Project 4: 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

- What decisions needs to be made?
  - The bank has received about 500 loan applications and these applications need to be processed, within one week, to determine the set of customers that are worthy to be granted loan based on prediction.

- What data is needed to inform those decisions?

  - We will be needing a dataset of customers' records from past applications to build a model which can make appropriate predictions to help arrive at the desired decision. Some of the important information needed would include age, current account balance, payment status for previous loans, employment status etc. All these factors, and more, would need to be considered in order to select those that would be granted loans.

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

  - Since the decision we're trying to make is determine whether a customer is creditworthy, there are only 2 possible outcomes (Creditworthy or Non-creditworthy). Therefore, using a binary model will be the most suitable approach to make the predictions to help inform our decision.

Step 2: Building the Training Set

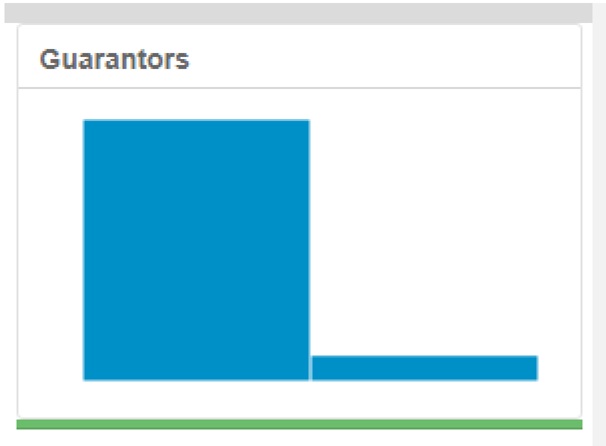
Build your training set given the data provided to you. The data has been cleaned up for you already so you shouldn't ***need to convert any data fields to the appropriate data types.***

Answer this question:

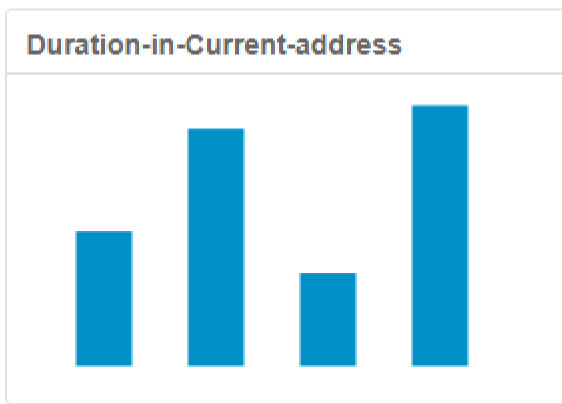
- In your cleanup process, which fields did you remove or impute? Please justify why you removed or imputed these fields. Visualizations are encouraged.

- During my cleanup process, I decided to remove 7 fields out of the 20 available fields. Also, I imputed values in the “Age-years” field using the median of the field dataset because only 2% of the data is missing and this imputation will help in having a more reliable model. The fields removed are as highlighted below;

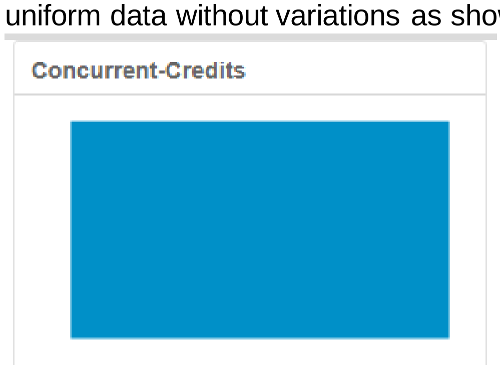
1. **Guarantors:** This field was removed because there is a low variability between the categories of the field as shown below.



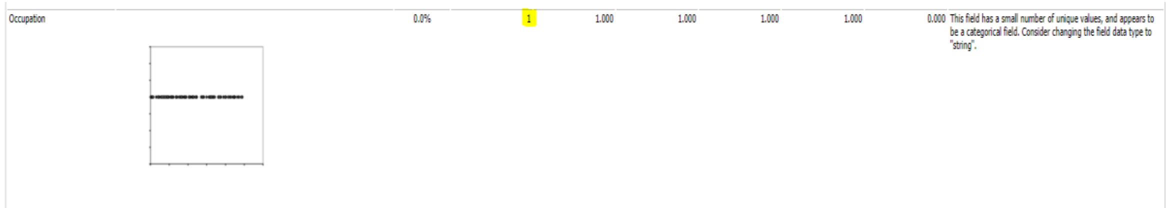
2. **Duration-in-Current-address:** This field was removed because about 69% of the overall data is missing as indicated below.



