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?
  We need to identify the creditworthy customers to give loan to. Due to a financial crisis, there has been an increase in the influx of customers asking for loan.

- What data is needed to inform those decisions?
  We need the following information for existing customers and the new customers whose creditworthy we are trying to find out. With the existing customer, we would identify the criteria that should be looked at to predict the creditworthiness.

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?
  We need binary model since we are predicting 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

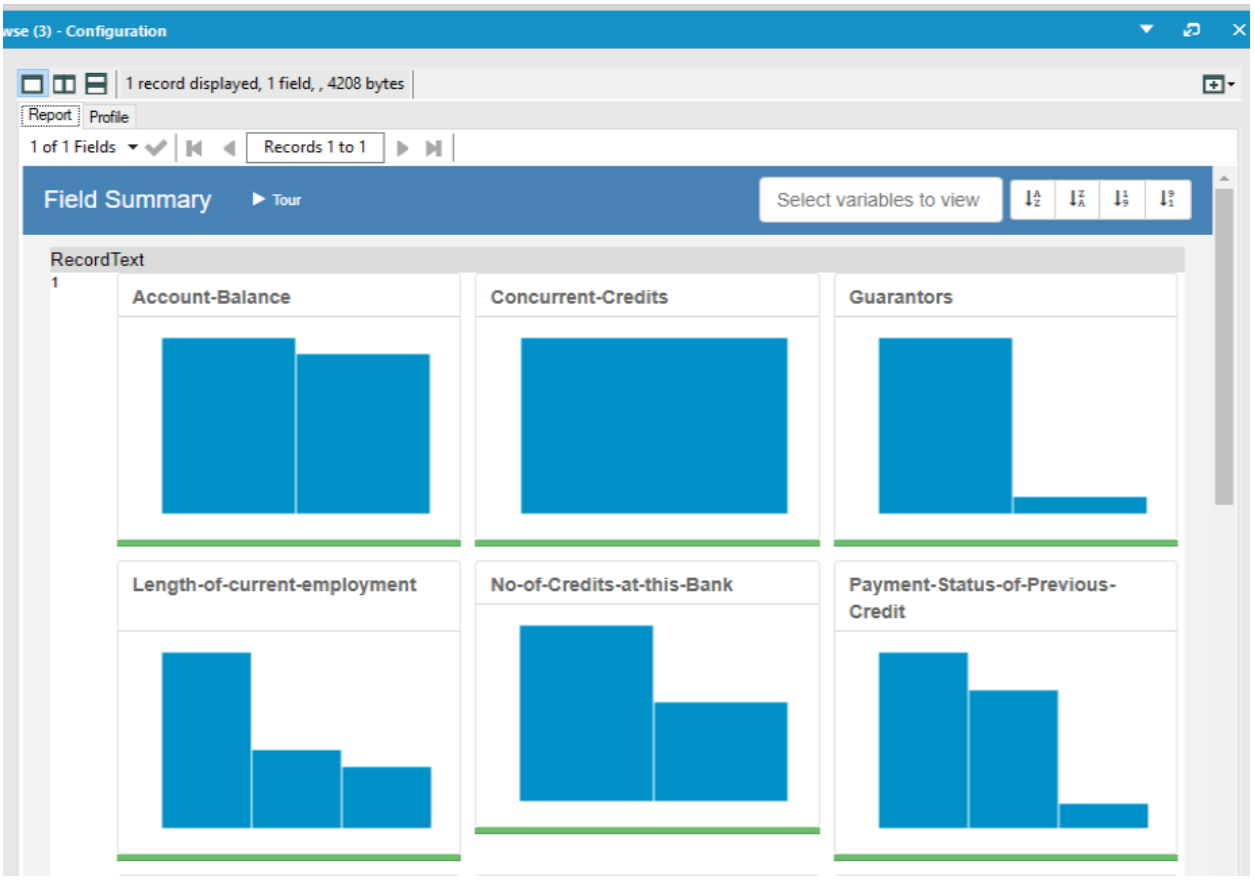
Duration-of-Credit-Month	Double
Payment-Status-of-Previous-Credit	String
Purpose	String
Credit-Amount	Double
Value-Savings-Stocks	String
Length-of-current-employment	String
Instalment-per-cent	Double
Guarantors	String
Duration-in-Current-address	Double
Most-valuable-available-asset	Double
Age-years	Double
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

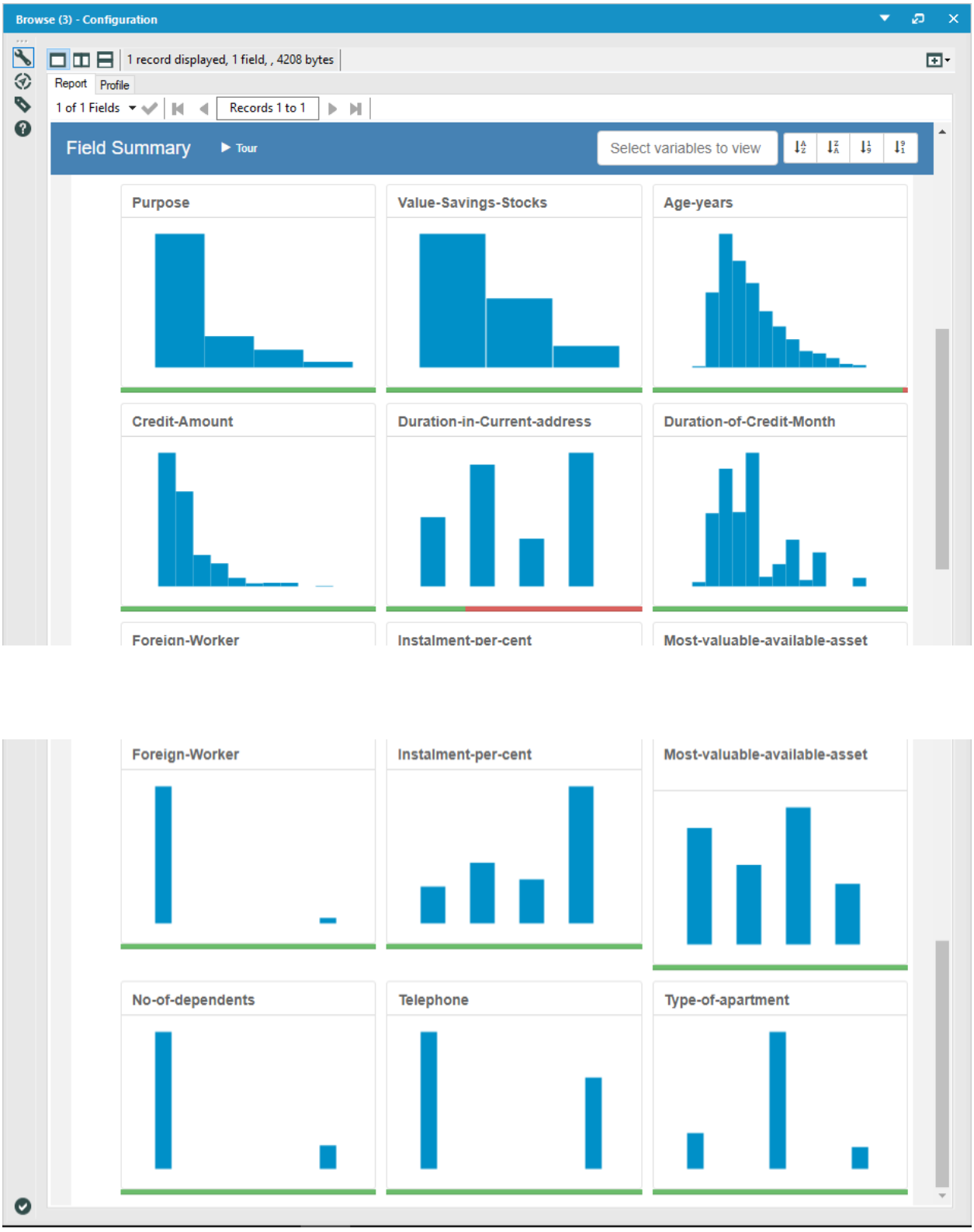
To achieve consistent results reviewers expect.

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.

