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?
  The best object model to identify a customer is a creditworthy for a loan or not.

- What data is needed to inform those decisions?
  1. The history data of the loans & customer information and the outcome
  2. the new data with similar fields without the results
     - 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
     - Occupation

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

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.

  These variables I removed:
  Concurrent Credits has one value
  Guarantors 91% of the data are None
  Foreign Worker 96% of the data are 1
  No of Dependents 85% of them have 1 dependent
  Telephone is not relevant
  Duration in Current Address because 69% of the data missing

  2% in **Age Years** have missing data so will replace the Nulls with the median because the data is skewed to the left.

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.

- 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)

Fit and error measures

Model	Accuracy	F1	AUC	Accuracy_Creditworthy	Accuracy_Non-Creditworthy
FM	0.8200	0.8841	0.7455	0.9810	0.4444
DT	0.6733	0.7721	0.6296	0.7905	0.4000
BM	0.7800	0.8584	0.7524	0.9524	0.3778
sw	0.7600	0.8364	0.7306	0.8762	0.4889

a) Stepwise model (SW)

Setting the Credit Application Result as the target variable. I used the Step Wise after the Logistic Regression. So the most significant factors are Account Balance, Purpose, Payment Status of Previous Credit, Length of current employment and Instalment per cent. These variables have p-vales less than 0.05 with R-Squared 0.2048.
Overall accuracy is 76%.
The model is biased because the different in accuracy for Creditworthy and Non-Creditworthy is more than 10%.

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	***
Credit.Amount	0.0001704	5.733e-05	2.9716	0.00296	**
Instalment.per.cent	0.3016731	1.350e-01	2.2340	0.02549	*
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	*
Most.valuable.available.asset	0.2650267	1.425e-01	1.8599	0.06289	.
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	.

Significance codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
(Dispersion parameter for binomial taken to be 1)

Null deviance: 413.16 on 349 degrees of freedom
Residual deviance: 328.55 on 338 degrees of freedom
McFadden R-Squared: 0.2048, Akaike Information Criterion 352.5

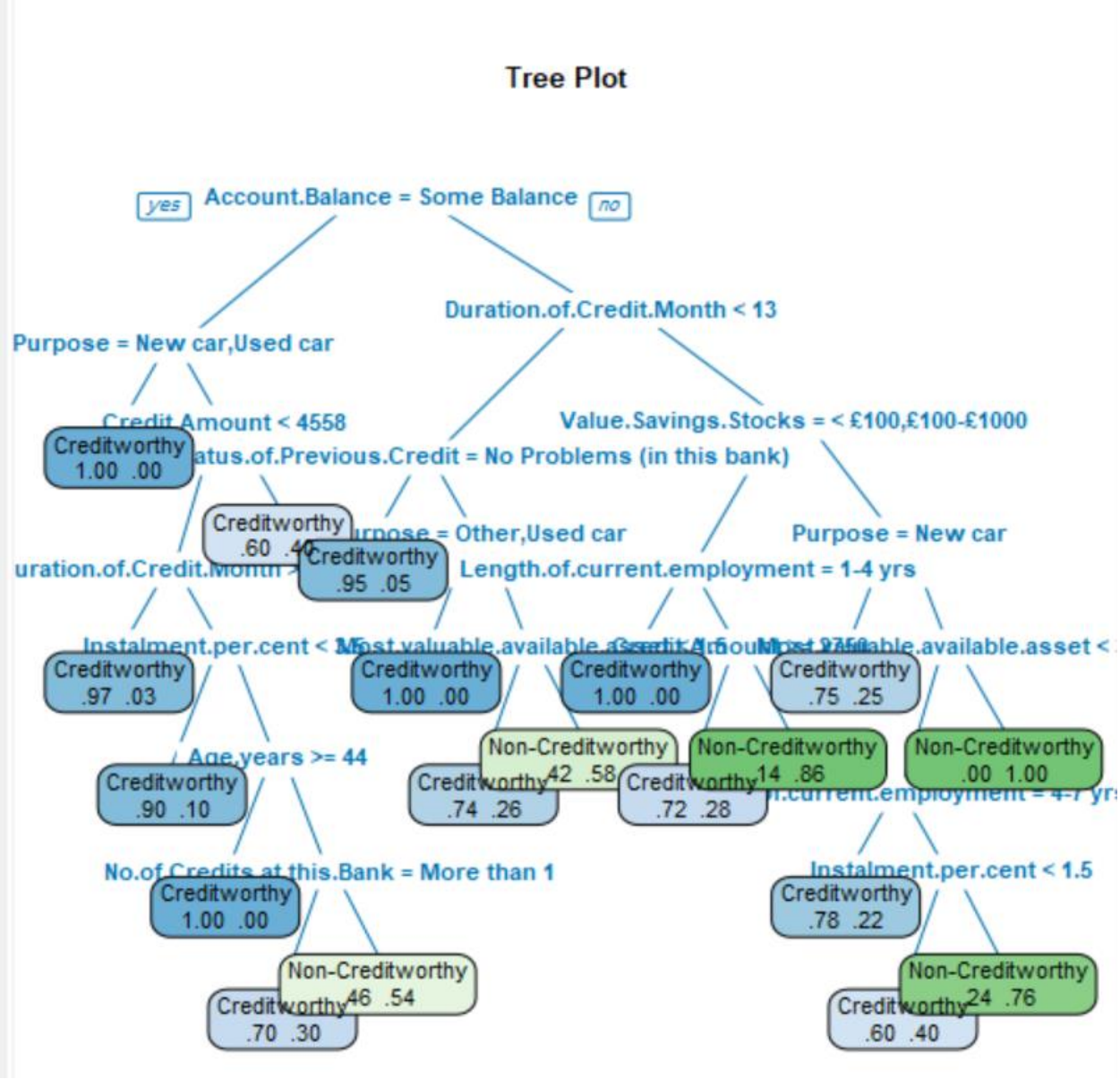
Decision Tree (DT)
Setting the Credit Application Result as the target variable. The most important variables

```
[1] Account.Balance Age.years
```

