Project: Creditworthiness

Complete each section. When you are ready, save your file as a PDF document and submit it here: https://classroom.udacity.com/nanodegrees/nd008/parts/11a7bf4c-2b69-47f3-9aec-108ce847f855/project

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?
  Decisions need to be made about which customers to give loans or not based on their (predicted) creditworthiness.

- What data is needed to inform those decisions?
  Data on old customers is needed to help build a model to predict the creditworthiness of new customers to enable the bank make a decision. Key data here include account balance, credit amount, purpose, and duration of credit month of both old customers and new customers.

- What kind of model (Continuous, Binary, Non-Binary, Time-Series) do we need to use to help make these decisions?
  We need a binary model to help make this decision as the expected result is a yes (creditworthy) or no (not creditworthy)

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:*

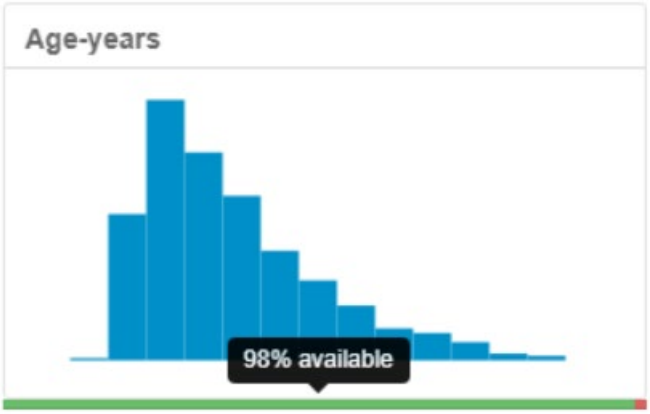
Variable	**Data Type**
Credit-Application-Result	String
Account-Balance	String
Duration-of-Credit-Month	Double
Payment-Status-of-Previous-Credit	String
Purpose	String
Credit-Amount	Double
Value-Savings-Stocks	String
Length-of-current-employment	String
Instalment-per-cent	Double
Guarantors	String
Duration-in-Current-address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

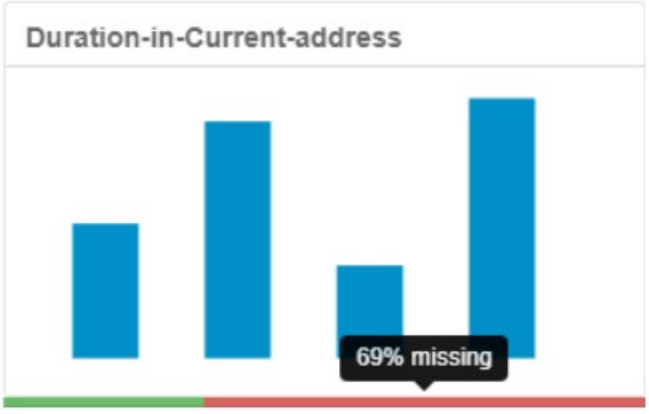
To achieve consistent results reviewers, expect.

Answer this question:

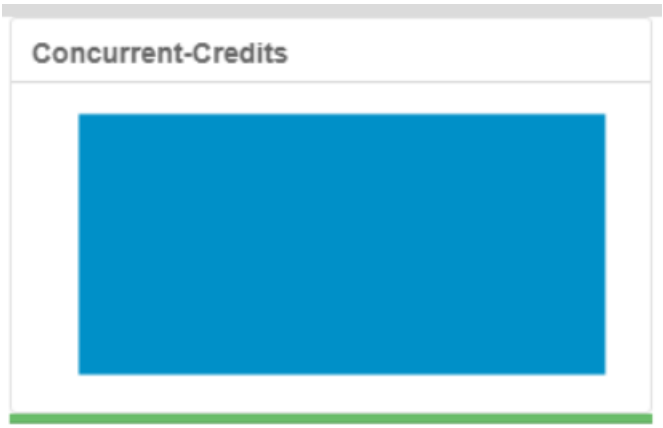
- In your cleanup process, which fields did you remove or impute? Please justify why you removed or imputed these fields. Visualizations are encouraged.
  Based on visualizations from the Field Summary tool, the following actions were taken:
  - Since 2% of **Age-Years** was missing (as shown below), the values were **imputed** using the median to eliminate nulls



