Project: Creditworthiness

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?
  - Decision whether to give loan to a customer or not.

- What data is needed to inform those decisions?
  - Information about the creditworthiness of a customer is required to make a sound decision. Creditworthiness of a customer can be calculated or predicted based on his past behaviour ex:
    - His present bank account balance
    - Repayment history
    - Frequency of defaults
    - Employment history

- What kind of model (Continuous, Binary, Non-Binary, Time-Series) do we need to use to help make these decisions?
  - This is a binary classification model, as we need to predict whether a customer is creditworthy or not.

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

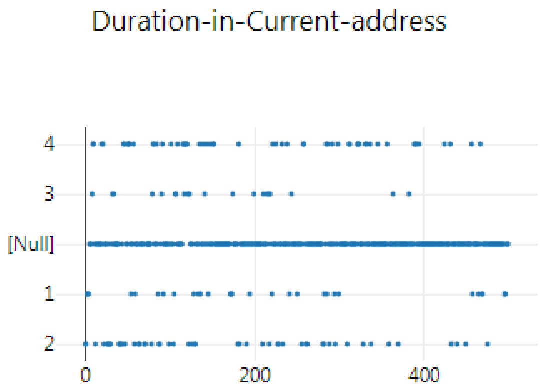
Note: *For students using software other than Alteryx, please format each variable as:*

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

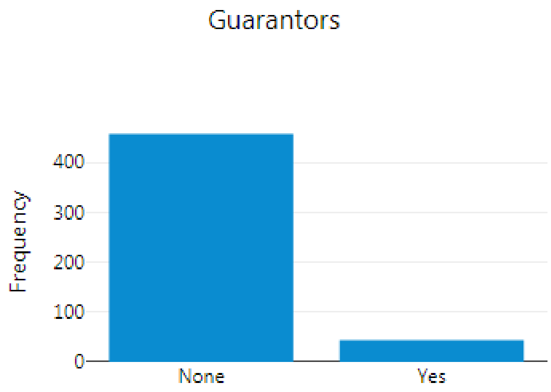
To achieve consistent results reviewers expect.

Answer this question:

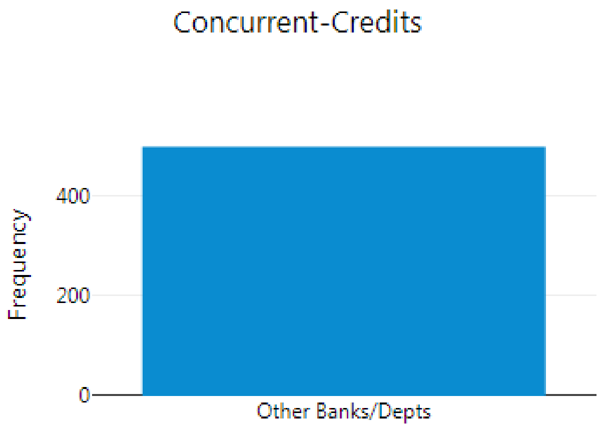
- In your cleanup process, which fields did you remove or impute? Please justify why you removed or imputed these fields. Visualizations are encouraged.
  - Fields removed
    - Guarantors
    - Duration-in-current-address
    - Concurrent-credits
    - Occupation
    - No-of-dependents
    - Foreign-Worker
    - Type-of-Apartment


Duration-in-Current-address

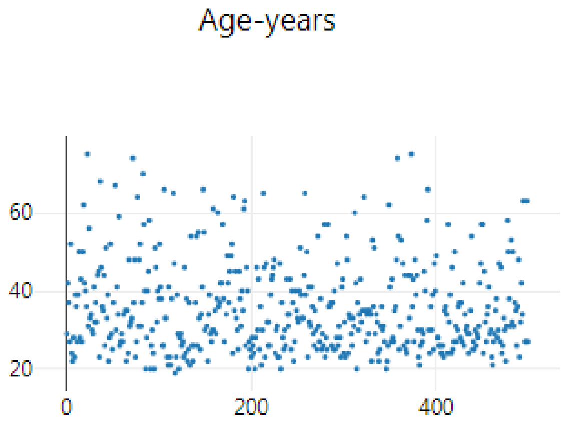
  - Removed because of high frequency of nulls


Guarantors

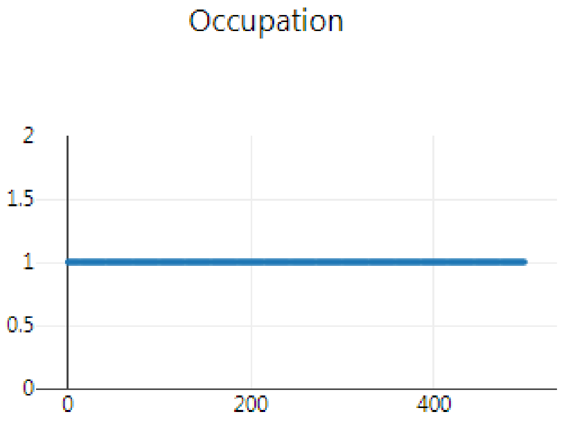
  - Removed Guarantors because of low variability of data (None – 457, Yes- 43)



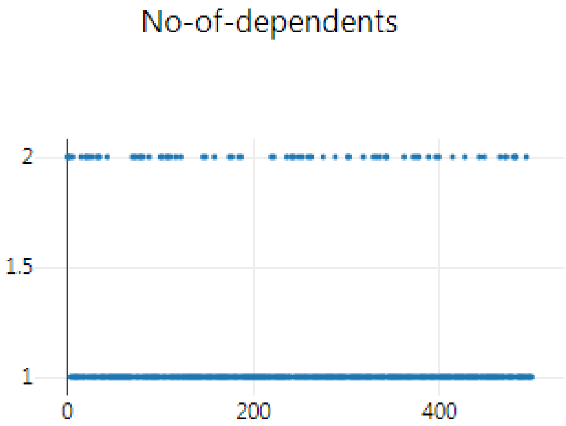
- Removed Concurrent-Credits because of low variability of data



- Imputed Nulls in Age-years with median, as number of nulls were 12.
- Imputing the value with median and not by mean so as to eliminate the effect of outlier.

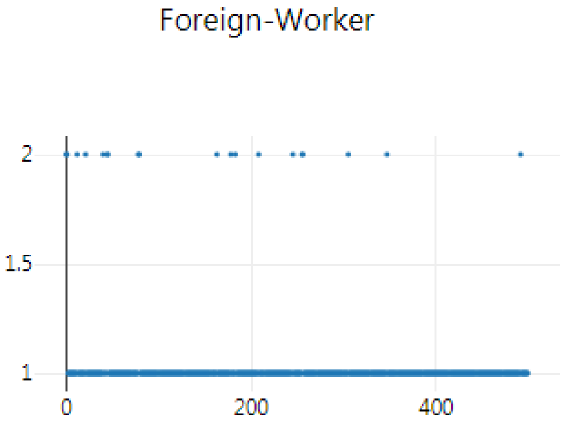


- Removed Occupation because of low variability of data



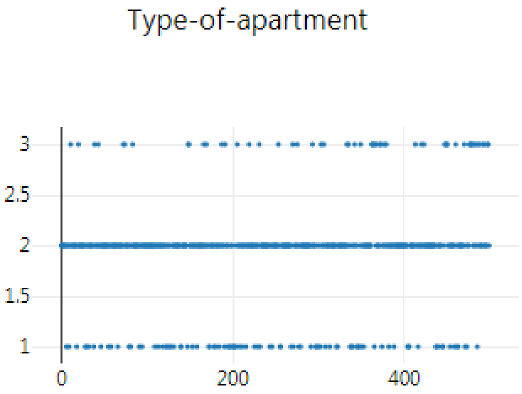
No-of-dependents

- Removed No-of-dependents because of low variability of data



Foreign-Worker

- Removed foreign-worker because of low variability of data



Type-of-apartment

- Removed Type-of-apartment because of low variability of data

Step 3: Train your Classification Models

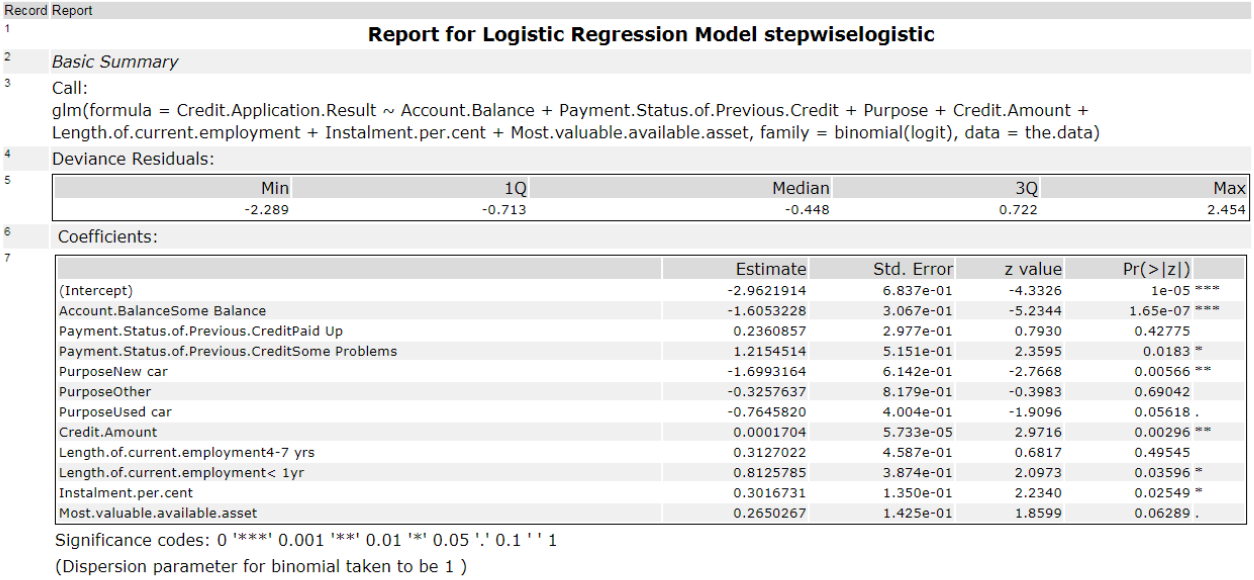
First, create your Estimation and Validation samples where 70% of your dataset should go to Estimation and 30% of your entire dataset should be reserved for Validation. Set the Random Seed to 1.

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.

Logistic regression



Record Report

1 **Report for Logistic Regression Model stepwiselogistic**

2 *Basic Summary*

3 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)

4 Deviance Residuals:

5

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

6 Coefficients:

7

	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	.

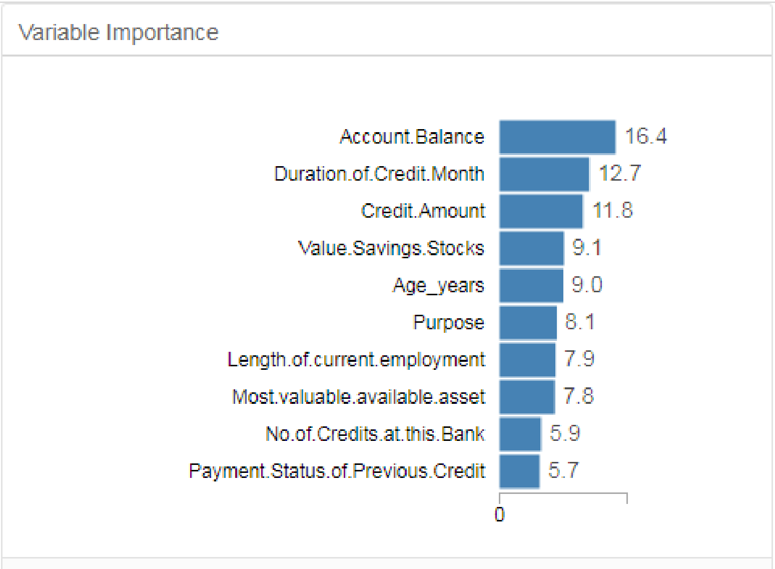
Significance codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial taken to be 1)

According to logistic model

- Account balance
- Purpose
- Credit Amount
- Length of current employment
- Instalment percent
- Most valuable available asset

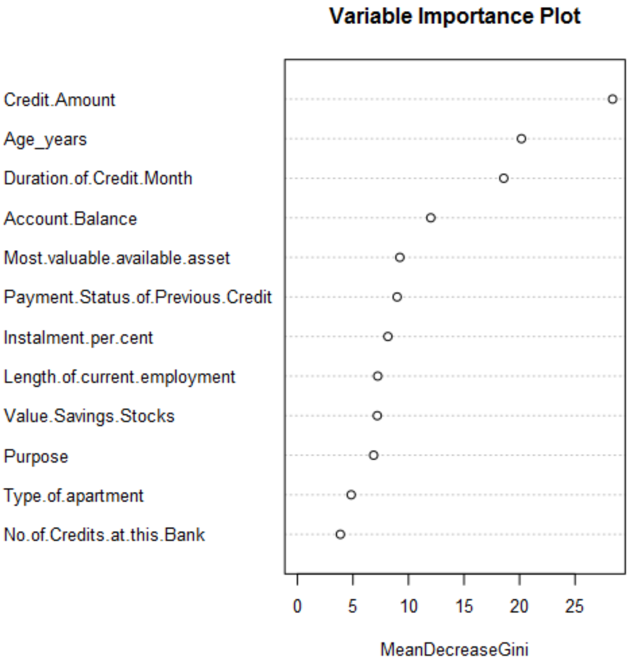
Decision Tree



According to decision model:

- Account balance
- Duration of credit month
- Credit Amount

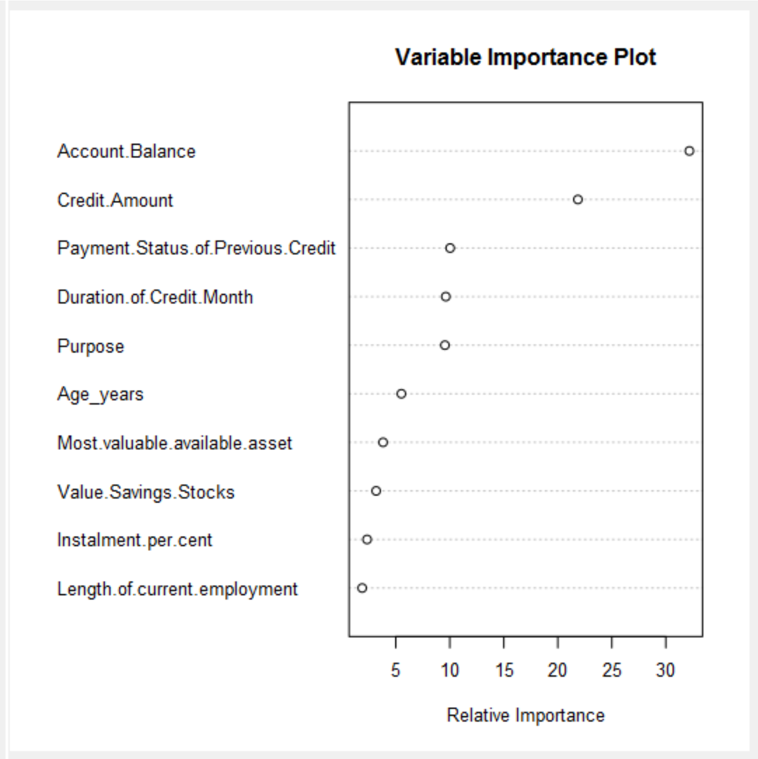
<u>Forest Model</u>



According to Forest model:

- Credit amount
- Age year
- Duration of credit month

Boosted Model

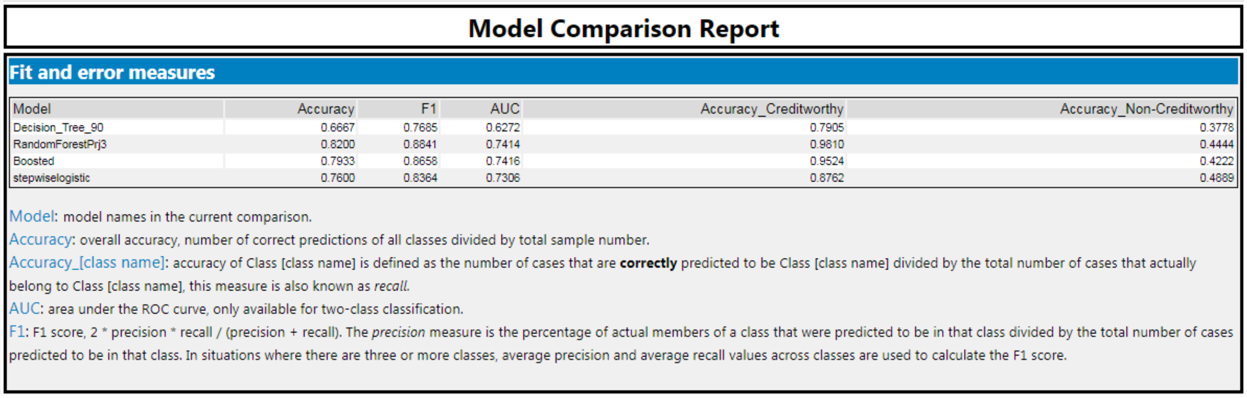


According to boosted model:

- Account balance
- Credit amount
- Payment status of previous credit

- 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?

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



Model Comparison Report

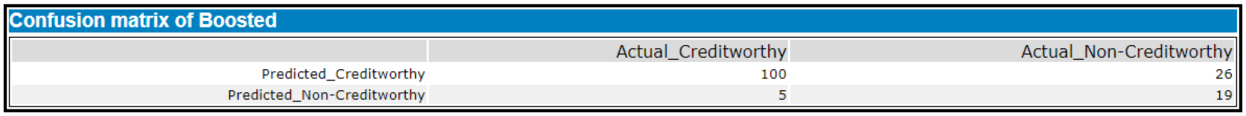
Fit and error measures

Model	Accuracy	F1	AUC	Accuracy_Creditworthy	Accuracy_Non-Creditworthy
Decision_Tree_90	0.6667	0.7685	0.6272	0.7905	0.3778
RandomForestPrj3	0.8200	0.8841	0.7414	0.9810	0.4444
Boosted	0.7933	0.8658	0.7416	0.9524	0.4222
stepwiselogistic	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.

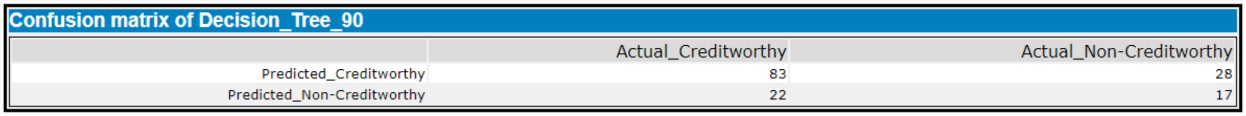
We can see the overall accuracy for these models as follows:

- Logistic model : 76%
- Decision tree model : 67%
- Random forest model : 82%
- Boosted model : 79%



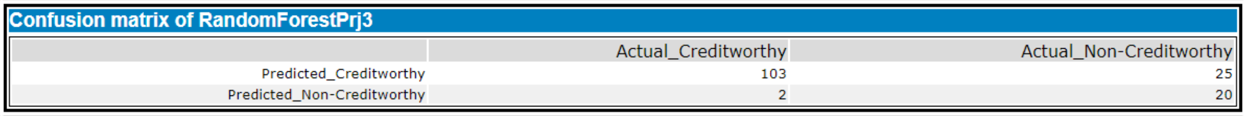
Confusion matrix of Boosted

	Actual_Creditworthy	Actual_Non-Creditworthy
Predicted_Creditworthy	100	26
Predicted_Non-Creditworthy	5	19



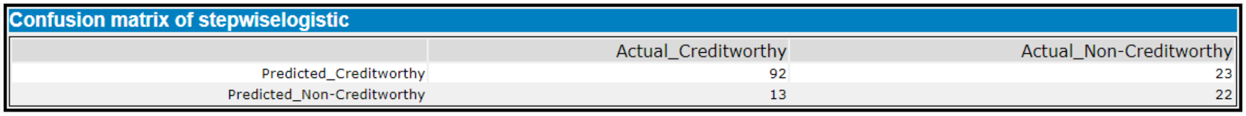
Confusion matrix of Decision_Tree_90

	Actual_Creditworthy	Actual_Non-Creditworthy
Predicted_Creditworthy	83	28
Predicted_Non-Creditworthy	22	17



Confusion matrix of RandomForestPrj3

	Actual_Creditworthy	Actual_Non-Creditworthy
Predicted_Creditworthy	103	25
Predicted_Non-Creditworthy	2	20



Confusion matrix of stepwiselogistic

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

In all the four models we can observe that the accuracy of predicting ‘creditworthiness’ is more than that of predicting ‘non_creditworthiness’.

i.e. accuracy is low for predicting ‘non_creditworthiness’

- Logistic model : 49%
- Decision tree model : 38%
- Random forest model : 44%
- Boosted model : 42%

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
  - Accuracies within “Creditworthy” and “Non-Creditworthy” segments
  - ROC graph
  - Bias in the Confusion Matrices

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

Overall Accuracy against your Validation set:

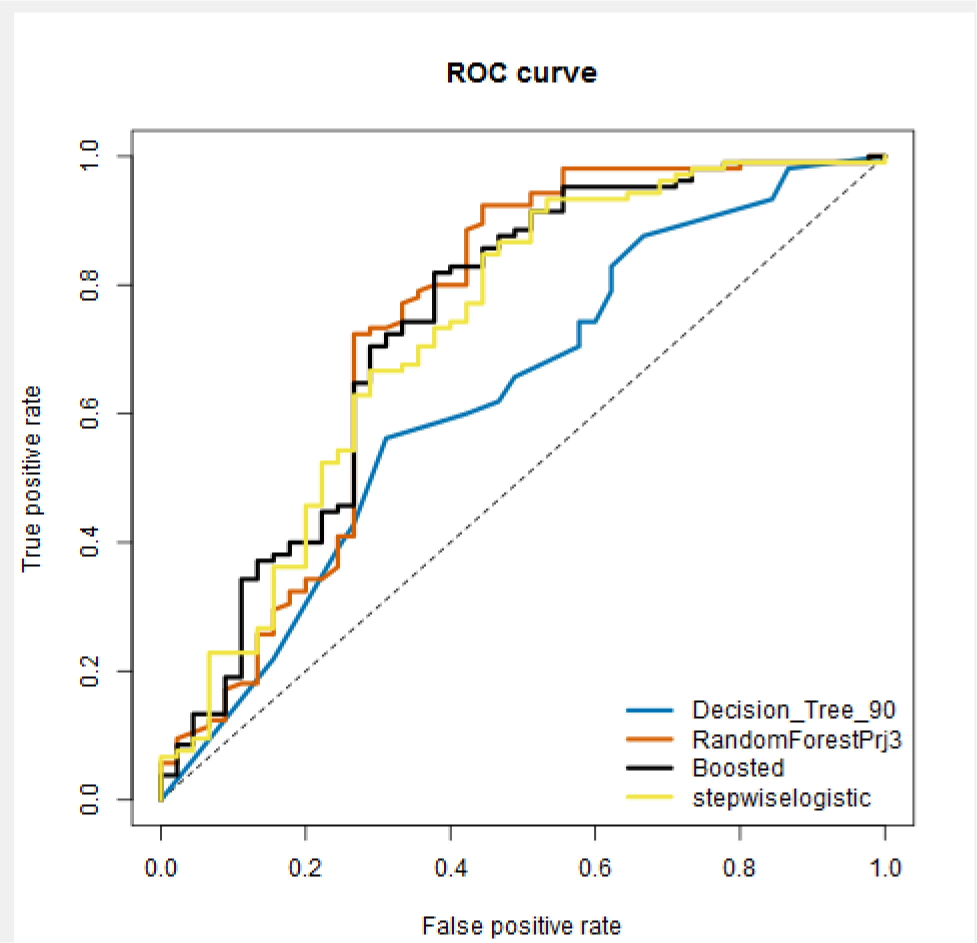
- Logistic model – 76%
- Decision tree model – 67%
- Random forest model – 82%
- Boosted model – 79%

Random forest model has the highest overall accuracy

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

When we compare accuracies with creditworthy and non creditworthy segment Random forest model has highest accuracy in predicting creditworthiness with 98% accuracy However logistic model has the highest accuracy in predicting non creditworthiness with 49% accuracy, random forest being second with 44 % accuracy

ROC graph



When we analyse the ROC curve, we can see the area covered by random forest and stepwise logistic model is maximum approximately similar

Bias in the Confusion Matrices

When we analyse the confusion matrix for all the four models.
Random forest and boosted model are good at predicting creditworthiness.
But boosted model is not good in predicting non credit worthiness.

Random and logistic model are better than other two model in predicting non creditworthiness.
All 4 model are biased in predicting creditworthiness more efficiently than predicting non creditworthiness.

Conclusion

When we compare the 4 models
We can see that overall accuracy of random forest model is highest at 82%
With accuracy for creditworthy approx. 98% and accuracy for non creditworthy 44%

Next best model seems to be logistic model with prediction accuracy for creditworthy and non creditworthy around 87% & 49% respectively but the overall accuracy of the model is low around 76%

Hence, I would choose random forest model

- How many individuals are creditworthy?

With Forest model, I can conclude that out of 500 customers, 410 are creditworthy and 90 are non 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.