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

Due to a financial scandal that hit a competitive bank last week new customers applying for loans for your bank instead of the other bank in your city .A new 500 loan applications for new customers need to be processed to determine if they are creditworthy or not just in one week .

Key Decisions:

Answer these questions

- What decisions needs to be made?

determine if customers are creditworthy to give a loan to.

- What data is needed to inform those decisions?

1. all credit approvals from past loan applicants the bank has ever completed.
2. new customers data

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

Binary model , we will predict a categorical variable (creditworthy vs non-creditworthy)

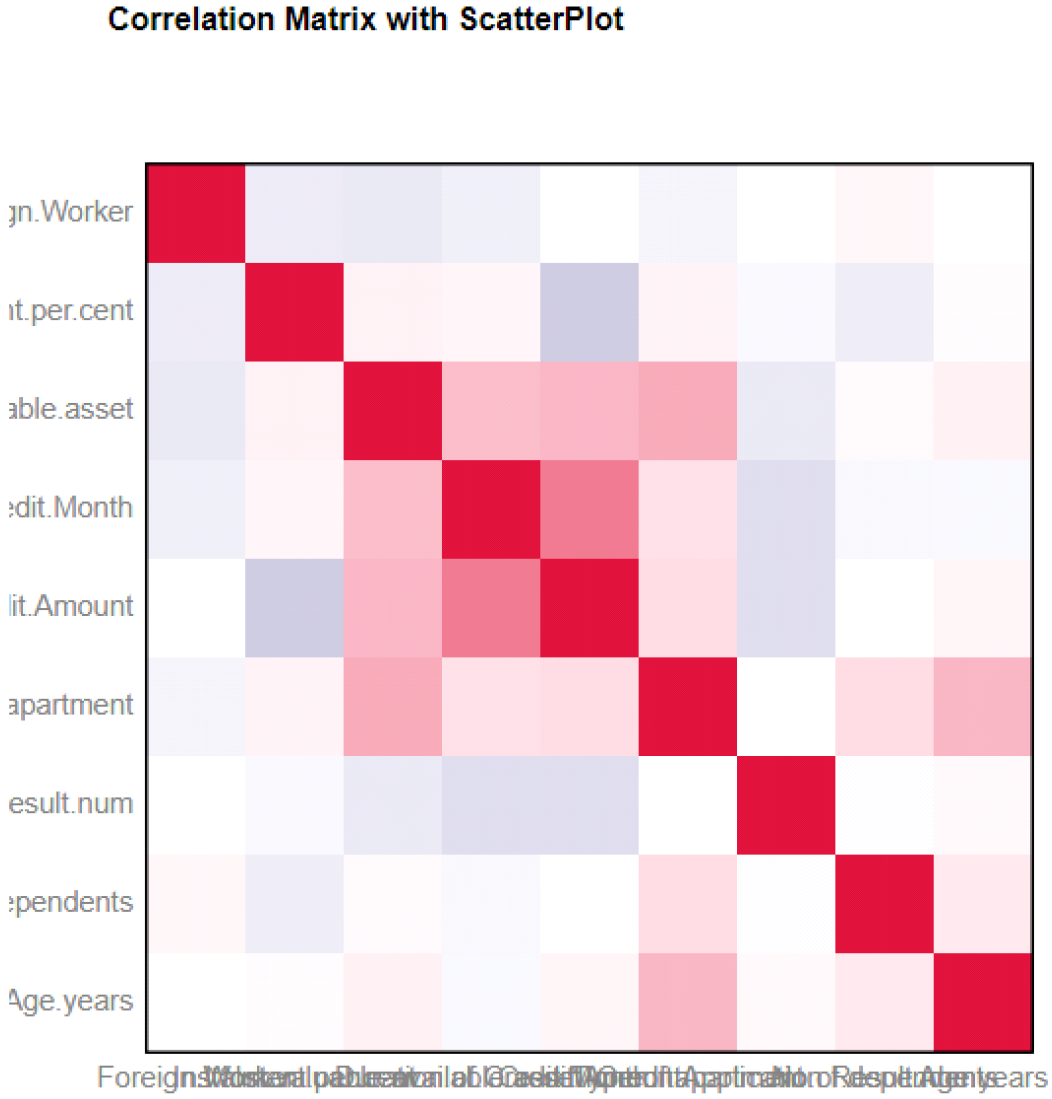
Step 2: Building the Training Set

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".

