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

----------All my answers are highlighted in yellow and non-italics----------------

Provide an explanation of the key decisions that need to be made. (250 word limit)

You work for a small bank and are responsible for determining if customers are creditworthy to give a loan to. Your team typically gets 200 loan applications per week and approves them by hand.

Due to a financial scandal that hit a competitive bank last week, you suddenly have an influx of new people applying for loans for your bank instead of the other bank in your city. All of a sudden you have nearly 500 loan applications to process this week!

Your manager sees this new influx as a great opportunity and wants you to figure out how to process all of these loan applications within one week.

Fortunately for you, you just completed a course in classification modeling and know how to systematically evaluate the creditworthiness of these new loan applicants.

For this project, you will analyze the business problem using the Problem Solving Framework and provide a list of creditworthy customers to your manager in the next two days.

You have the following information to work with:

1. Data on all past applications
2. The list of customers that need to be processed in the next few days

Key Decisions:

Answer these questions

- *What decisions needs to be made?*

  We need to decide whether the new customers are creditworthy, based on the data provided

- *What data is needed to inform those decisions?*

  - 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
- Guarantors
- Duration-in-Current-address
- Most-valuable-available-asset
- Age-years
- Concurrent-Credits
- Type-of-apartment
- No-of-Credits-at-this-Bank
- Occupation
- No-of-dependents
- Telephone
- Foreign-Worker

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

  Since we are determining is a customer is creditworthy or not, a binary model is needed.

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:

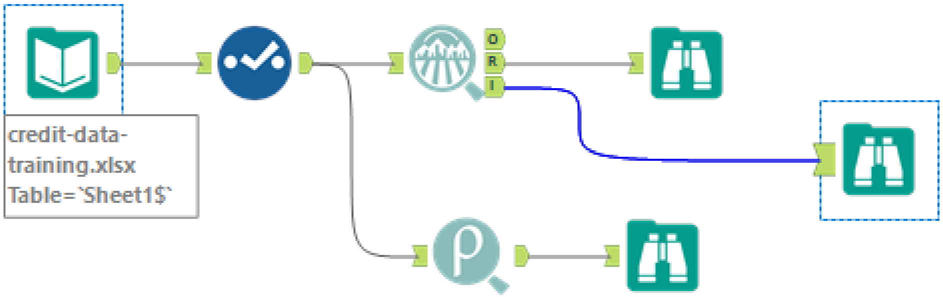
- *For numerical data fields, are there any fields that highly-correlate with each other? The correlation should be at least .70 to be considered "high".*
- *Are there any missing data for each of the data fields? Fields with a lot of missing data should be removed*
- *Are there only a few values in 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)*

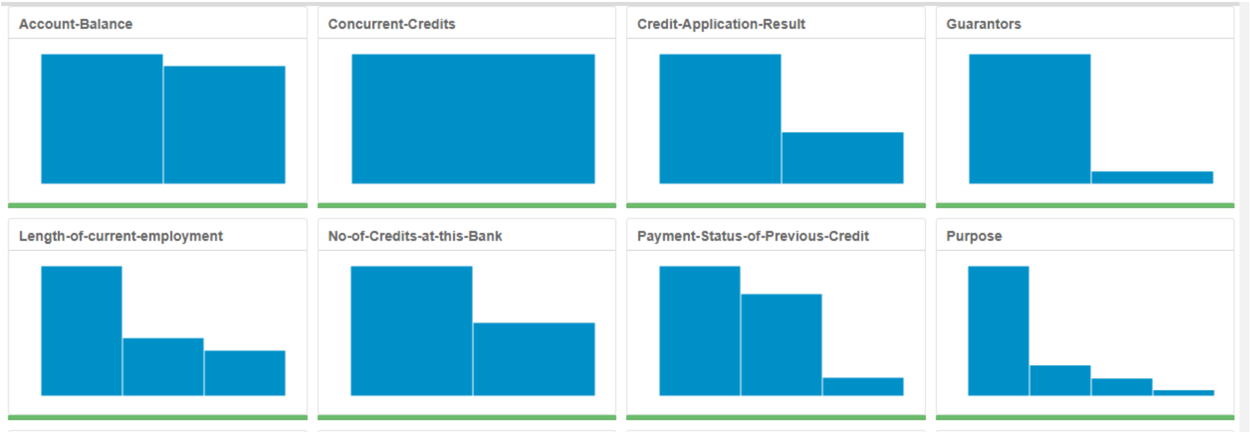
Note: *For the sake of consistency in the data cleanup process, impute data using the median of the entire data field instead of removing a few data points. (100 word limit)*

