Project: Creditworthiness

Step 1: Business and Data Understanding

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

You work for a small bank and are responsible for determining if customers are creditworthy to give a loan to. Your team typically gets 200 loan applications per week and approves them by hand.

Due to a financial scandal that hit a competitive bank last week, you suddenly have an influx of new people applying for loans for your bank instead of the other bank in your city. Suddenly you have nearly 500 loan applications to process this week!

Your manager sees this new influx as a great opportunity and wants you to figure out how to process all these loan applications within one week.

Fortunately for you, you just completed a course in classification modelling and know how to systematically evaluate the creditworthiness of these new loan applicants.

Key Decisions:

Answer these questions.

- What decisions needs to be made?

  The key decision is evaluating the creditworthiness of loan applicants and providing the list of creditworthy customers to the Manager.

- What data is needed to inform those decisions?

  Data on all past applications
  The list of customers that need to be processed in the next few days.

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

  Binary: it's a Yes or No type of decision. i.e., Identifying people who qualify or do not qualify for a loan.

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 should not ***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 clean-up process, which fields did you remove or impute? Please justify why you removed or imputed these fields. Visualizations are encouraged.

Analysing the result after applying the “Field summary” tool on the dataset, I decided to

- Remove “**Duration-Current-address**” column seeing ~70% of its data is missing.
- Remove “**Concurrent-Credits”** and “**Occupation**” data field as all the data were the same (completely uniform).

Record	Name	Field Category	Min	Max	Median	Std. Dev.	Percent Missing	Unique Values	Mean
1	Age-years	Numeric	19	75	33	11.501522	2.4	54	35.6372
2	Credit-Amount	Numeric	276	18,424	2,236.5	2,831.386861	0	464	3,199.98
3	Duration-in-Current-address	Numeric	1	4	2	1.150017	68.8	5	2.66025
4	Duration-of-Credit-Month	Numeric	4	60	18	12.30742	0	30	21.434
5	Foreign-Worker	Numeric	1	2	1	0.191388	0	2	1.038
6	Instalment-per-cent	Numeric	1	4	3	1.113724	0	4	3.01
7	Most-valuable-available-asset	Numeric	1	4	3	1.064268	0	4	2.36
8	No-of-dependents	Numeric	1	2	1	0.35346	0	2	1.146
9	Occupation	Numeric	1	1	1	0	0	1	1
10	Telephone	Numeric	1	2	1	0.490389	0	2	1.4
11	Type-of-apartment	Numeric	1	3	2	0.539814	0	3	1.928
12	Account-Balance	String	[Null]	[Null]	[Null]	[Null]	0	2	[Null]
13	Concurrent-Credits	String	[Null]	[Null]	[Null]	[Null]	0	1	[Null]
14	Credit-Application-Result	String	[Null]	[Null]	[Null]	[Null]	0	2	[Null]
15	Guarantors	String	[Null]	[Null]	[Null]	[Null]	0	2	[Null]
16	Length-of-current-employment	String	[Null]	[Null]	[Null]	[Null]	0	3	[Null]
17	No-of-Credits-at-this-Bank	String	[Null]	[Null]	[Null]	[Null]	0	2	[Null]
18	Payment-Status-of-Previous-Credit	String	[Null]	[Null]	[Null]	[Null]	0	3	[Null]
19	Purpose	String	[Null]	[Null]	[Null]	[Null]	0	4	[Null]
20	Value-Savings-Stocks	String	[Null]	[Null]	[Null]	[Null]	0	3	[Null]

- Remove “**Guarantor**”, “**Foreign worker**”, “**No of dependents**” for low variability.
- Missing data in **“Age”** was replaced using the median of the entire data field according to the suggestion in the project details.
- Remove “**Telephone”** from the data with the reasoning that there is no logical reason for including the variable.

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:*

- 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: based on the report below, the significant predictive variables are Account Balance, Payment Status of Previous Credit, Purpose, Credit Amount, Length of Current Employment, and Installment per Cent.