using Association analysis tool
no inner correlation higher than 0.70 between variables



- Are there any missing data for each of the data fields? Fields with a lot of missing data should be removed

delete **Duration-in-Current-address** **field with 69% missing value**
replace records with missing value with median in **Age-years** **field**

- 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.

delete
Occupation
Concurrent-Credits
becouse they are low variablity fields

- Your clean data set should have 13 columns where the Average of **Age Years** should be 36 (rounded up)

delete Telephone as recommended
delete Foreign-Worker ,No-of-dependents ,Guarantors

Table

Guarantors

Value	Frequency	Percent	Cumulative Frequency	Cumulative Percent
None	457	91.40	457	91.40
Yes	43	8.60	500	100.00

No-of-dependents

Value	Frequency	Percent	Cumulative Frequency	Cumulative Percent
1	427	85.40	427	85.40
2	73	14.60	500	100.00

Foreign-Worker

Value	Frequency	Percent	Cumulative Frequency	Cumulative Percent
1	481	96.20	481	96.20
2	19	3.80	500	100.00

above variables skewed to one value, with high frequncy percentage,
it wouldn't be very heloful as predictor variables.

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

Step 3: Train your Classification Models

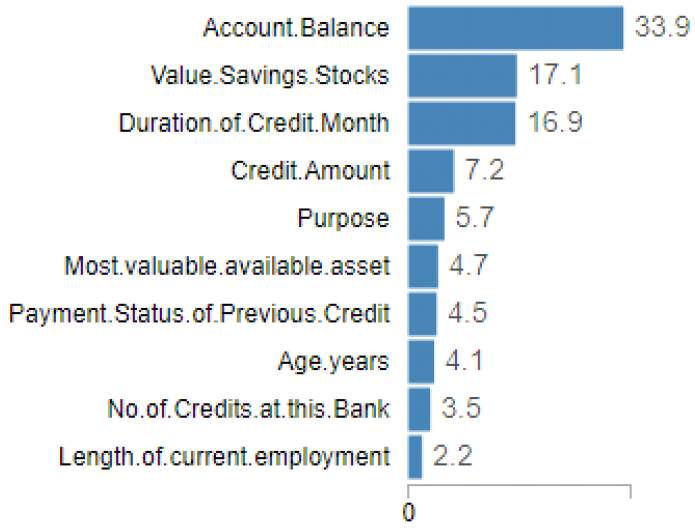
Create all of the following models: Logistic Regression, Decision Tree, Forest Model, Boosted Model

Logistic Regression Model :
using stepwise tool to choose most important predictor variables

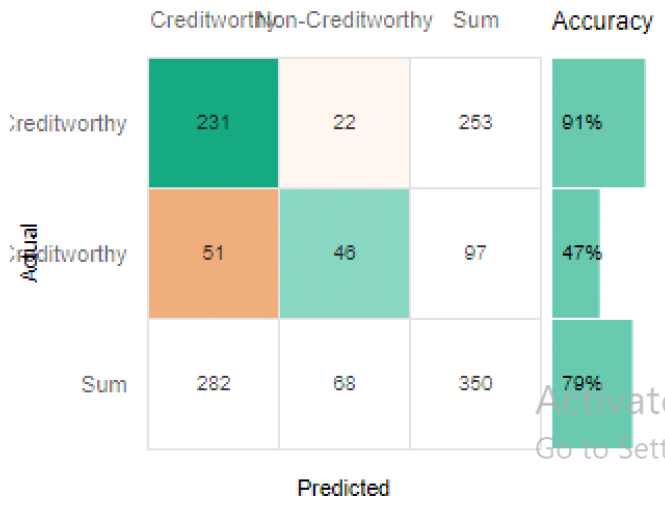
	LR Chi-Sq	DF	Pr(>Chi-Sq)	
Account.Balance	31.129	1	2.41e-08	***
Payment.Status.of.Previous.Credit	5.687	2	0.05823	.
Purpose	12.225	3	0.00665	**
Credit.Amount	9.882	1	0.00167	**
Length.of.current.employment	5.522	2	0.06324	.
Instalment.per.cent	5.198	1	0.02261	*
Most.valuable.available.asset	3.509	1	0.06104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

Decision Tree Model:

Variable Importance

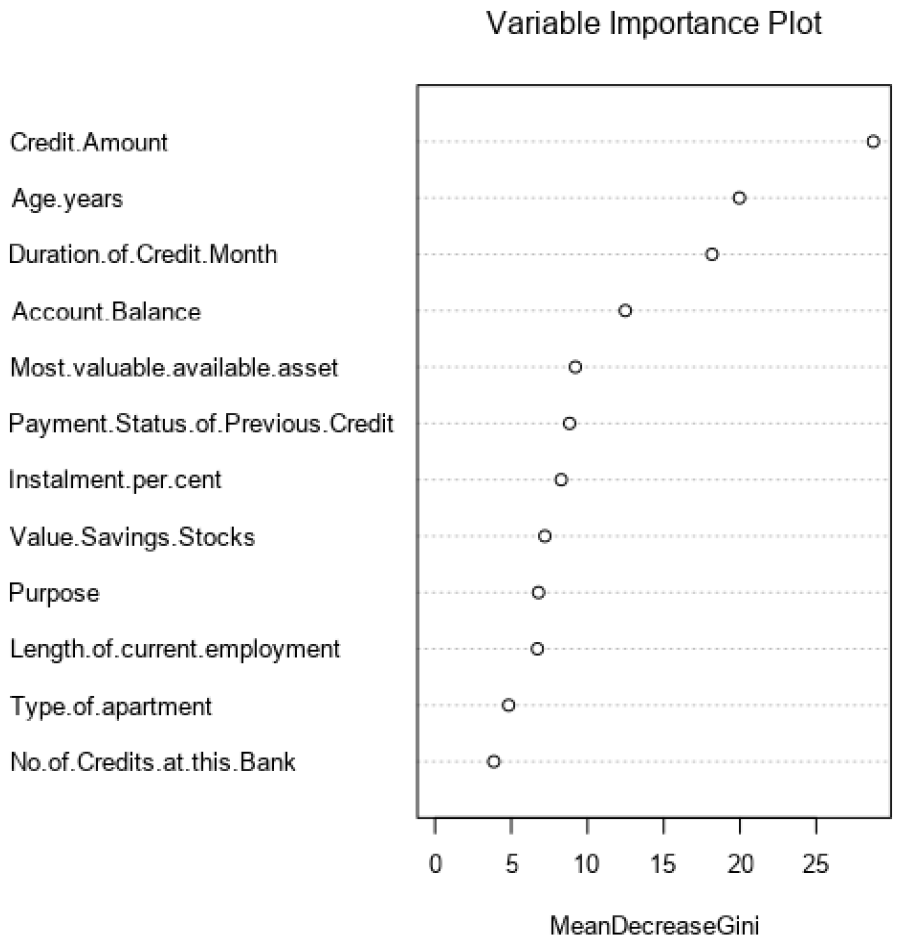


Confusion Matrix



Activate Windows
Go to Settings to activate Windows.

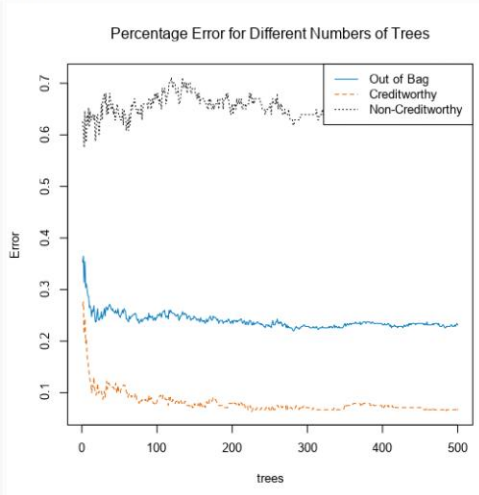
Forest Model:



Confusion Matrix:

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

Plots



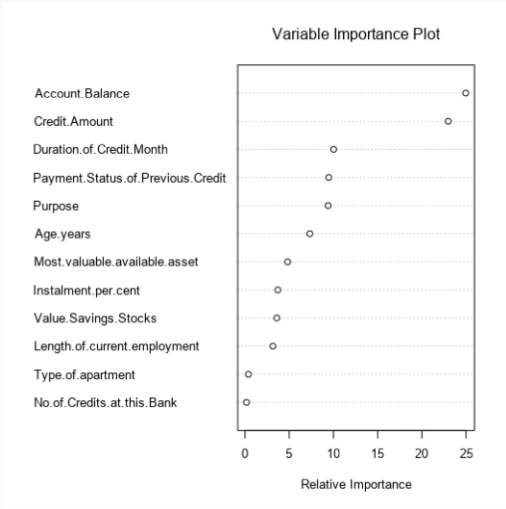
Activate Windows

OOB estimate of the error rate: 23.1%

we can see classification error Non_Creditworthy is really high with 0.66

Boosted Model:

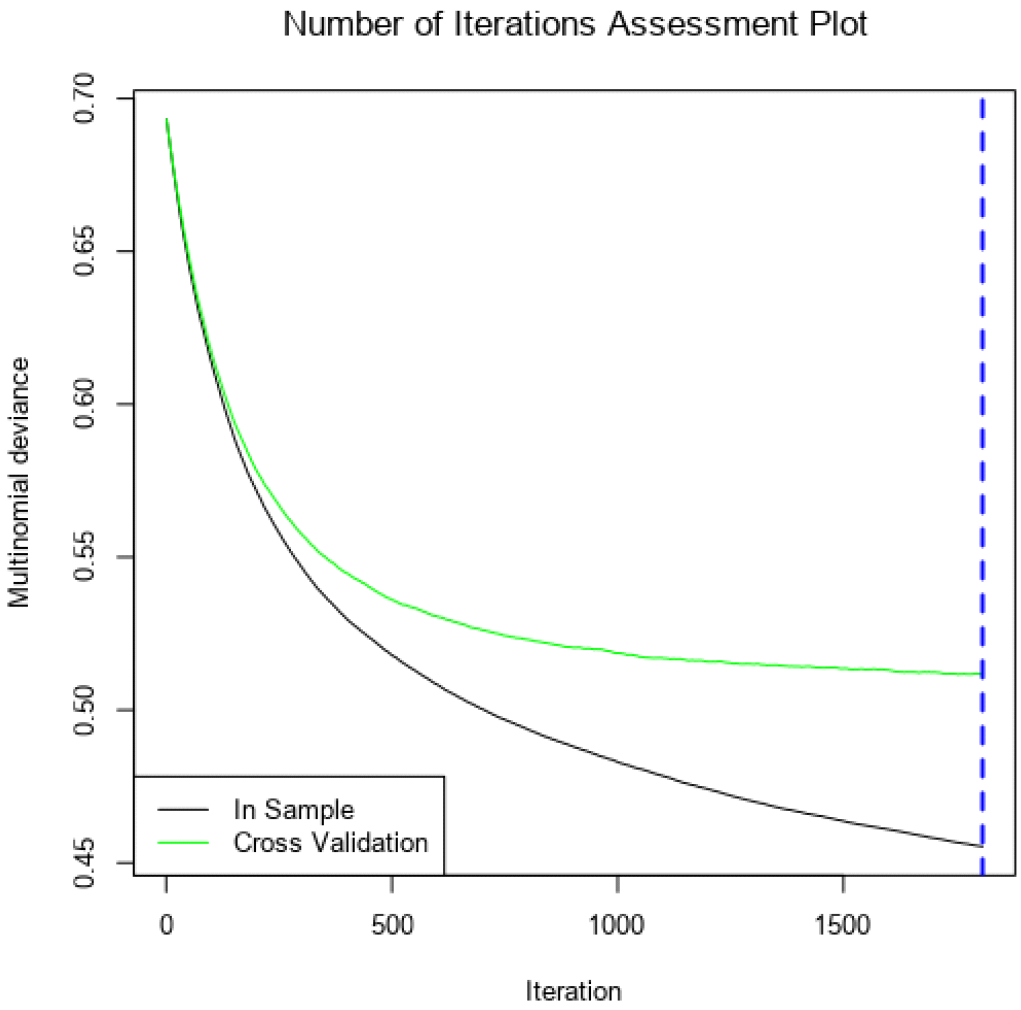
Plots:



The Variable Importance Plot provides information about the relative importance of each predictor field. The measures are normalized to sum to 100, and the value for each field gives the relative percentage importance of that field to the overall model.