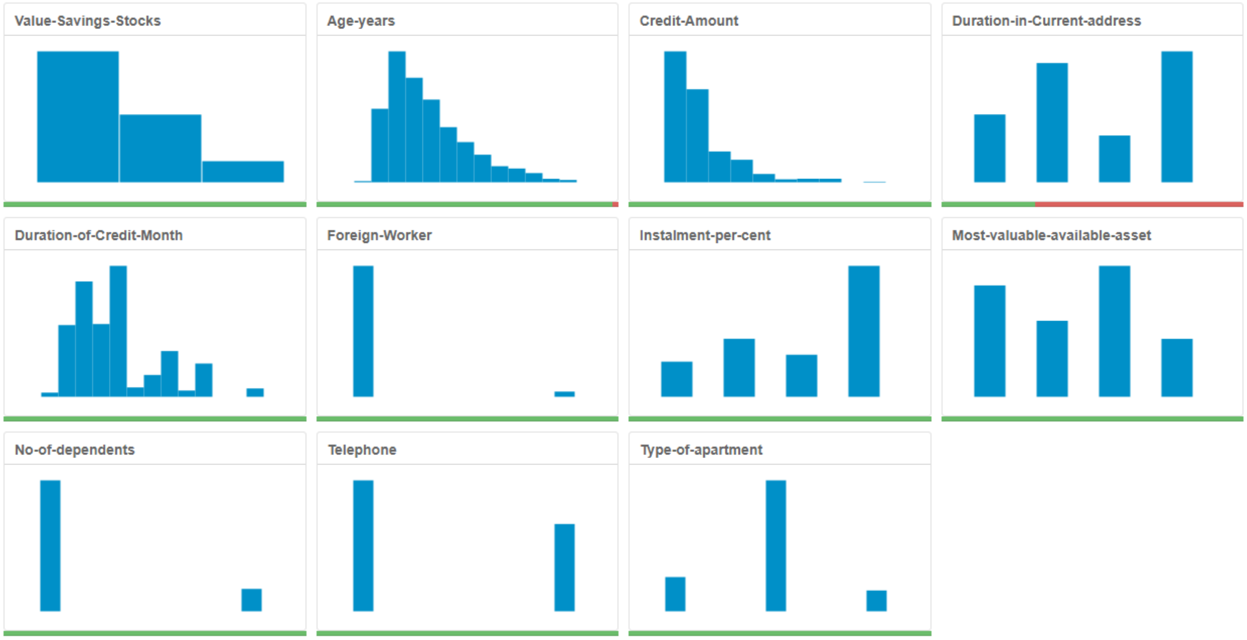
To achieve consistent results reviewers expect.

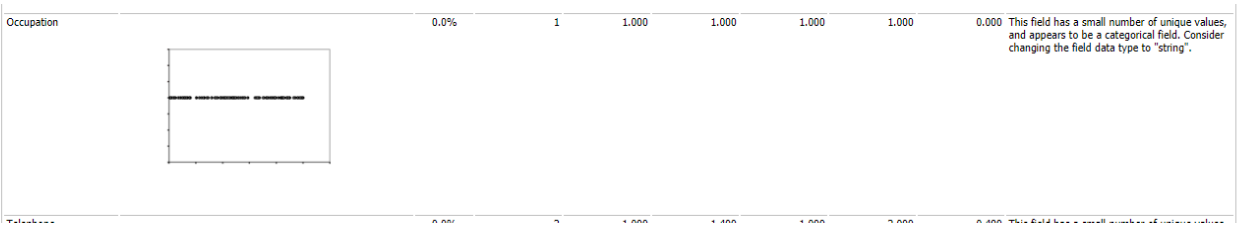
Answer this question:

- *In your cleanup process, which fields did you remove or impute? Please justify why you removed or imputed these fields. Visualizations are encouraged.*









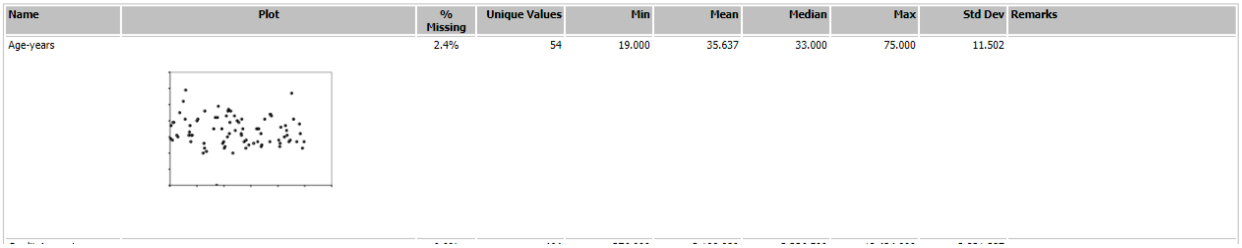
Occupation		0.0%	1	1.000	1.000	1.000	1.000	0.000	This field has a small number of unique values, and appears to be a categorical field. Consider changing the field data type to "string".

©

Based on the above I am removing 7 fields:

1. Concurrent-Credits…low variability and only one value
2. Occupation… low variability and only on value
3. Duration-in-current-address…Missing data
4. Guarantors…low variability
5. Foreign Worker…low variability
6. No-of-dependents…low variability
7. Telephone…low variability and not relevant