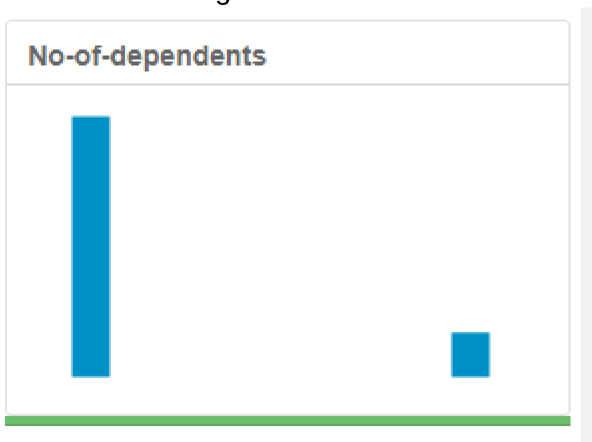
3. **Concurrent-Credits:** This field was removed because it only contains uniform data without variations as shown below.



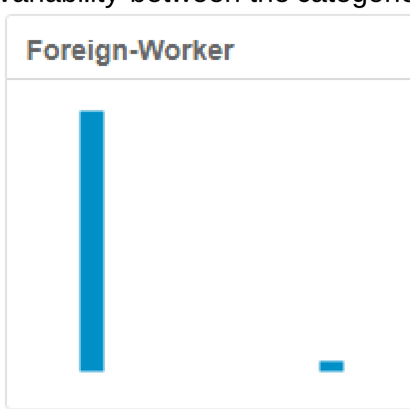
4. **Occupation:** This field was removed because it only contains uniform numeric data without variations as shown in the plot below.



5. **No-of-dependents:** This field was removed because there is a low variability between the categories of the field as shown below.



6. **Telephone:** This field was removed because it is not a logical variable that can have any effect on the model that will be created.
7. **Foreign-WorkerStep:** This field was removed because there is a low variability between the categories of the field as shown below.



3: Train your Classification Models

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

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

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

- 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?

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

- ***Logistic Regression Model**:*
  - From my Logistic + Stepwise regression model, the most significant predictor variables are- Account-Balance, Purpose and Credit-amount. The report showing the P-values can be seen below as it highlights all predictor variables considered.

Deviance Residuals:

Min	1Q	Median	3Q	Max
-2.289	-0.713	-0.448	0.722	2.454

Coefficients:

	Estimate	Std. Error	z value	Pr(>|z|)
(Intercept)	-2.9621914	6.837e-01	-4.3326	1e-05 ***
Account.BalanceSome Balance	-1.6053228	3.067e-01	-5.2344	1.65e-07 ***
Payment.Status.of.Previous.CreditPaid Up	0.2360857	2.977e-01	0.7930	0.42775
Payment.Status.of.Previous.CreditSome Problems	1.2154514	5.151e-01	2.3595	0.0183 *
PurposeNew car	-1.6993164	6.142e-01	-2.7668	0.00566 **
PurposeOther	-0.3257637	8.179e-01	-0.3983	0.69042
PurposeUsed car	-0.7645820	4.004e-01	-1.9096	0.05618 .
Credit.Amount	0.0001704	5.733e-05	2.9716	0.00296 **
Length.of.current.employment4-7 yrs	0.3127022	4.587e-01	0.6817	0.49545
Length.of.current.employment< 1yr	0.8125785	3.874e-01	2.0973	0.03596 *
Instalment.per.cent	0.3016731	1.350e-01	2.2340	0.02549 *
Most.valuable.available.asset	0.2650267	1.425e-01	1.8599	0.06289 .

