounded up)

Note: *For the sake of consistency in the data cleanup process, impute data using the median of the entire data field instead of removing a few data points. (100 word limit)*

Note: *For students using software other than Alteryx, please format each variable as:*

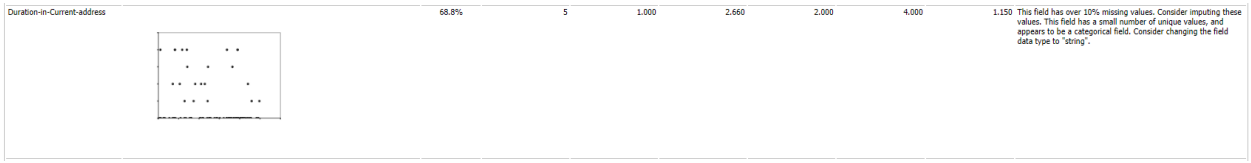
Variable	**Data Type**
Credit-Application-Result	String
Account-Balance	String
Duration-of-Credit-Month	Double
Payment-Status-of-Previous-Credit	String
Purpose	String
Credit-Amount	Double
Value-Savings-Stocks	String
Length-of-current-employment	String
Instalment-per-cent	Double
Guarantors	String
Duration-in-Current-address	Double
Most-valuable-available-asset	Double
Age-years	Double
Concurrent-Credits	String
Type-of-apartment	Double
No-of-Credits-at-this-Bank	String
Occupation	Double
No-of-dependents	Double
Telephone	Double
Foreign-Worker	Double

To achieve consistent results reviewers expect.

Answer this question:

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

I imputed Age-years with median, to prevent huge impact of age outliers, using median instead of average is appropriate. Additionally, I excluded **No-of-dependents**, **Occupation**, **Foreign-Worker**, **Concurrent-Credits**, **Guarantor**, and **Type-of-apartment** due to low variability. If we forced our model to consider these data fields, it may result in over-fitting. **Duration-in-current-address** is also removed due to 68.8% missing values.



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.

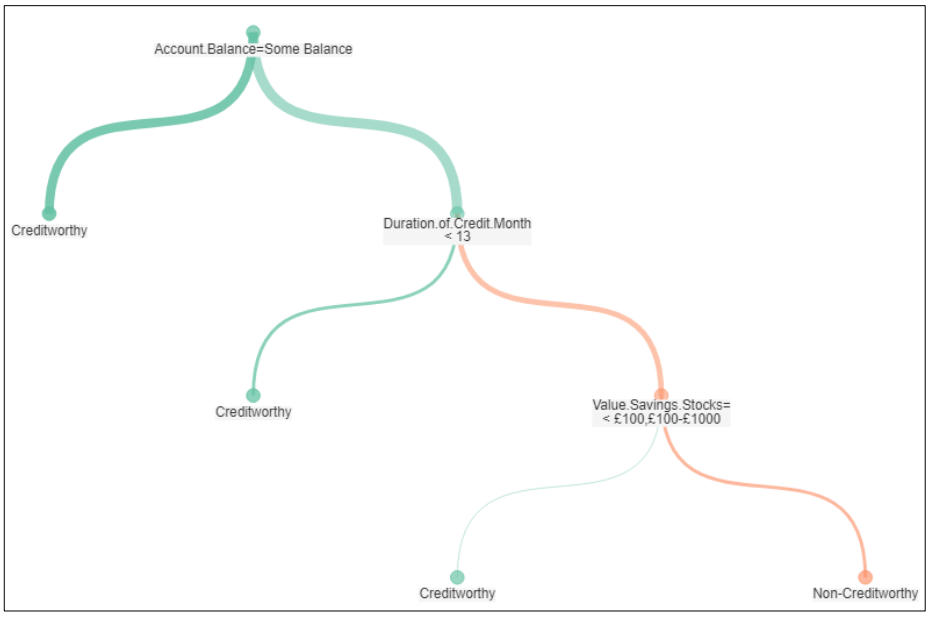
(a.) Logistic Regression Model (stepwise not used)