Below are the analysis of all the fields:


wse (3) - Configuration
1 record displayed, 1 field, , 4208 bytes
Report
Profile
1 of 1 Fields
Records 1 to 1
Field Summary
Tour
Select variables to view
RecordText
1
Account-Balance
Concurrent-Credits
Guarantors
Length-of-current-employment
No-of-Credits-at-this-Bank
Payment-Status-of-Previous-Credit



Low variability variables were removed since they had only one value for the entire field.

1. Removed Concurrent-Credits as it had only one value.
2. Removed Number of dependents usually tended to be one value.
3. Removed Foreign workers had 2 unique value but data was distributed to one.
4. Removed telephone as that does not contribute towards figuring credit worthiness.
5. Removed guarantors as it had only one value.
6. Removed duration in current address as it had lots of missing data.
7. Removed Occupation as it had only one data.
8. Null values in Age-year was replaced by the median value.
9. Categorized account-balance, Value-Savings-Stocks, Length-of-current-employment into 3 distinct values in a new field.

Step 3: Train your Classification Models

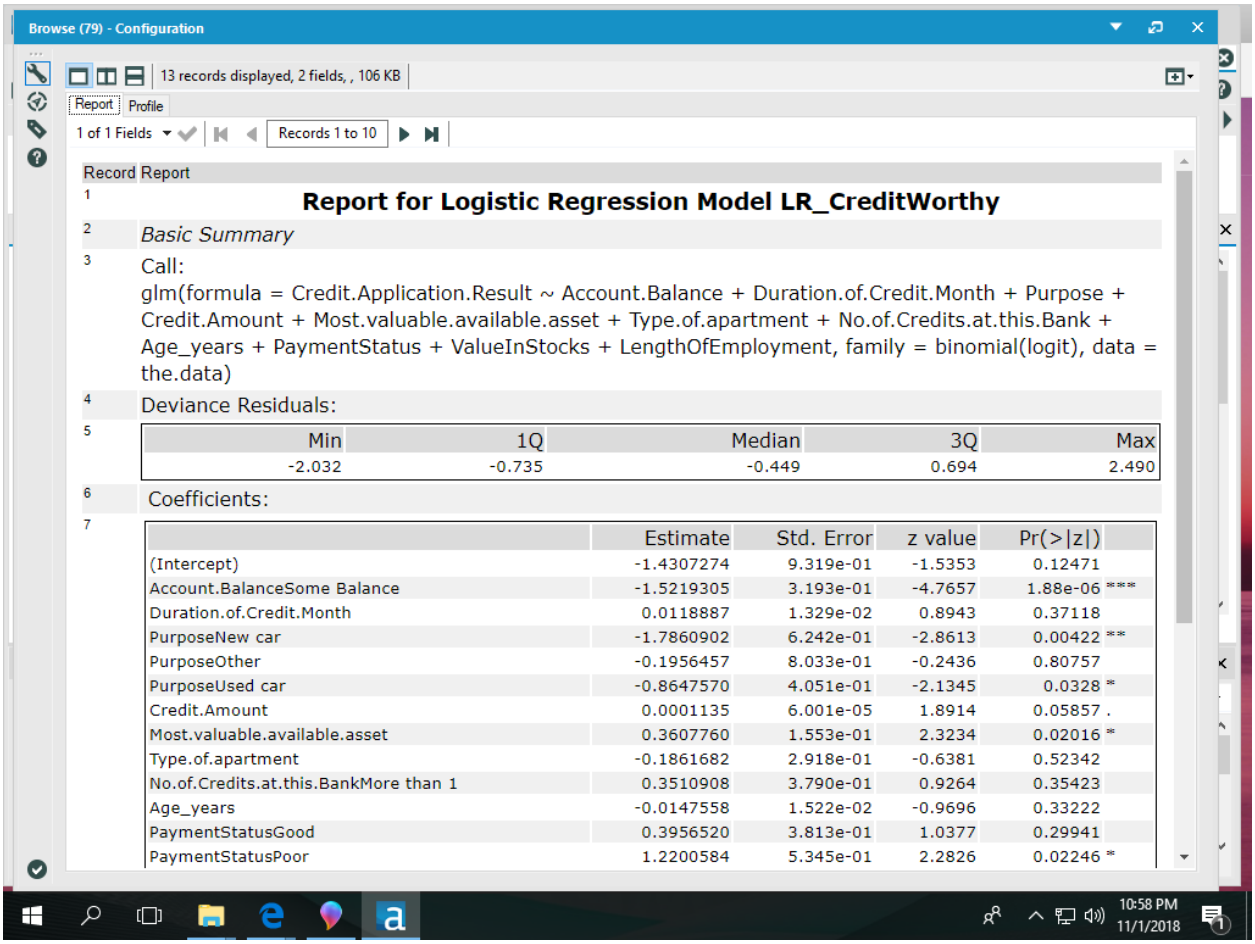
First, create your Estimation and Validation samples where 70% of your dataset should go to Estimation and 30% of your entire dataset should be reserved for Validation. Set the Random Seed to 1.

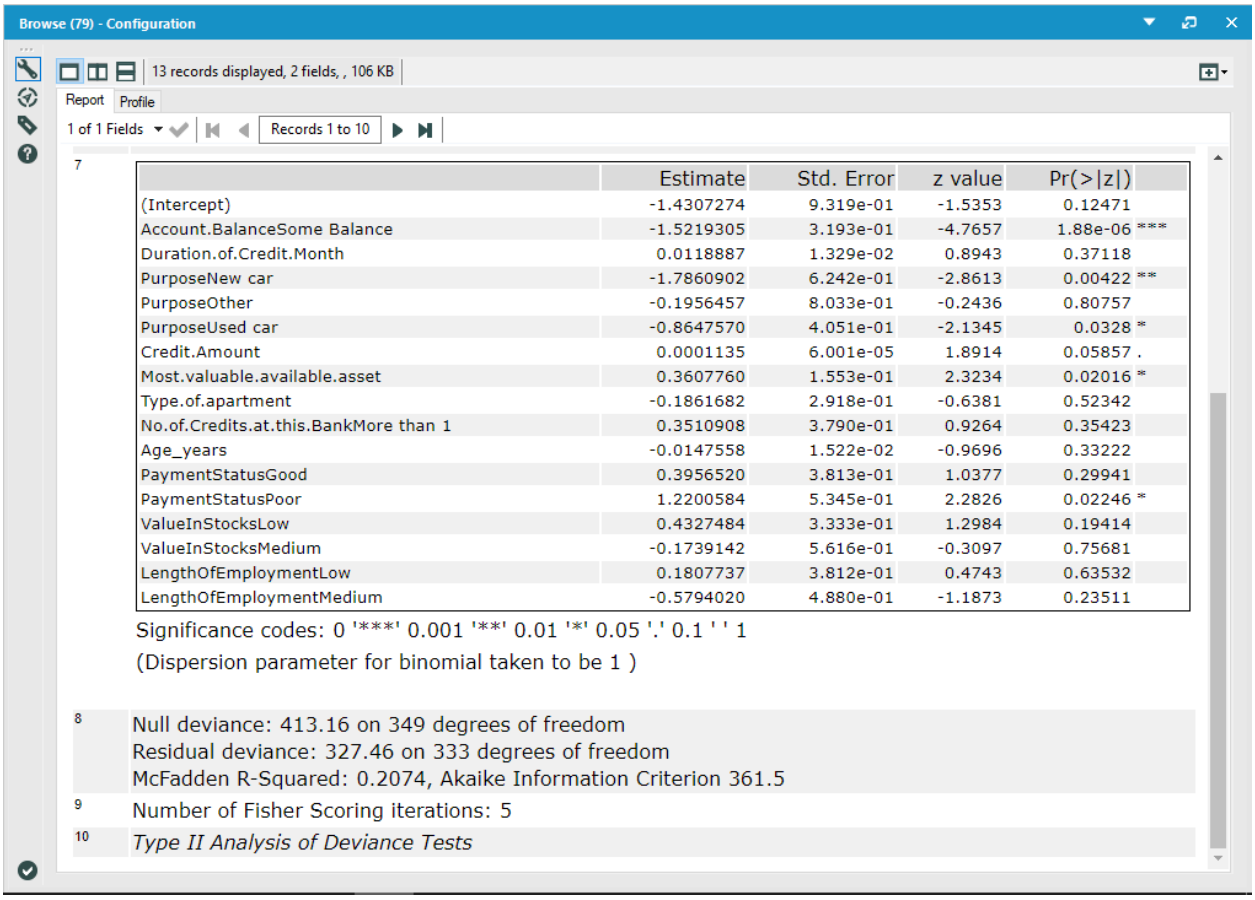
Create all of the following models: Logistic Regression, Decision Tree, Forest Model, Boosted Model

*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.

