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

1. What decisions needs to be made?

   The company wants to classify new customers for loan approvals.

2. What data is needed to inform those decisions?

   For Loan approvals, we usually need data like

   Account balance
   Age of the customers
   Credit information
   Total Debt information, previous loans from other banks
   Income information
   Investments in terms of stock, assets
   Employment information
   Surety information

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

   Typically, when a Yes or No decision we use Binary models but for a two-variable classification we can use both Binary models and Non-Binary models with two variables.

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 average of the entire data field instead of removing a few data points. (100 word limit)*

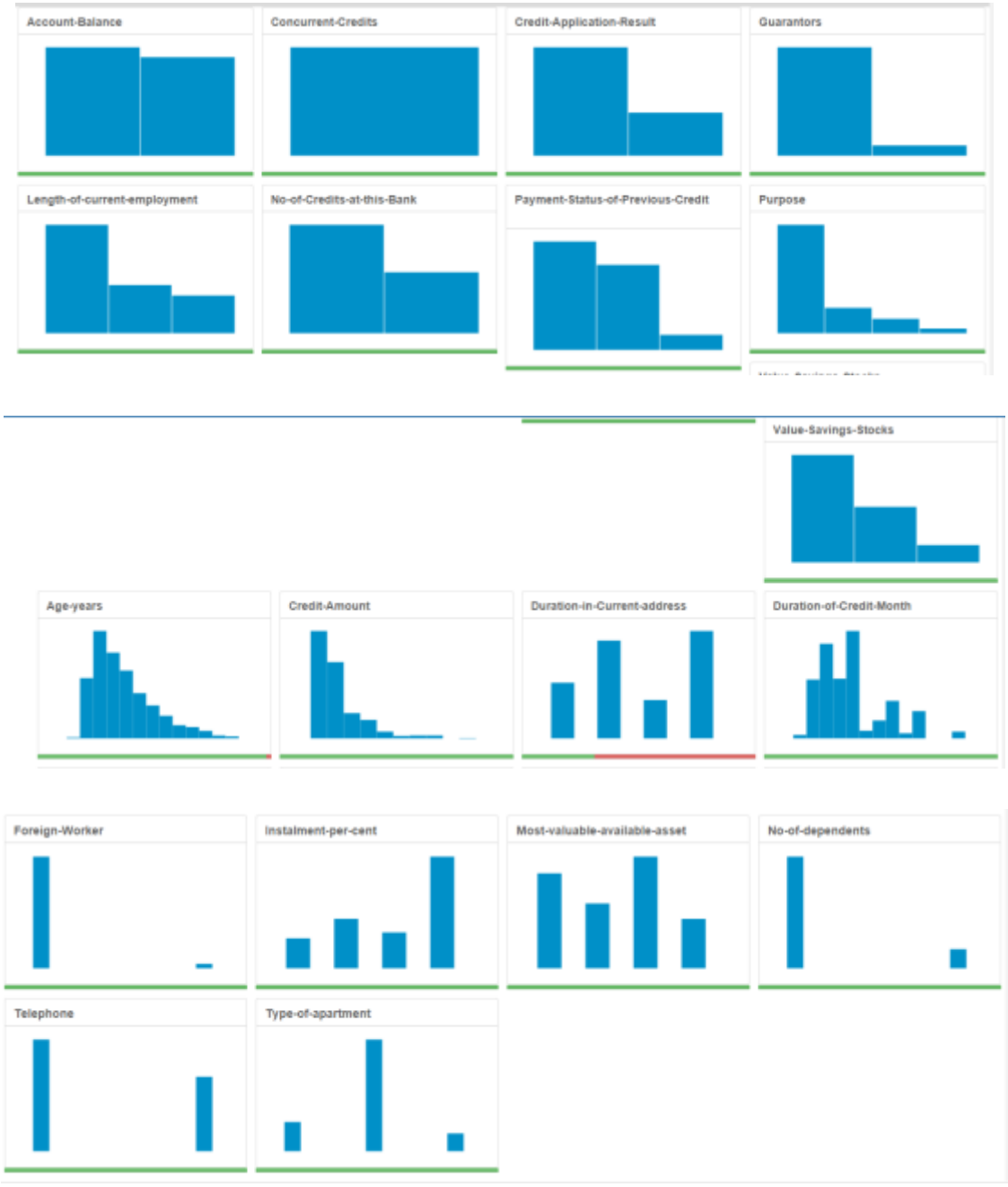
***Note**: For students using software other than Alteryx, please format each variable as:*

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:

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



Variables Removed –

Concurrent Credits: Reason – Low variability.
Guarantors: Reason – Low Variability.