  - The **Duration-in-Current-Address** column was **removed** since 69% of the values were missing as shown below



  - The **Concurrent-Credits** and **Occupation** columns were **removed** since there were no classifications in their values (they had only one value)



- **Guarantors, Foreign-Worker,** and **No-of-dependents** columns were also removed since they had low variability (most of their values belonged to one class)

- The **Telephone** column was removed as it is unlikely to tell us anything about the creditworthiness of customers

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.

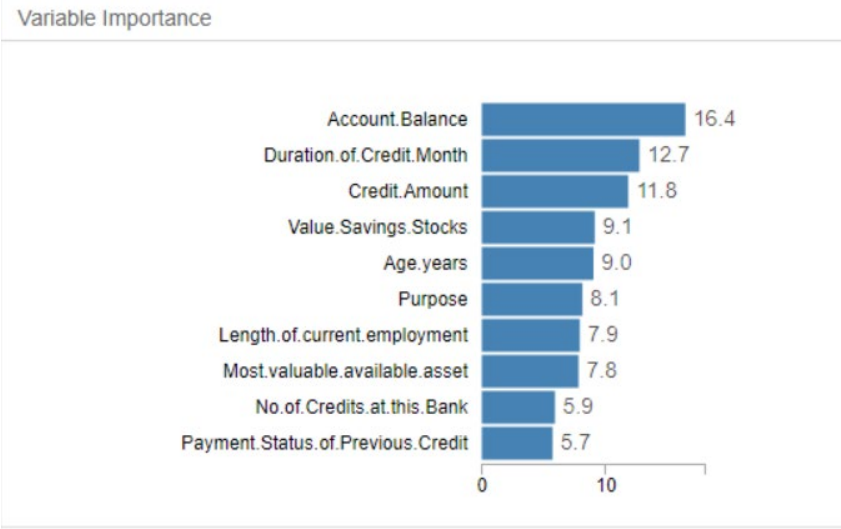
Create all of the following models: Logistic Regression, Decision Tree, Forest Model, Boosted Model

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

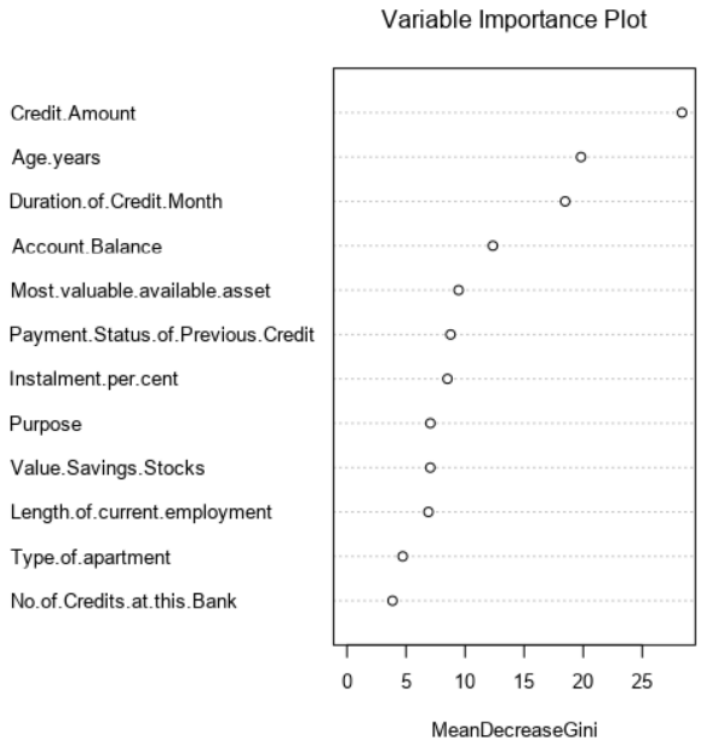
- Which predictor variables are significant or the most important? Please show the p-values or variable importance charts for all of your predictor variables.
  The top 5 most important predictor variables per the various models are as follows:
  a. Logistic Regression + Stepwise model: Account Balance, Credit Amount, Purpose, Payment status of previous credit, and Installment percent