1. **Logistic Regression**: This took account balance, purpose, most valuable asset and payment status to be the strong predictor variables.


Browse (79) - Configuration
13 records displayed, 2 fields, , 106 KB
Report
Profile
1 of 1 Fields
Records 1 to 10
Record Report
Report for Logistic Regression Model LR_CreditWorthy
Basic Summary
Call:
glm(formula = Credit.Application.Result ~ Account.Balance + Duration.of.Credit.Month + Purpose + Credit.Amount + Most.valuable.available.asset + Type.of.apartment + No.of.Credits.at.this.Bank + Age_years + PaymentStatus + ValueInStocks + LengthOfEmployment, family = binomial(logit), data = the.data)
Deviance Residuals:
Min 1Q Median 3Q Max
-2.032 -0.735 -0.449 0.694 2.490
Coefficients:
Estimate Std. Error z value Pr(>|z|)
(Intercept) -1.4307274 9.319e-01 -1.5353 0.12471
Account.BalanceSome Balance -1.5219305 3.193e-01 -4.7657 1.88e-06 ***
Duration.of.Credit.Month 0.0118887 1.329e-02 0.8943 0.37118
PurposeNew car -1.7860902 6.242e-01 -2.8613 0.00422 **
PurposeOther -0.1956457 8.033e-01 -0.2436 0.80757
PurposeUsed car -0.8647570 4.051e-01 -2.1345 0.0328 *
Credit.Amount 0.0001135 6.001e-05 1.8914 0.05857 .
Most.valuable.available.asset 0.3607760 1.553e-01 2.3234 0.02016 *
Type.of.apartment -0.1861682 2.918e-01 -0.6381 0.52342
No.of.Credits.at.this.BankMore than 1 0.3510908 3.790e-01 0.9264 0.35423
Age_years -0.0147558 1.522e-02 -0.9696 0.33222
PaymentStatusGood 0.3956520 3.813e-01 1.0377 0.29941
PaymentStatusPoor 1.2200584 5.345e-01 2.2826 0.02246 *
10:58 PM
11/1/2018



	Estimate	Std. Error	z value	Pr(>|z|)	
(Intercept)	-1.4307274	9.319e-01	-1.5353	0.12471	
Account.BalanceSome Balance	-1.5219305	3.193e-01	-4.7657	1.88e-06	***
Duration.of.Credit.Month	0.0118887	1.329e-02	0.8943	0.37118	
PurposeNew car	-1.7860902	6.242e-01	-2.8613	0.00422	**
PurposeOther	-0.1956457	8.033e-01	-0.2436	0.80757	
PurposeUsed car	-0.8647570	4.051e-01	-2.1345	0.0328	*
Credit.Amount	0.0001135	6.001e-05	1.8914	0.05857	.
Most.valuable.available.asset	0.3607760	1.553e-01	2.3234	0.02016	*
Type.of.apartment	-0.1861682	2.918e-01	-0.6381	0.52342	
No.of.Credits.at.this.BankMore than 1	0.3510908	3.790e-01	0.9264	0.35423	
Age_years	-0.0147558	1.522e-02	-0.9696	0.33222	
PaymentStatusGood	0.3956520	3.813e-01	1.0377	0.29941	
PaymentStatusPoor	1.2200584	5.345e-01	2.2826	0.02246	*
ValueInStocksLow	0.4327484	3.333e-01	1.2984	0.19414	
ValueInStocksMedium	-0.1739142	5.616e-01	-0.3097	0.75681	
LengthOfEmploymentLow	0.1807737	3.812e-01	0.4743	0.63532	
LengthOfEmploymentMedium	-0.5794020	4.880e-01	-1.1873	0.23511	

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

(Dispersion parameter for binomial taken to be 1)

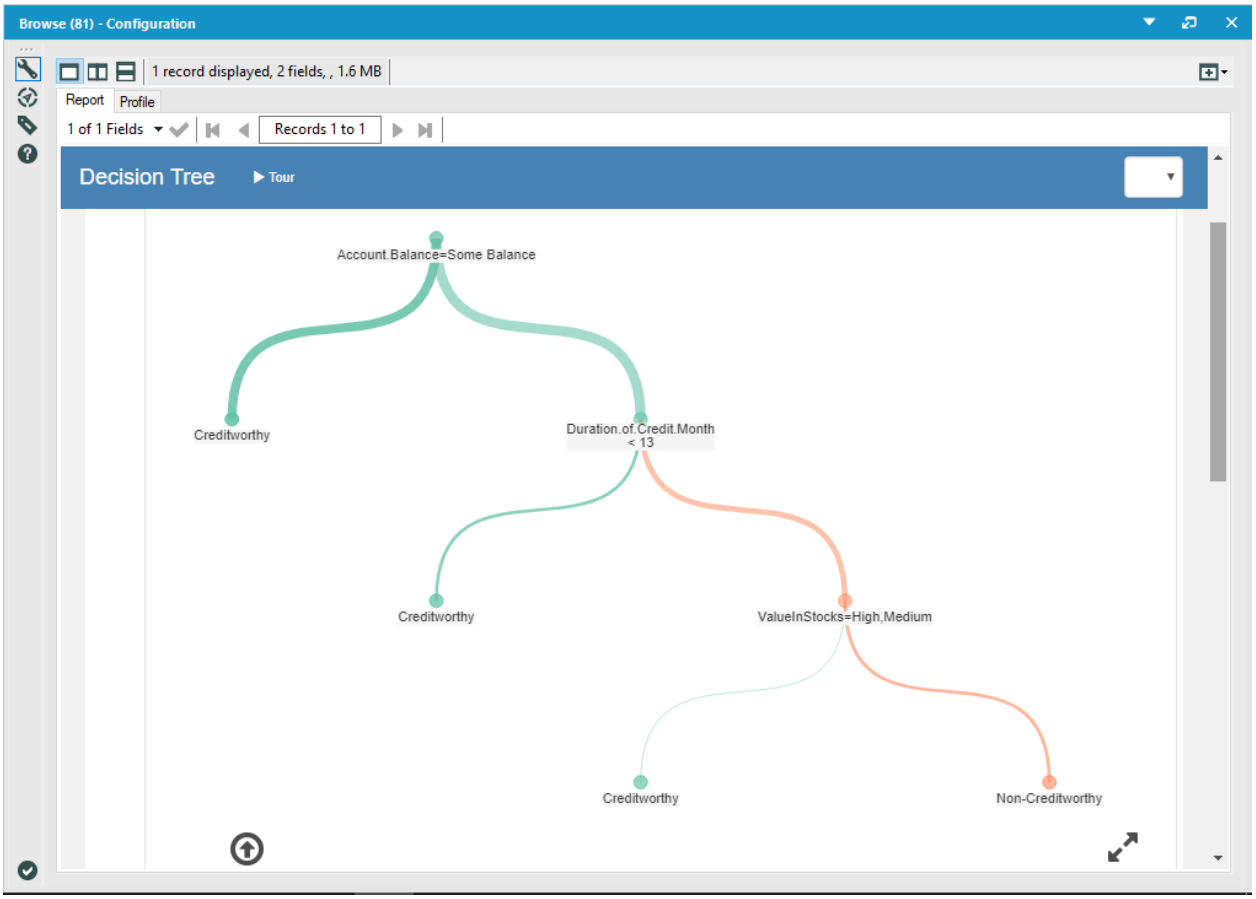
Null deviance: 413.16 on 349 degrees of freedom
Residual deviance: 327.46 on 333 degrees of freedom
McFadden R-Squared: 0.2074, Akaike Information Criterion 361.5