Age-Years..missing values..median is 33, imputed 33 in all missing values



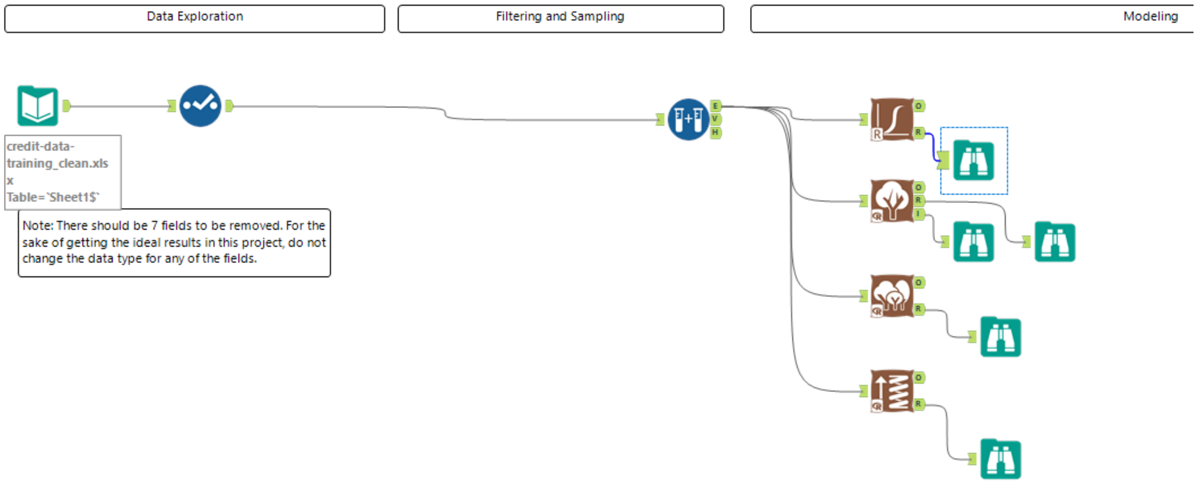
Name	Plot	% Missing	Unique Values	Min	Mean	Median	Max	Std Dev	Remarks
Age-years		2.4%	54	19.000	35.637	33.000	75.000	11.502	

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

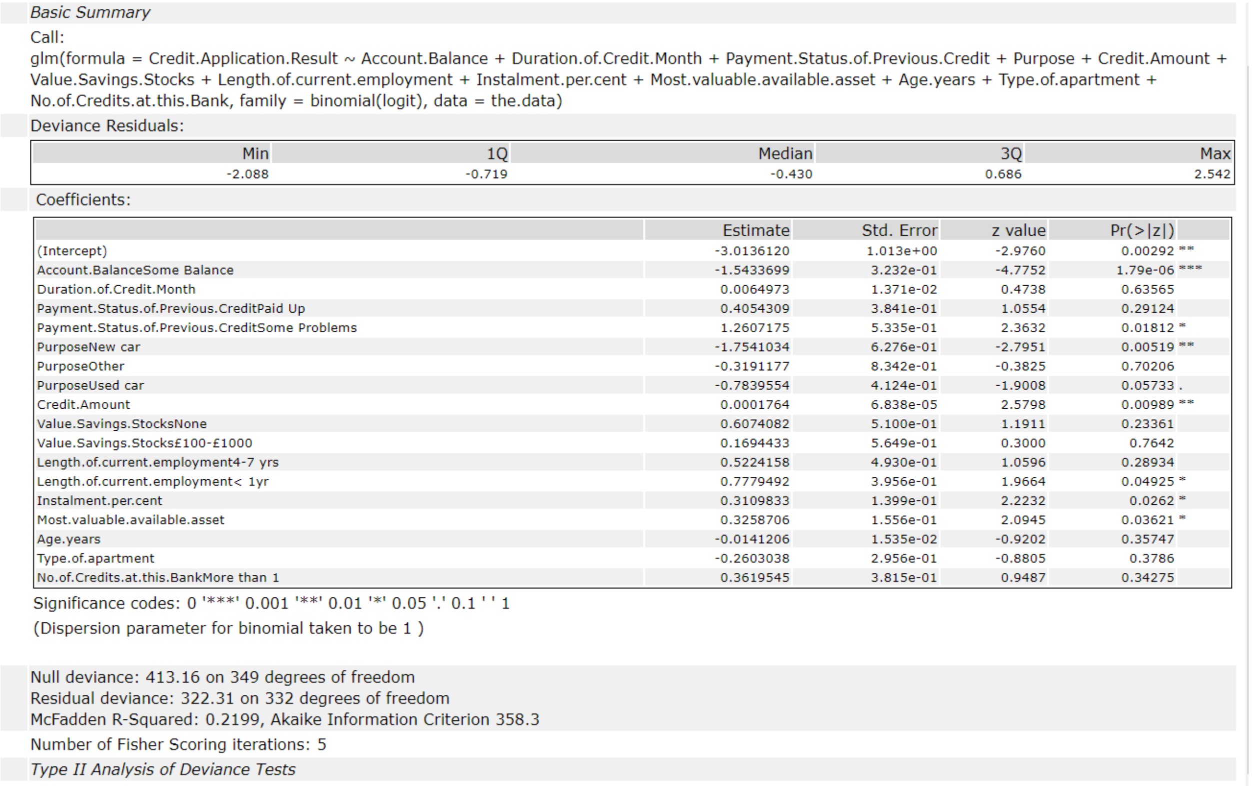
Answer these questions for ***each model*** *you created:*



credit-data-training_clean.xlsx Table=`Sheet1$`

Note: There should be 7 fields to be removed. For the sake of getting the ideal results in this project, do not change the data type for any of the fields.