Significance codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial taken to be 1)

Null deviance: 413.16 on 349 degrees of freedom
Residual deviance: 328.55 on 338 degrees of freedom
McFadden R-Squared: 0.2048, Akaike Information Criterion 352.5

Number of Fisher Scoring iterations: 5

Type II Analysis of Deviance Tests

  - After validating my Logistic + Stepwise regression model against the Validation set, the overall accuracy of the model is 76% which is looking good. From the confusion matrix, the rate of prediction of the Creditworthy category seems to be higher than that of the Non-creditworthy. From the result obtained from this model, I think it has the ability to make a fair prediction when used. The model comparison report is as shown below.

Model Comparison Report

Fit and error measures

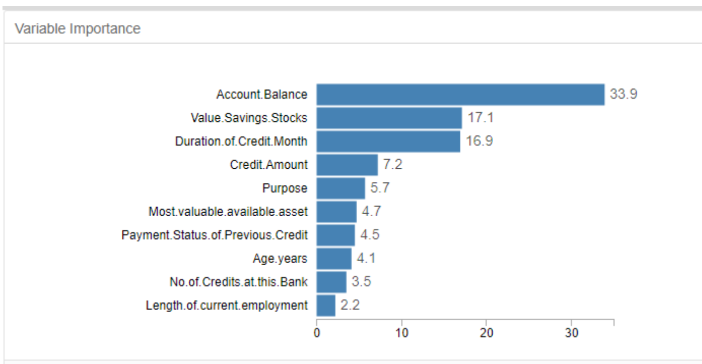
Model	Accuracy	F1	AUC	Accuracy_Creditworthy	Accuracy_Non-Creditworthy
SW_Creditworthiness	0.7600	0.8364	0.7306	0.8762	0.4889

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 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 SW_Creditworthiness

	Actual_Creditworthy	Actual_Non-Creditworthy
Predicted_Creditworthy	92	23
Predicted_Non-Creditworthy	13	22

- ***Decision Tree Model:***
  - From my Decision Tree model, the most important predictor variables are- Account-Balance, Value-savings-stock, Duration-of-credit-month and Purpose. The variable importance chart can be seen below as it highlights all predictor variables considered.



  - After validating my Decision Tree model against the Validation set, the overall accuracy of the model is 74.67% which is quite good. But from the confusion matrix, it seems that the Non-Creditworthy was quite difficult to predict as compared to the Creditworthy category. From the result obtained from this model, I think it has the ability to make a fair prediction when used. The model comparison report is as shown below.

Model Comparison Report

Fit and error measures

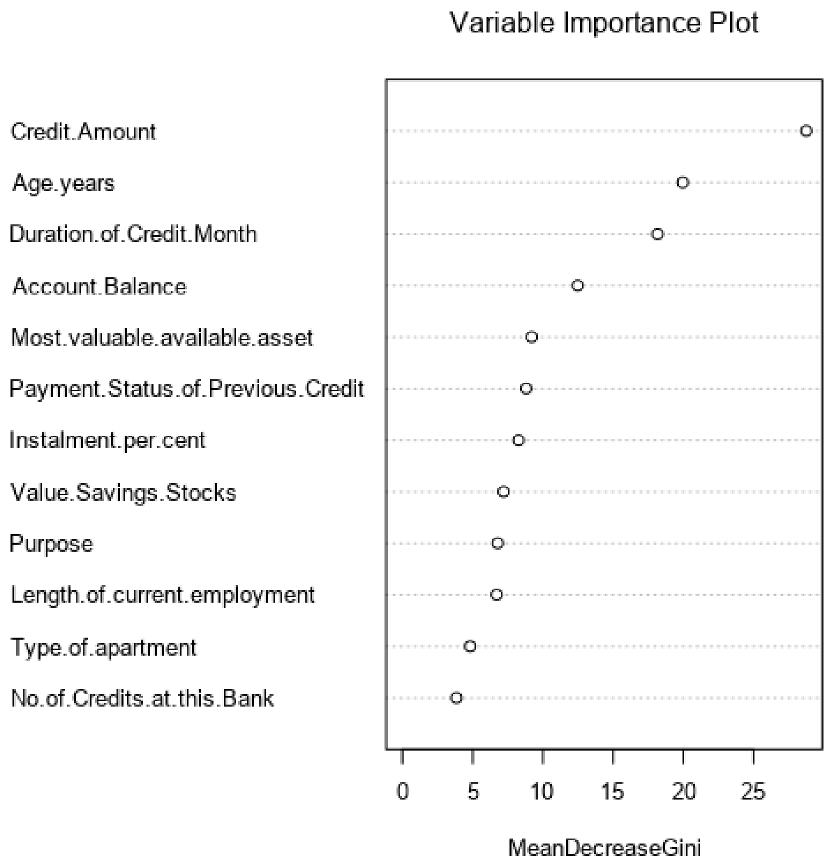
Model	Accuracy	F1	AUC	Accuracy_Creditworthy	Accuracy_Non-Creditworthy
DT_Creditworthy	0.7467	0.8304	0.7035	0.8857	0.4222

