Project: Creditworthiness

Complete each section. When you are ready, save your file as a PDF document

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 needs to check if the new customers are creditworthy to give a loan to.
- What data is needed to inform those decisions?
    - ✓ Data needed to inform the decision are:
        - Non-financial (age, occupation, dependents)
        - Financial (assets, stocks)
        - Information on past applications
- What kind of model (Continuous, Binary, Non-Binary, Time-Series) do we need to use to help make these decisions?
    - ✓ Considering that the predicted outcome falls into 2(two) categories (“Creditworthy” vs “Non-Creditworthy”), we need to use Binary Classification Models.

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:*

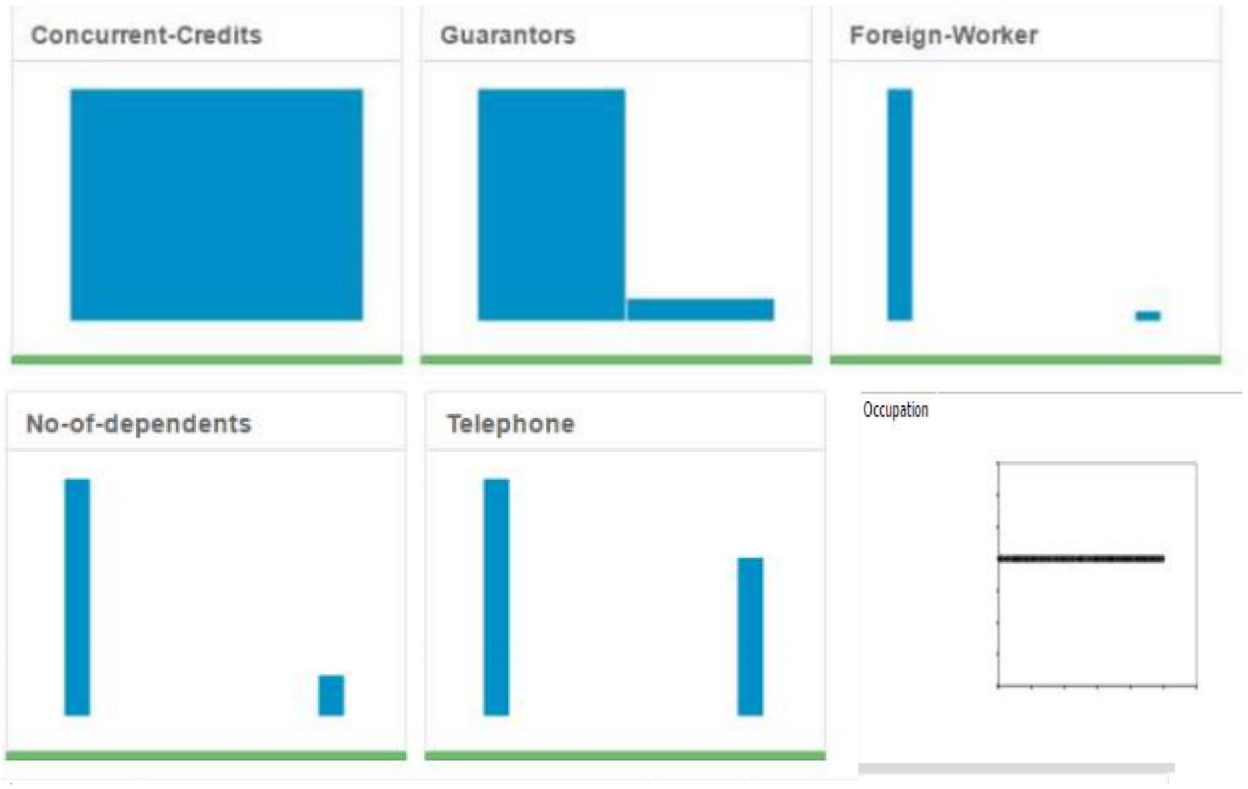
Variable	**Data Type**
Credit-Application-Result	String
Account-Balance	String
Duration-of-Credit-Month	Double
Payment-Status-of-Previous-Credit	String
Purpose	String
Credit-Amount	Double
Value-Savings-Stocks	String
Length-of-current-employment	String
Instalment-per-cent	Double
Guarantors	String
Duration-in-Current-address	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

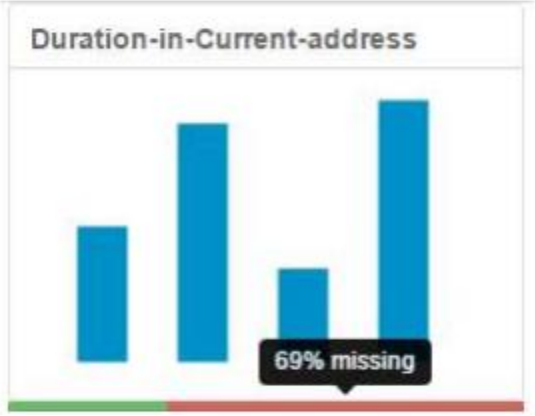
To achieve consistent results reviewers expect.