Regression model:



Basic Summary

Call:
glm(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, family = binomial(logit), data = the.data)

Deviance Residuals:

Min	1Q	Median	3Q	Max
-2.088	-0.719	-0.430	0.686	2.542

Coefficients:

	Estimate	Std. Error	z value	Pr(>|z|)	
(Intercept)	-3.0136120	1.013e+00	-2.9760	0.00292	**
Account.BalanceSome Balance	-1.5433699	3.232e-01	-4.7752	1.79e-06	***
Duration.of.Credit.Month	0.0064973	1.371e-02	0.4738	0.63565	
Payment.Status.of.Previous.CreditPaid Up	0.4054309	3.841e-01	1.0554	0.29124	
Payment.Status.of.Previous.CreditSome Problems	1.2607175	5.335e-01	2.3632	0.01812	*
PurposeNew car	-1.7541034	6.276e-01	-2.7951	0.00519	**
PurposeOther	-0.3191177	8.342e-01	-0.3825	0.70206	
PurposeUsed car	-0.7839554	4.124e-01	-1.9008	0.05733	.
Credit.Amount	0.0001764	6.838e-05	2.5798	0.00989	**
Value.Savings.StocksNone	0.6074082	5.100e-01	1.1911	0.23361	
Value.Savings.Stocks£100-£1000	0.1694433	5.649e-01	0.3000	0.7642	
Length.of.current.employment4-7 yrs	0.5224158	4.930e-01	1.0596	0.28934	
Length.of.current.employment< 1yr	0.7779492	3.956e-01	1.9664	0.04925	*
Instalment.per.cent	0.3109833	1.399e-01	2.2232	0.0262	*
Most.valuable.available.asset	0.3258706	1.556e-01	2.0945	0.03621	*
Age.years	-0.0141206	1.535e-02	-0.9202	0.35747	
Type.of.apartment	-0.2603038	2.956e-01	-0.8805	0.3786	
No.of.Credits.at.this.BankMore than 1	0.3619545	3.815e-01	0.9487	0.34275	

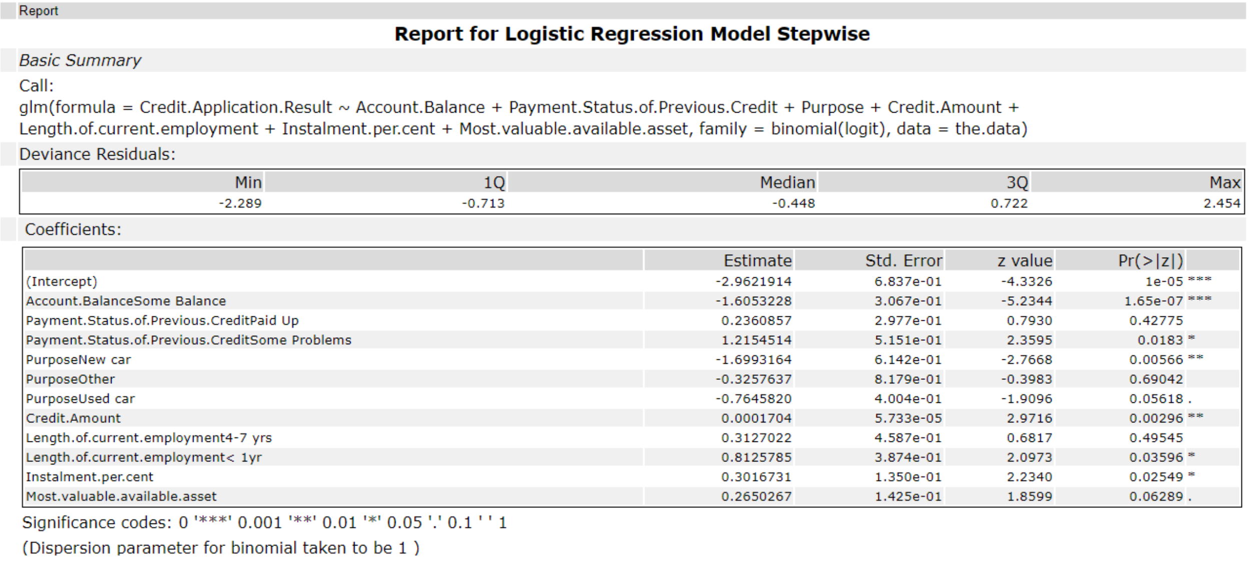
Significance codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial taken to be 1)

Null deviance: 413.16 on 349 degrees of freedom
Residual deviance: 322.31 on 332 degrees of freedom
McFadden R-Squared: 0.2199, Akaike Information Criterion 358.3

