Project: Creditworthiness

Step 1: Business and Data Understanding

Key Decisions:

Answer these questions

- What decisions needs to be made?

The primary decision that need to be made is that whether the individual is Creditworthy or not. Based on this decision the individual will be issued a loan or not.

- What data is needed to inform those decisions?

The data that is needed to inform these decisions is as follows. Firstly, we need the data about the past customers that were issued loan or not. This data can include the following variables that will help us predict the result. Age can be a good variable to base our decision on, so is account balance that will help us understand whether the customer asking for loan has an account with the bank and if yes how much money is in the account. The payment status of previous credit can also be an excellent variable to base our decision on as if the customer has timely Paid Up his previous payments than he/she may do so for this loan as well. The purpose of the loan might also help us further base our decision. Secondly, we require the data of all the variables for the new customers whose decisions that we have to predict.

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

The model is **Binary** as the final prediction can either be Creditworthy or Non-Creditworthy

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.

The following steps were performed for the cleanup process:

1. First we removed **Duration in Current Address** as it has 69% **missing values**.
2. **Guarantors** has been removed due to **low variability** which can be seen from the plot above in which there are 457 instances of None and only 43 instances of Yes.
3. **Foreign Workers** has been removed due to **low variability** which can be seen from the plot above in which there are 481 instances of 1.0 to 1.1 and only 19 instances of 2.0 to 2.1.
4. **Occupation** is removed due to **uniform value** across data.
5. **Concurrent Credits** has been removed due to **uniformity** which can be seen from the plot above in which all of the 500 instances are of the same value.
6. **Number of Dependents** has been removed due to **low variability** which can be seen from the plot above. 427 instances are between 1.0 and 1.1 whereas only 73 instances are between 2.0 to 2.1.
7. There is no logical connection between telephone and creditworthiness and hence **Telephone** has been removed.
8. **Age** has some missing values (2%) that are solved by **imputation** with the **median** data.

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.

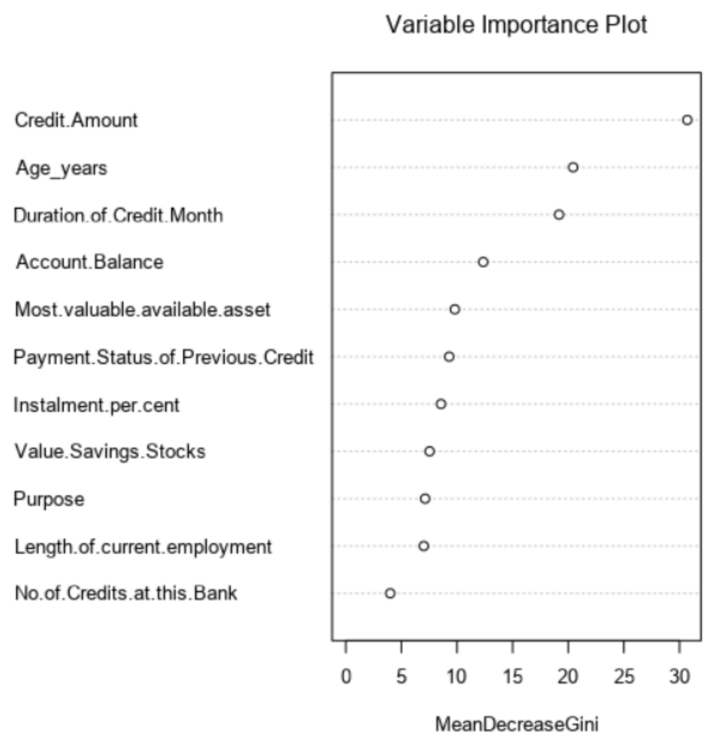
1. Logistic Regression:
   Account Balance, Payment Status of Previous Credit, Credit Amount, Instalment percent, Length of Current Employment, Purpose

	Estimate	Std. Error	z value	Pr(>|z|)	
(Intercept)	-3.2290394	9.845e-01	-3.2800	0.00104	**
Account.BalanceSome Balance	-1.5843791	3.200e-01	-4.9511	7.38e-07	***
Duration.of.Credit.Month	0.0058321	1.365e-02	0.4272	0.6692	
Payment.Status.of.Previous.CreditPaid Up	0.4306851	3.847e-01	1.1195	0.26294	
Payment.Status.of.Previous.CreditSome Problems	1.2872278	5.339e-01	2.4109	0.01591	*
PurposeNew car	-1.7472435	6.271e-01	-2.7862	0.00533	**
PurposeOther	-0.2780516	8.305e-01	-0.3348	0.73778	
PurposeUsed car	-0.7651003	4.108e-01	-1.8624	0.06255	.
Credit.Amount	0.0001734	6.833e-05	2.5375	0.01116	*
Value.Savings.StocksNone	0.5996934	5.065e-01	1.1840	0.2364	
Value.Savings.Stocks£100-£1000	0.1818563	5.621e-01	0.3236	0.74628	
Length.of.current.employment4-7 yrs	0.5259720	4.934e-01	1.0660	0.28642	
Length.of.current.employment< 1yr	0.7776684	3.951e-01	1.9681	0.04906	*
Instalment.per.cent	0.2969774	1.384e-01	2.1457	0.0319	*
Most.valuable.available.asset	0.2877408	1.488e-01	1.9337	0.05315	.
No.of.Credits.at.this.BankMore than 1	0.3918288	3.812e-01	1.0280	0.30397	
Age_years	-0.0180861	1.475e-02	-1.2259	0.22022	

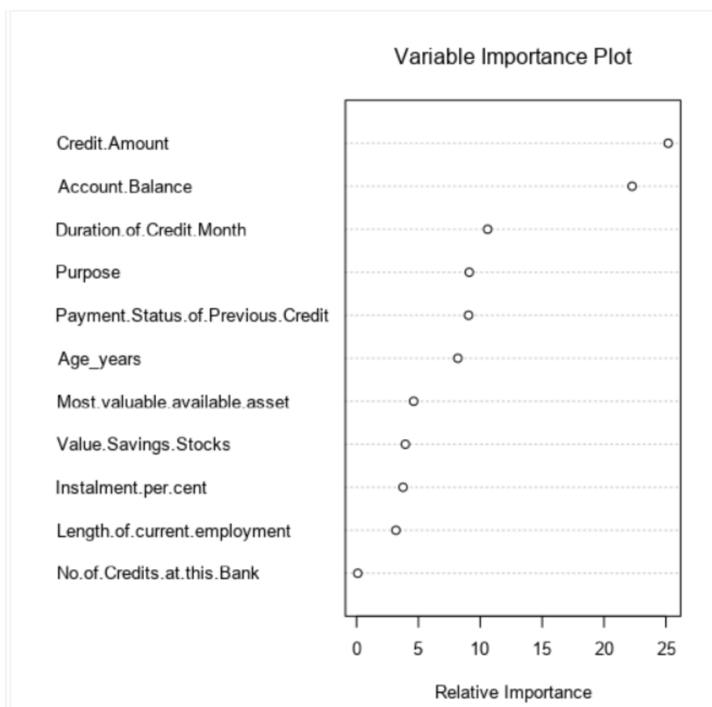
2. Decision Trees
   Account Balance, Duration of Credit Month, Value Savings Stock

```
Leaf Summary
node), split, n, loss, yval, (yprob)
      * denotes terminal node
 1) root 350 97 Creditworthy (0.7228571 0.2771429)
   2) Account.Balance=Some Balance 166 20 Creditworthy (0.8795181 0.1204819) *
   3) Account.Balance=No Account 184 77 Creditworthy (0.5815217 0.4184783)
     6) Duration.of.Credit.Month< 13 74 18 Creditworthy (0.7567568 0.2432432) *
     7) Duration.of.Credit.Month>=13 110 51 Non-Creditworthy (0.4636364 0.5363636)
      14) Value.Savings.Stocks=< £100,£100-£1000 34 11 Creditworthy (0.6764706 0.3235294) *
      15) Value.Savings.Stocks=None 76 28 Non-Creditworthy (0.3684211 0.6315789) *
```

3. Forest



4. Boosted Model



- 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?

Model	Accuracy	F1	AUC	Accuracy_Creditworthy	Accuracy_Non-Creditworthy
Logistic_Regression_Credit	0.8000	0.8661	0.7371	0.9238	0.5111
Decision_Tree_Credit	0.7467	0.8273	0.7054	0.8667	0.4667
Forest_Credit	0.8200	0.8841	0.7414	0.9810	0.4444
Boosted_Credit	0.7800	0.8584	0.7524	0.9524	0.3778

From the below matrices we can see that for:

1. **Boosted Model**: Accuracy: 78%. The model is not biased. (PPV = 100/128 = .78 NPV = 17/22 = .77)

2. **Decision Tree**: Accuracy: 74.67%. The model is biased towards Creditworthy as the accuracy is way higher in this segment. (PPV = 91/118 = .79 NPV = 21/35 = .6)

3. **Forest Model**: Accuracy: 82%. The model is very slightly biased towards Non Creditworthy as the accuracy is higher(PPV = 103/128 = .80 NPV = .90)

4. **Logistic Regression Model**: Accuracy: 80% The model is not biased (PPV = 97/119 = .81 NPV = 23/31 = .74)

[PPV = TP/(TP+FP) NPV = TN/(TN+FN)]

Confusion matrix of Boosted_Credit

	Actual_Creditworthy	Actual_Non-Creditworthy
Predicted_Creditworthy	100	28
Predicted_Non-Creditworthy	5	17

Confusion matrix of Decision_Tree_Credit

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

Confusion matrix of Forest_Credit

	Actual_Creditworthy	Actual_Non-Creditworthy
Predicted_Creditworthy	103	25
Predicted_Non-Creditworthy	2	20

Confusion matrix of Logistic_Regression_Credit

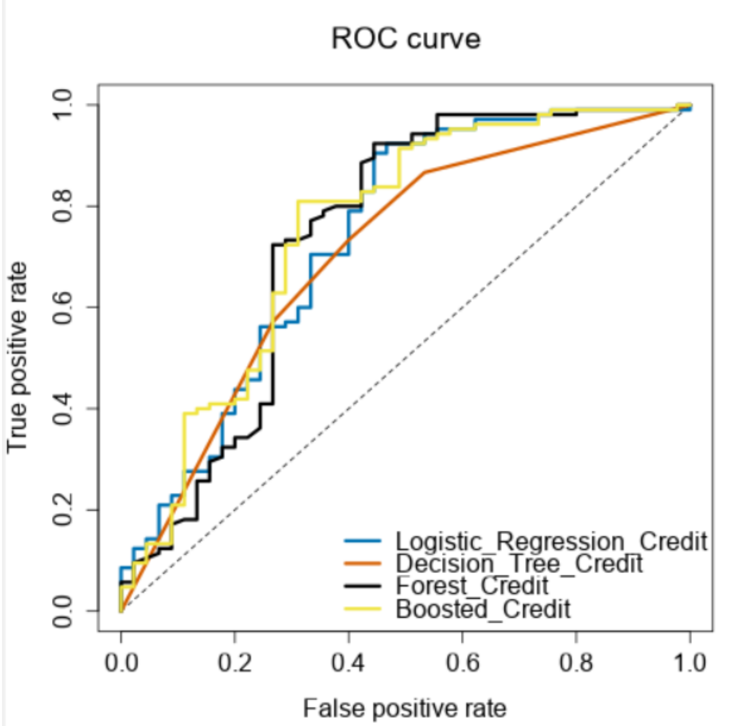
	Actual_Creditworthy	Actual_Non-Creditworthy
Predicted_Creditworthy	97	22
Predicted_Non-Creditworthy	8	23

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

The final model chosen is **Forest**. This was due to the following reasons:

1. The overall accuracy for Forest was 82% which was greater than Boosted, Decision Tree model and Logistic Regression Model.
2. The accuracies for Creditworthy was greater for Forest model (98.10%) when compared to any other model. However, the accuracy for Non-Creditworthy was second highest for Forest Model (44.44%) behind Logistic Regression Model at 51.11%.
3. From the ROC curve we can see that the Forest hugs or is closer to the upper left corner of the plot as compared to the Logistic Regression Model.

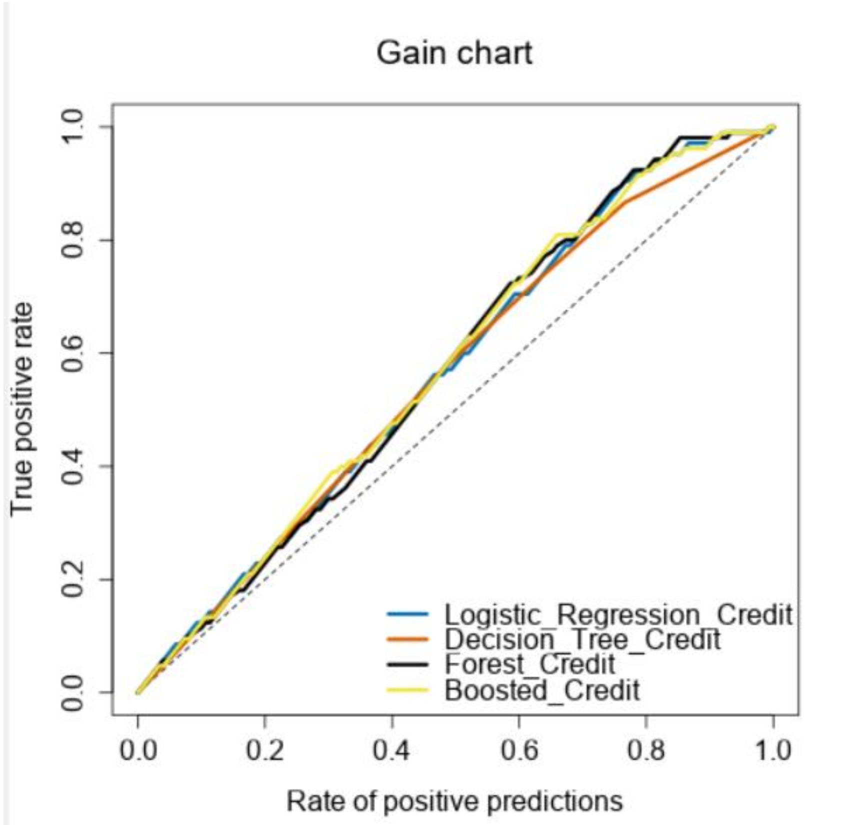


4. From the confusion matrix we can see that there is a high number of Non-Creditworthy values that are predicted Creditworthy.

Confusion matrix of Forest_Credit

	Actual_Creditworthy	Actual_Non-Creditworthy
Predicted_Creditworthy	103	25
Predicted_Non-Creditworthy	2	20

5. In the gain chart we can see that the Forest Model reaches the highest and hence is a better model when compared to others.



- How many individuals are creditworthy?
  409 individuals are Creditworthy.