		Pr(>|z|)	
(Intercept)		1e-05	***
Account.BalanceSome Balance	1	1.65e-07	***
Payment.Status.of.Previous.CreditPaid Up		0.42775	
Payment.Status.of.Previous.CreditSome Problems	4	0.0183	*
PurposeNew car	3	0.00566	**
PurposeOther		0.69042	
PurposeUsed car		0.05618	.
Credit.Amount	2	0.00296	**
Length.of.current.employment4-7 yrs		0.49545	
Length.of.current.employment< 1yr		0.03596	*
Instalment.per.cent	5	0.02549	*
Most.valuable.available.asset		0.06289	.

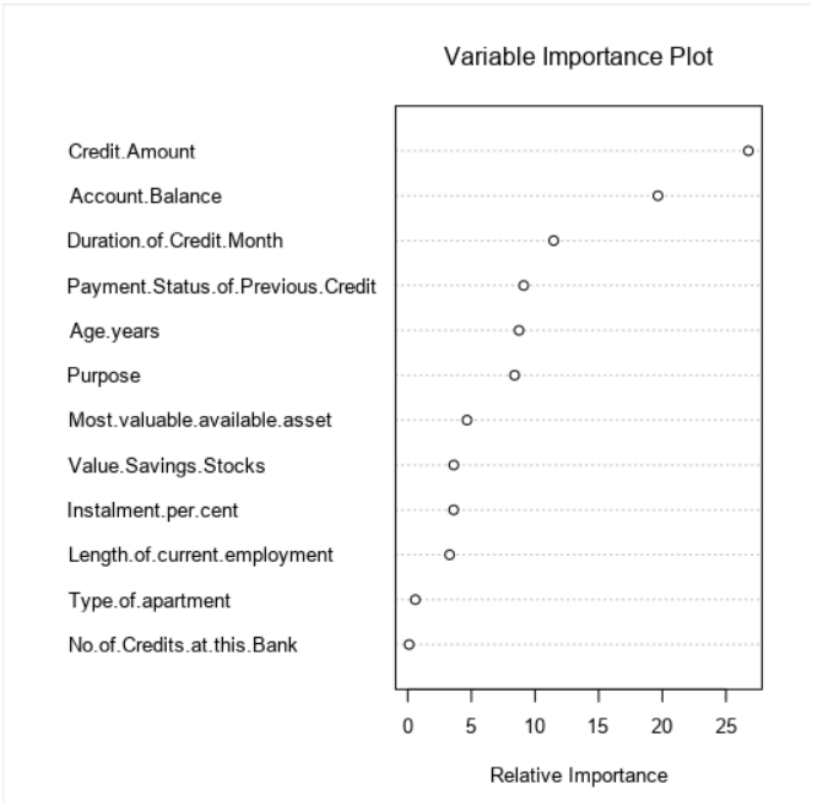
  b. Decision Tree: Credit Amount, Duration of Credit Month, Credit Amount, Value Savings Stocks, and Age-Years



c. Random Forest Model: Credit Amount, Age-Years, Duration of Credit Month, Account Balance, and Most valuable available asset



d. Boosted Model: Credit Amount, Account Balance, Duration of Credit Month, Payment Status of Previous Credit, and Age-Years