Number of Fisher Scoring iterations: 5

Type II Analysis of Deviance Tests



Report

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

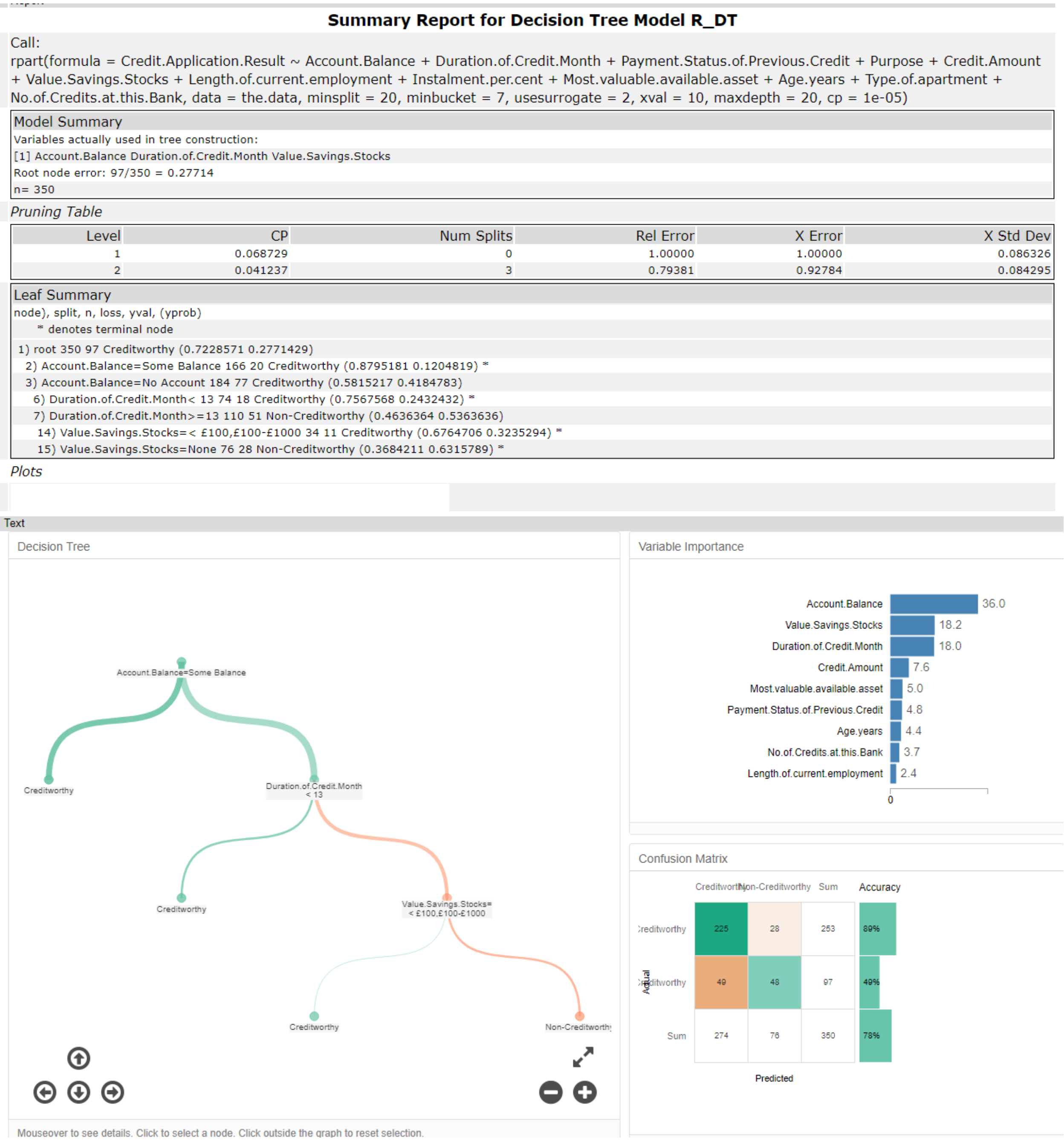
Coefficients:

	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)

Decision Tree model:



Summary Report for Decision Tree Model R_DT

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 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	0.92784	0.084295

Leaf Summary

```
node), split, n, loss, yval, (yprob)
      * denotes terminal node
 1) root 350 97 Creditworthy (0.7228571 0.2771429)
   2) Account.Balance=Some Balance 166 20 Creditworthy (0.8795181 0.1204819) *
   3) Account.Balance=No Account 184 77 Creditworthy (0.5815217 0.4184783)
     6) Duration.of.Credit.Month< 13 74 18 Creditworthy (0.7567568 0.2432432) *
     7) Duration.of.Credit.Month>=13 110 51 Non-Creditworthy (0.4636364 0.5363636)
      14) Value.Savings.Stocks=< £100,£100-£1000 34 11 Creditworthy (0.6764706 0.3235294) *
      15) Value.Savings.Stocks=None 76 28 Non-Creditworthy (0.3684211 0.6315789) *
```

Plots

Text

Decision Tree

Mouseover to see details. Click to select a node. Click outside the graph to reset selection.

Variable Importance

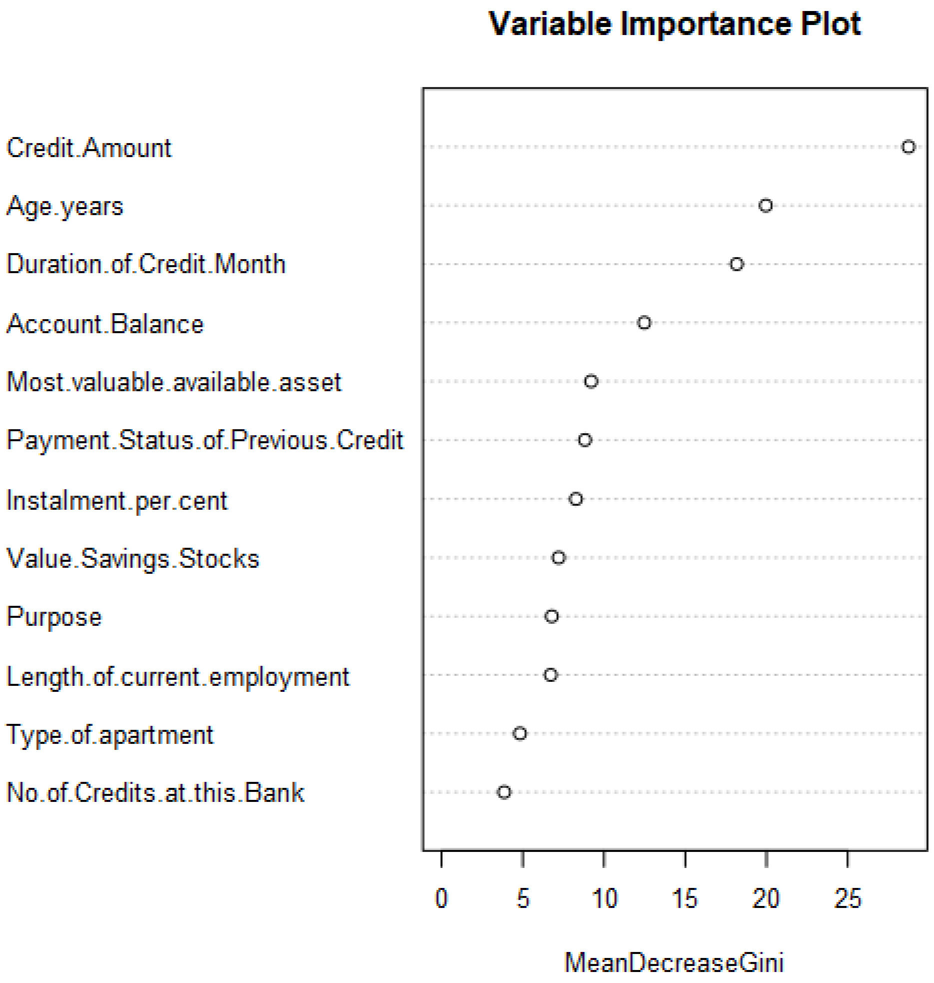
Account.Balance 36.0
Value.Savings.Stocks 18.2
Duration.of.Credit.Month 18.0
Credit.Amount 7.6
Most.valuable.available.asset 5.0
Payment.Status.of.Previous.Credit 4.8
Age.years 4.4
No.of.Credits.at.this.Bank 3.7
Length.of.current.employment 2.4
0