Answer this question:

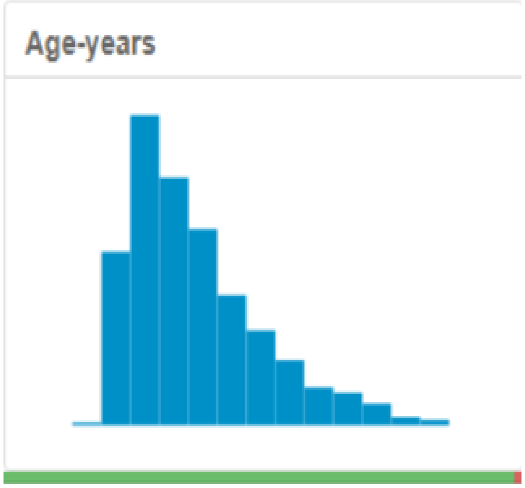
- In your cleanup process, which fields did you remove or impute? Please justify why you removed or imputed these fields. Visualizations are encouraged.

    - ✓ First, I checked each variable variability using Alteryx *Fields Summary Tool,* and as we can se from the figure below**, Concurrent Credits, Foreign worker, No of dependents, Guarantors,** and **Occupation fields** have low variability so I decided to not include them in the model. **Telephone** field should also be removed due to its irrelevancy to the customer creditworthy.





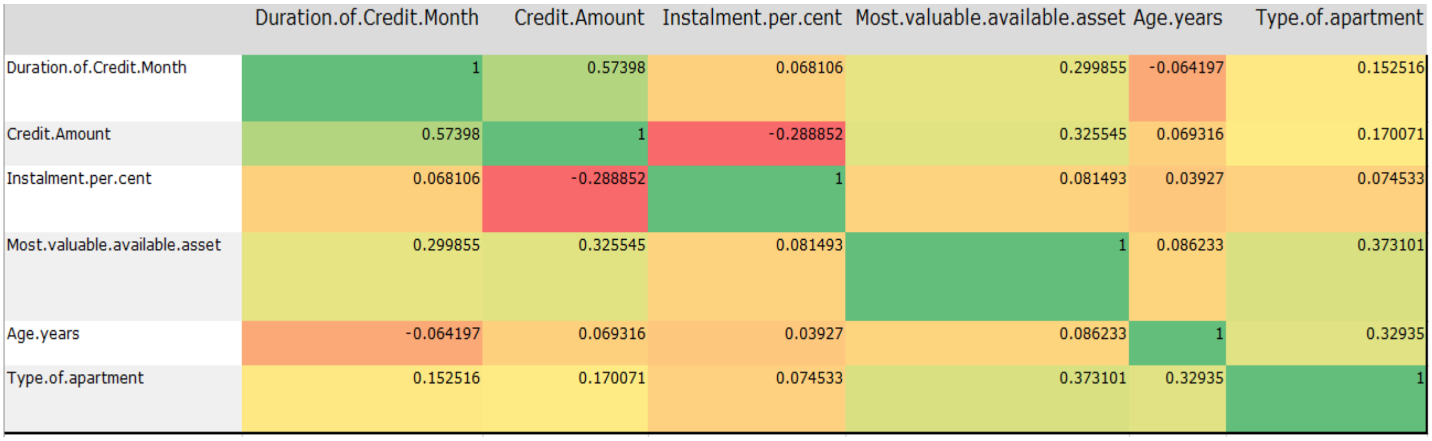
I also noticed that **Duration-in-Current-address** has 68.8% missing data. It would add unnecessary bias to the model if I imputed such a huge number of data so I decided to also not include this variable in my prediction model



Next, I decided to impute Age years because it is only missing 2.4% of its values. The distribution of this variable is asymmetric. I imputed using the median age of 33, considering that the median is less affected by outliers and skewed data.

I avoided using the mean to impute because as the data becomes skewed, the mean loses its ability to provide the best central location.

- ✓ Finally, to be completely confident on the remaining variables that will be used for prediction, I used the Pearson correlation coefficient to see if there is a strong connection on numerical variables. We can see that **Credit.Amount** is positively correlated with **Duration.of.Credit.Month** but still the coefficient is less than .70 so no risk for damaging multicollinearity.



	Duration.of.Credit.Month	Credit.Amount	Instalment.per.cent	Most.valuable.available.asset	Age.years	Type.of.apartment
Duration.of.Credit.Month	1	0.57398	0.068106	0.299855	-0.064197	0.152516
Credit.Amount	0.57398	1	-0.288852	0.325545	0.069316	0.170071
Instalment.per.cent	0.068106	-0.288852	1	0.081493	0.03927	0.074533
Most.valuable.available.asset	0.299855	0.325545	0.081493	1	0.086233	0.373101
Age.years	-0.064197	0.069316	0.03927	0.086233	1	0.32935
Type.of.apartment	0.152516	0.170071	0.074533	0.373101	0.32935	1