Duration in Current Address: Reason – 70% of the data missing.
Occupation: Reason – Not an insightful variable.
No of dependents: Reason – Low Variability
Telephone – Low Variability
Foreign Worker – Low Variability

Imputation –

Age – 2% of the data was missing. Because the data was less than 5% we can remove those fields. But I have imputed this 2% of this data with the average of the Age field.

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:*

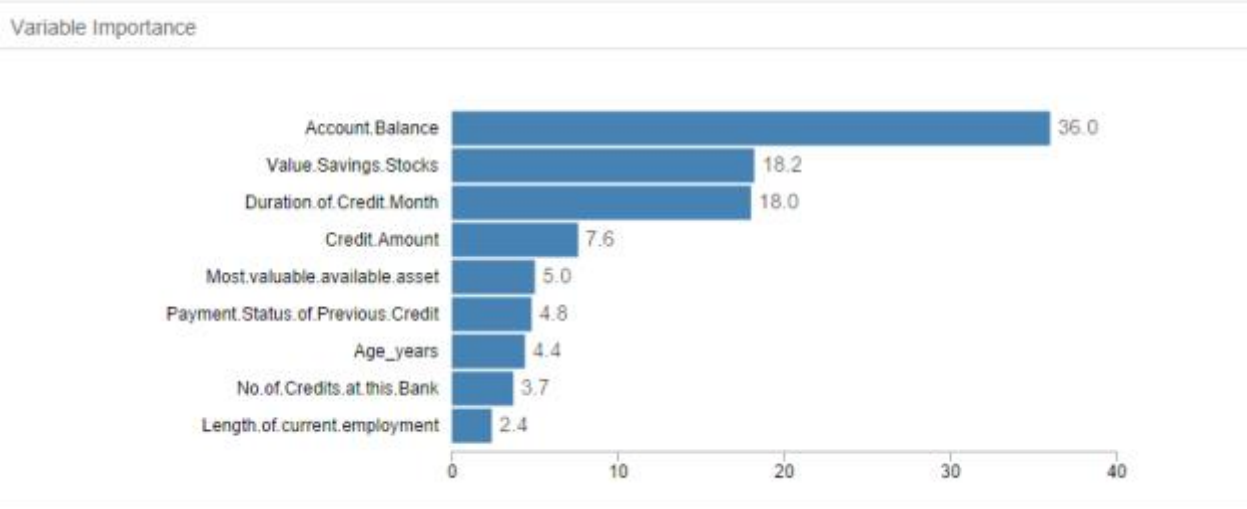
1. Which predictor variables are significant or the most important? Please show the p-values or variable importance charts for all of your predictor variables.

Logistic Regression Significant Variables –

	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 to 7 yrs	0.3127022	4.587e-01	0.6817	0.49545	
Length.of.current.employmentLess Than 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	.

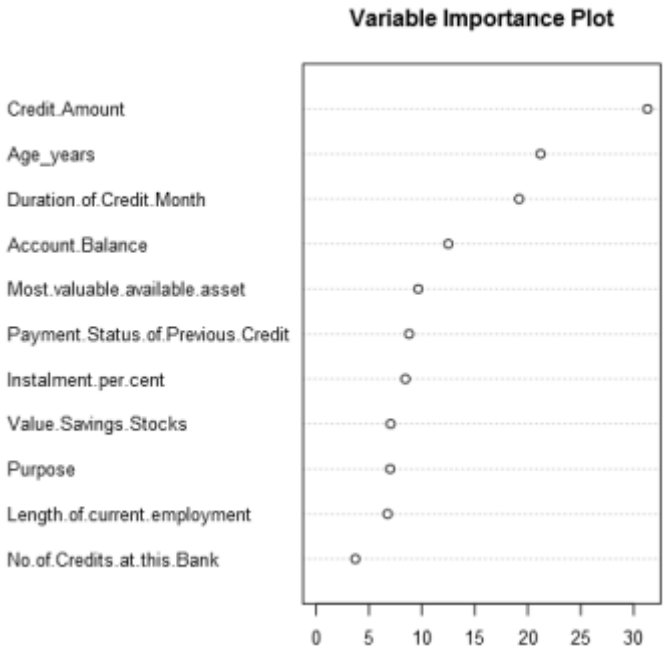
Account balance
Purpose – New Car, Other
Credit Amount, Installments
Current Employment – Less than 1 year.

