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?

An efficient solution needs to be implemented to classify new customers on whether they can be approved for a loan or not. determining if customers are creditworthy to give a loan to.

- What data is needed to inform those decisions?

Avaluable data to make this decision are two lists:

1. Data on all past applications which will be used to create and train the model of the loan applicants
2. The list of customers that need to be processed in the next few days, and it will be used to validate and test the model.

Specific data that will be used to built, train, validate and test the model: Purpose for which loan is being taken, Account balance of the applicant, Credit Amount applied for, age years of the applicant, duration of credit month, the length of the current employment of the loaner etc., which will be required to make the creditworthiness decision of the applicants.

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

Taken in consideration we need to find out whether the loan applicant is creditworthy or non creditworthy we need to use a Binary model to predict the outcome of our analysis.

If we look at the model accuracy we see that the Forest Model performed best.

The Forest Model provides the highest accuracy, and there are now 2 reasons to pick the Forest Model:

1. It has the highest overall accuracy

2. The averaged results from the forest model helps deal with the decision tree model's bias to overfit the data

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.*** *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 that was imputed is the Age-years with the median value, because the median is less affected by outliers and shows the value towards which the data have central tendency to move. The fields that were removed were the ones with low variability, skewedness and one with a lot of missing data. Low variability removed fields: (1) Guarantors, (2) Concurrent-Credits, (3) Foreign-worker, (4) Occupation, (5) No-of-dependents, (6) Telephone and missing data removed field: (7) Duration-in-Current-address.

There was no high correlation between any of the fields.

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.

    1. *Logistic Regression*- significant predictor variables that are most important are Account-Balance, Purpose and Credit-Amount.

Report for Logistic Regression Model Stepwise

Basic Summary

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)

Deviance Residuals:

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

Response: Credit.Application.Result

	LR Chi-Sq	DF	Pr(>Chi-Sq)
Account.Balance	31.129	1	2.41e-08***
Payment.Status.of.Previous.Credit	5.687	2	0.05823.
Purpose	12.225	3	0.00665**
Credit.Amount	9.882	1	0.00167**
Length.of.current.employment	5.522	2	0.06324.
Instalment.per.cent	5.198	1	0.02261*
Most.valuable.available.asset	3.509	1	0.06104.

2. *Decision Tree* - significant predictor variables that are most important are Account Balance, Duration-of-Credit-Month, Value-Saving-Stocks and Purpose

Summary Report for Decision Tree Model Decision_Tree

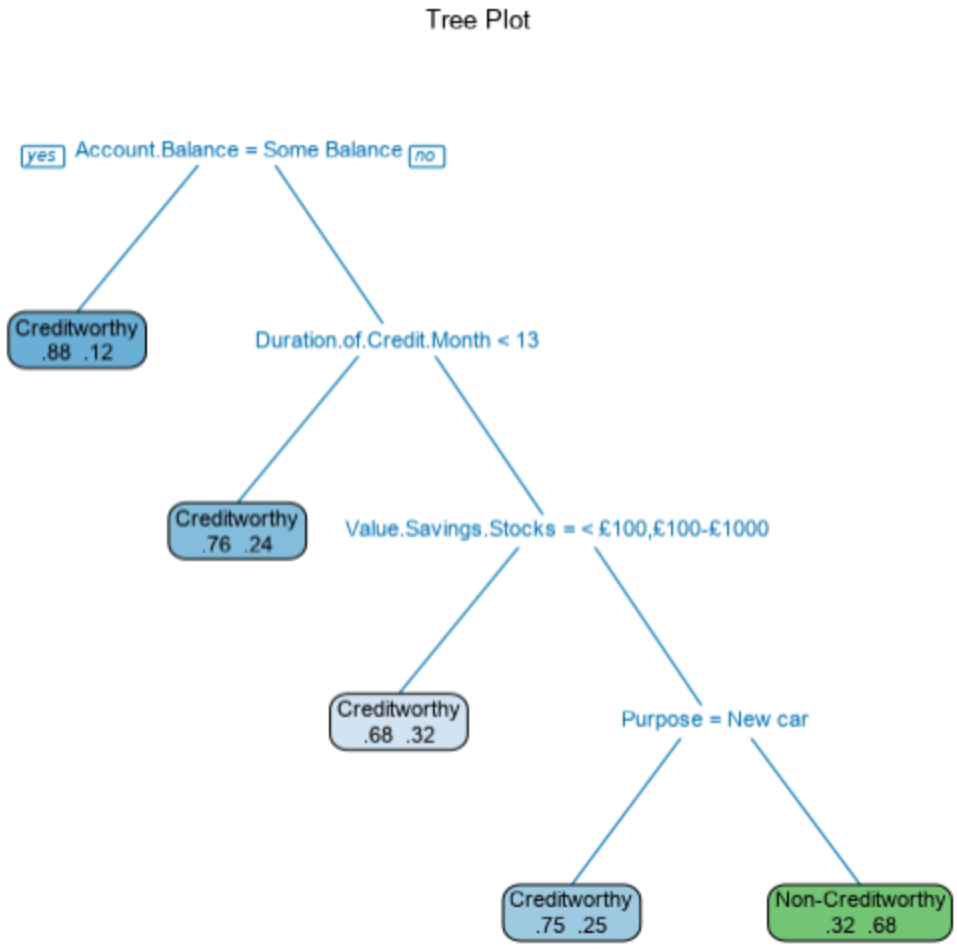
Call:

rpart(formula = Credit.Application.Result ~ Account.Balance + Duration.of.Credit.Month + Payment.Status.of.Previous.Credit + Purpose + Credit.Amount + Value.Savings.Stocks + Length.of.current.employment + Instalment.per.cent + Most.valuable.available.asset + Age.years + Type.of.apartment + No.of.Credits.at.this.Bank, data = the.data, minsplit = 20, minbucket = 7, usesurrogate = 2, xval = 10, maxdepth = 20, cp = 1e-05)

Model Summary
Variables actually used in tree construction:
[1] Account.Balance Duration.of.Credit.Month Purpose
[4] Value.Savings.Stocks
Root node error: 97/350 = 0.27714
n= 350

Pruning Table

Level	CP	Num Splits	Rel Error	X Error	X Std Dev
1	0.068729	0	1.00000	1.00000	0.086326
2	0.041237	3	0.79381	0.94845	0.084898
3	0.025773	4	0.75258	0.88660	0.083032



3. *Forest Model* - significant predictor variables that are most important are Credit-Amount, Age-years and Duration-of-Credit-Month.

Basic Summary

Call:

randomForest(formula = Credit.Application.Result ~ Account.Balance + Duration.of.Credit.Month + Payment.Status.of.Previous.Credit + Purpose + Credit.Amount + Value.Savings.Stocks + Length.of.current.employment + Instalment.per.cent + Most.valuable.available.asset + Age.years + Type.of.apartment + No.of.Credits.at.this.Bank, data = the.data, ntree = 500, replace = TRUE)