- 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?
  a. Logistic Regression + Stepwise Model
  This model had an overall accuracy of **76%**.
  The predictions showed a bias toward accurately predicting positive results (creditworthy) as shown below where it has an 80% positive prediction accuracy whilst the negative prediction accuracy lags behind at 63%
    - Positive Prediction Accuracy = true positives/(true positives + false positives)
      = 92/(92+23) = 0.80
    - Negative Prediction Accuracy = true negatives/(true negatives + false negatives)
      = 22/(22+13) = 0.63

Confusion matrix of LogisticR_Class

	Actual_No	Actual_Yes
Predicted_No	22	13
Predicted_Yes	23	92

Kindly take note of the following abbreviations going forward:

- *PPA: Positive prediction accuracy*
- *TP: True positives*
- *FP: False positives*
- *NPA: Negative prediction accuracy*
- *TN: True negatives*
- *FN: False negatives*

  b. Decision Tree Model
  This model had an overall accuracy of **66.67%**, accurately predicting 75% of positive results and 44% of negative responses as shown below. This also exhibits a bias toward predicting positive creditworthiness.
    - PPA = TP/(TP + FP)
      = 83/(83+28) = 0.75
    - NPA = TN/(TN + FN)
      = 17/(17+22) = 0.44

Confusion matrix of Tree_Class

	Actual_No	Actual_Yes
Predicted_No	17	22
Predicted_Yes	28	83

  c. Random Forest Model
  This model showed an overall accuracy of **80%**, predicting 86% of negative and 79% of positive responses accurately (see below). This model shows close to no bias in predictions.
    - PPA = TP/(TP + FP)
      = 102/(102+27) = 0.79
    - NPA = TN/(TN + FN)
      = 18/(18+3) = 0.86

Confusion matrix of Forest_Class

	Actual_No	Actual_Yes
Predicted_No	18	3
Predicted_Yes	27	102

d. Boosted Model

This model's overall accuracy was **77.33%.** It had a 78% positive prediction accuracy and a 74% negative prediction accuracy. This model may be said to have no bias, as the negative and positive prediction accuracies are close

- PPA = TP/(TP + FP)
  = 99/(99+28) = 0.78
- NPA = TN/(TN + FN)
  = 17/(17+6) = 0.74

Confusion matrix of Boosted_Class

	Actual_No	Actual_Yes
Predicted_No	17	6
Predicted_Yes	28	99

The above is re-presented here for ease of comparison:

Model	Accuracy
Tree_Class	0.6667
Forest_Class	0.8000
Boosted_Class	0.7733
LogisticR_Class	0.7600

Confusion matrix of Boosted_Class

	Actual_No	Actual_Yes
Predicted_No	17	6
Predicted_Yes	28	99

Confusion matrix of Forest_Class

	Actual_No	Actual_Yes
Predicted_No	18	3
Predicted_Yes	27	102

Confusion matrix of LogisticR_Class

	Actual_No	Actual_Yes
Predicted_No	22	13
Predicted_Yes	23	92

Confusion matrix of Tree_Class

	Actual_No	Actual_Yes
Predicted_No	17	22
Predicted_Yes	28	83

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

Step 4: Writeup

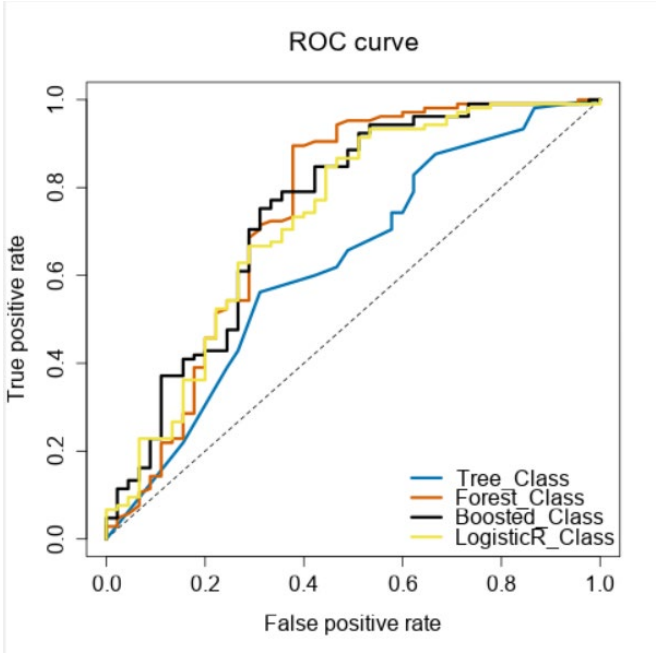
Decide on the best model and score your new customers. For reviewing consistency, if Score_Creditworthy is greater than Score_NonCreditworthy, the person should be labeled as "Creditworthy"