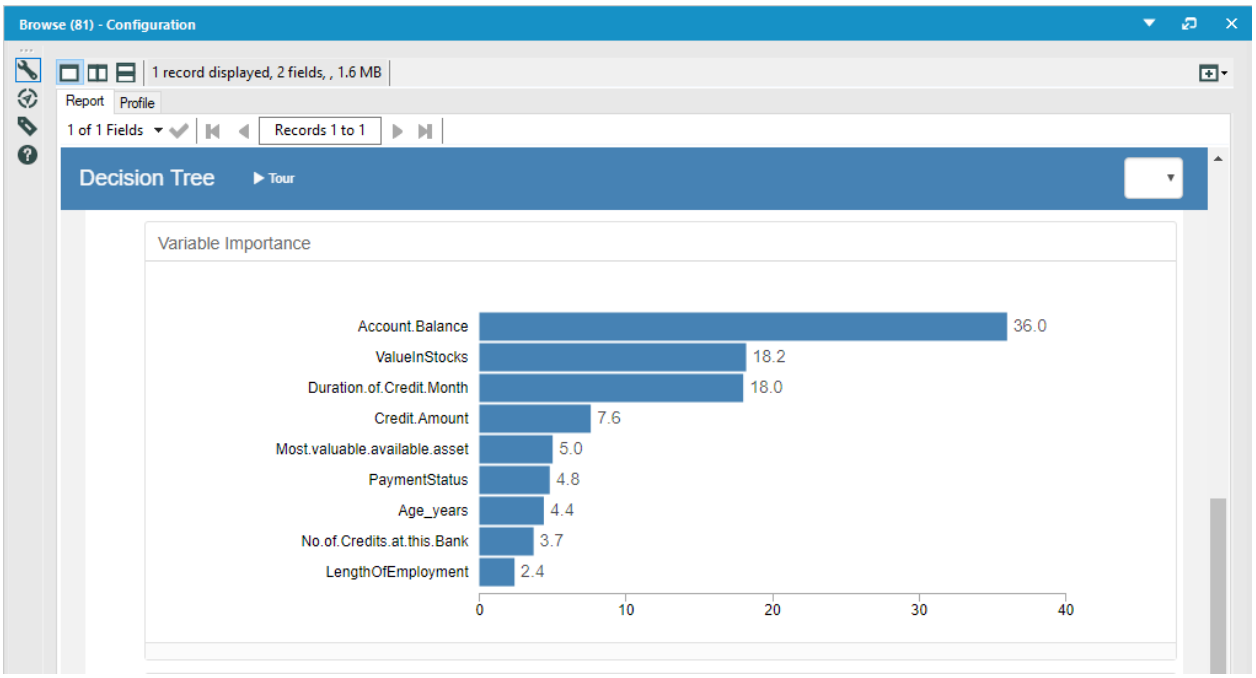
Number of Fisher Scoring iterations: 5

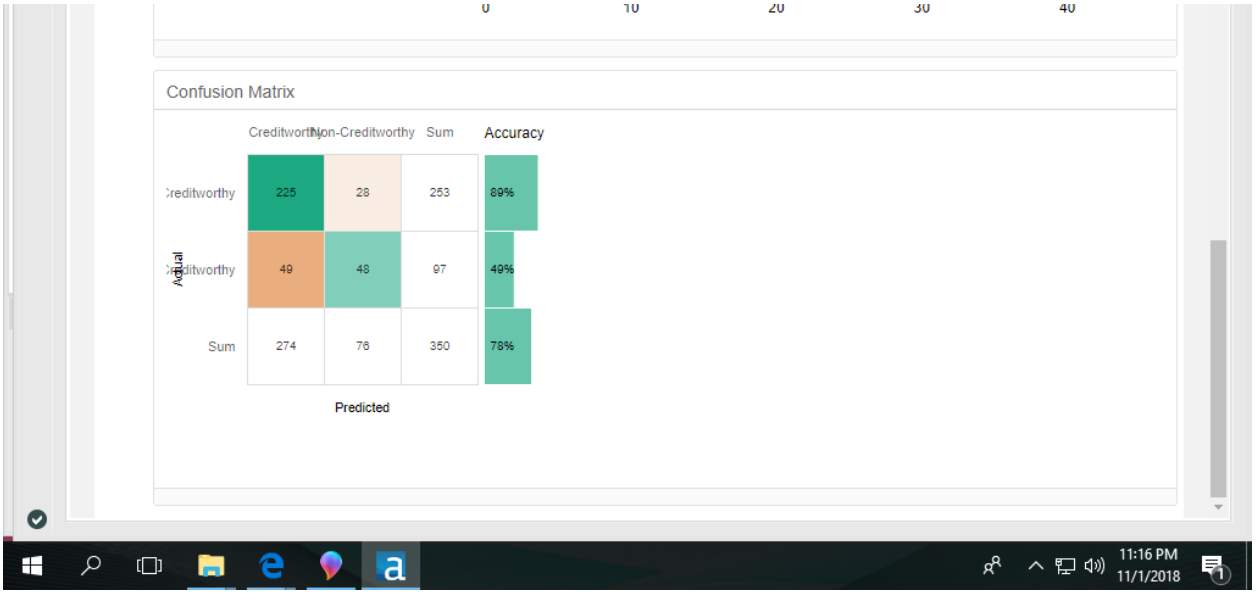
Type II Analysis of Deviance Tests

It had an error rate of 20.74 % as indicated by the R-squared value. This implies it classified 79.26 % correctly.

2. **Decision Tree**: This model took account balance, duration of credit month and value in stocks as the top predictor variables.

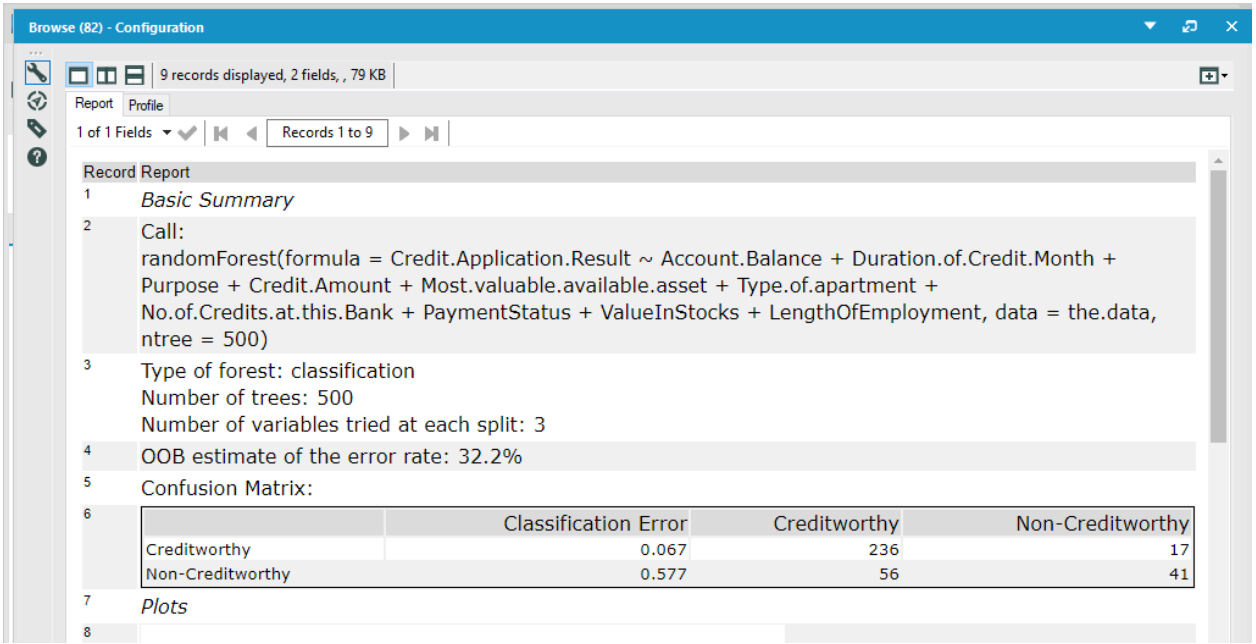

Browse (81) - Configuration
1 record displayed, 2 fields, , 1.6 MB
Report
Profile
1 of 1 Fields
Records 1 to 1
Decision Tree
Tour
Account.Balance=Some Balance
Creditworthy
Duration.of.Credit.Month
< 13
Creditworthy
ValueInStocks=High,Medium
Creditworthy
Non-Creditworthy