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 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 DT_Creditworthy

	Actual_Creditworthy	Actual_Non-Creditworthy
Predicted_Creditworthy	93	26
Predicted_Non-Creditworthy	12	19

- ***Forest Model:***
    - From my Forest model, the most important predictor variables are- Credit-Amount, Age-years and Duration-of-credit-month. The variable importance chart can be seen below as it highlights all predictor variables considered.



    - After validating my Forest model against the Validation set, the overall accuracy of the model is 79.67% which is also looking good. The confusion matrix also indicates how effectively the model predicted the Creditworthy category compared to the Non-creditworthy. From the result obtained from this model, I think it has the ability to make a good prediction when used. The model comparison report is as shown below.

Model Comparison Report

Fit and error measures

Model	Accuracy	F1	AUC	Accuracy_Creditworthy	Accuracy_Non-Creditworthy
FM_Creditworthy	0.7933	0.8681	0.7368	0.9714	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 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 FM_Creditworthy

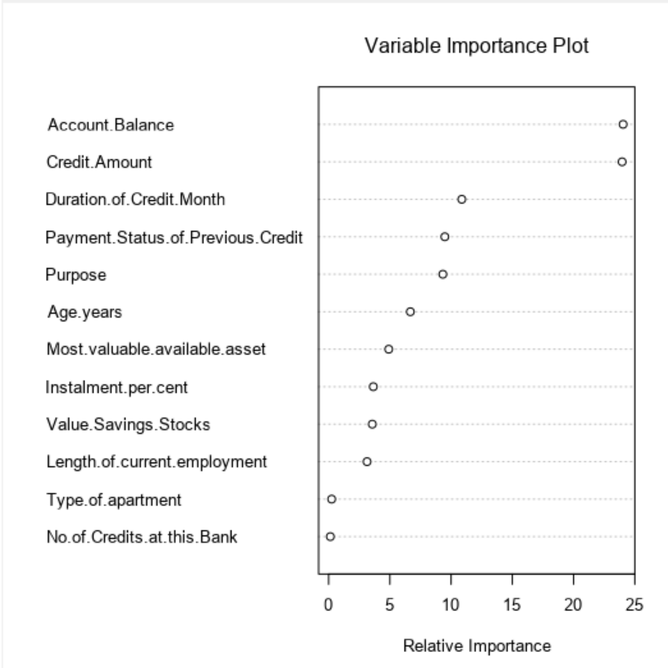
	Actual_Creditworthy	Actual_Non-Creditworthy
Predicted_Creditworthy	102	28
Predicted_Non-Creditworthy	3	17

- ***Boosted Model:***
  - From my Boosted model, the most important predictor variables are- Account-Balance and Credit-Amount. The variable importance chart can be seen below as it highlights all predictor variables considered.

Basic Summary:

Loss function distribution: Bernoulli
Total number of trees used: 4000
Best number of trees based on 5-fold cross validation: 3710

Plots:



  - After validating my Boosted model against the Validation set, the overall accuracy of the model is 79.33% which is also looking good. I can see that it was quite difficult predicting for the Non-creditworthy category compared to the Creditworthy. From the result obtained from this model, I believe it also has the ability to make a good prediction when used. The model comparison report is as shown below.

