Project: Predicting Default Risk

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

1. **What decisions needs to be made?**

   The major decision that needs to be made here to determine the credit worthiness of the consumers and whether the bank should process the credit application for a particular consumer

2. **What data is needed to inform those decisions?**

   We need the following data to get to know the outcome
   - Account balance
   - Payment status of previous credit
   - Credit amount
   - Instalment percent

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

We need to use the binary model since, we have only two outcomes, whether to give credit to a particular customer or not.

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 average of the entire data field instead of removing a few data points. (100 word limit)*

Answer this question:

1. 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 are removed
   - Duration in Current address: This field consists of too many null values and imputing those would add bias for the model
   - Concurrent Credits: This field is a string and has only one value which cannot be used for the analysis
   - Guarantors: This field has only yes or no and the value cannot be used for the analysis, since it does get included in any of the model
   - Occupation:
   - No-of-dependents
   - Telephone
   - Foreign-Worker

These fields have values which are either 1\2 and also there are no explanations as to what the value denotes and this does not have any impact on the model
The above variables are removed after considering the p-values,

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.

Create all of the following models: Logistic Regression, Decision Tree, Forest Model, Boosted Model

*Answer these questions for **each model** you created:*

1. Which predictor variables are significant or the most important? Please show the p-values or variable importance charts for all of your predictor variables.

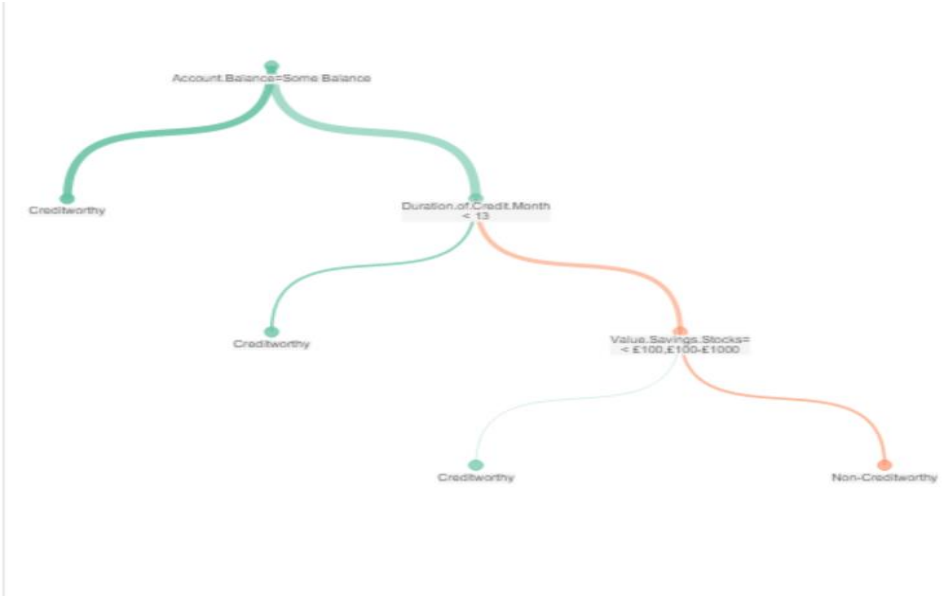
   Logistic Regression:
   The following predictor variables are most important
   - Account balance some balance
   - Credit amount
   - Instalment percent
   - Purpose new car

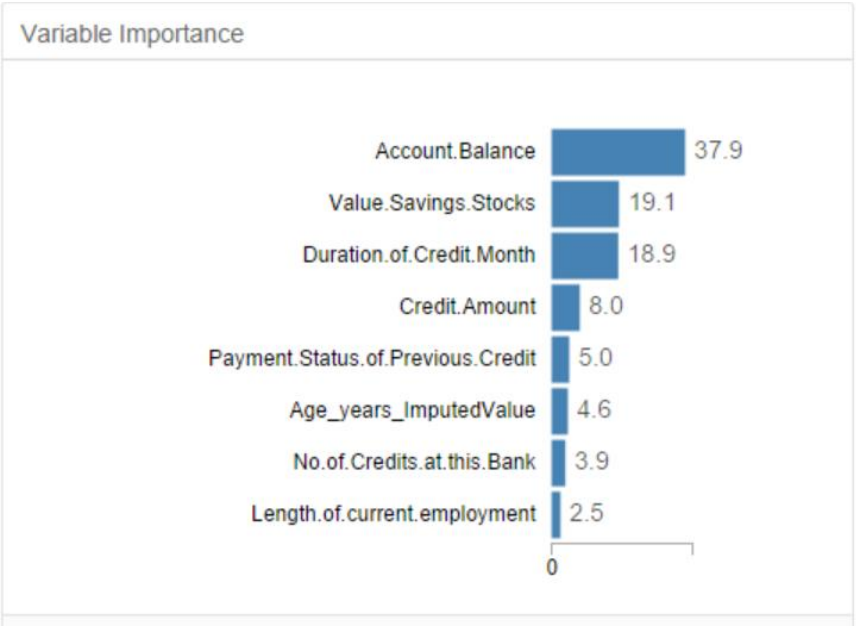
Min	1Q	Median	3Q	Max
-2.044	-0.744	-0.425	0.713	2.581