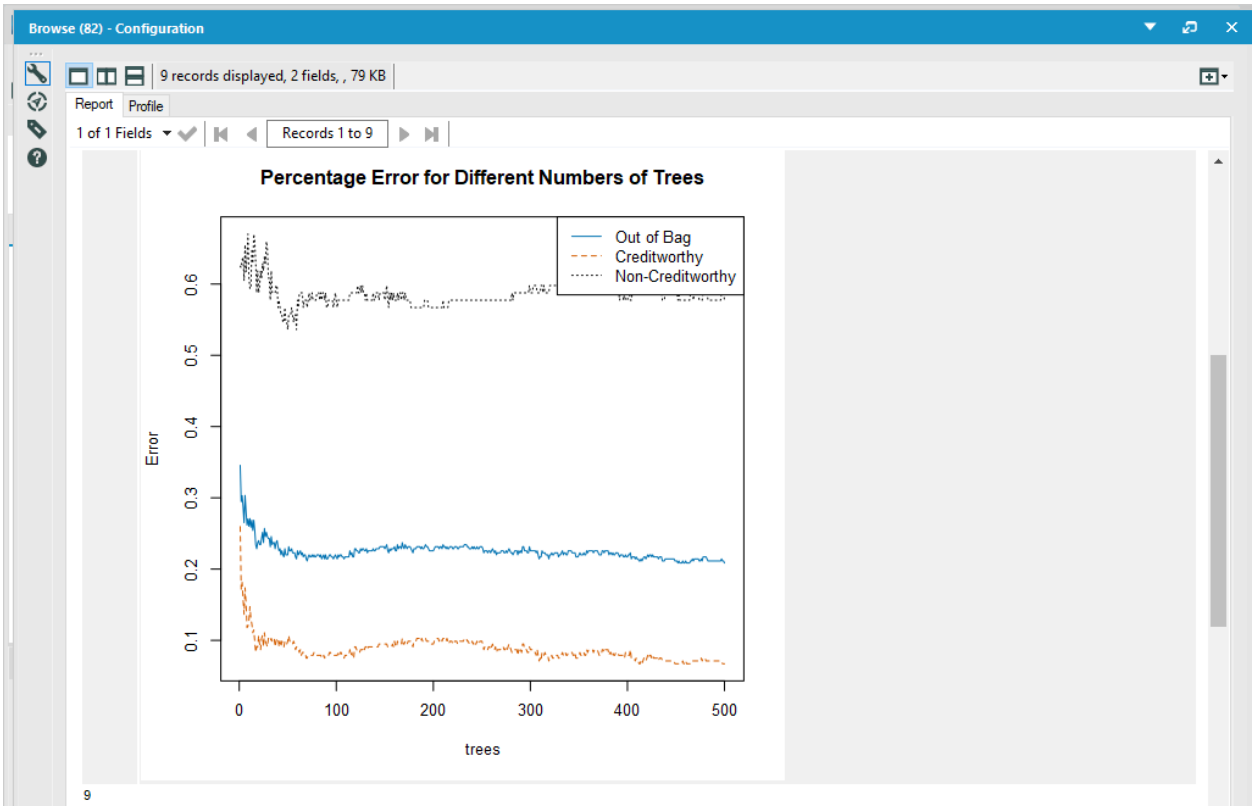

Browse (81) - Configuration
1 record displayed, 2 fields, , 1.6 MB
Report
Profile
1 of 1 Fields
Records 1 to 1
Decision Tree
Tour
Variable Importance
Account.Balance
36.0
ValueInStocks
18.2
Duration.of.Credit.Month
18.0
Credit.Amount
7.6
Most.valuable.available.asset
5.0
PaymentStatus
4.8
Age_years
4.4
No.of.Credits.at.this.Bank
3.7
LengthOfEmployment
2.4
0
10
20
30
40

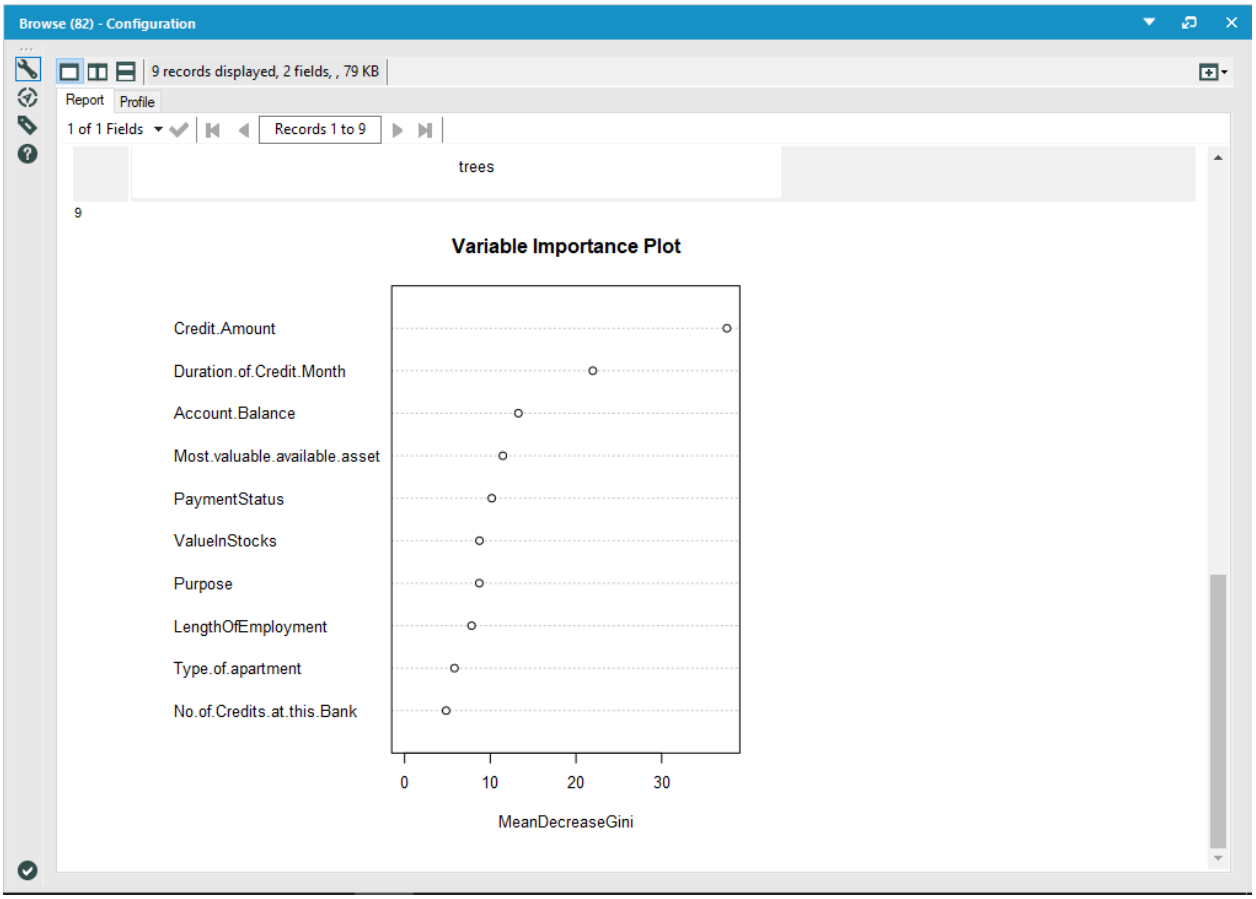


It classified 89% correctly as worthy while incorrectly classified 49% who were not credit worthy as worthy. It got an overall 78% accuracy for classifying the data.

3. **Random Forrest:** This model took credit_amount, duration of credit month and account balance as top 3 predictors.



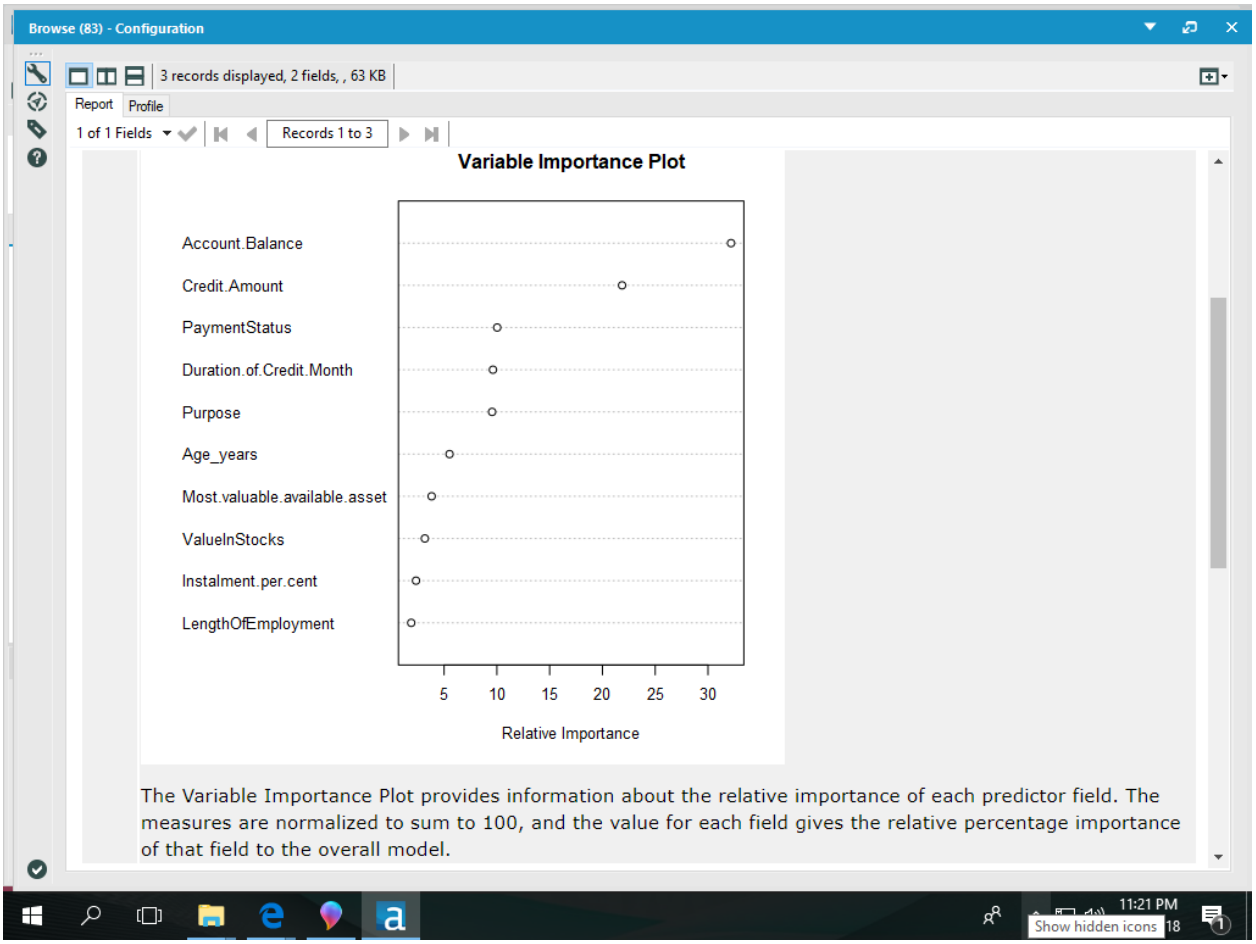

Browse (82) - Configuration
9 records displayed, 2 fields, , 79 KB
Report
Profile
1 of 1 Fields
Records 1 to 9
Percentage Error for Different Numbers of Trees
Out of Bag
Creditworthy
Non-Creditworthy
Error
0.1
0.2
0.3
0.4
0.5
0.6
0
100
200
300
400
500
trees
9

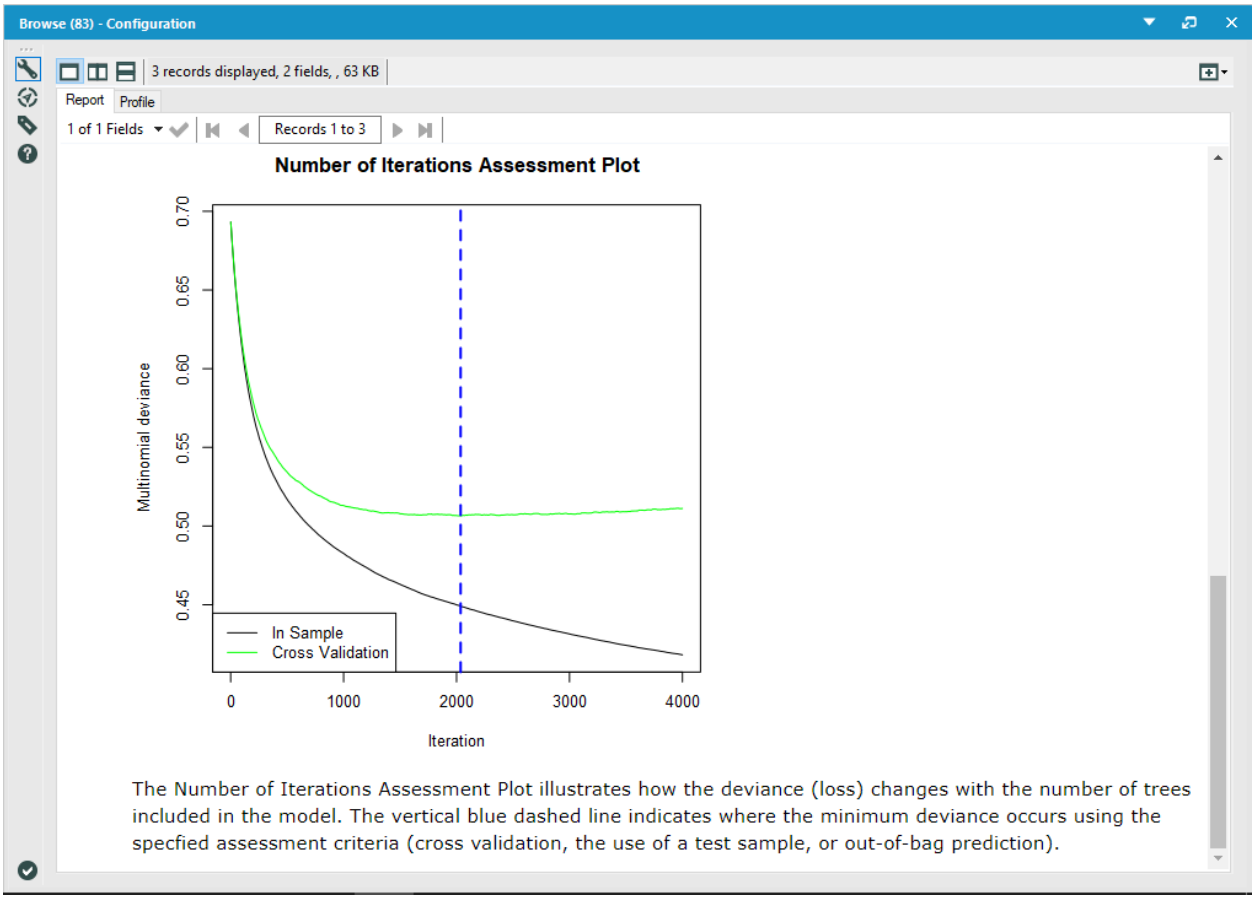


It had a classification error for credit worthy as 6.7% . This implies it classified 93.3 % times correctly. It had a higher error rate in classifying the non creditworthy users at 57.7%.

4. **Boosted Model**: Boosted model took account balance, credit amount as the top predictor variables. The duration of credit month, payment status and purpose camenext with almost equal significance.