	Estimate	Std. Error	z value	Pr(>|z|)	
(Intercept)	-2.990817	1.013e+00	-2.9527	0.00315	**
Payment.Status.of.Previous.CreditPaid Up	0.402974	3.843e-01	1.0487	0.2943	
Payment.Status.of.Previous.CreditSome Problems	1.259683	5.334e-01	2.3616	0.0182	*
PurposeNew car	-1.755074	6.278e-01	-2.7954	0.00518	**
PurposeOther	-0.290165	8.359e-01	-0.3471	0.72848	
PurposeUsed car	-0.785627	4.124e-01	-1.9049	0.05679	.
Type.of.apartment	-0.254565	2.958e-01	-0.8605	0.38949	
Value.Savings.StocksNone	0.609298	5.099e-01	1.1949	0.23213	
Value.Savings.Stocks£100-£1000	0.172241	5.649e-01	0.3049	0.76046	
No.of.Credits.at.this.BankMore than 1	0.362688	3.816e-01	0.9505	0.34184	
Credit.Amount	0.000177	6.841e-05	2.5879	0.00966	**
Account.BalanceSome Balance	-1.543669	3.233e-01	-4.7745	1.80e-06	***
Age.years	-0.015092	1.539e-02	-0.9809	0.32666	
Length.of.current.employment4-7 yrs	0.530959	4.932e-01	1.0767	0.28163	
Length.of.current.employment< 1yr	0.777372	3.957e-01	1.9646	0.04946	*
Most.valuable.available.asset	0.325606	1.557e-01	2.0918	0.03645	*
Duration.of.Credit.Month	0.006391	1.371e-02	0.4660	0.6412	
Instalment.per.cent	0.310524	1.399e-01	2.2197	0.02644	*

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: 322.19 on 332 degrees of freedom
McFadden R-Squared: 0.2202, Akaike Information Criterion 358.2

Number of Fisher Scoring iterations: 5

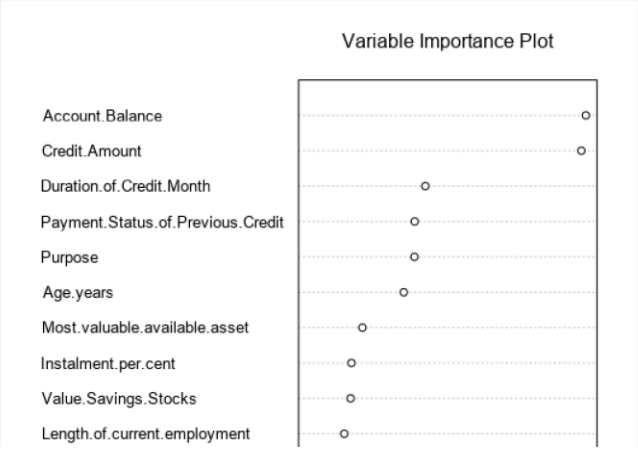
Boosted Model: based on the variable importance plot below, the top 3 important predictive variables are Amount Balance, Credit Amount and Duration of Credit Month.

Report for Boosted Model Bosted_Results

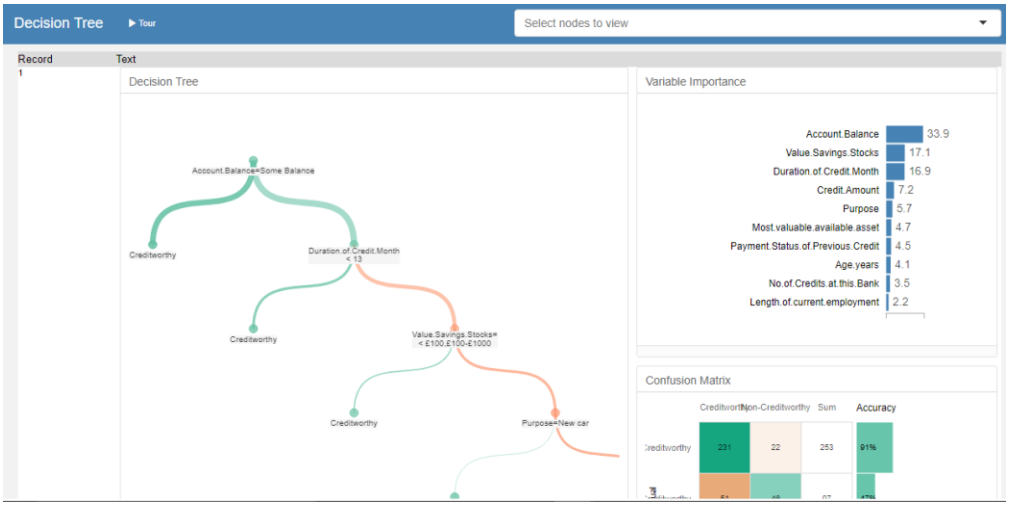
Basic Summary:

Loss function distribution: Bernoulli
Total number of trees used: 4000
Best number of trees based on 5-fold cross validation: 1955