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

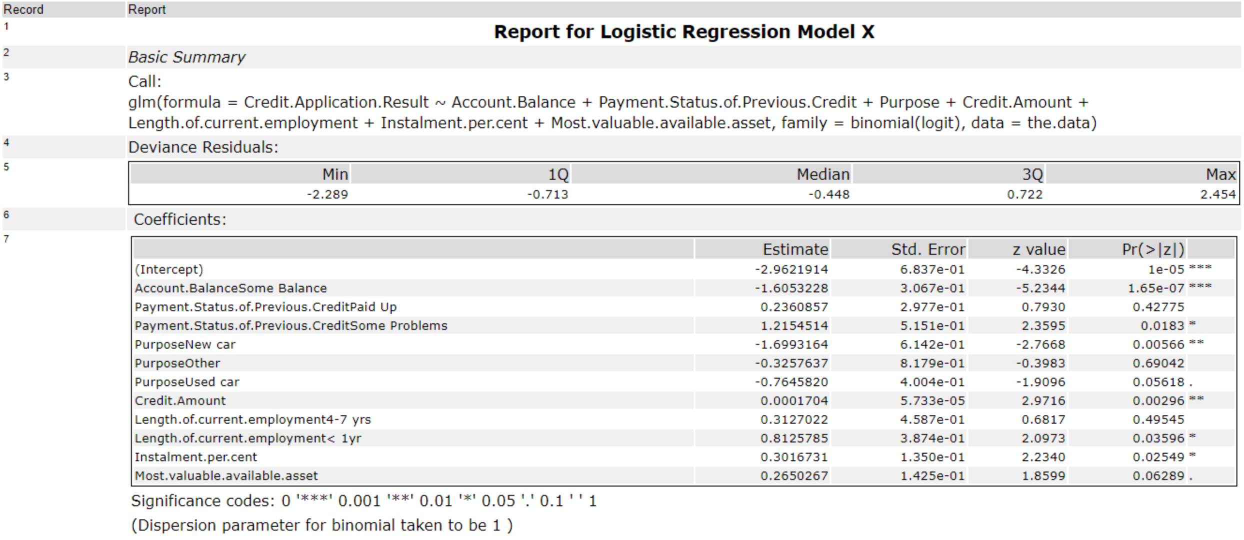
*Answer these questions for **each model** you created:*

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

- 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)

1. Logistic Regression



Record | Report

1 **Report for Logistic Regression Model X**

2 *Basic Summary*

3 Call:
glm(formula = Credit.Application.Result ~ Account.Balance + Payment.Status.of.Previous.Credit + Purpose + Credit.Amount + Length.of.current.employment + Instalment.per.cent + Most.valuable.available.asset, family = binomial(logit), data = the.data)

4 Deviance Residuals:

5

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

6 Coefficients:

7

	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	.

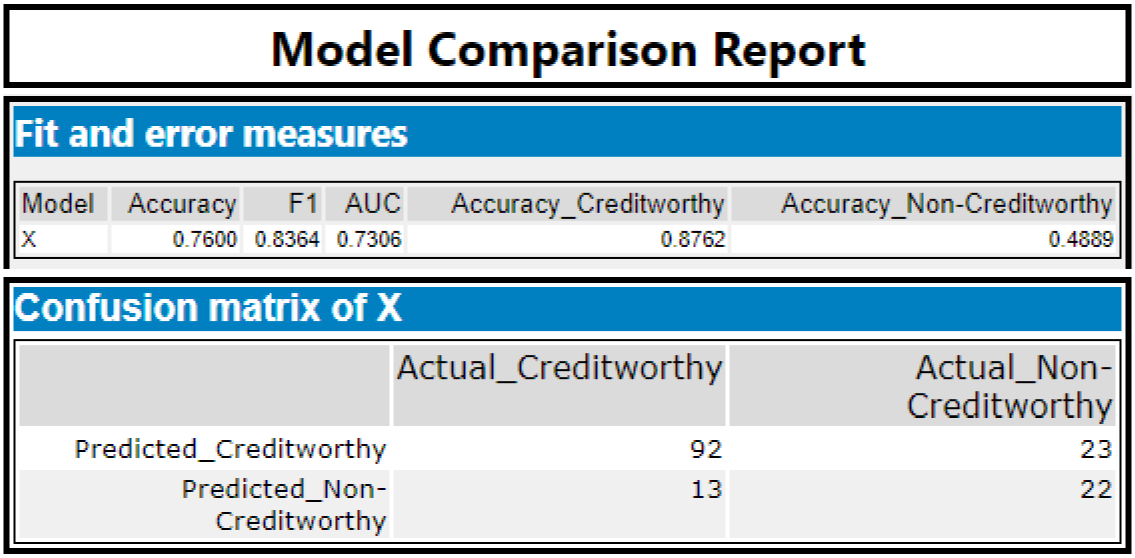
Significance codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial taken to be 1)

R-Squared value = 0.2048 which is not a good value.

✓ From the report window, we see that the following variables are significant:

- Account-Balance
- Payment-Status-of-Previous-Credit
- Purpose
- Credit-Amount
- Length-of-current-employment
- Instalment-per-cent