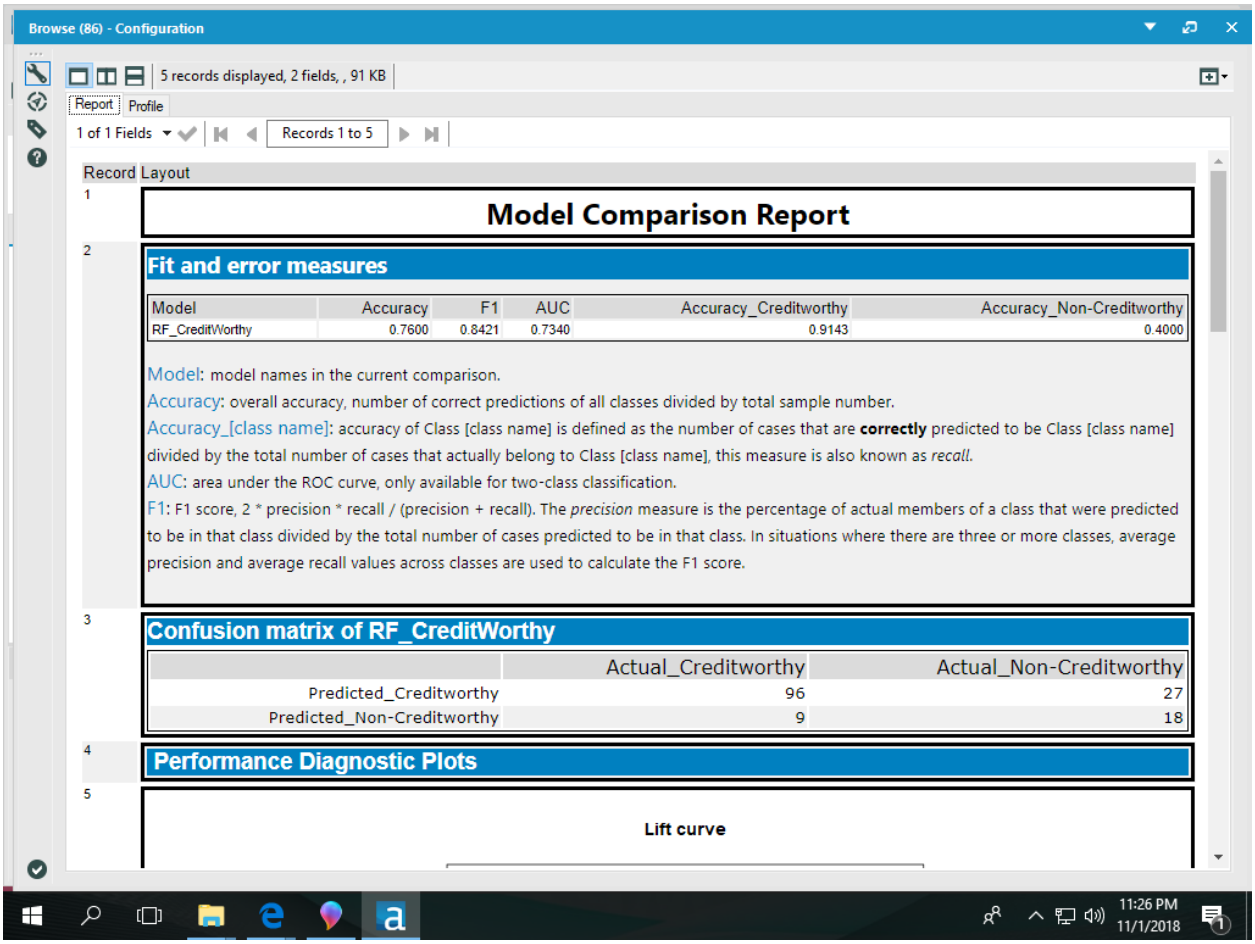
5.


Browse (83) - Configuration
3 records displayed, 2 fields, , 63 KB
Report
Profile
1 of 1 Fields
Records 1 to 3
Variable Importance Plot
Account.Balance
Credit.Amount
PaymentStatus
Duration.of.Credit.Month
Purpose
Age_years
Most.valuable.available.asset
ValueInStocks
Instalment.per.cent
LengthOfEmployment
5
10
15
20
25
30
Relative Importance
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.
11:21 PM
Show hidden icons



Based on the ROC and error rates, I used Random Forrest model as that classified the creditworthy users most accurately.

- 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?



It classified 91.43% correctly as worthy while incorrectly classified 4% who were not credit worthy as worthy. It got an overall 76% accuracy for classifying the data.

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

Step 4: Writeup

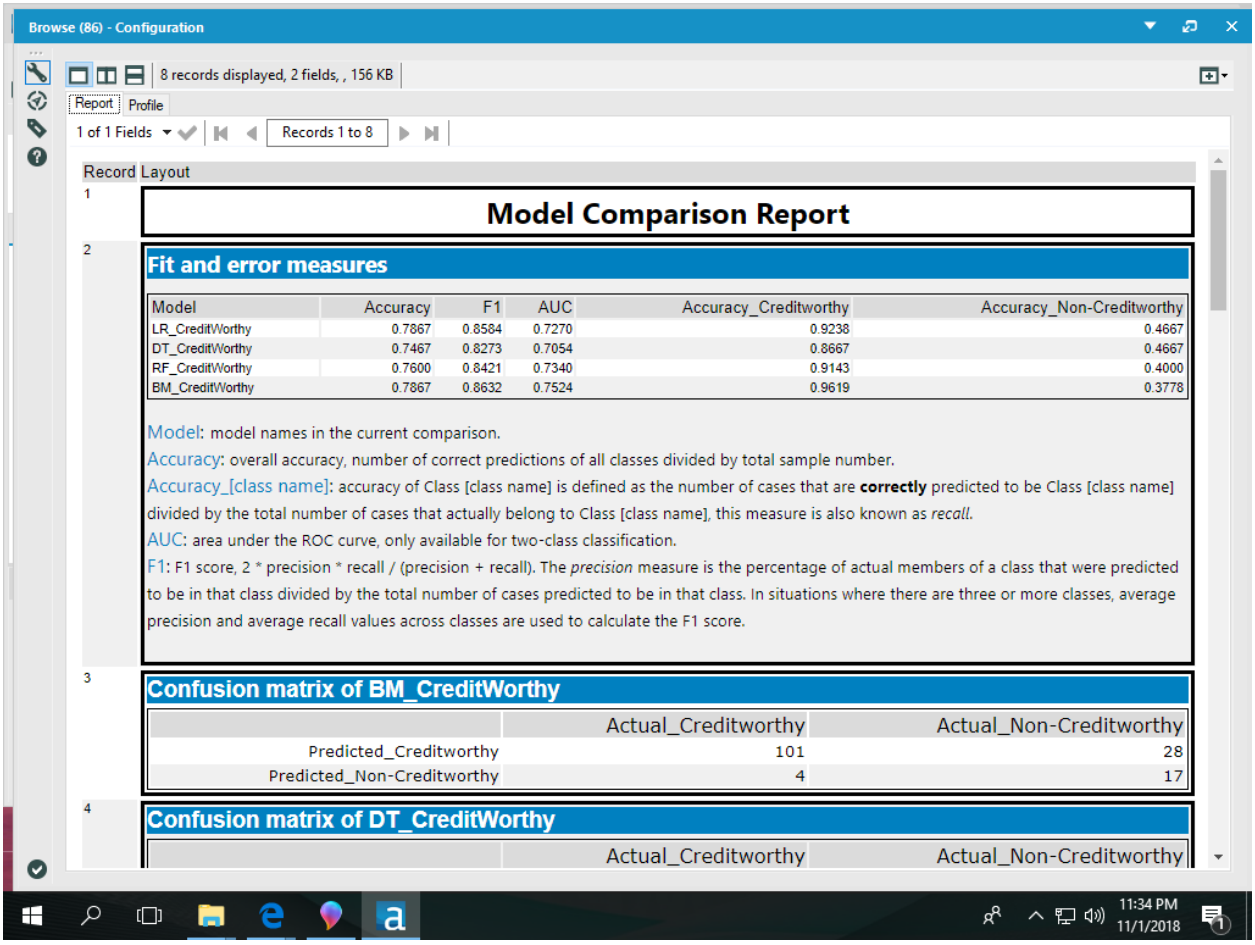
Decide on the best model and score your new customers. For reviewing consistency, if Score_Creditworthy is greater than Score_NonCreditworthy, the person should be labeled as "Creditworthy"

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)