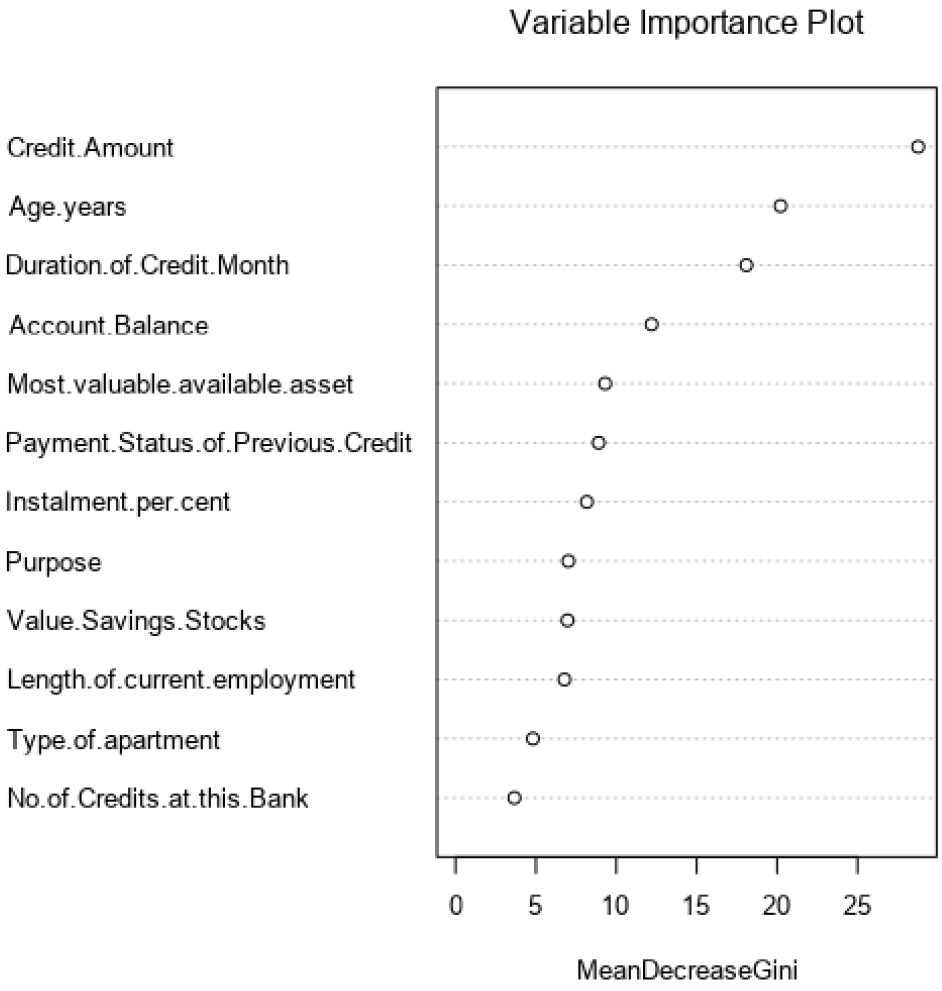
Plots:



Decision Tree: based on the variable importance report below, the top 3 predictive variables are Account Balance, Value Savings Stocks, and Duration of Credit Month.



Forest Model: based on the variable importance plot below, the top 3 predictive variables are Credit Amount, Age Years, and Duration of Credit Month.



- 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?

  Below image shows Model comparison report for the four models applied to the dataset, which shows side by side comparison of their accuracy.

Model Comparison Report

Fit and error measures

Model	Accuracy	F1	AUC	Accuracy_Creditworthy	Accuracy_Non-Creditworthy
Logistic_Result	0.7800	0.8520	0.7310	0.9048	0.4889
DecisionTree_Results	0.7467	0.8304	0.7035	0.8857	0.4222
ForestModel_Results	0.8200	0.8831	0.7447	0.9714	0.4667
Bosted_Results	0.7933	0.8670	0.7539	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.

  Below image shows the confusion matrix for the models

Confusion matrix of Bosted_Results

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

Confusion matrix of DecisionTree_Results

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

Confusion matrix of ForestModel_Results

	Actual_Creditworthy	Actual_Non-Creditworthy
Predicted_Creditworthy	102	24
Predicted_Non-Creditworthy	3	21

Confusion matrix of Logistic_Result

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

  The Decision Tree model achieved an overall accuracy of 75% in the validation set, comparing to the other model, this is the one with the less accurate. This model again shows to be biased to predict that the customer is creditworthy.