Model Comparison Report

Fit and error measures

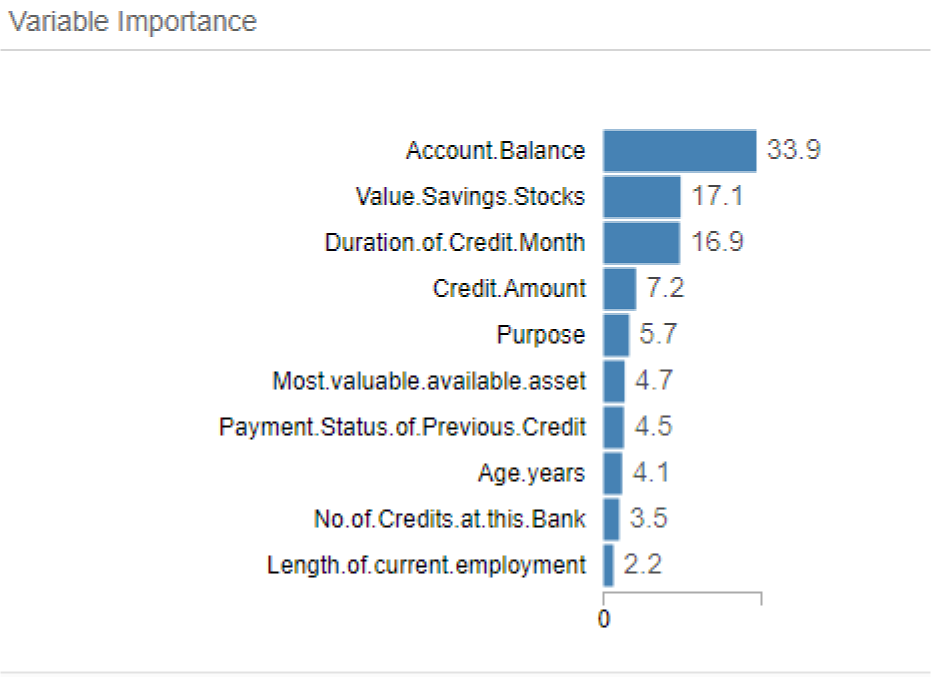
Model	Accuracy	F1	AUC	Accuracy_Creditworthy	Accuracy_Non-Creditworthy
X	0.7600	0.8364	0.7306	0.8762	0.4889

Confusion matrix of X

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

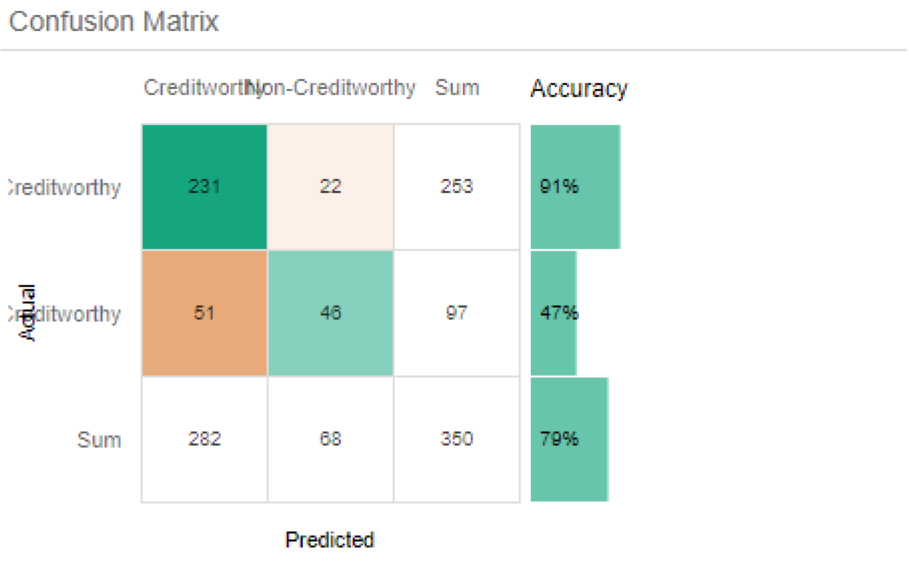
Now, let’s check the performance perform with the Validation data (30% of the dataset). We see an overall accuracy of 0.7600, which is not an excellent value, but it still a good prediction. If we look at confusion matrix, we see that the model is correctly predicting Creditworthy individuals better than its predicting non- creditworthy (81% vs 63%) This means that there is a bias towards Creditworthiness.

2. ***Decision Tree Model***



✓ The Variable Importance plot shows the following variables as the most important:

- Account-Balance
- Value-Saving-Stocks
- Duration-of-Credit-Month

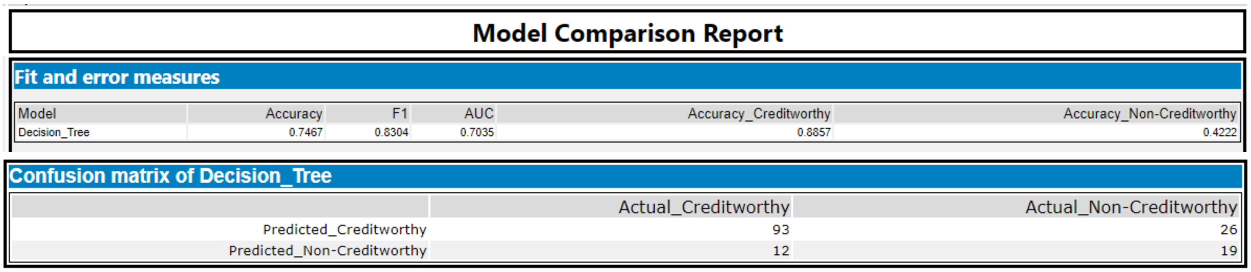


Confusion Matrix

	Creditworthy	Non-Creditworthy	Sum	Accuracy
reditworthy	231	22	253	91%
ditworthy	51	46	97	47%
Sum	282	68	350	79%