Coefficients:

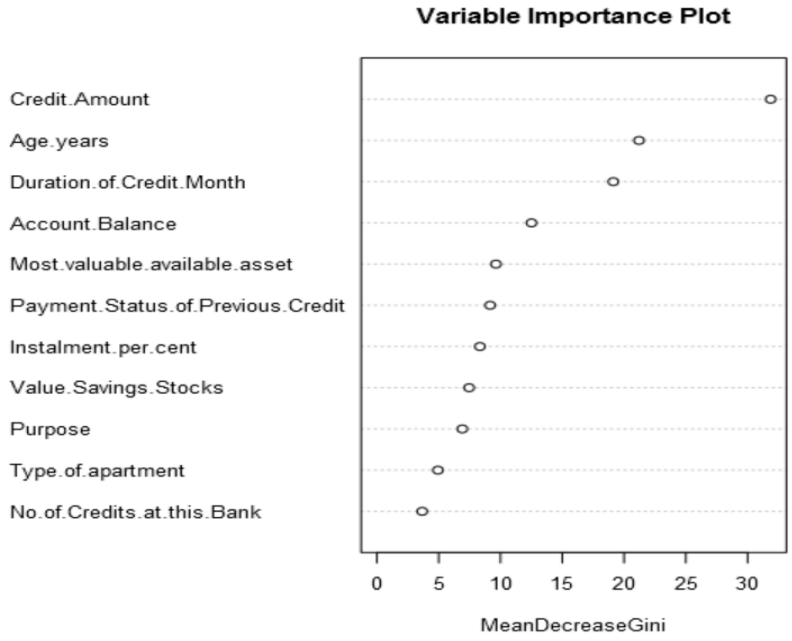
	Estimate	Std. Error	z value	Pr(>|z|)	
(Intercept)	-2.4330771	9.443e-01	-2.5766	0.00998	**
Account.BalanceSome Balance	-1.4917254	3.191e-01	-4.6749	2.94e-06	***
Duration.of.Credit.Month	0.0051942	1.374e-02	0.3781	0.70532	
Payment.Status.of.Previous.CreditPaid Up	0.3550467	3.809e-01	0.9322	0.35126	
Payment.Status.of.Previous.CreditSome Problems	1.1906226	5.268e-01	2.2601	0.02381	*
PurposeNew car	-1.8005507	6.235e-01	-2.8880	0.00388	**
PurposeOther	-0.2829132	8.258e-01	-0.3426	0.73191	
PurposeUsed car	-0.8012336	4.081e-01	-1.9632	0.04962	*
Credit.Amount	0.0001975	6.832e-05	2.8917	0.00383	**
Value.Savings.StocksNone	0.7348247	5.017e-01	1.4646	0.14302	
Value.Savings.Stocks£100-£1000	0.3009189	5.545e-01	0.5427	0.58732	
Instalment.per.cent	0.3048055	1.394e-01	2.1863	0.02879	*
Most.valuable.available.asset	0.3058936	1.532e-01	1.9972	0.0458	*
Type.of.apartment	-0.2590562	2.934e-01	-0.8831	0.3772	
No.of.Credits.at.this.BankMore than 1	0.3400567	3.761e-01	0.9041	0.36596	
Age_years_ImputedValue	-0.0160860	1.415e-02	-1.1366	0.25572	

Decision Tree:

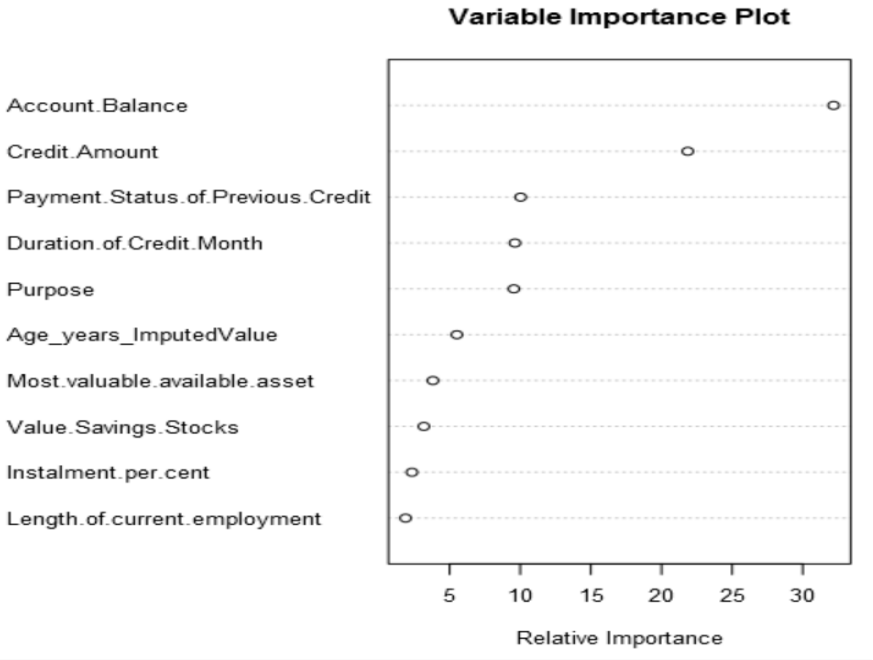




Forest Model:



Boosted Tree:



2. 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?

Logistic Regression:

The accuracy of the logistic regression model is 0.7600

Fit and error measures

Model	Accuracy	F1	AUC	Accuracy_Creditworthy	Accuracy_Non-Creditworthy
X	0.7600	0.8393	0.7230	0.7899	0.6452

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], number of samples that are **correctly** predicted to be Class [class name] divided by number of samples predited to be Class [class name]
AUC: area under the ROC curve, only available for two-class classification.
F1: F1 score, precision * recall / (precision + recall)

Confusion matrix of X

	Actual_Creditworthy	Actual_Non-Creditworthy
Predicted_Creditworthy	94	25
Predicted_Non-Creditworthy	11	20

Performance Diagnostic Plots

Attached is the model comparison chart for logistic regression

Logistic_regression.pdf

Decision Tree:

The accuracy of the decision tree model is 0.7467

Fit and error measures

Model	Accuracy	F1	AUC	Accuracy_Creditworthy	Accuracy_Non-Creditworthy
Decision_Tree	0.7467	0.8273	0.7054	0.7913	0.6000

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], number of samples that are **correctly** predicted to be Class [class name] divided by number of samples predited to be Class [class name]

AUC: area under the ROC curve, only available for two-class classification.

F1: F1 score, precision * recall / (precision + recall)

Confusion matrix of Decision_Tree

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

Attached is the model comparison chart for decision tree

Decision_Tree.pdf

Forest Model:

The accuracy of the Forest model is 0.8000

Fit and error measures

Model	Accuracy	F1	AUC	Accuracy_Creditworthy	Accuracy_Non-Creditworthy
X	0.8000	0.8707	0.7315	0.7953	0.8261

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], number of samples that are **correctly** predicted to be Class [class name] divided by number of samples predited to be Class [class name]

AUC: area under the ROC curve, only available for two-class classification.

F1: F1 score, precision * recall / (precision + recall)

Confusion matrix of X

	Actual_Creditworthy	Actual_Non-Creditworthy
Predicted_Creditworthy	101	26
Predicted_Non-Creditworthy	4	19

Attached is the model comparison chart for forest model

Forest_Model.pdf

Boosted Tree:

The accuracy of the boosted model is 0.7867.

Fit and error measures

Model	Accuracy	F1	AUC	Accuracy_Creditworthy	Accuracy_Non-Creditworthy
X	0.7867	0.8632	0.7524	0.7829	0.8095

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], number of samples that are **correctly** predicted to be Class [class name] divided by number of samples predited to be Class [class name]

AUC: area under the ROC curve, only available for two-class classification.