[3] Credit.Amount Duration.of.Credit.Month
[5] Instalment.per.cent Length.of.current.employment
[7] Most.valuable.available.asset No.of.Credits.at.this.Bank
[9] Payment.Status.of.Previous.Credit Purpose
[11] Value.Savings.Stocks

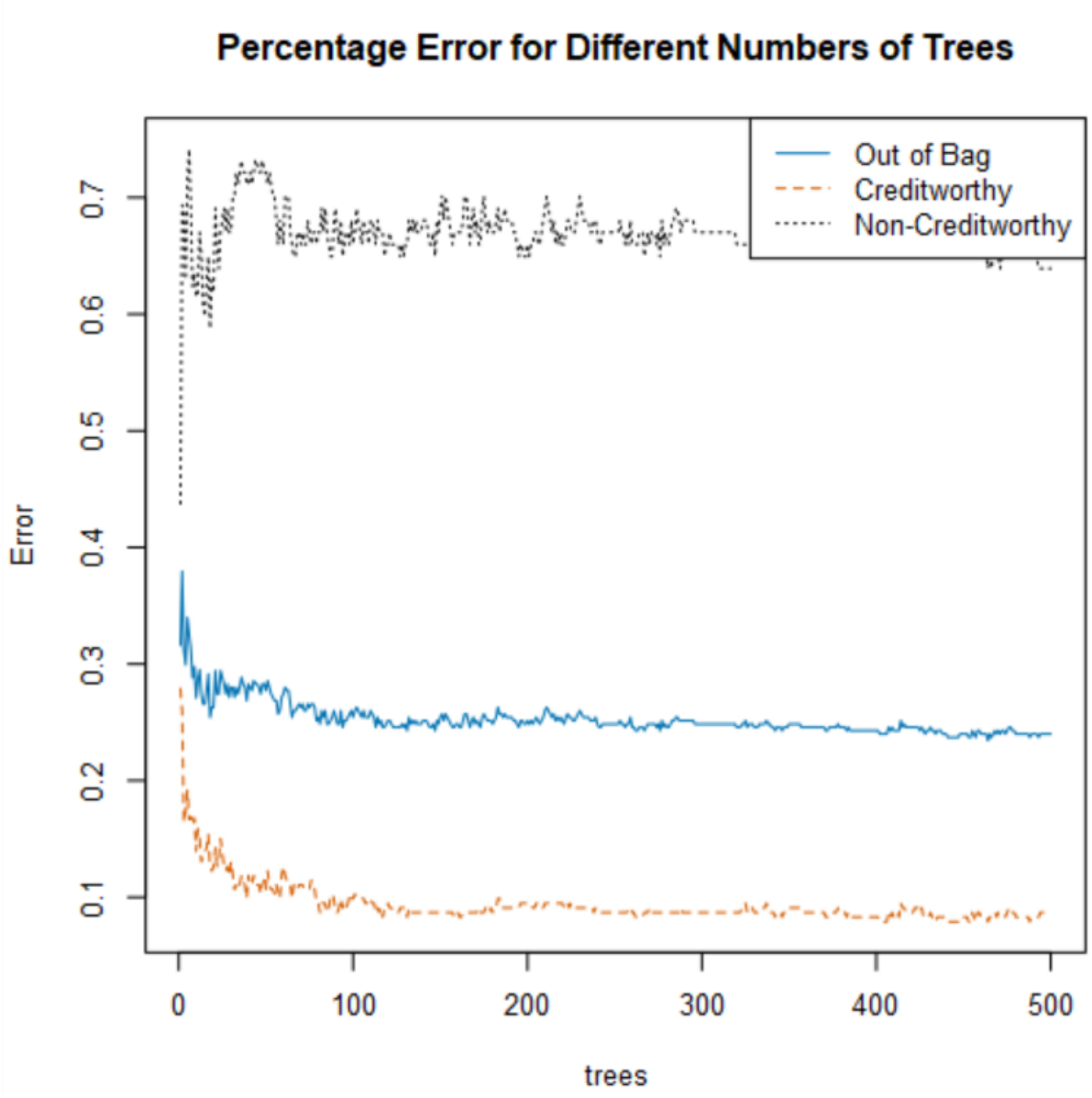
.

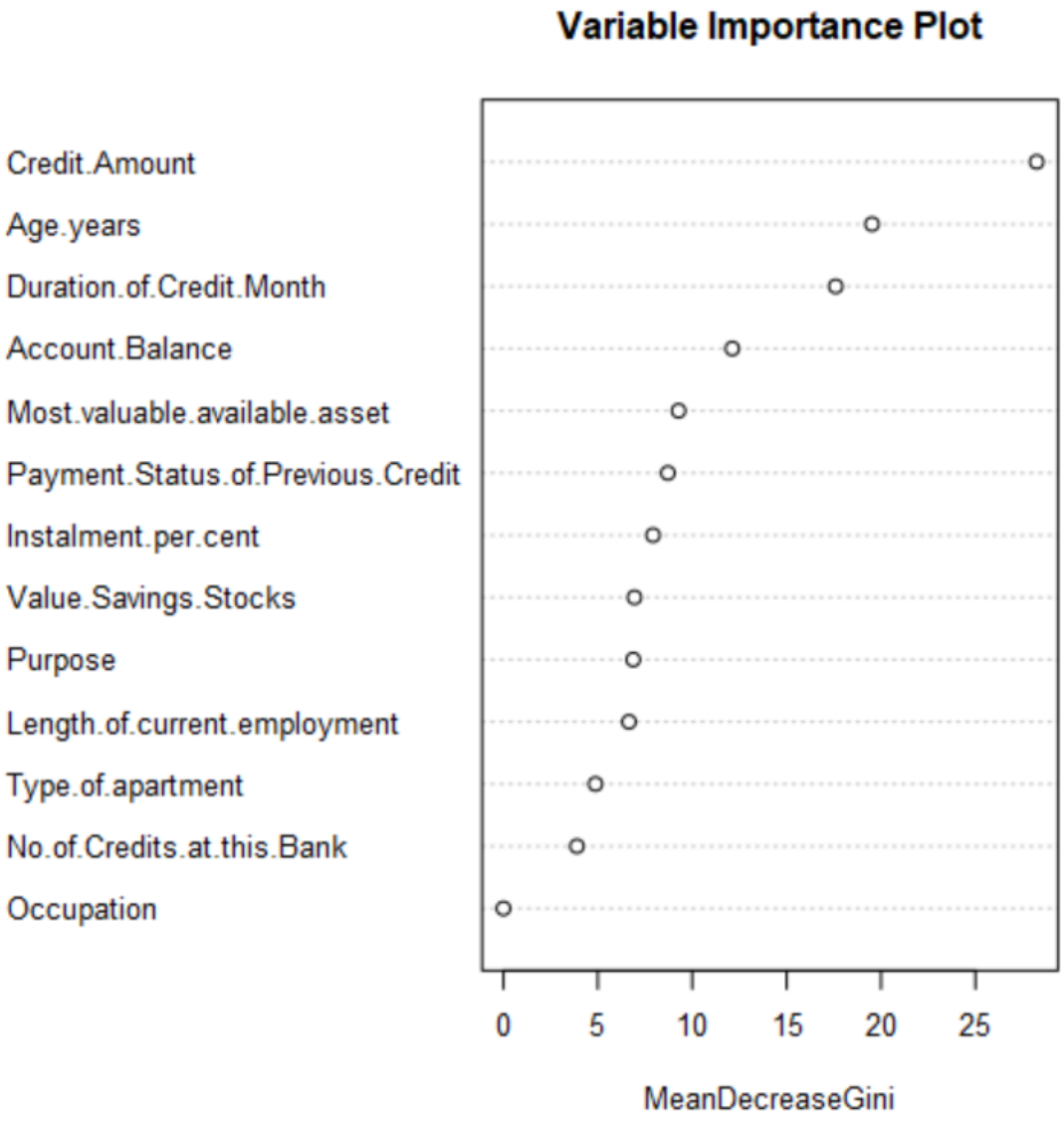
Total accuracy is 67%. The model is biased because the different in accuracy for Creditworthy and Non-Creditworthy is more than 10%.



Forest Model (FM)

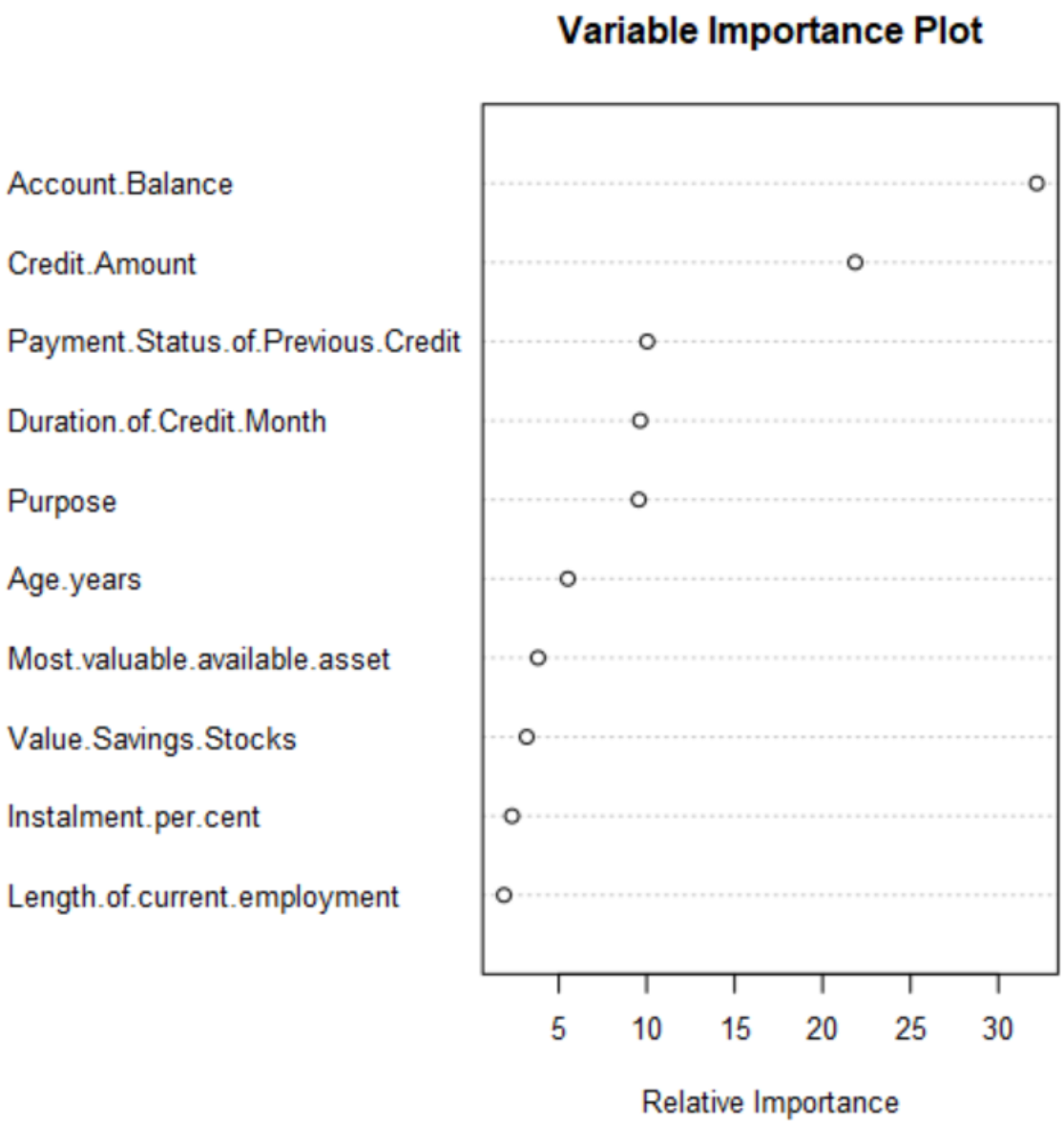
Setting the Credit Application Result as the target variable. The top 3 most important variables are Credit Amount, Age Years and Duration of Credit Month. The total accuracy is 82%. The model is not biased because the different in accuracy for Creditworthy and Non-Creditworthy is less than 10%.


Percentage Error for Different Numbers of Trees
Out of Bag
Creditworthy
Non-Creditworthy
Error
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0
100
200
300
400
500
trees



d) Boosted Model (BM)

Setting the Credit Application Result as the target variable. The top 3 most important variables are Account Balance, Credit Amount and Payment Statues of previous credit. The total accuracy is 78%. The model is not biased because the different in accuracy for Creditworthy and Non-Creditworthy is less than 10%.



Confusion matrix of BM

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

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:

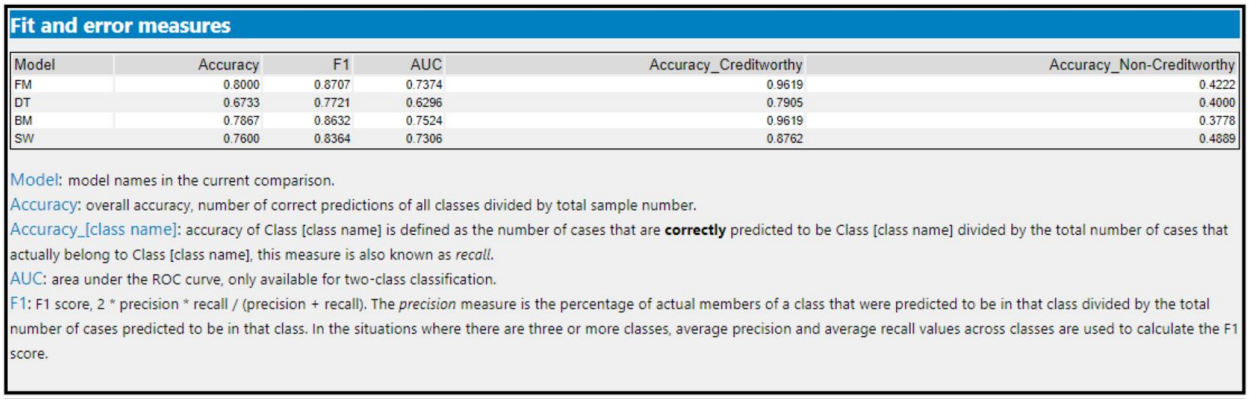
- Which model did you choose to use? Please justify your decision using **all** of the following techniques. Please only use these techniques to justify your decision:
    - Overall Accuracy against your Validation set
    - Accuracies within "Creditworthy" and "Non-Creditworthy" segments
    - ROC graph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

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

- How many individuals are creditworthy?

I choosed the FM, because it has the highest in total accuracy, reaches the true positive rate at faster rate than the other models and it has high accuracy in both Creditworthy and Non-Creditworthy. The accuracy difference between Creditworthy and Non-Creditworthy are comparable which makes it least bias towards any decisions

There are 412 Creditworthy customers using forest models to score new customers.



Fit and error measures

Model	Accuracy	F1	AUC	Accuracy_Creditworthy	Accuracy_Non-Creditworthy
FM	0.8000	0.8707	0.7374	0.9619	0.4222
DT	0.6733	0.7721	0.6296	0.7905	0.4000
BM	0.7867	0.8632	0.7524	0.9619	0.3778
SW	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 the 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 BM

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

Confusion matrix of DT

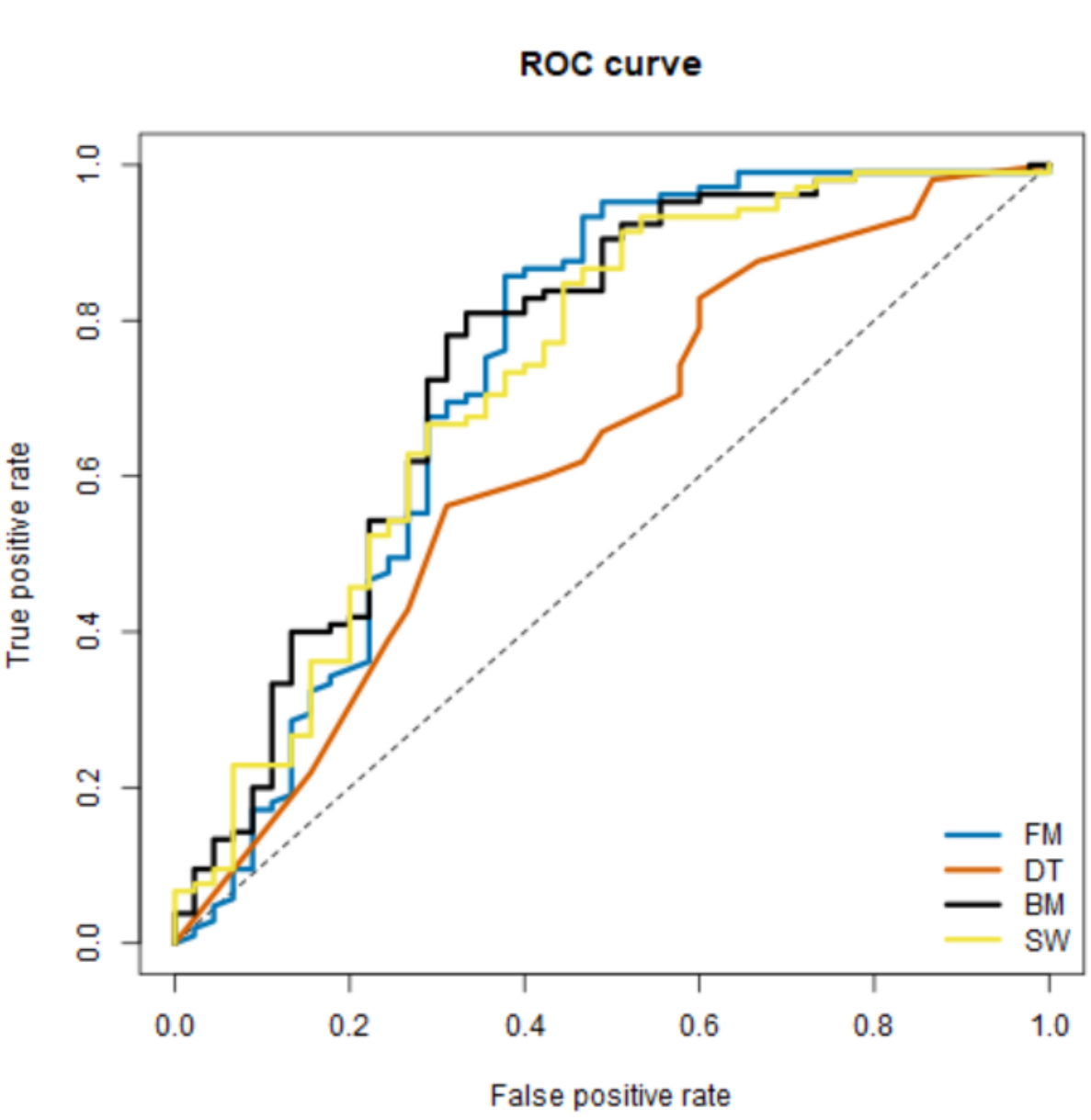
	Actual_Creditworthy	Actual_Non-Creditworthy
Predicted_Creditworthy	83	27
Predicted_Non-Creditworthy	22	18

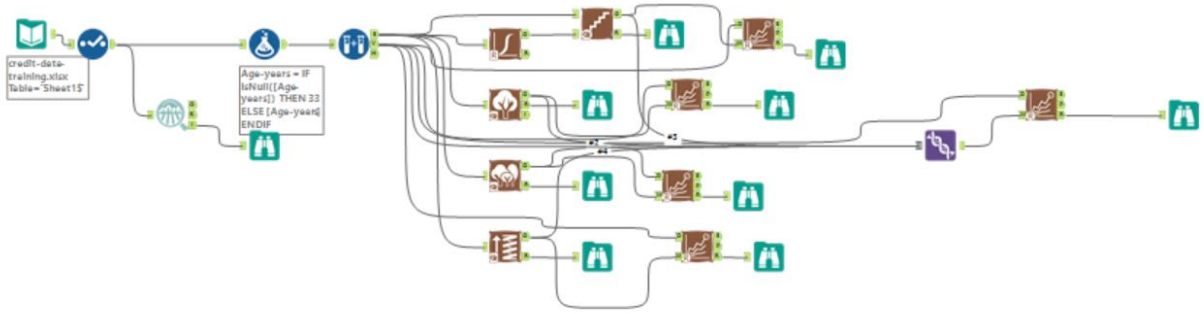
Confusion matrix of FM

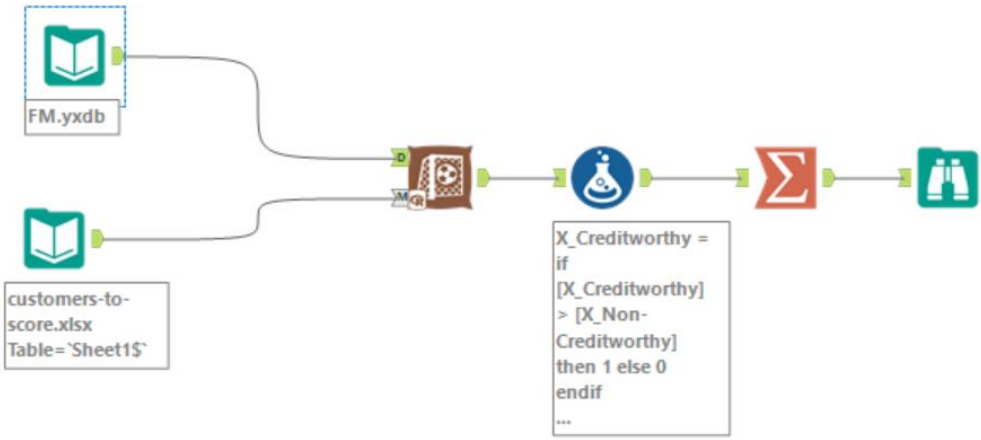
	Actual_Creditworthy	Actual_Non-Creditworthy
Predicted_Creditworthy	101	26
Predicted_Non-Creditworthy	4	19

Confusion matrix of SW

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


ROC curve
True positive rate
False positive rate
0.0
0.2
0.4
0.6
0.8
1.0
FM
DT
BM
SW


credit-data-training.xlsx Table= 'Sheet1$'
Age-years = IF IsNull([Age-years]) THEN 33 ELSE [Age-years] ENDIF



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.