Confusion Matrix

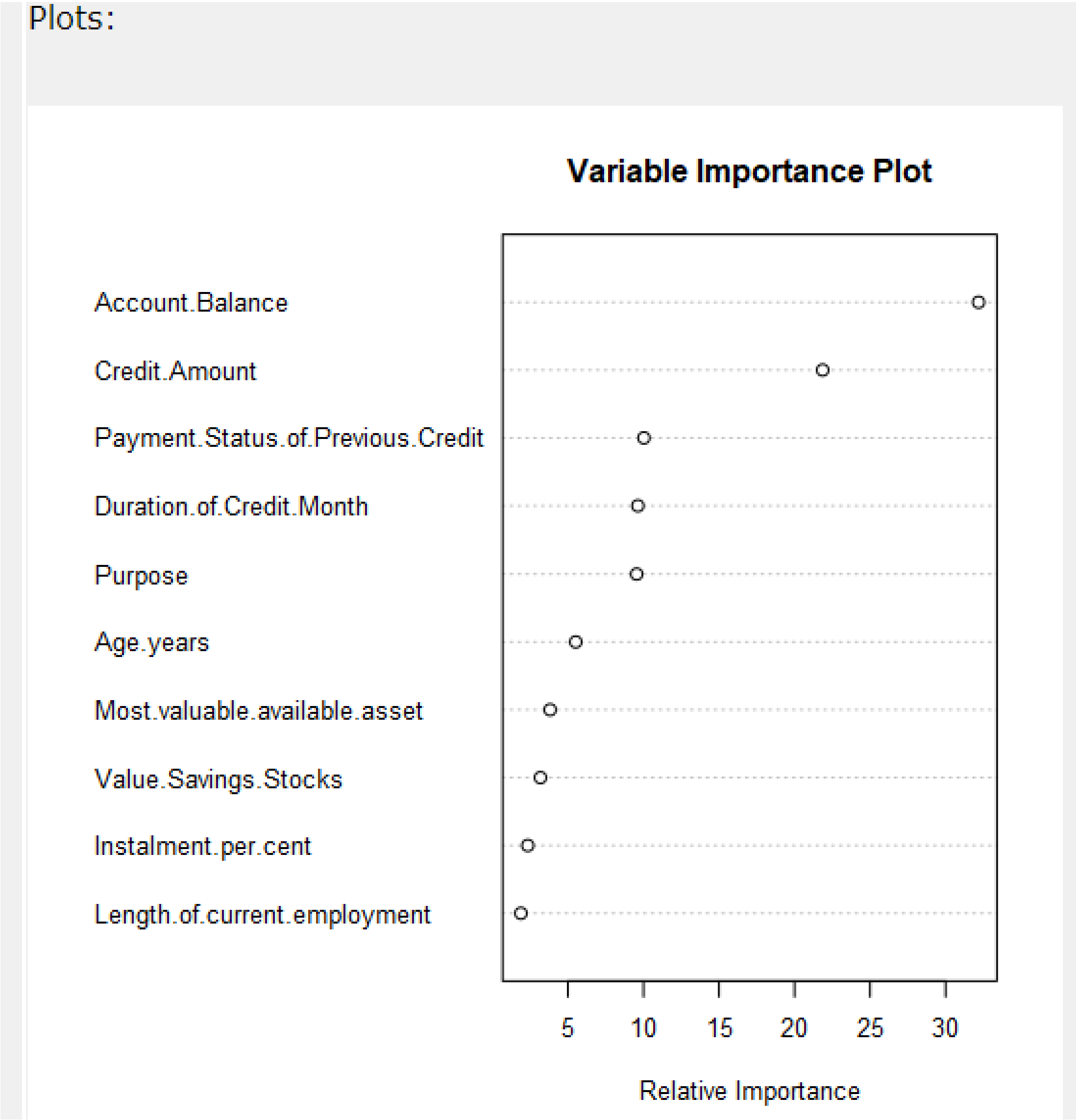
Actual	Creditworthy	Non-Creditworthy	Sum	Accuracy
reditworthy	225	28	253	89%
reditworthy	49	48	97	49%
Sum	274	76	350	78%

Predicted

Forest model:



Boosted model:



Plots:

- *Which predictor variables are significant or the most important? Please show the p-values or variable importance charts for all of your predictor variables.*

Summary of significant variables for each model: Most important on top, Ascending thereafter.

