Project: Creditworthiness - Kevin Raj Submission 2

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?
    - Determining which new customers can be approved for a loan or not.
    - Determining which model is the most accurate in predicting which new customers can be approved for a loan or not.
    - Determining how to process all the new 500 loan applications in one week.

- What data is needed to inform those decisions?
    - Data on past applications
    - List of new customer applications
    - Some of the variables which could influence our decision to determine if a customer is credit worthy or not are their current length of employment, income, credit score, if the customer carries a credit balance from month to month, age, and their current savings. Useful variables could be payment status of previous credit to predict if the customer would be credit worthy based on past history and duration of credit month to see how long a particular customer's credit history and if they've had any credit problems in the past.

- What kind of model (Continuous, Binary, Non-Binary, Time-Series) do we need to use to help make these decisions?
    - Binary (loan will be approved or not approved).

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". There are no numerical data fields that are highly-correlated, the closest one is credit-amount and duration-of-credit-month which has a correlation of .57398.
- Are there any missing data for each of the data fields? Fields with a lot of missing data should be removed
    - Yes, the duration-in-current-address field has 69% of the data missing, and the age-years field has 2% of the data missing. The duration-in-current address field 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.
    - Foreign worker, concurrent credits, no of dependents, occupation, and guarantors all have a low number of unique values and have a high concentration of values in a single value and should be removed from the data set (low variability). Telephone should be removed from the data set due to there is no logical connection with this variable to the predicted outcome.
- 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:*

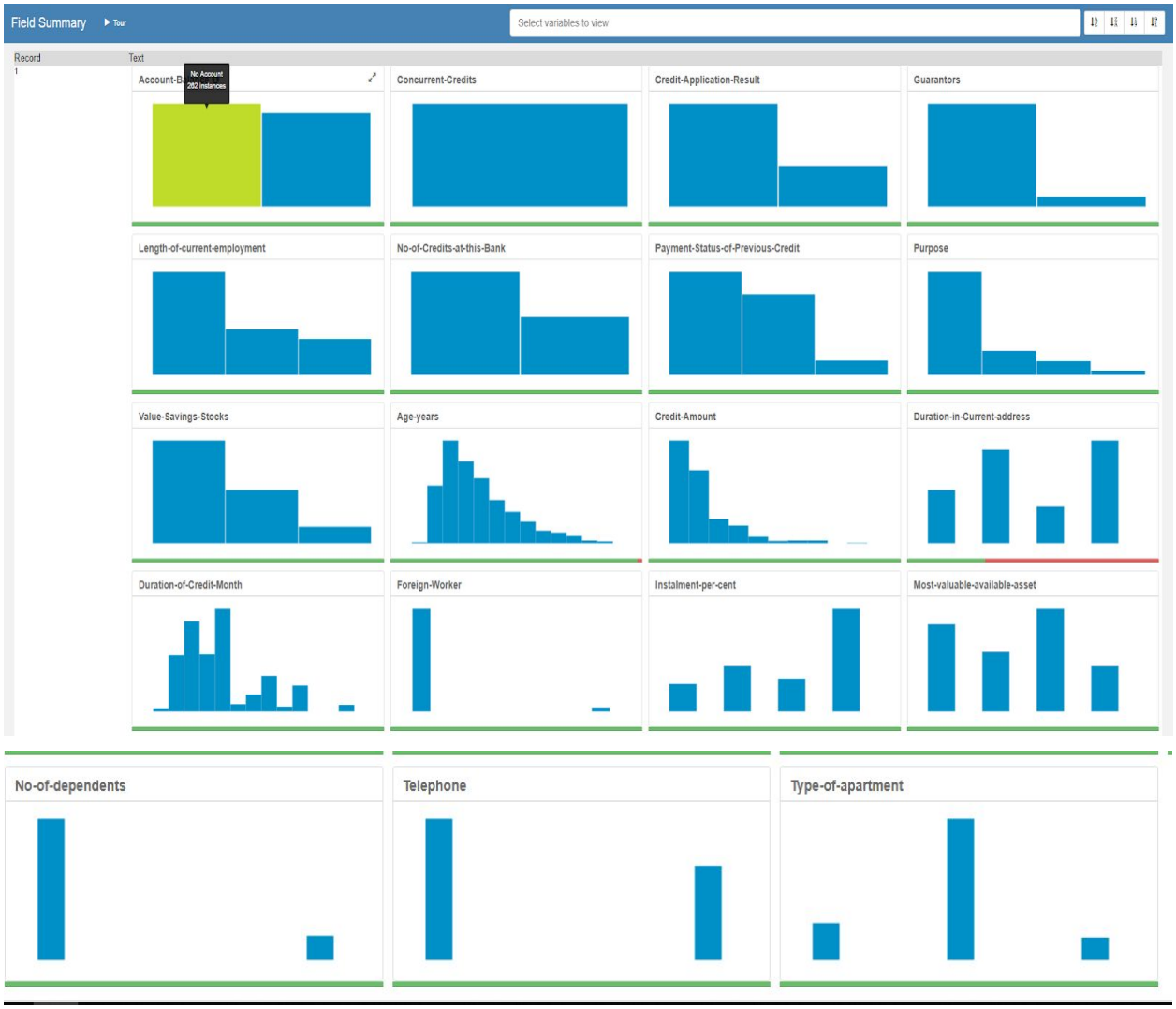
Variable	**Data Type**
Credit-Application-Result	String
Account-Balance	String
Duration-of-Credit-Month	Double
Payment-Status-of-Previous-Credit	String

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.
    - Foreign worker, concurrent credits, no of dependents, occupation, and guarantors all have a low number of unique values and have a high concentration of values in a single value and should be removed from the data set (low variability). Telephone should be removed from the data set due to there is no logical connection with this variable to the predicted outcome. The duration-in-current-address field has 69% of the data missing and should be removed. The age-in-years variable has 2% data missing, and I imputed the field with the median age due to age being skewed to the left (younger age) and not normal.



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.

- Stepwise Logistic regression:

Record | Report

Report for Logistic Regression Model stepwise_regression

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

Coefficients:

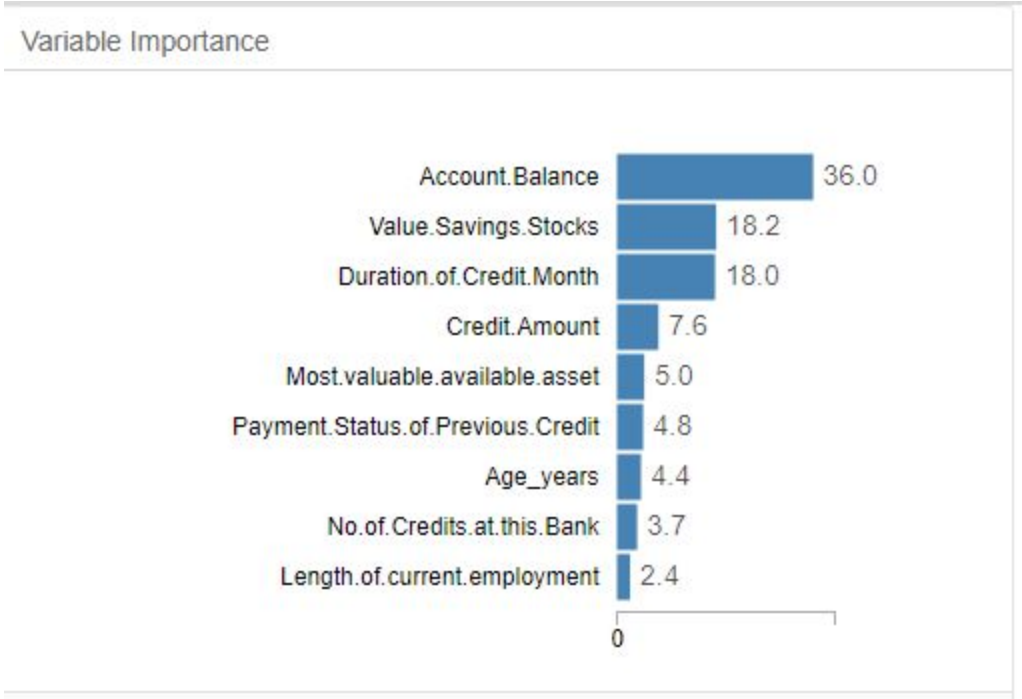
	Estimate	Std. Error	z value	Pr(>|z|)	
(Intercept)	-2.9621914	6.837e-01	-4.3326	1e-05	***
Account.BalanceSome Balance	-1.6053228	3.067e-01	-5.2344	1.65e-07	***
Payment.Status.of.Previous.CreditPaid Up	0.2360857	2.977e-01	0.7930	0.42775	
Payment.Status.of.Previous.CreditSome Problems	1.2154514	5.151e-01	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.employment4-7 yrs	0.3127022	4.587e-01	0.6817	0.49545	
Length.of.current.employment< 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	.

Significance codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

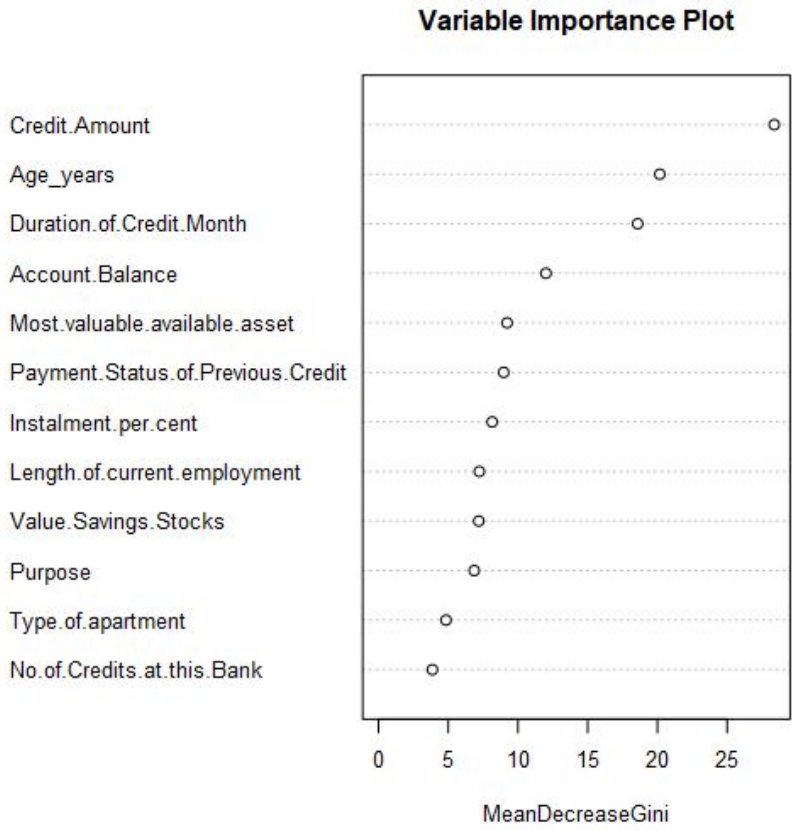
(Dispersion parameter for binomial taken to be 1)

These variables were the most significant with low p-values (less than .05): account balance, payment status of previous credit, purpose, credit amount, length of current employment (less than 1 year), instalment percent.

- Decision tree (account balance, value savings, and duration of credit month were significant predictor variables):

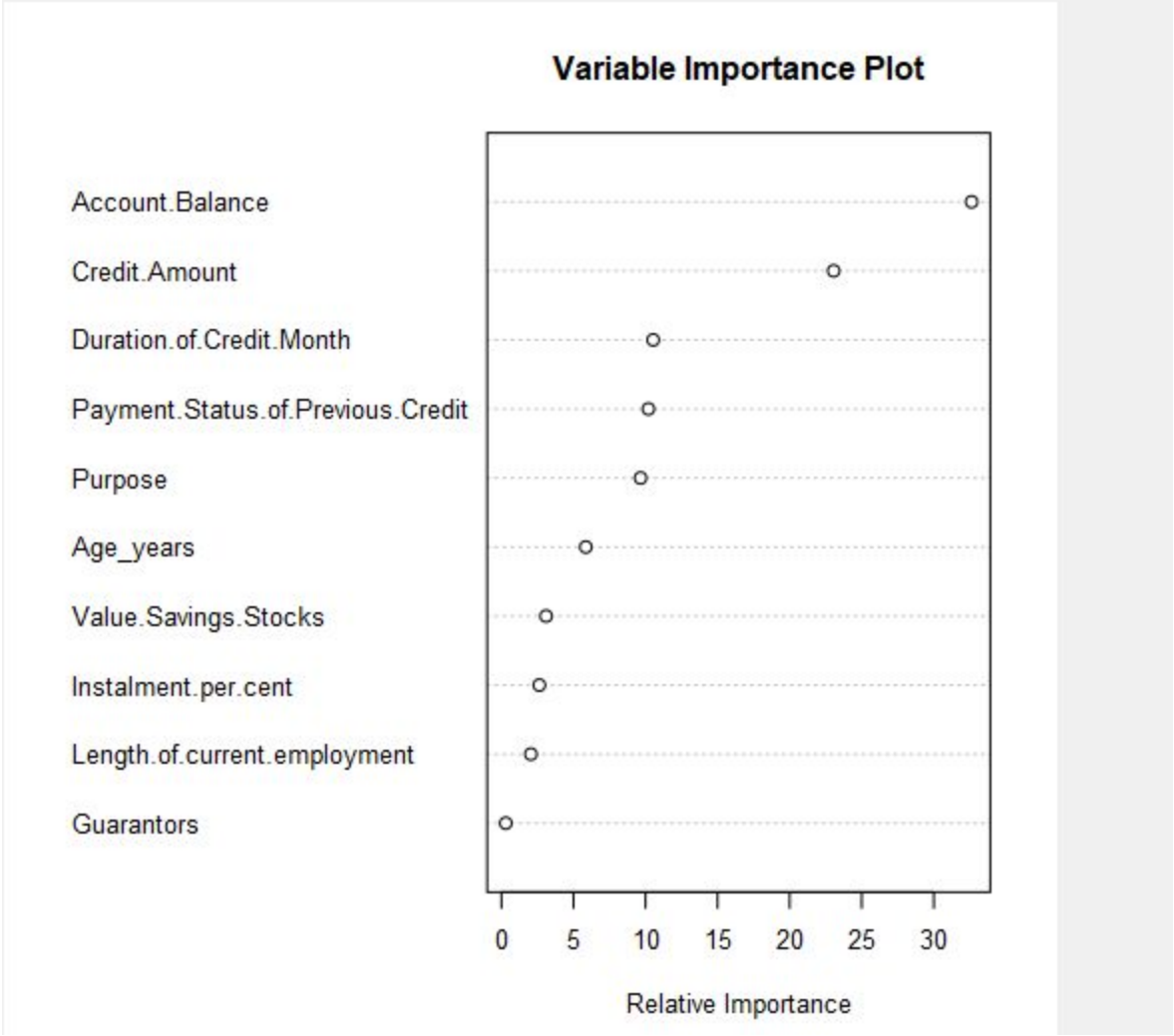


Forest model (credit amount, age, and duration of credit month were significant predictor variables):



Boosted model (account balance and credit amount were significant predictor variables):

Plots:



- 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?
    - Stepwise logistic regression:

Model Comparison Report

Fit and error measures

Model	Accuracy	F1	AUC	Accuracy_Creditworthy	Accuracy_Non-Creditworthy
Decision_tree	0.7467	0.8273	0.7054	0.8667	0.4667
forest_model	0.8000	0.8707	0.7361	0.9619	0.4222
boosted_model	0.7867	0.8632	0.7524	0.9619	0.3778
stepwise_regression	0.7600	0.8364	0.7306	0.8762	0.4889

Overall accuracy was .7600

Confusion matrix of stepwise_regression

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

Stepwise Logistic regression - misclassification rate: 36/150 = .24 (error rate), indicating there is bias in predicting creditworthiness - the model is not 100% accurate.

Decision tree:

Model Comparison Report

Fit and error measures

Model	Accuracy	F1	AUC	Accuracy_Creditworthy	Accuracy_Non-Creditworthy
Decision_tree	0.7467	0.8273	0.7054	0.8667	0.4667
forest_model	0.8000	0.8707	0.7361	0.9619	0.4222
boosted_model	0.7867	0.8632	0.7524	0.9619	0.3778
stepwise_regression	0.7600	0.8364	0.7306	0.8762	0.4889

Overall accuracy: .7467

Confusion matrix of Decision_tree

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

Decision tree - misclassification rate: 38/150 = .25 (error rate), indicating there is bias in predicting creditworthiness - the model is not 100% accurate.

Forest model:

Overall accuracy: .80

Confusion matrix of forest_model

	Actual_Creditworthy	Actual_Non-Creditworthy
Predicted_Creditworthy	101	26
Predicted_Non-Creditworthy	4	19

Forest model - misclassification rate: 30/150 = .20 (error rate), indicating there is bias in predicting credit worthiness - the model is not 100% accurate.

Boosted model:

Model Comparison Report

Fit and error measures

Model	Accuracy	F1	AUC	Accuracy_Creditworthy	Accuracy_Non-Creditworthy
Decision_tree	0.7467	0.8273	0.7054	0.8667	0.4667
forest_model	0.8000	0.8707	0.7361	0.9619	0.4222
boosted_model	0.7867	0.8632	0.7524	0.9619	0.3778
stepwise_regression	0.7600	0.8364	0.7306	0.8762	0.4889

Overall accuracy: .7867

Confusion matrix of boosted_model

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

Boosted model - misclassification rate: 32/150 = .21 (error rate), indicating there is bias in predicting credit worthiness - the model is not 100% accurate.
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 - forest model has the highest accuracy with .80