Type of forest: classification

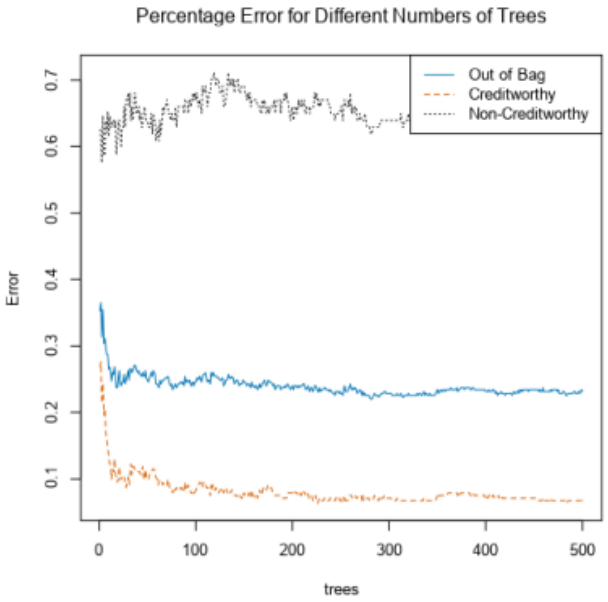
Number of trees: 500

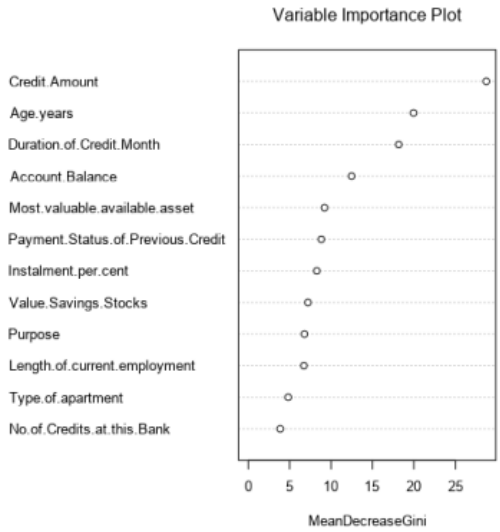
Number of variables tried at each split: 3

OOB estimate of the error rate: 23.1%

Confusion Matrix:

	Classification Error	Creditworthy	Non-Creditworthy
Creditworthy	0.067	236	17
Non-Creditworthy	0.66	64	33



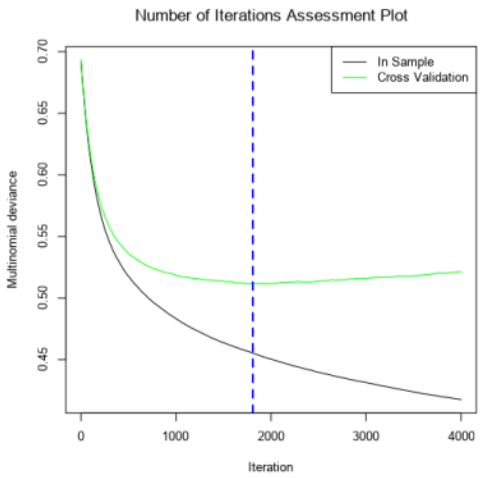


4. *Boosted Model* - significant predictor variables that are most important are Account-Balance and Credit-Amount

Report for Boosted Model Boosted_Model

Basic Summary:

Loss function distribution: Bernoulli
Total number of trees used: 4000
Best number of trees based on 5-fold cross validation: 1808



- 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?

After the validation is made the results are as follows:
Accuracy for the Decision Tree model is 0.746, for the Forest Model is 0.793, for the Boosted Model is 0.787, and for the Logistic Regression Stepwise 0.76.

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
Forest_Model	0.7933	0.8681	0.7368	0.9714	0.3778
Boosted_Model	0.7867	0.8632	0.7490	0.9619	0.3778
Stepwise	0.7600	0.8364	0.7306	0.8762	0.4889

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] is defined as the number of cases that are **correctly** predicted to be Class [class name] divided by the total number of cases that actually belong to Class [class name], this measure is also known as *recall*.

AUC: area under the ROC curve, only available for two-class classification.

F1: F1 score, 2 * precision * recall / (precision + recall). The *precision* measure is the percentage of actual members of a class that were predicted to be in that class divided by the total number of cases predicted to be in that class. In situations where there are three or more classes, average precision and average recall values across classes are used to calculate the F1 score.

Confusion matrix of Boosted_Model

	Actual_Creditworthy	Actual_Non-Creditworthy
Predicted_Creditworthy	101	28
Predicted_Non-Creditworthy	4	17

Confusion matrix of Decision_Tree

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

Confusion matrix of Forest_Model

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

Confusion matrix of Stepwise

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

The overall percent accuracy of the Logistic model is 76% which is strong.
PPV= true positives \ (true positives + false positives) = 92 / (92+23) =.80

NPV= true negatives\ (true negatives + false negatives) =22/ (22+13)= .63
So after checking the confusion matrix there is bias seen in the model's prediction to Creditworthy.

The accuracy of the Decision Tree model is 75% which is strong
PPV= true positives \ (true positives + false positives) = 93 / (93+26) =.78
NPV= true negatives \ (true negatives + false negatives) = 19/ (19+12) = .61
So after checking the confusion matrix there is bias seen in the model's prediction to Creditworthy.

The accuracy of the Forest model is 79% which is strong
PPV= true positives \ (true positives + false positives) = 102 / (102+28) =.78
NPV= true negatives \ (true negatives + false negatives) = 17/ (17+3) = .85
So after checking the confusion matrix there is no bias seen in the model's prediction.