F1: F1 score, precision * recall / (precision + recall)

Confusion matrix of X

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

Performance Diagnostic Plots

Attached is the model comparison chart for Boosted model

Boosted_model.pdf

There are biases in the models which were used for predicting the credit worthiness,
We see that there are different accuracies in predicting creditworthy vs non creditworthy customers

All the reports have been embedded in this document to give more information on the biases

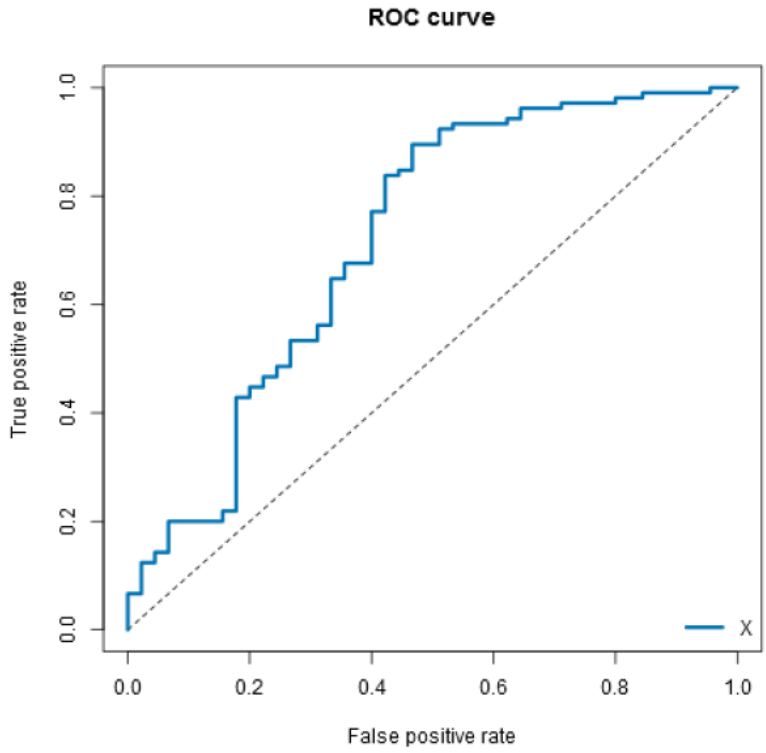
Step 4: Writeup

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"

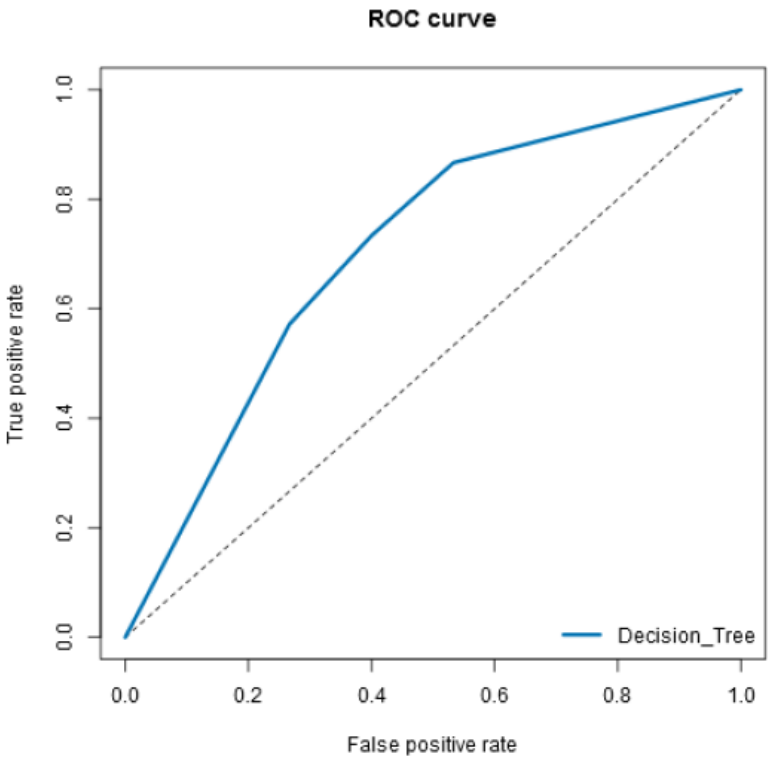
Write a brief report on how you came up with your classification model and write down how many of the new customers would qualify for a loan. (250 word limit)

1. Which model did you choose to use? Please justify your decision using only the following techniques:
    a. Overall Accuracy against your Validation set
       The overall accuracy of the forest model is 0.8000
    b. Accuracies within "Creditworthy" and "Non-Creditworthy" segments
       Accuracy_Creditworthy is 0.7953
       Accuracy_Non-Creditworthy is 0.8261
    c. ROC