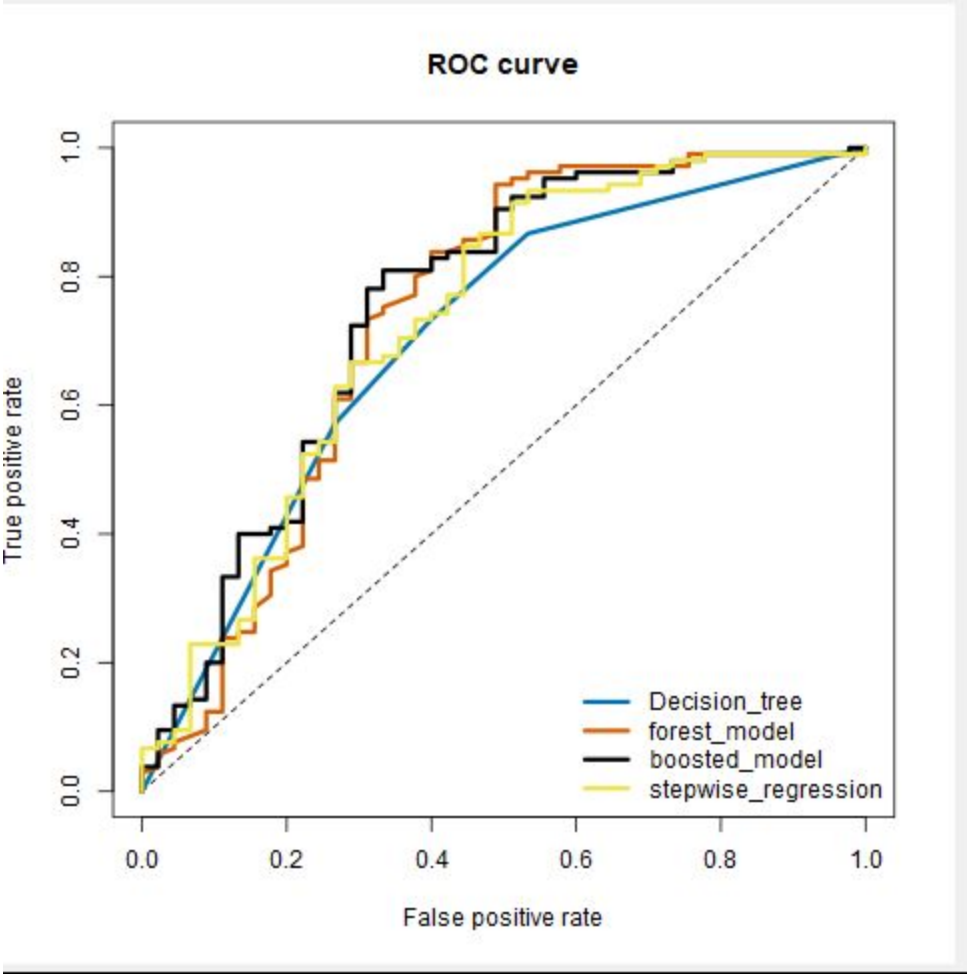
Model Comparison Report

Fit and error measures

Model	Accuracy	F1	AUC	Accuracy_Creditworthy	Accuracy_Non-Creditworthy
Decision_tree	0.7467	0.8273	0.7054	0.8667	0.4667
forest_model	0.8000	0.8707	0.7361	0.9619	0.4222
boosted_model	0.7867	0.8632	0.7524	0.9619	0.3778
stepwise_regression	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.

- Accuracies within “Creditworthy” and “Non-Creditworthy” segments
    - Forest model is the most accurate for the creditworthy segment (.9619 accuracy). ROC graph - the AUC for forest model is at .7361 and is the second highest indicating the model is a better predictive model than the others. From wikipedia: The ROC curve is created by plotting the true positive rate (TPR) against the false positive rate (FPR) at various threshold settings. The true-positive rate is also known as sensitivity, recall or probability of detection[1] in machine learning. The false-positive rate is also known as the fall-out or probability of false alarm[1] and can be calculated as (1 − specificity). It can also be thought of as a plot of the Power as a function of the Type I Error of the decision rule (when the performance is calculated from just a sample of the population, it can be thought of as estimators of these quantities). The ROC curve is thus the sensitivity as a function of fall-out. When using normalized units, the area under the curve (often referred to as simply the AUC) is equal to the probability that a classifier will rank a randomly chosen positive instance higher than a randomly chosen negative one (assuming 'positive' ranks higher than 'negative'). A higher AUC can indicate the model is a better predictive model.



- Bias in the Confusion Matrices

Confusion matrix of forest_model		
	Actual_Creditworthy	Actual_Non-Creditworthy
Predicted_Creditworthy	101	26
Predicted_Non-Creditworthy	4	19

- Forest model - misclassification rate: 30/150 = .20 (error rate), indicating there is bias in predicting credit worthiness - the model is not 100% accurate.

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

- How many individuals are creditworthy?

Confusion matrix of Decision_tree

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

Confusion matrix of boosted_model

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

Confusion matrix of forest_model

	Actual_Creditworthy	Actual_Non-Creditworthy
Predicted_Creditworthy	101	26
Predicted_Non-Creditworthy	4	19

Confusion matrix of stepwise_regression

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

Using the forest model, I used the “Score” tool to calculate the probabilities of the credit-worthiness and non-creditworthiness for the “customers to score” file. I used a formula to say if the probability of predicted creditworthy is higher than the probability of the predicted non-credit worthy based on the score tool then score 1 for that customer. I summed up all the customers and got **406** customers who are credit worthy.

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.