  The Logistic Regression Model achieved an overall accuracy of 78%. The model is bit biased to predict that the customer is creditworthy, comparing to the other models, it shows the best Accuracy for the Non-Creditworthy customers.

The Boosted Model achieved an overall accuracy of 79% on the validation set, its Accuracy to predict Non Creditworthy customers is bias, it means the model is biased to predict that the customer is creditworthy.

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:

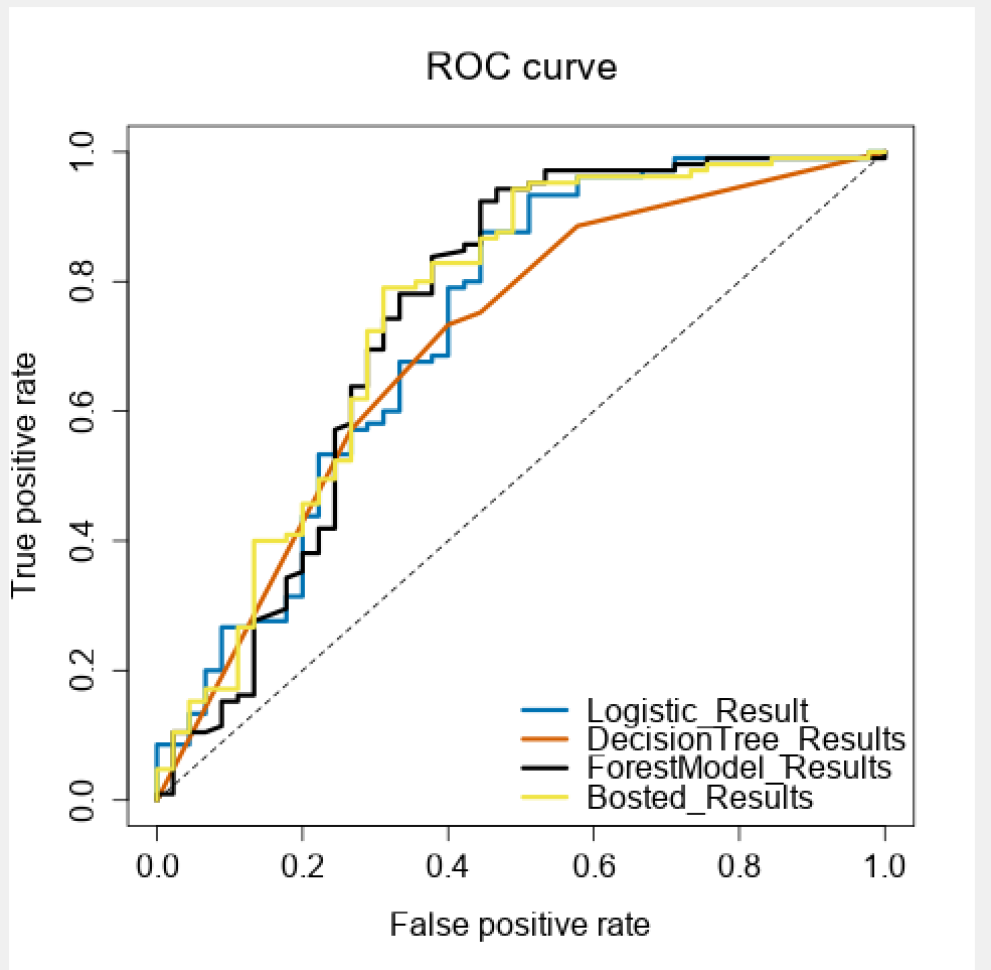
- Which model did you choose to use? Please justify your decision using **all** of the following techniques. Please only use these techniques to justify your decision:
    - Overall Accuracy against your Validation set

      Forest model is chosen as it outperformed other models by offering the highest accuracy at 82% against validation set.

    - Accuracies within "Creditworthy" and "Non-Creditworthy" segments

      Forest Model accuracies for creditworthy and non-creditworthy are among the highest.

- **ROC graph**



- Bias in the Confusion Matrices.

  The accuracy difference between creditworthy and non-creditworthy are comparable which makes it least bias towards any decisions. This is crucial in avoiding lending money to customers with high probability of defaulting while ensuring opportunities are not overlooked by not loaning to creditworthy customers.

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

- How many individuals are creditworthy?

  408 Applicants are creditworthy.