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?
  Whether or not a loan applicant is credit worthy.

- What data is needed to inform those decisions?
  We need to find the credit application report based on the predictor variables of the applicant from old customer's data, including: Account balance, Payment Status of Previous Credit, Purpose, Credit amount, Value Saving Stocks, Length of Current Employment, Insalment Percent, Guarantors Most valuable available asset, Age, Number of Credit at this Bank and number of dependents.

- What kind of model (Continuous, Binary, Non-Binary, Time-Series) do we need to use to help make these decisions?
  Since the decision needed to be made is Credit Application Result has two result: Credit worthy or Non-Credit worthy, we need to apply the Binary model to predict the outcome.

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.

  - The field **Duration in Current Address** contains 69% Null values, then we should remove this field out of the analysis, while the 2.4% missing in **Age (years)** can be replaced by the mean value.
  - **Occupation** and **Concurrent credit** fields only have one value, then we eliminate these fields out of the prediction model.

  - I alsdo removed **Foreign Worker, Guarantors** and **No of dependents** since they had low variabiltity. Test running the models showed **Telephone** is not revelant to the prediction, then this field can be removed.

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.

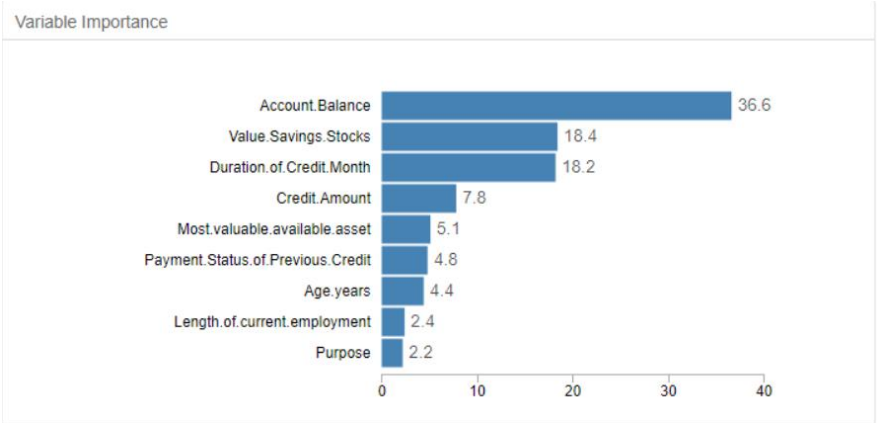
  The most important predictor variables of each model:

  1. Logistic Regression Model: **Account Balance, Payment Satus of Previous Credit, Purpose, Credit Amount, Installment percent** and **Credit Amount**

Coefficients:

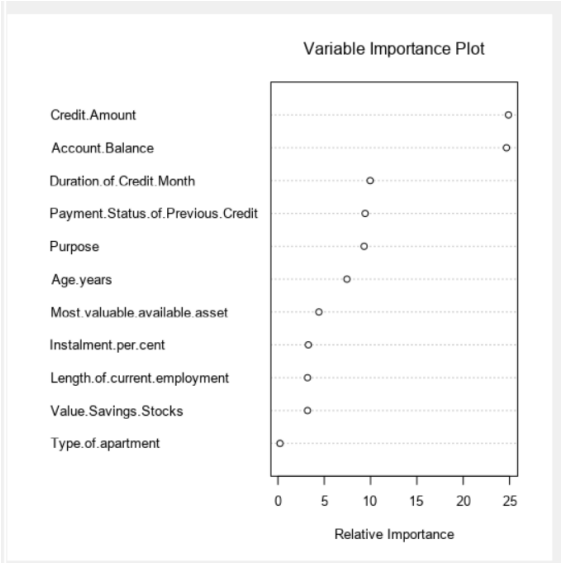
	Estimate	Std. Error	z value	Pr(>|z|)
(Intercept)	-2.9621914	6.837e-01	-4.3326	1e-05 ***
Account.BalanceSome Balance	-1.6053228	3.067e-01	-5.2344	1.65e-07 ***
Payment.Status.of.Previous.CreditPaid Up	0.2360857	2.977e-01	0.7930	0.42775
Payment.Status.of.Previous.CreditSome Problems	1.2154514	5.151e-01	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 .

  2. Decision tree Model: **Account Balance, Value savings Stocks, Credit Months, Credit Amount** and **Duration of Credit Month**



3. Forest Model: **Account Balance, Age years, Credit Amount** and **Duration of Credit Month** and **Most valuable available asset**

4. Boosted Model: **Account Balance, Payment of Previous Credit, Credit Amount, Purpose** 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?

  The overall accuraccy of each model:

Model	Accuracy
Tree_Model	0.7467
Forest_Model	0.7933
Boosted_Model	0.7933
Log_Reg_Model	0.7600

  1. Logistic Regression model:

  Accuraccy: 76%
  Confusion matrix:

Confusion matrix of Log_Reg_Model

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

  Positive accuracy: TP/(TP+FP) = 92/(92+23) = 80%
  Negative accuracy: TN/(TN+FN) = 22/(22+13) = 63%

  with: - TP: True Positive outcomes
  - FP: False Positive outcomes
  - TN: True Negative outcomes
  - FN: False Negative outcomes

  Bias: This model has the bias toward accurately predicting creditworthy result (positive), as the accuracy of positive prediction is 80% compared to 63% of the negative.