Logistics Regression:

- Account.BalanceSome Balance
- Payment.Status.of.Previous.CreditPaid Up
- PurposeNew car
- Credit.Amount
- Length.of.current.employment< 1yr
- Installment.per.cent

Decision Tree:

- Account.Balance=Some Balance 166 20 Creditworthy
- Duration.of.Credit.Month< 13 74 18 Creditworthy
- Value.Savings.Stocks=< £100,£100-£1000 34 11 Creditworthy

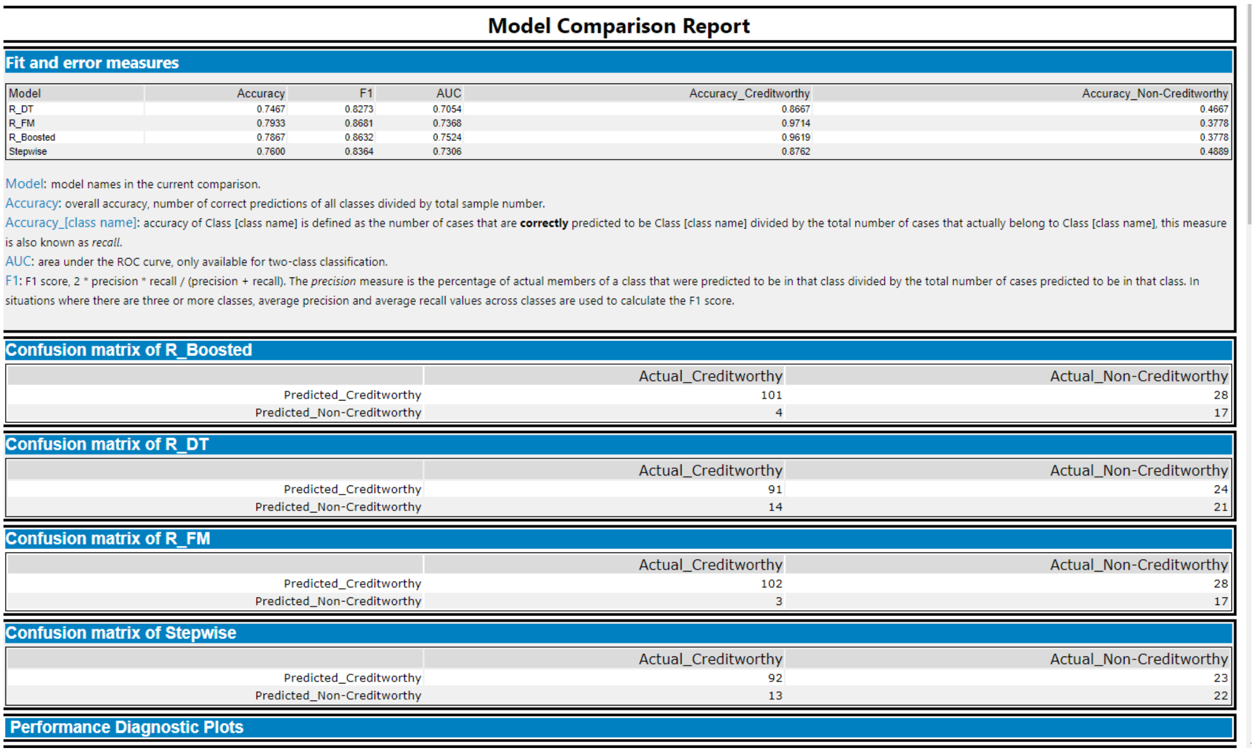
Forest Model:

- Credit.Amount
- Age.years
- Duration.of.Credit.Month
- Account.Balance

Boosted Model:

- Account.Balance
- Credit.Amount
- Payment.Status.of.Previous.Credit
- Duration.of.Credit.Month

- *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?*



Model Comparison Report

Fit and error measures

Model	Accuracy	F1	AUC	Accuracy_Creditworthy	Accuracy_Non-Creditworthy
R_DT	0.7467	0.8273	0.7054	0.8667	0.4667
R_FM	0.7933	0.8681	0.7368	0.9714	0.3778
R_Boosted	0.7867	0.8632	0.7524	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 R_Boosted

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

Confusion matrix of R_DT

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

Confusion matrix of R_FM

	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

Performance Diagnostic Plots

- For Decision Tree model overall accuracy is 75%, credit worthiness is good @ 87% but non credit worthiness is 47%
- For Forest model overall accuracy is 79%, credit worthiness is great @ 97% but non credit worthiness is only 38%
- For Boosted model overall accuracy is 79%, credit worthiness is great @ 96% but non credit worthiness is only 38%
- For Stepwise (logistics regression) model overall accuracy is 76%, credit worthiness is good 87% but non credit worthiness is 49%

  There is a bias due to less sample for non-credit worthy customers.