Model Comparison Report

Fit and error measures

Model	Accuracy	F1	AUC	Accuracy_Creditworthy	Accuracy_Non-Creditworthy
B_Creditworthiness	0.7933	0.8670	0.7505	0.9619	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 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 B_Creditworthiness

	Actual_Creditworthy	Actual_Non-Creditworthy
Predicted_Creditworthy	101	27
Predicted_Non-Creditworthy	4	18

Step 4: Writeup

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

- For the purpose of this predictive analysis, 4 model were created using the Creditworthiness data, which are, Logistic Regression Model, Decision Tree Model, Forest Tree and Boosted Model. Model comparison was carried out on the models and the result is as shown below.

Model Comparison Report

Fit and error measures

Model	Accuracy	F1	AUC	Accuracy_Creditworthy	Accuracy_Non-Creditworthy
SW_Creditworthiness	0.7600	0.8364	0.7306	0.8762	0.4889
DT_Creditworthy	0.7467	0.8304	0.7035	0.8857	0.4222
FM_Creditworthy	0.7933	0.8681	0.7368	0.9714	0.3778
B_Creditworthiness	0.7933	0.8670	0.7505	0.9619	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 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 B_Creditworthiness

	Actual_Creditworthy	Actual_Non-Creditworthy
Predicted_Creditworthy	101	27
Predicted_Non-Creditworthy	4	18

Confusion matrix of DT_Creditworthy

	Actual_Creditworthy	Actual_Non-Creditworthy
Predicted_Creditworthy	93	26
Predicted_Non-Creditworthy	12	19

Confusion matrix of FM_Creditworthy

	Actual_Creditworthy	Actual_Non-Creditworthy
Predicted_Creditworthy	102	28
Predicted_Non-Creditworthy	3	17

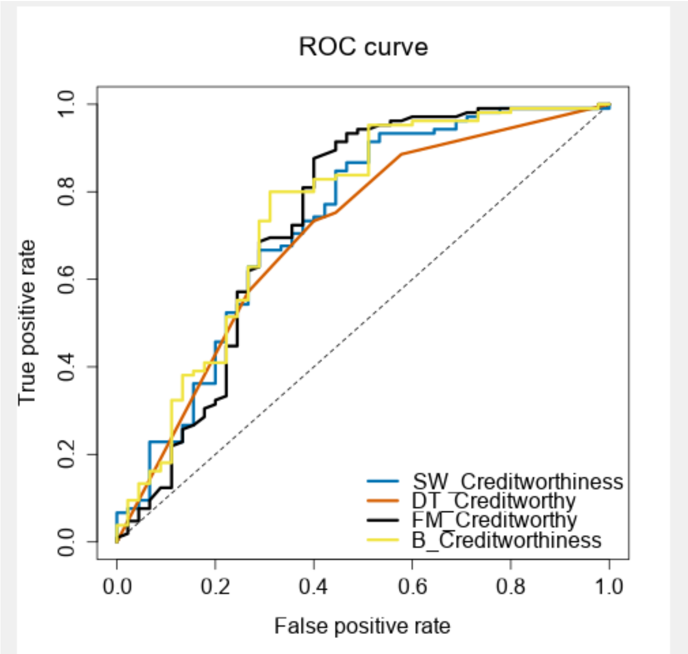
Confusion matrix of SW_Creditworthiness

	Actual_Creditworthy	Actual_Non-Creditworthy
Predicted_Creditworthy	92	23
Predicted_Non-Creditworthy	13	22

From the result shown above, the overall accuracy of all four models looks quite good but the Forrest Model and the Boosted Model show the highest accuracy of 79.33%.

Only one of these two models will have to be selected for the analysis, so looking at the individual accuracy within the Creditworthy and Non-creditworthy for both the Forrest Model and the Boosted Model, I observed that the Forest Model has a higher accuracy of 97.14% compared to 96.19% of the Boosted Model which makes it a better choice to go with.

Also, looking at the ROC curve below, the Forest Model tends to appear to perform better on the True positive rate compared to other models.



This sentiment is also reflected on the confusion matrix as the Forest Model has the highest number of correctly predicted Creditworthy category compared to other models.

Based on the analysis above, I have chosen the Forest Model as the most suitable for the prediction of the 500 new customers.

- How many individuals are creditworthy?
  - After applying the Forest Model to the new customers, 408 out of the 500 customers are creditworthy.