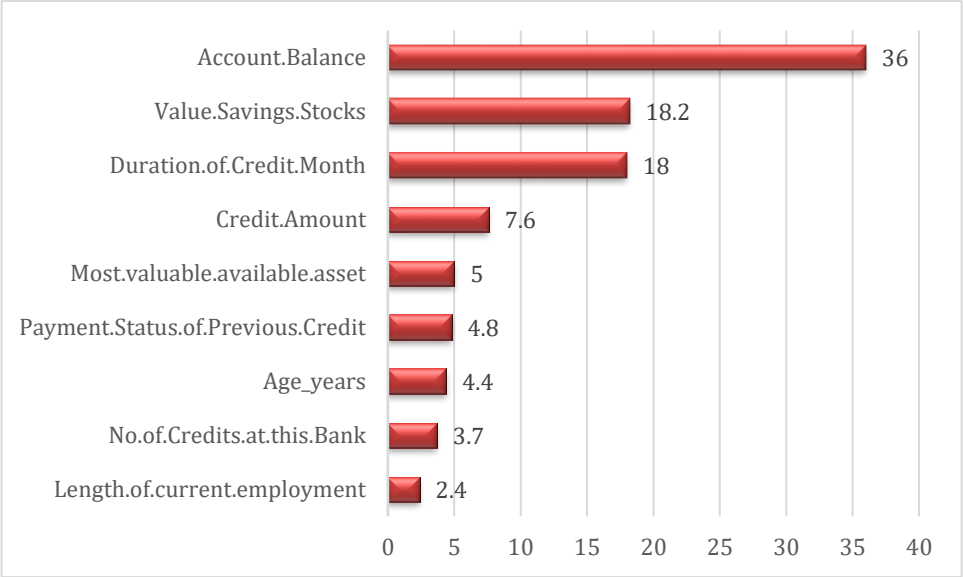
Amount-Balance is the most important predictor, which-value is 5.68e-07, quite small.

	Estimate	Std. Error	z value	Pr(>|z|)	
(Intercept)	-3.6041138	1.036e+00	-3.4786	5e-04	***
Account.BalanceSome Balance	-1.6152718	3.229e-01	-5.0016	5.68e-07	***
Credit.Amount	0.0001507	7.096e-05	2.1240	0.03367	*
Duration.of.Credit.Month	0.0072250	1.369e-02	0.5276	0.59777	
Instalment.per.cent	0.2882110	1.393e-01	2.0683	0.03861	*
Length.of.current.employment4-7 yrs	0.5313580	4.916e-01	1.0809	0.27973	
Length.of.current.employment< 1yr	0.8040089	3.939e-01	2.0411	0.04124	*
Most.valuable.available.asset	0.2671762	1.498e-01	1.7840	0.07442	.
No.of.Credits.at.this.BankMore than 1	0.3897906	3.826e-01	1.0188	0.30828	
Payment.Status.of.Previous.CreditPaid Up	0.4475591	3.863e-01	1.1587	0.24658	
Payment.Status.of.Previous.CreditSome Problems	1.3374204	5.356e-01	2.4972	0.01252	*
PurposeNew car	-1.7349564	6.274e-01	-2.7654	0.00569	**
PurposeOther	-0.1926841	8.355e-01	-0.2306	0.8176	
PurposeUsed car	-0.7804912	4.126e-01	-1.8915	0.05856	.
Telephone	0.3786710	3.138e-01	1.2068	0.22752	
Value.Savings.StocksNone	0.6188301	5.067e-01	1.2213	0.22199	
Value.Savings.Stocks £ 100- £ 1000	0.1726049	5.623e-01	0.3070	0.75887	
Age_years	-0.0199363	1.491e-02	-1.3375	0.18107	

(b.) Decision Tree Model

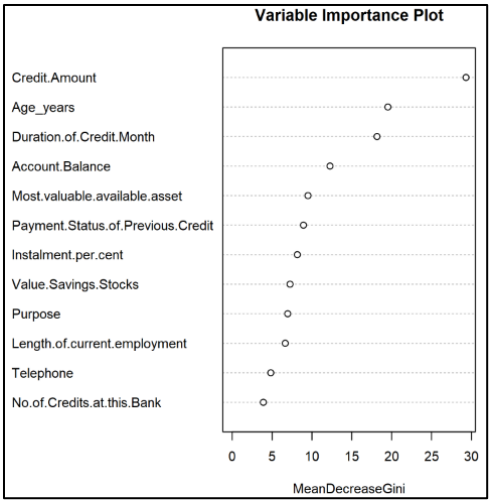
In this decision tree model, the most significant feature is **Amount-Balance**.





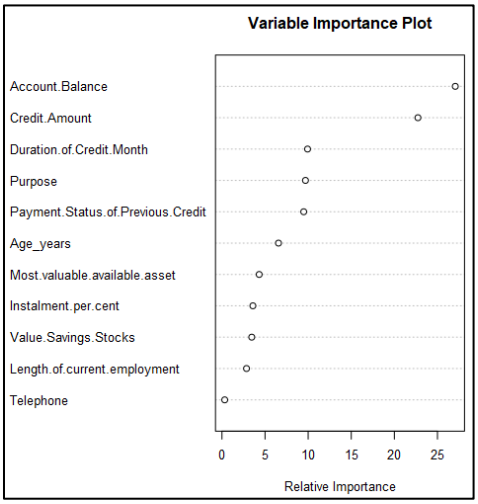
(c.) Forest Model

In Forest Model, the most important variable is **Credit-Amount**.