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*
    - ○ *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.*

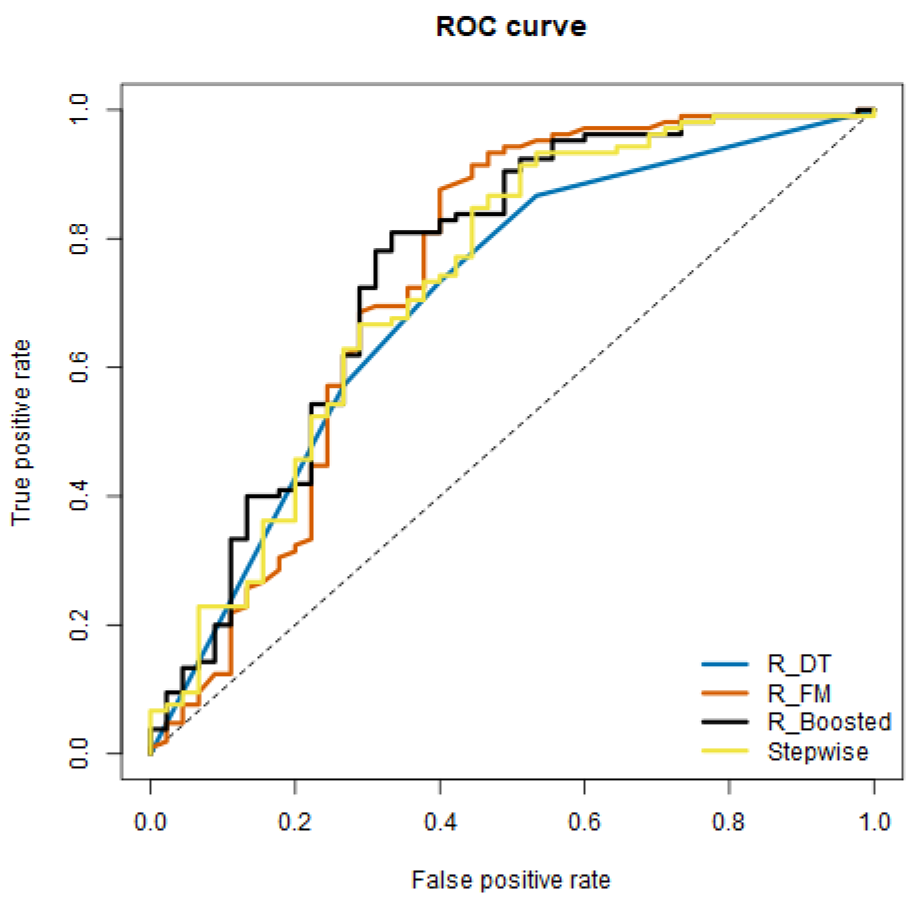
○ *Overall Accuracy against your Validation set*

The Forest model has the highest overall accuracy of 79.33%

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

The Forest tree model has the highest creditworthy accuracy @ 97.14%, it does not however have the highest non creditworthy accuracy. It is 37.78% however it's the best model overall.

○ *ROC graph*

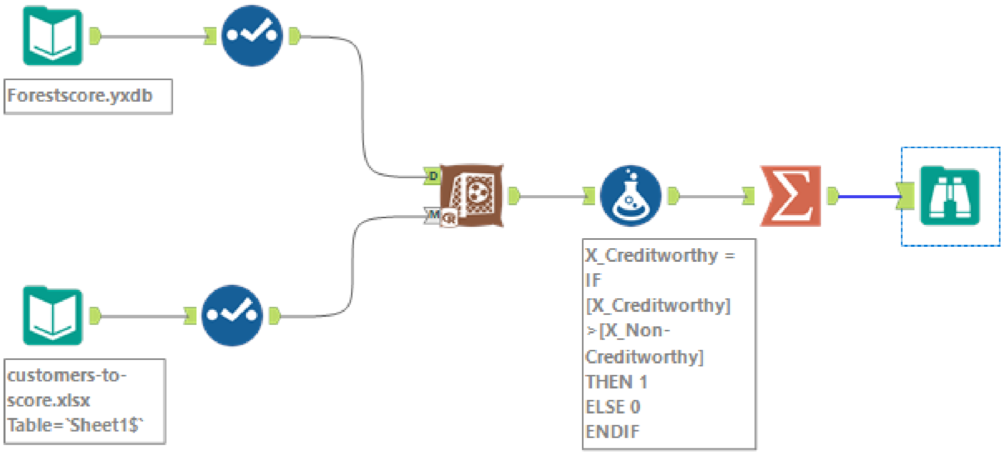


According to the ROC curve, and using the rule that the best model is the one that reaches the top first, the forest model is the best model.

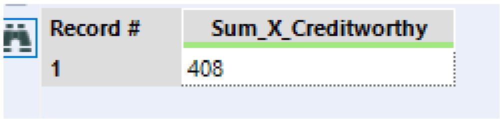
- *Bias in the Confusion Matrices*

On the confusion matrix above, we see that Forest & Boosted models have more non creditworthy customers, compared to Stepwise and decision tree has less.
Forest model still has the highest accuracy overall.

- *How many individuals are creditworthy?*



```
X_Creditworthy =
IF
[X_Creditworthy]
>[X_Non-
Creditworthy]
THEN 1
ELSE 0
ENDIF
```



Record #	Sum_X_Creditworthy
1	408

Using the score tool, 408 customers out of 500 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.