Actual

Predicted



Model Comparison Report

Fit and error measures

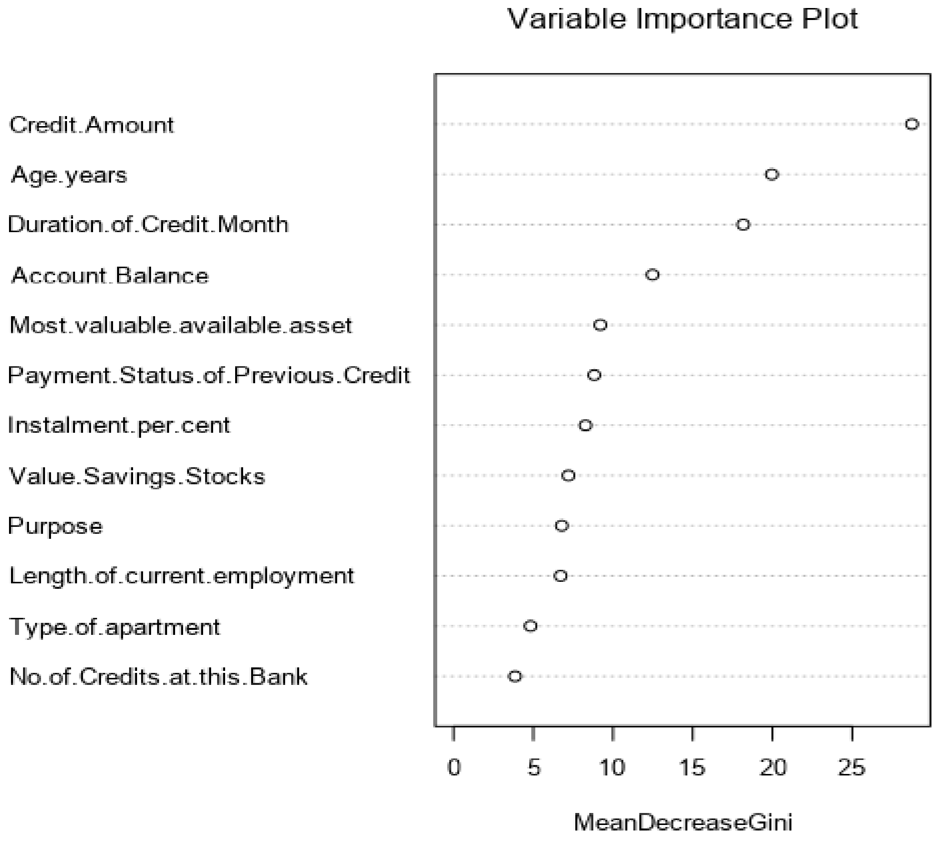
Model	Accuracy	F1	AUC	Accuracy_Creditworthy	Accuracy_Non-Creditworthy
Decision_Tree	0.7467	0.8304	0.7035	0.8857	0.4222

Confusion matrix of Decision_Tree

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

Looking at how well this model would perform with the Validation data, we see that the overall percent accuracy is lower than the Logistic Regression Accuracy – 0.7467. If we look at confusion matrix, the analysis is similar to the Logistic Regression model, the model is correctly predicting Creditworthy individuals better than its predicting non- creditworthy. This means that there is a bias towards Creditworthiness.

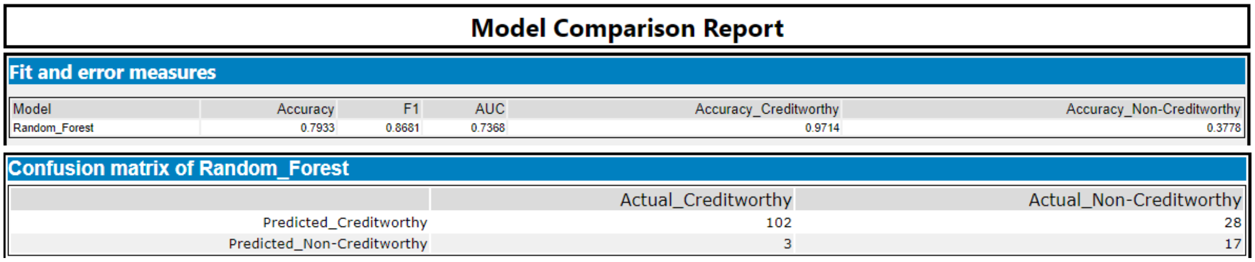
3. ***Random Forest Model***



✓ The Variable Importance plot shows the following variables as the most important:

- Credit.Amount
- Age.years
- Duration.of.Credit.Month

Forest Model – Validation Data



Model Comparison Report

Fit and error measures

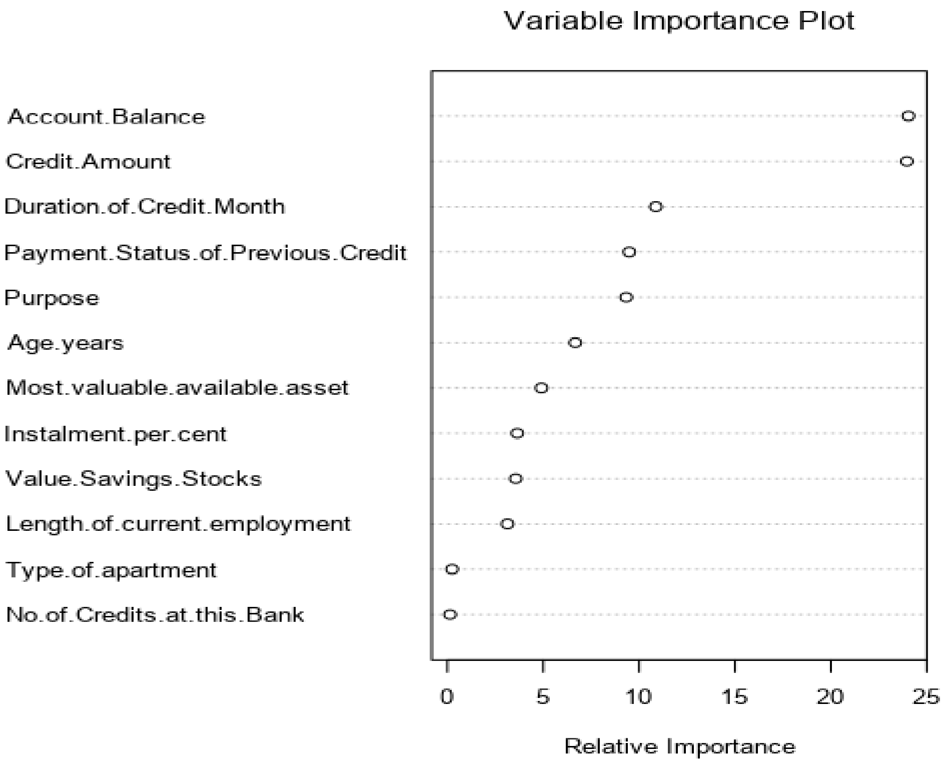
Model	Accuracy	F1	AUC	Accuracy_Creditworthy	Accuracy_Non-Creditworthy
Random_Forest	0.7933	0.8681	0.7368	0.9714	0.3778

Confusion matrix of Random_Forest

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

The accuracy of predicting creditworthiness in this model is about 78%(102/130) and the accuracy of predicting non-creditworthiness is 85%(17/20). In this case, we can say that this model is almost not biased at all, because the difference between those accuracies is very small.

4. <u>*Boosted Model*</u>



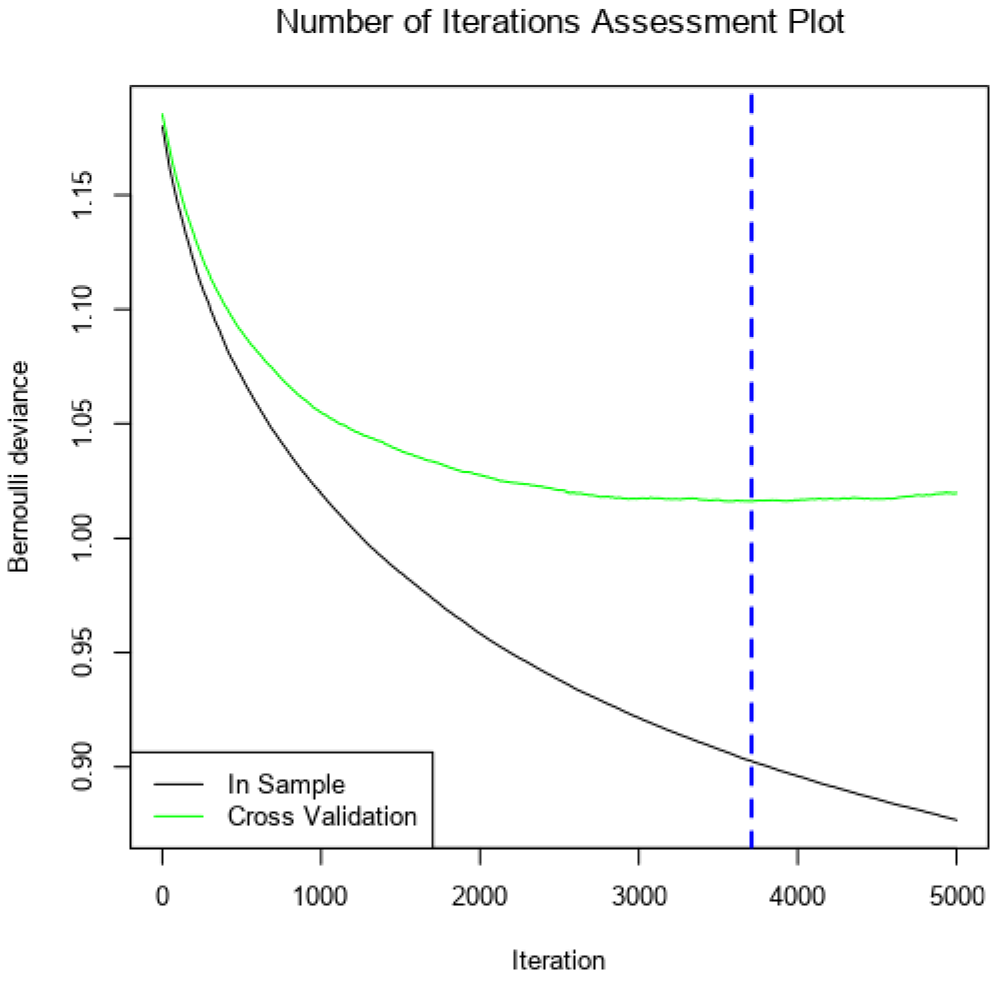
✓ The Variable Importance plot shows the following variables as the most important:

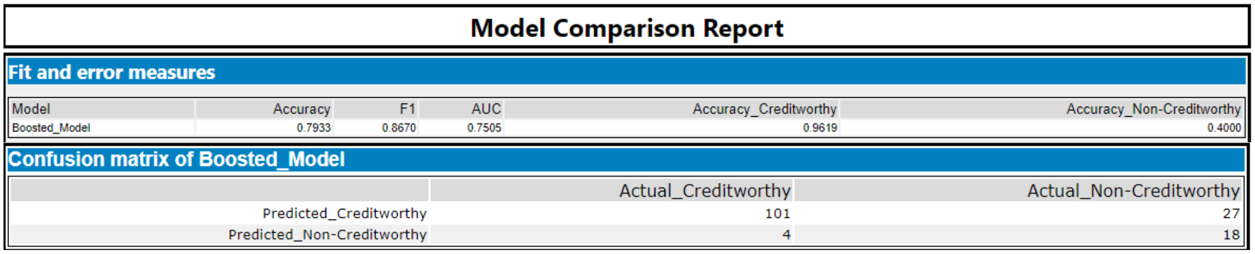
- Account-Balance
- Credit-Amount

Report for Boosted Model Boosted_Model

Basic Summary:

Loss function distribution: Bernoulli
Total number of trees used: 5000
Best number of trees based on 5-fold cross validation: 3710





Model Comparison Report

Fit and error measures

Model	Accuracy	F1	AUC	Accuracy_Creditworthy	Accuracy_Non-Creditworthy
Boosted_Model	0.7933	0.8670	0.7505	0.9619	0.4000

Confusion matrix of Boosted_Model

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

The accuracy of predicting creditworthiness in this model is 79% and the accuracy of predicting non-creditworthiness is 82%. In this case, we can say that this model also is almost not biased since the difference between those accuracies is very small.

Step 4: Writeup

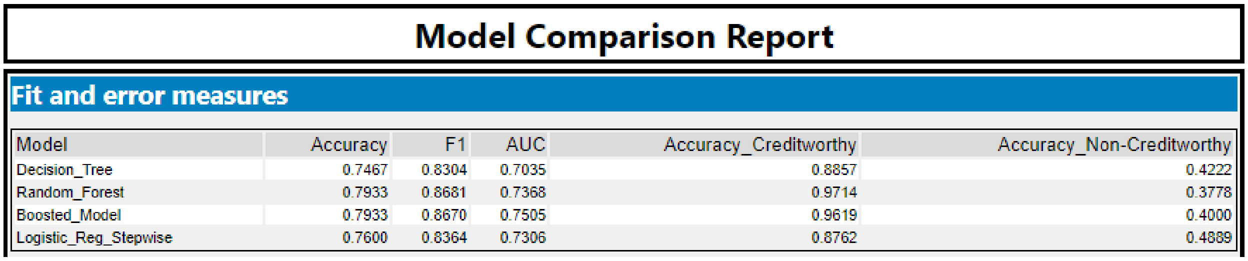
Decide on the best model and score your new customers. For reviewing consistency, if Score_Creditworthy is greater than Score_NonCreditworthy, the person should be labeled as "Creditworthy"

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

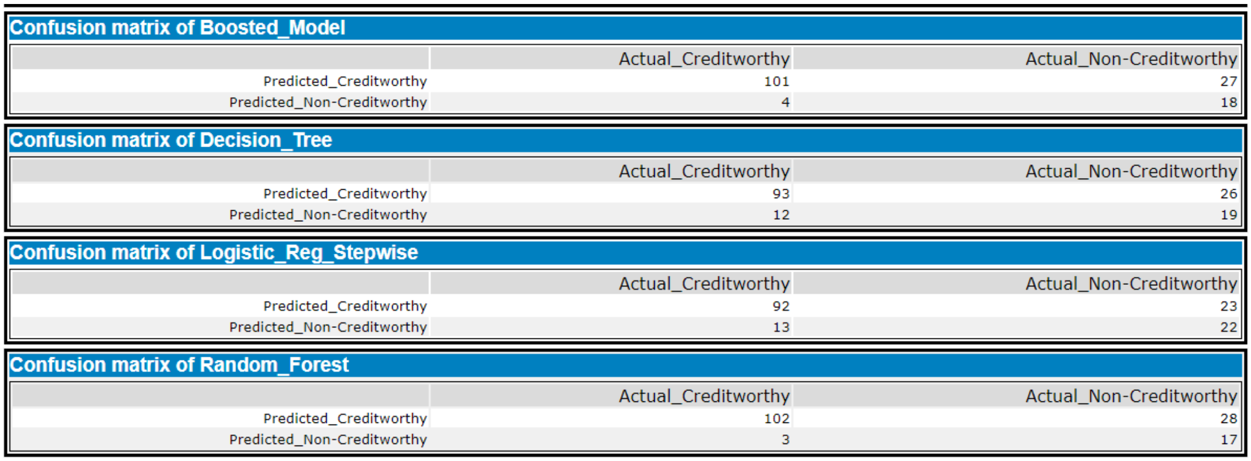
Note: Remember that your boss only cares about prediction accuracy for Creditworthy and Non-Creditworthy segments.