(d.) Boosted Model

In this Boosted Model, the most important features are **Account-Balance** and **Credit-Amount**.



- Validate your model against the Validation set. What was the overall percent accuracy? Show the confusion matrix. Is there any bias seen in the model's predictions?

  The **Forest model** and the **Boosted model** presented the best accuracy for validation data. From confusion matrix, we can calculate PPV (Positive Predictive Value, also called Precision) and NPV (Negative Predictive Value.) Based on the PPV and NPV, we found that the Forest and the Boosted Models are unbiased, because PPV and NPV are comparable for these models. On the contrary, the Logistic and Decision Tree Models are biased.

Model	Accuracy
Logistic	78.67%
Decision_Tree	74.67%

Forest	79.33%
Boosted	79.33%

Model	Accuracy	Accuracy_Creditworthy	Accuracy_Non-Creditworthy
Logistic	0.7866667	0.8119658	0.6969697
Decision_Tree	0.7466667	0.7913043	0.6000000
Forest	0.7933333	0.7846154	0.8500000
Boosted	0.7933333	0.7936508	0.7916667

Confusion matrix of Logistic		
	Actual_Creditworthy	Actual_Non-Creditworthy
Predicted_Creditworthy	95	22
Predicted_Non-Creditworthy	10	23
Confusion matrix of Decision_Tree		
	Actual_Creditworthy	Actual_Non-Creditworthy
Predicted_Creditworthy	91	24
Predicted_Non-Creditworthy	14	24
Confusion matrix of Forest		
	Actual_Creditworthy	Actual_Non-Creditworthy
Predicted_Creditworthy	102	28
Predicted_Non-Creditworthy	3	17
Confusion matrix of Boosted		
	Actual_Creditworthy	Actual_Non-Creditworthy
Predicted_Creditworthy	100	26
Predicted_Non-Creditworthy	5	19

Confusion matrix of Logistic	
PPV	81.20%
NPV	69.70%
Confusion matrix of Decision_Tree	
PPV	79.13%
NPV	60.00%
Confusion matrix of Forest	
PPV	78.46%
NPV	85.00%
Confusion matrix of Boosted	
PPV	79.37%
NPV	79.17%

You should have four sets of questions answered. (500 word limit)

Step 4: Writeup

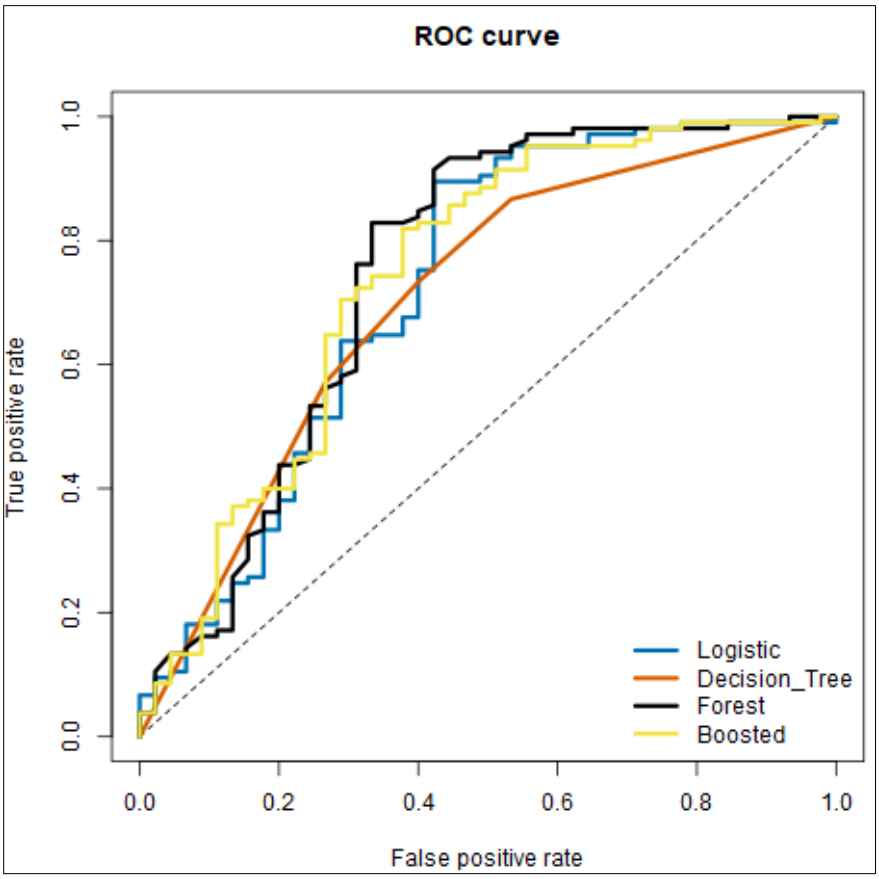
Decide on the best model and score your new customers. For reviewing consistency, if Score_Creditworthy is greater than Score_NonCreditworthy, the person should be labeled as "Creditworthy"