Decision Tree Significant Variables –



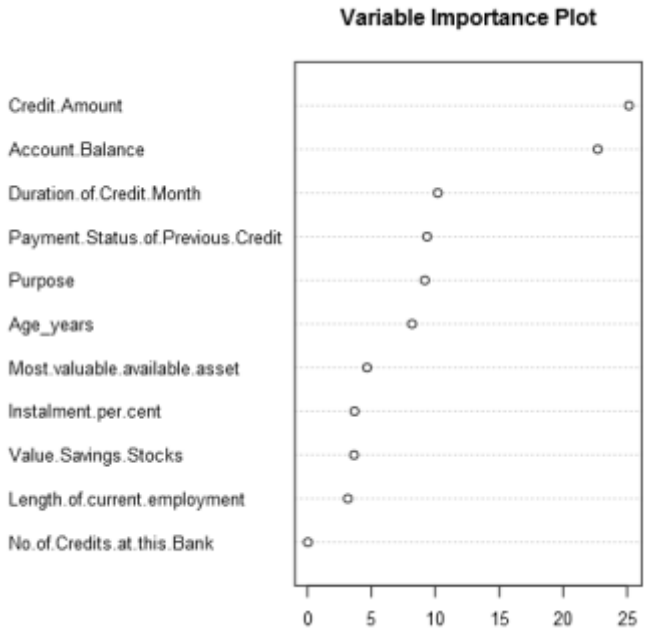
Account Balance
Stocks
Credit Months

Forest Model Significant Variables –



Credit Amount, Age, Credit Months, Account Balance.

Boosted Model Significant Variables –



Credit Amount
Account Balance

2. 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 below image shows the overall percent accuracy. I see bias in all the models towards the creditworthy variable.

Model Comparison Report

Fit and error measures

Model	Accuracy	F1	AUC	Accuracy_Creditworthy	Accuracy_NonCreditworthy
SLR	.7600	.8364	.7306	.8000	.6286
DT	.7467	.8273	.7054	.7913	.6000
FM	.8067	.8755	.7401	.7969	.8636
BM	.7933	.8670	.7541	.7891	.8182

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], number of samples that are **correctly** predicted to be Class [class name] divided by number of samples predited to be Class [class name]
AUC: area under the ROC curve, only available for two-class classification.
F1: F1 score, precision * recall / (precision + recall)

Confusion matrix of BM

	Actual_Creditworthy	Actual_NonCreditworthy
Predicted_Creditworthy	101	27
Predicted_NonCreditworthy	4	18

Confusion matrix of DT

	Actual_Creditworthy	Actual_NonCreditworthy
Predicted_Creditworthy	91	24
Predicted_NonCreditworthy	14	21

Confusion matrix of FM

	Actual_Creditworthy	Actual_NonCreditworthy
Predicted_Creditworthy	102	26
Predicted_NonCreditworthy	3	19

Confusion matrix of SLR

	Actual_Creditworthy	Actual_NonCreditworthy
Predicted_Creditworthy	92	23
Predicted_NonCreditworthy	13	22

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:

1. Which model did you choose to use? Please justify your decision using only the following techniques:

   I chose the Forest Model based on the below

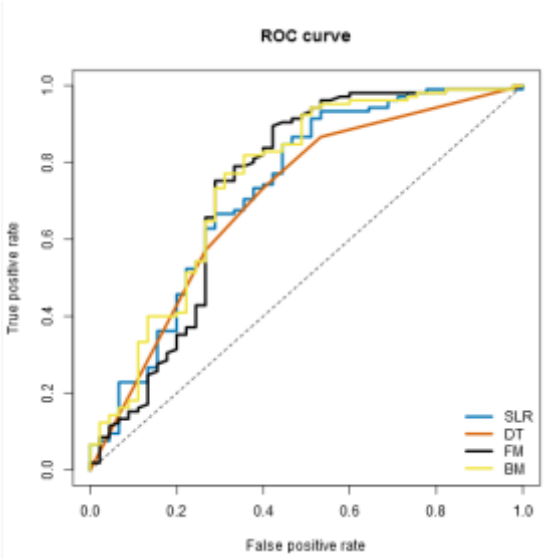
   a. Overall Accuracy against your Validation set

      The model had an accuracy of 81%

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

      Creditworthy Accuracy was 80%
      Non-Creditworthy Accuracy was 86%

c. ROC graph



The ROC curve above shows the accuracy of the Forest Model (FM)

d. Bias in the Confusion Matrices

Confusion matrix of FM

	Actual_Creditworthy	Actual_NonCreditworthy
Predicted_Creditworthy	102	26
Predicted_NonCreditworthy	3	19

The model Seems to have a bias towards the Creditworthy fields

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

2. How many individuals are creditworthy?

406 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.