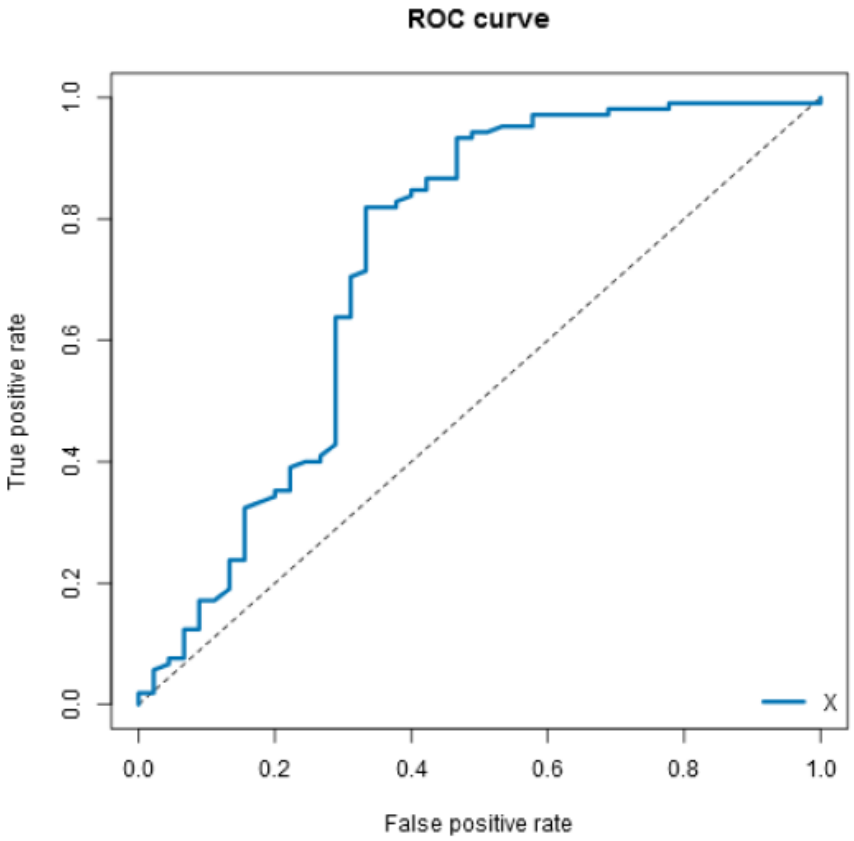
Graph for logistic regression



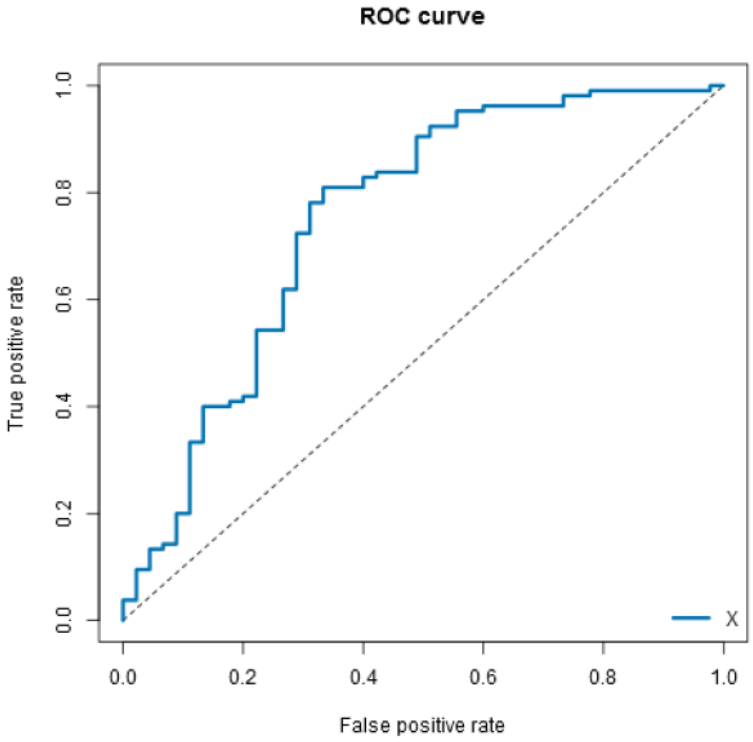
Graph for Decision tree model



Graph for Forest model



Graph for Boosted model



d. Bias in the Confusion Matrices

Confusion matrix of X		
	Actual_Creditworthy	Actual_Non-Creditworthy
Predicted_Creditworthy	101	26
Predicted_Non-Creditworthy	4	19

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

2. How many individuals are creditworthy?
   A total of 411 members are creditworthy