Activate Windows

the maximum number of trees used in the model 1808



Accuracy change with model comparison tool:

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	0.7933	0.8681	0.7368	0.9714	0.3778
boosted	0.7867	0.8632	0.7515	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

	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

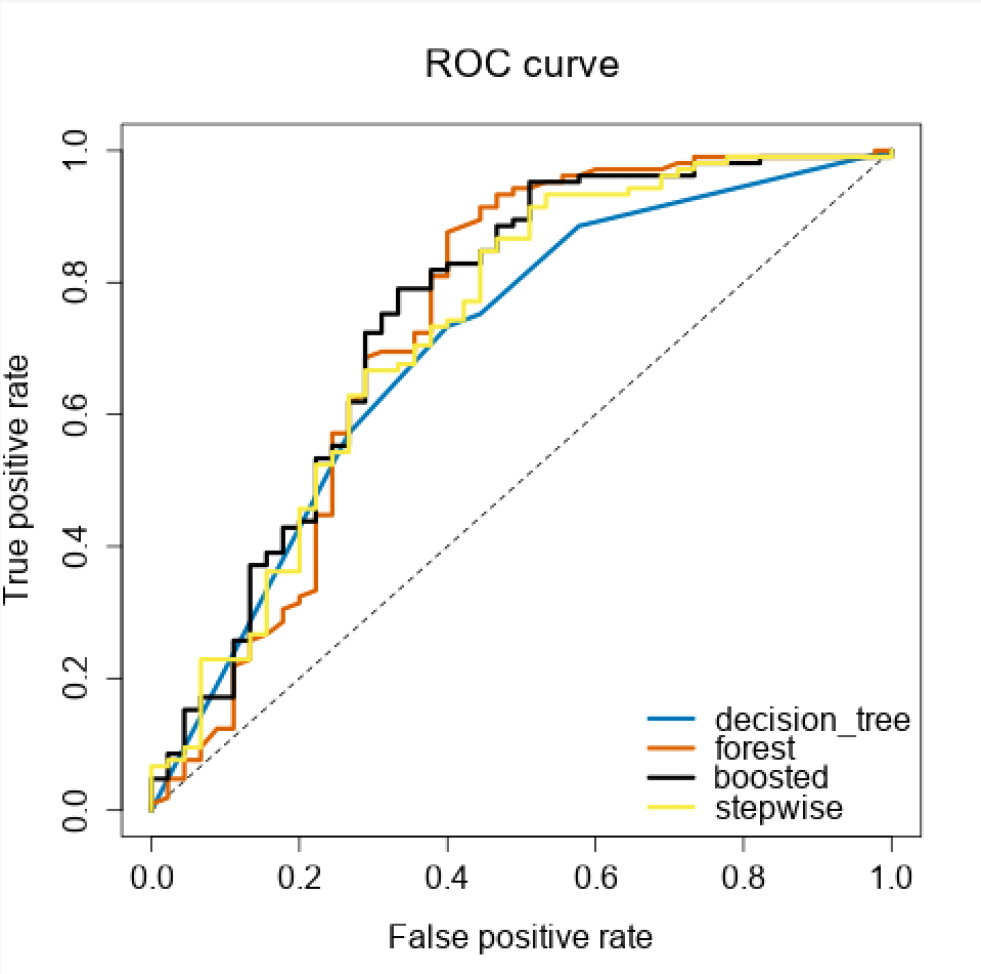
	Actual_Creditworthy	Actual_Non-Creditworthy
Predicted_Creditworthy	102	28
Predicted_Non-Creditworthy	3	17

Confusion matrix of stepwise

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

Activate Windows
Go to Settings to activate Windows.

- **the forest model has the highest Creditworty segment and overall accuracy**
- **all models predictions bias toward Creditworthy**
- **the logistic regression predict more Non-Creditworthy than other models**



In ROC curve forest model is the closest to 1 in true positive rate (top left corner), and decisin tree model show a bad performance (bais)

Step 4: Writeup

- **I will use Forest model with highest accuracy 0.79**
- **and the highest number of Creditworty customers predicted correct (102) accuracy (0.9714)**
- **correctly pridict Non-Creditworthy (17)customers accuracy (0.3778)**
- **With a clear bias toward Creditworthy**
- **ROC curve show that the forest model has the best performance**

because of these points we will use the forest model to predict if the new customers are Creditworthy.

- How many individuals are creditworthy?

There are 408 new customer creditworthey

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.