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

  *Per the diagrams above, I chose **Forest Model** since it had the highest overall accuracy (80%), combining the highest negative prediction and second highest positive prediction accuracies level at 86% and 79% respectively.*
  Observing the ROC graph (below) shows the Forest model reaching the top quickest and having the highest curve overall. It is can also be seen to have the largest area under curve, implying that the curve covers the largest area. It is followed closely by the Boosted and Logistic Models.
  Again, the Confusion Matrices (also below) show the Forest Model correctly predicting the highest number of positive (creditworthy) responses. It is second to the Logistic + Stepwise Model in correctly predicting negative (not creditworthy) responses, thus keeping its performance high.
  *Overall, these qualify my selection of the **Forest Model** for predicting customer creditworthiness*



Confusion matrix of Boosted_Class

	Actual_No	Actual_Yes
Predicted_No	17	6
Predicted_Yes	28	99

Confusion matrix of Forest_Class

	Actual_No	Actual_Yes
Predicted_No	18	3
Predicted_Yes	27	102

Confusion matrix of LogisticR_Class

	Actual_No	Actual_Yes
Predicted_No	22	13
Predicted_Yes	23	92

Confusion matrix of Tree_Class

	Actual_No	Actual_Yes
Predicted_No	17	22
Predicted_Yes	28	83

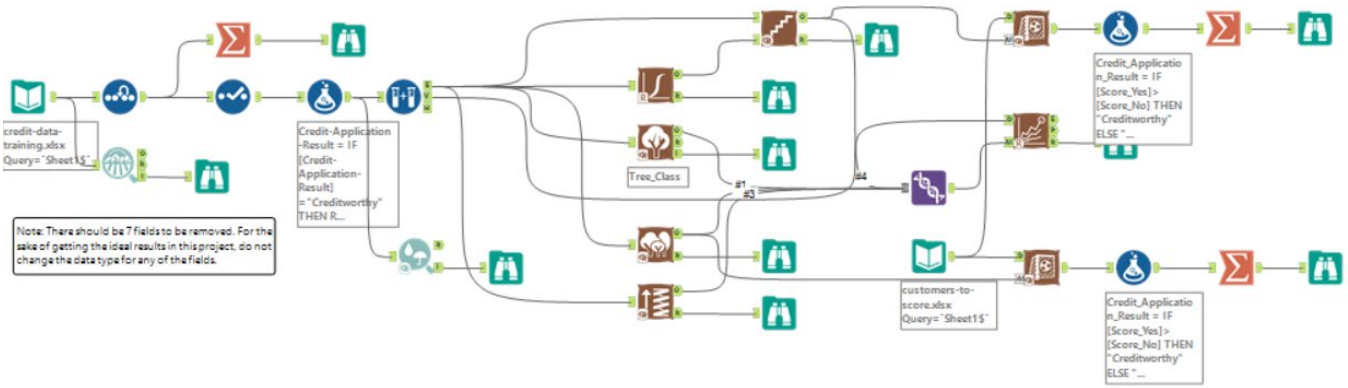
Note: Remember that your boss only cares about prediction accuracy for Creditworthy and Non-Creditworthy segments.

- How many individuals are creditworthy?
  Going by my choice of model, 405 individuals 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.

A visual of my Alteryx workflow



Reference

- Udacity Project Reviewer recommendations