The accuracy of the Boosted model is 79% which is strong
PPV= true positives \ (true positives + false positives) = 101 / (101+28) =.78
NPV= true negatives \ (true negatives + false negatives) = 17/ (17+4) = .81
So after checking the confusion matrix there is no bias seen in the model's prediction.

When compared the Forest Tree model and the Boosted model have very similar values, but the Forest model showed slightly better results for the NPV.

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

Step 4: Writeup

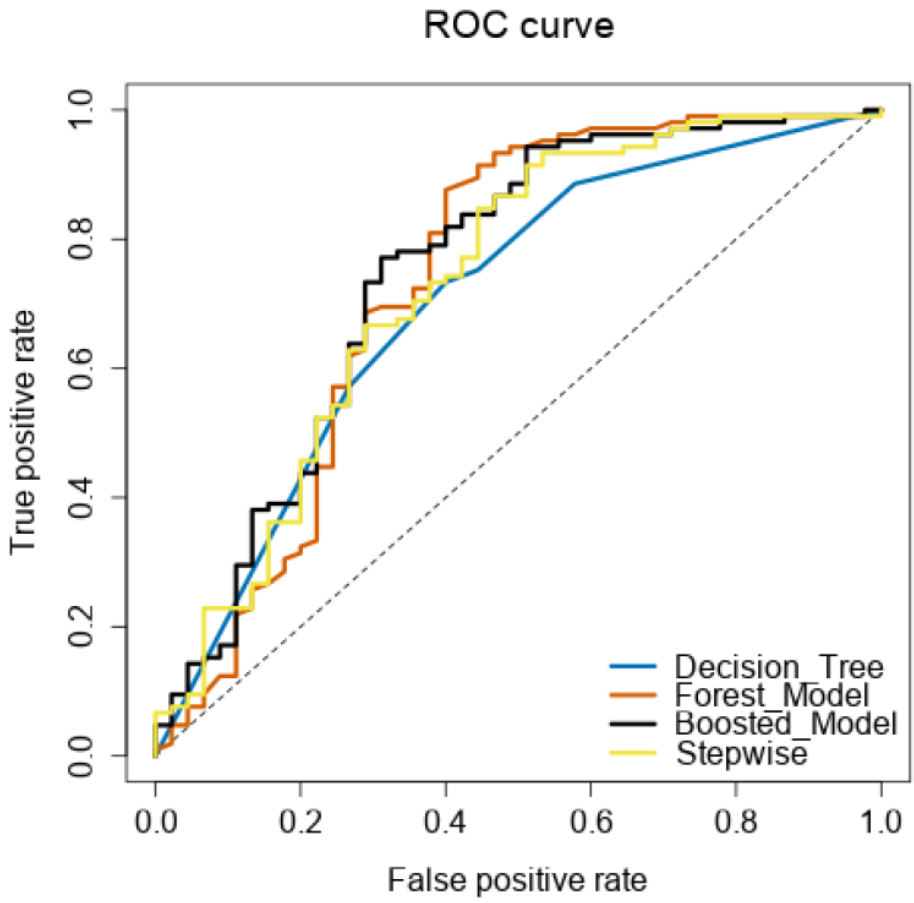
Decide on the best model and score your new customers. For reviewing consistency, if Score_Creditworthy is greater than Score_NonCreditworthy, the person should be labeled as "Creditworthy"

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:
  I chose to use the Forest Tree model because:
  - Overall Accuracy against your Validation was highest compared to the other models' accuracy 0.793
  - Accuracies within "Creditworthy" and "Non-Creditworthy" segments: Accuracy_Creditworthy is 0.971 and Accuracy_Non-Creditworthy is 0.378

- ROC graph –is a probability curve that illustrates how good our binary classification is in classifying classes based on true-positive and false-positive rates. The Forest Tree model out of the Four models reaches the top the quickest of all.



- Bias in the Confusion Matrices

Bias was noticed in the Logistic and Decision Tree models. Тхере њас но биас инт хе Форест Трее анд Боостед Модел.

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

- How many individuals are creditworthy?
  The Creditworthy individuals are considered the ones that have a greater score in creditworthiness from their respective score of non-creditworthiness. The total number of Creditworthy individuals is 408, while Non-Creditworthy are 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.