Model Comparison Report

Fit and error measures

Model	Accuracy	F1	AUC	Accuracy_Creditworthy	Accuracy_Non-Creditworthy
Decision_Tree	0.7467	0.8304	0.7035	0.8857	0.4222
Random_Forest	0.7933	0.8681	0.7368	0.9714	0.3778
Boosted_Model	0.7933	0.8670	0.7505	0.9619	0.4000
Logistic_Reg_Stepwise	0.7600	0.8364	0.7306	0.8762	0.4889



Confusion matrix of Boosted_Model

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

Confusion matrix of Decision_Tree

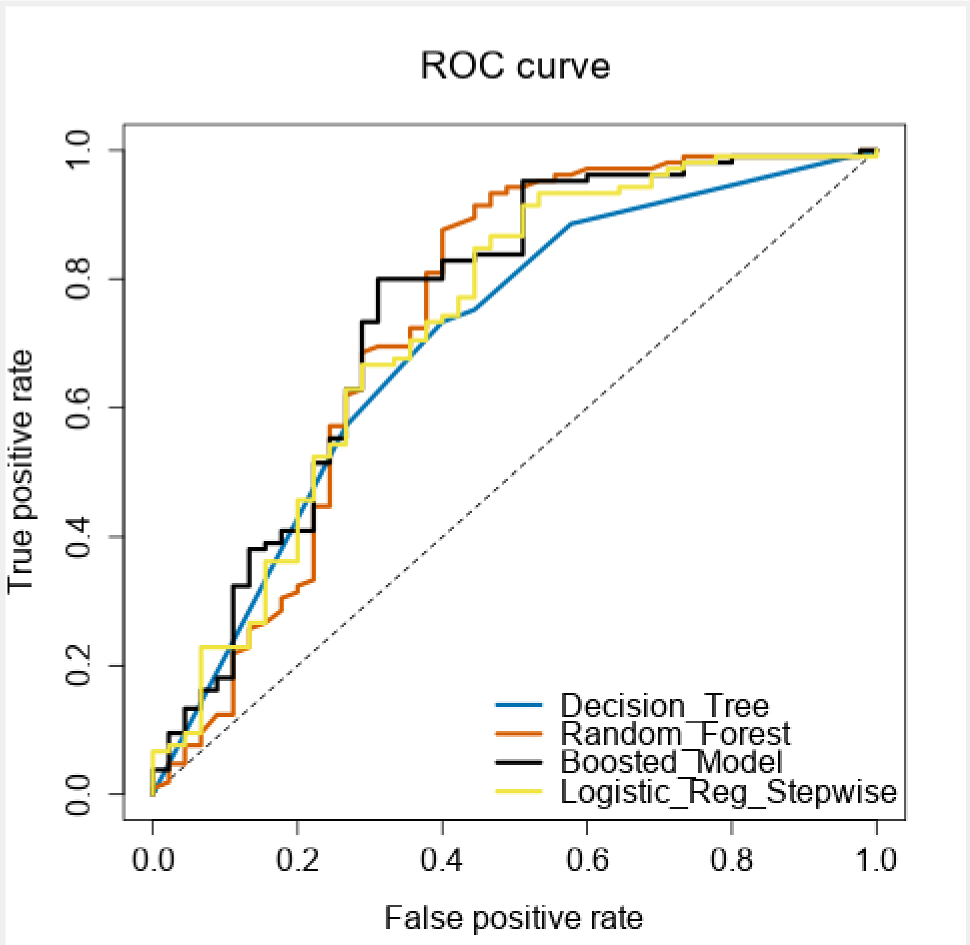
	Actual_Creditworthy	Actual_Non-Creditworthy
Predicted_Creditworthy	93	26
Predicted_Non-Creditworthy	12	19

Confusion matrix of Logistic_Reg_Stepwise

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

Confusion matrix of Random_Forest

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



- ✓ Considering the overall accuracy to predict the best fit model, we notice two best models: Forest Model and Boosted Model, both with the highest overall accuracy of 79.33% So we have to choose the best between the two of them. Let’s check the ROC curve:

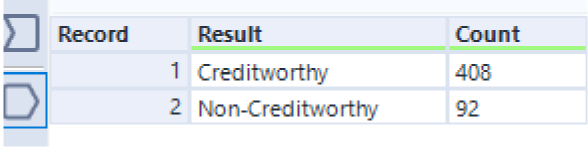
✓ We know that the best model is the one that reaches the top fast and is positioned on the top left. The Forest Model fulfills this condition.

✓ Also, the F1 score (harmonic mean between Precision and Recall) is the highest of all models. F1 is usually more useful than accuracy, especially if you have an uneven class distribution which is our case(more creditworthy than non-creditworthy)

- How many individuals are creditworthy?

✓ Using the random forest model for prediction, adding the new data that we want to predict, and the adding Alteryx score tool, I got the following result:

✓ 408 clients are ‘Creditworthy’



Record	Result	Count
1	Creditworthy	408
2	Non-Creditworthy	92

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.