Write a brief report on how you came up with your classification model and write down how many of the new customers would qualify for a loan. (250 word limit)

Answer these questions:

- 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

From the overall accuracy table shown able, we should not consider Decision Tree Model. Additionally, we found that Decision Tree Model and Logistic Model are biased. Judged from accuracy of creditworthy and non-creditworthy, which are our business will focus on. Accuracy of creditworthy of the Forest and the Boosted model are comparable, whereas the former model performed better accuracy on non-creditworthy obviously. Furthermore, the Forest model has higher AUC, which indicates a better classifier for this dataset. From the other viewpoint, the forest model rises the fastest, meaning that we are getting a higher rate of true positive rates versus false positives. This is important because we do not want to extend loans to people who are not creditworthy. Consequently, we can conclude the **Forest Model** is recommended.

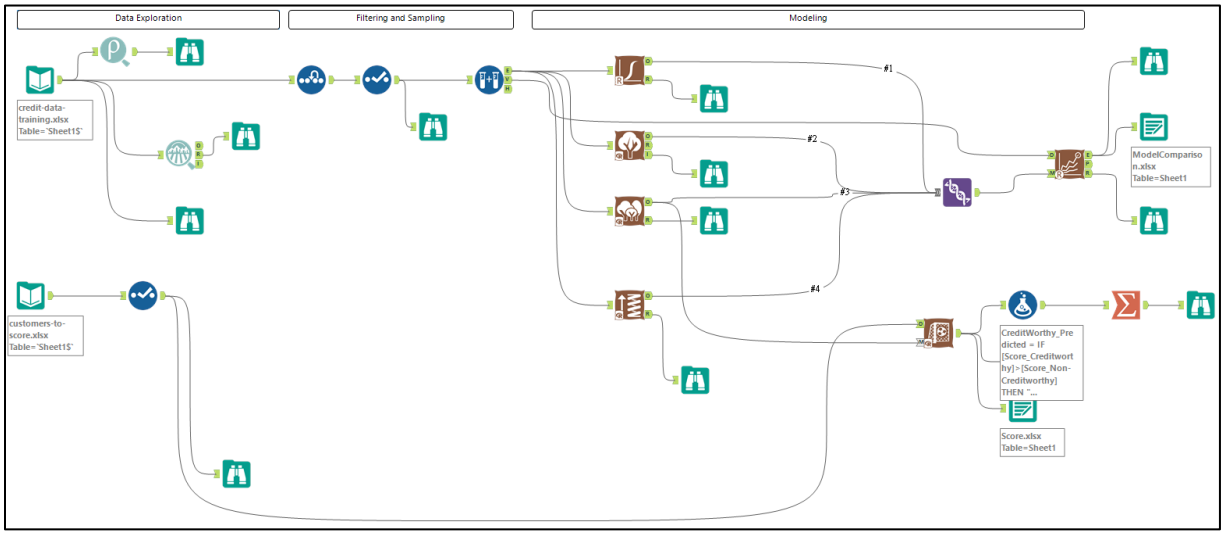


Model	AUC
Logistic	0.724444
Decision_Tree	0.705397
Forest	0.749524
Boosted	0.741587

Note: Remember that your boss only cares about prediction accuracy for Creditworthy and Non-Creditworthy segments.

- How many individuals are creditworthy?
  Based on the forest model, **408** individuals are creditworthy.

Appendix


Data Exploration
Filtering and Sampling
Modeling
credit-data-training.xlsx Table='Sheet1$'
customers-to-score.xlsx Table='Sheet1$'
#1
#2
#3
#4
ModelComparison.xlsx Table=Sheet1
CreditWorthy_Predicted = IF [Score_Creditworthy]>[Score_Non-Creditworthy] THEN "...
Score.xlsx Table=Sheet1