  2. Decision Tree model:

  Accuraccy: 74.7%
  Confusion matrix:

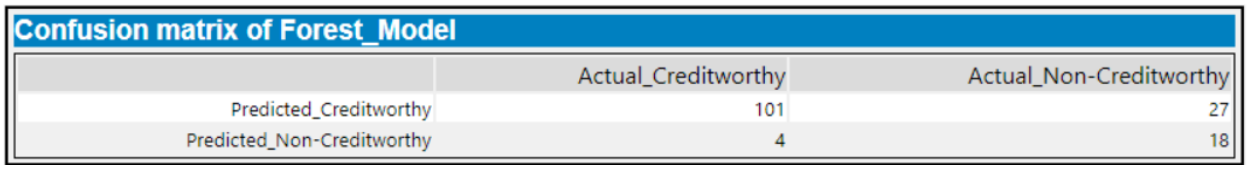
Confusion matrix of Tree_Model

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

  Positive accuracy: TP/(TP+FP) = 91/(91+24) = 79.1%
  Negative accuracy: TN/(TN+FN) = 21/(21+14) = 60%
  Bias: This model has the bias toward accurately predicting creditworthy result (positive), as the accuracy of positive prediction is 79.1% compared to 60% of the negative.

3. Forest model:

Accuracy: 79.33%
Confusion matrix:

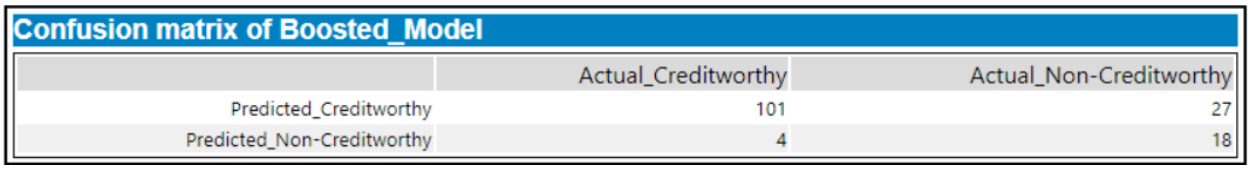


Confusion matrix of Forest_Model		
	Actual_Creditworthy	Actual_Non-Creditworthy
Predicted_Creditworthy	101	27
Predicted_Non-Creditworthy	4	18

Positive accuracy: TP/(TP+FP) = 101/(101+27) = 78.9%
Negative accuracy: TN/(TN+FN) = 18/(18+4) = 81.8%
Bias: This model doesn't show bias since its positive prediction accuracy and negative's are close.

4. Boosted model:

Accuracy: 79.33%
Confusion matrix:



Confusion matrix of Boosted_Model		
	Actual_Creditworthy	Actual_Non-Creditworthy
Predicted_Creditworthy	101	27
Predicted_Non-Creditworthy	4	18

Positive accuracy: TP/(TP+FP) = 101/(101+27) = 78.9%
Negative accuracy: TN/(TN+FN) = 18/(18+4) = 81.8%
Bias: This model doesn't show bias since its positive prediction accuracy and negative's are close.

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"

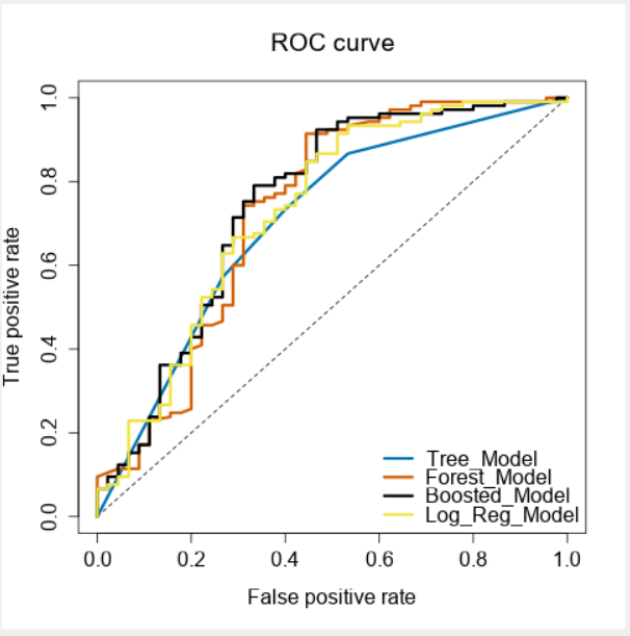
Write a brief report on how you came up with your classification model and write down how many of the new customers would qualify for a loan. (250 word limit)

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

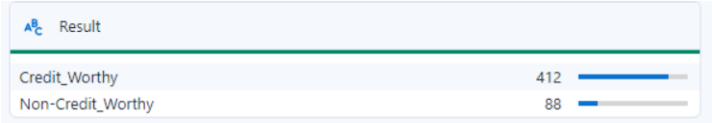
  The model I decide to use to score: Forest model. But we can either choose Forest and Boosted model since these two have the same prediction accuracies. Besides, because of the following reasons below:

    - Overall accuracy: 79.33% (highest)
    - Accuracies within “Creditworthy” and “Non-Creditworthy” segments: 78.9% and 81.9%, respectively. The model’s accuracy of predicting “Non-Creditworthy” is another factor, as it has higher chance of predicting the “Non-Creditworthy” than the others. Therefore, reducing the posibility of False-Positive outcomes, then reduce the risk of approving the non-worthy risky loans.
    - The model does not show bias toward any outcomes.
    - ROC graph of the Forest model (the Orange curve):



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

- How many individuals are creditworthy? 412 applicants are creditworthy:



Before you Submit

Please check your answers against the requirements of the project dictated by the rubric here. Reviewers will use this rubric to grade your project.

Alteryx workflow: