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?
   For each new customer, whether the bank should approve the loan application or not.

2. What data is needed to inform those decisions?
   a. A training dataset: credit-data-training.xlsx
   b. A testing dataset: customer-to-score.xlsx
   c. After data preprocessing, the following fields will be used in the analysis:

Variable	**Data Type**
Credit-Application-Result(target)	String
Account-Balance	String
Duration-of-Credit-Month	Double
Payment-Status-of-Previous-Credit	String
Credit-Amount	Double
Value-Savings-Stocks	String
Length-of-current-employment	String
Instalment-per-cent	Double
Most-valuable-available-asset	Double
Age-years	Double
Type-of-apartment	Double
No-of-Credits-at-this-Bank	String

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

For each customer or each row in the data sheet, the output shall be yes or no indicating whether to provide the loan or not. So this is a binary classification problem, and we shall use 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".*
  No, all absolute correlations are under .70.
- *Are there any missing data for each of the data fields? Fields with a lot of missing data should be removed*
  Two fields contain missing data.
  1. Age-years: 2.4% missing, will be imputed with median since it has a very strong skew towards left side.
  2. Duration-in-Current-address: 68.8% missing, will be removed entirely.
- *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.*
  Remove the following 7 fields since they look very uniform: Current-Credits, Guarantors, Foreign-Worker, No-of-dependents, Purpose, Duration-in-Current-address, Concurrent-Credits
- *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.
   Two fields contain missing data.
   1. Age-years: 2.4% missing, will be imputed with average.
   2. Duration-in-Current-address: 68.8% missing, will be removed entirely.
   Remove the following 7 fields since they look very uniform: Current-Credits, Guarantors, Foreign-Worker, No-of-dependents, Purpose, Duration-in-Current-address, Concurrent-Credits.
   Moreover, Telephone field is removed as indicated in the project description.

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:*

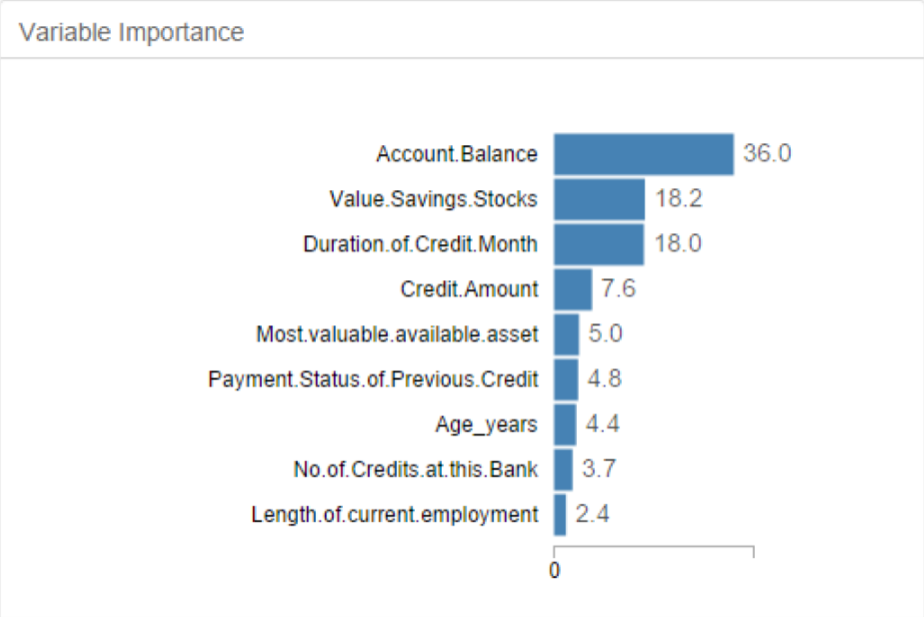
1. Which predictor variables are significant or the most important? Please show the p-values or variable importance charts for all of your predictor variables.

Logistic model: the most important predictor variables are Account-Balance, Credit-Amount, Payment-Status-of-Previous-Credit, Length-of-current-employment, Installment-per-cent.

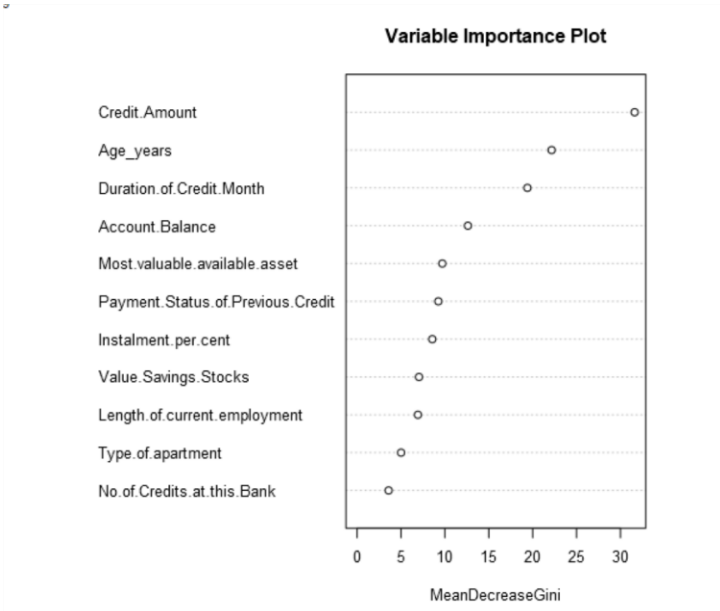
Coefficients:

	Estimate	Std. Error	z value	Pr(>|z|)	
(Intercept)	-3.1569568	6.747e-01	-4.6793	2.87e-06	***
Account.BalanceSome Balance	-1.6261462	2.991e-01	-5.4362	5.44e-08	***
Payment.Status.of.Previous.CreditPaid Up	0.2317475	2.905e-01	0.7978	0.42499	
Payment.Status.of.Previous.CreditSome Problems	1.2330655	5.004e-01	2.4642	0.01373	*
Credit.Amount	0.0001392	5.244e-05	2.6549	0.00793	**
Length.of.current.employment4-7 yrs	0.3756741	4.461e-01	0.8421	0.39971	
Length.of.current.employment< 1yr	0.8481799	3.780e-01	2.2440	0.02483	*
Instalment.per.cent	0.3351614	1.323e-01	2.5340	0.01128	*
Most.valuable.available.asset	0.2249825	1.384e-01	1.6256	0.10404	

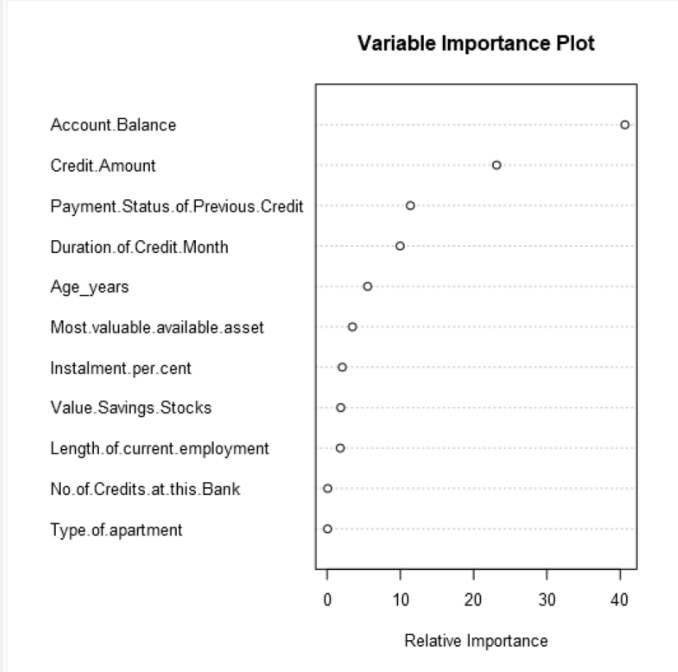
Decision Tree model: most important predictor variables are Account-Balance, Value-Savings-Stocks and Duration-of-Credit-Month



Forest model: most important predictor variables are Credit-Amount, Age-years, Duration-of-Credit-Month and Account-Balance.



Boosted model: most important predictor variables are Account-Balance, Credit-Amount, Payment-Status-of-Previous-Credit and Duration-of-Credit-Month.



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 overall percent accuracy and confusion matrix for each model is given in the following picture.
   Logistic, decision tree and forest models tend to have more false negatives, .i.e. misclassifying actual creditworthy as non-creditworthy. However, boosted model does not have this bias as indicated by its 85% accuracy on non-creditworthy.

Fit and error measures

Model	Accuracy	F1	AUC	Accuracy_Creditworthy	Accuracy_Non-Creditworthy
credit_tree	0.7467	0.8273	0.7054	0.7913	0.6000
credit_forest	0.7867	0.8609	0.7560	0.7920	0.7600
credit_boosted	0.7933	0.8681	0.7560	0.7846	0.8500
log_stepwise	0.8000	0.8649	0.7575	0.8205	0.7273

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 credit_boosted

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

Confusion matrix of credit_forest

	Actual_Creditworthy	Actual_Non-Creditworthy
Predicted_Creditworthy	99	26
Predicted_Non-Creditworthy	6	19

Confusion matrix of credit_tree

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

Confusion matrix of log_stepwise

	Actual_Creditworthy	Actual_Non-Creditworthy
Predicted_Creditworthy	96	21
Predicted_Non-Creditworthy	9	24

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"

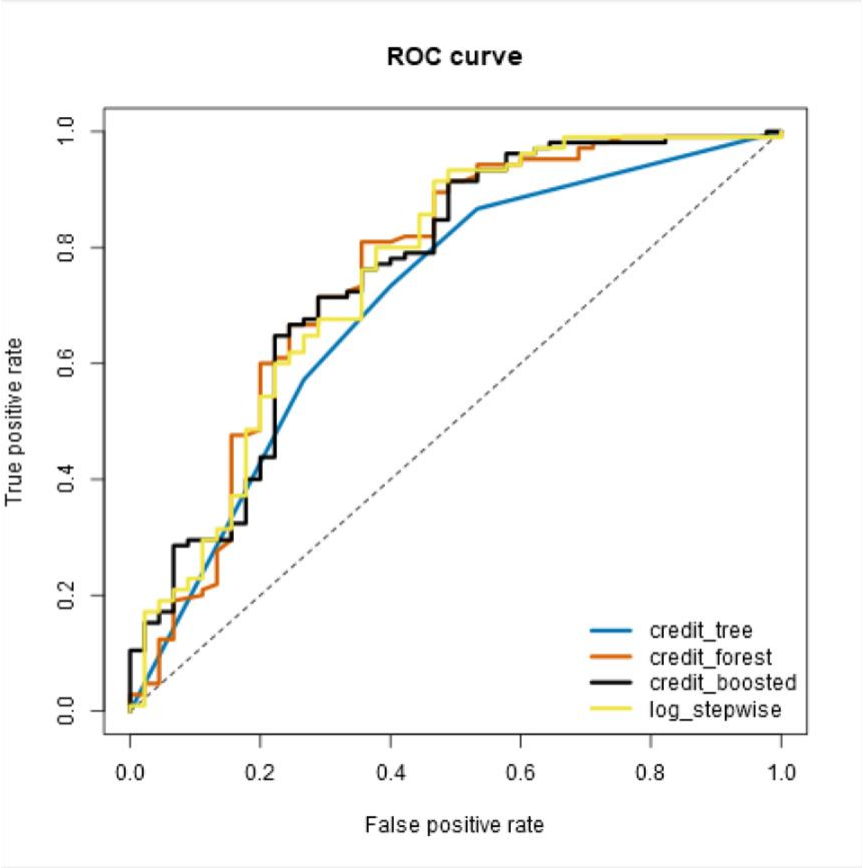
Write a brief report on how you came up with your classification model and write down how many of the new customers would qualify for a loan. (250 word limit)

Answer these questions:

1. Which model did you choose to use? Please justify your decision using only the following techniques:
    a. Overall Accuracy against your Validation set
    b. Accuracies within "Creditworthy" and "Non-Creditworthy" segments
    c. ROC graph
    d. Bias in the Confusion Matrices

    I chose the boosted model to use, because it has the second highest overall accuracy(79.33%) which is only slightly lower than the highest one (80%). In addition, it has the highest accuracy on non-creditworthy (85%) which is much larger than those of the other models. Meanwhile, its accuracy on creditworthy is still close to those of other models. Its ROC curve is comparable to logistic and forest models, much better than

decision trees. In addition, it does not have the false negative bias as the other three models.



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

2. How many individuals are creditworthy?
   Among 500 individuals, the boosted model predicts 442 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.