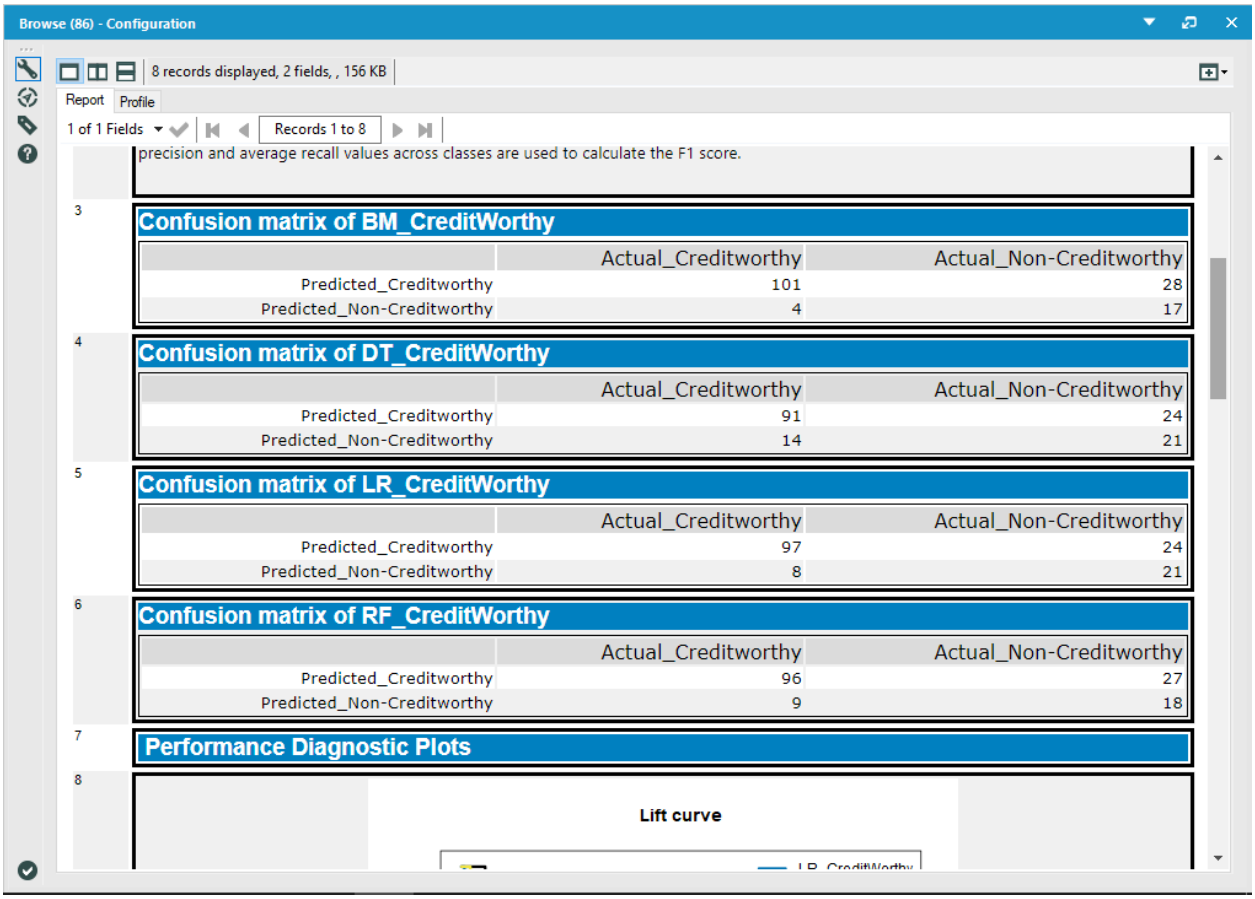
Answer these questions:

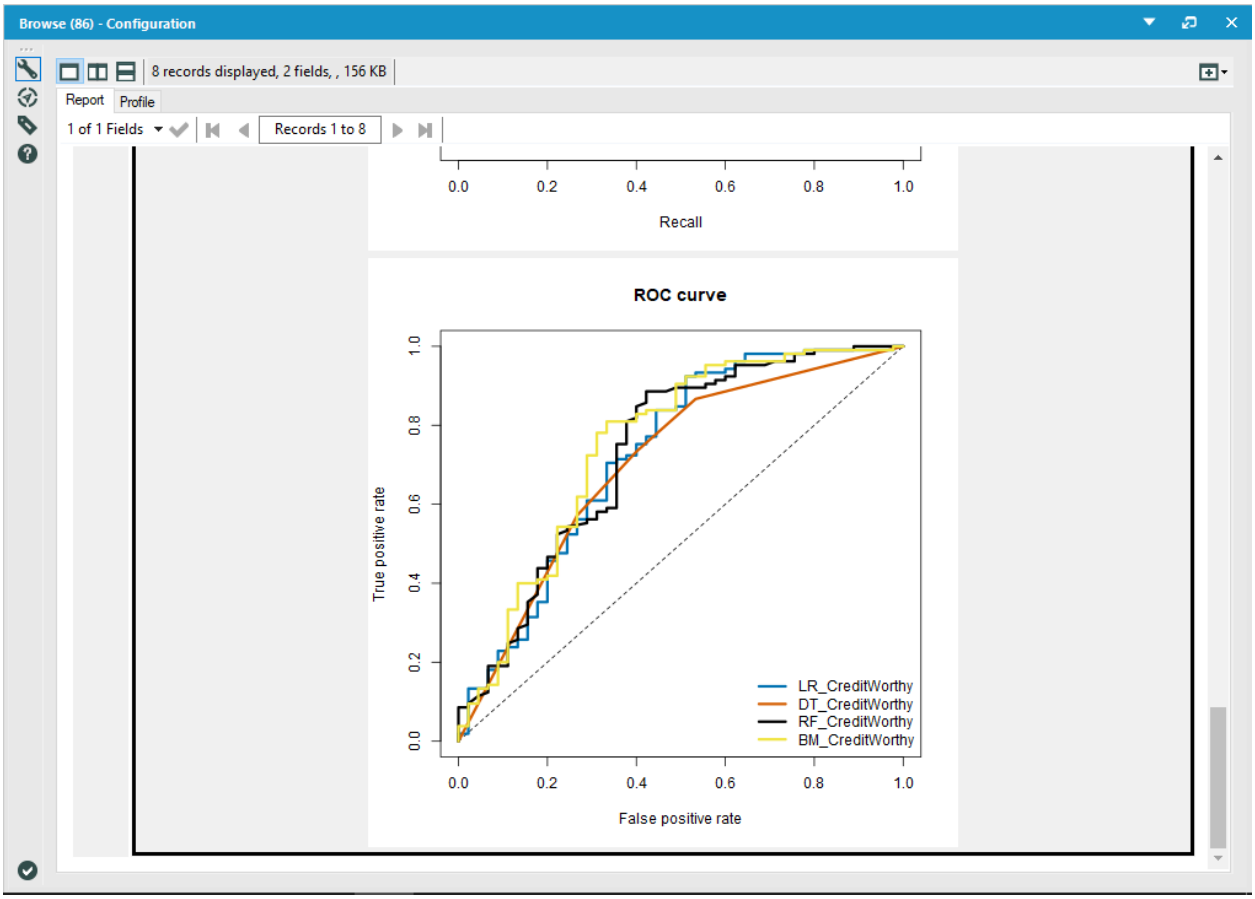
- Which model did you choose to use? Please justify your decision using **all** of the following techniques. Please only use these techniques to justify your decision:
    - Overall Accuracy against your Validation set
    - Accuracies within “Creditworthy” and “Non-Creditworthy” segments
    - ROC graph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

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



Model	Accuracy	F1	AUC	Accuracy_Creditworthy	Accuracy_Non-Creditworthy
LR_CreditWorthy	0.7867	0.8584	0.7270	0.9238	0.4667
DT_CreditWorthy	0.7467	0.8273	0.7054	0.8667	0.4667
RF_CreditWorthy	0.7600	0.8421	0.7340	0.9143	0.4000
BM_CreditWorthy	0.7867	0.8632	0.7524	0.9619	0.3778


Browse (86) - Configuration
8 records displayed, 2 fields, , 156 KB
Report
Profile
1 of 1 Fields
Records 1 to 8
precision and average recall values across classes are used to calculate the F1 score.
3
Confusion matrix of BM_CreditWorthy
Actual_Creditworthy
Actual_Non-Creditworthy
Predicted_Creditworthy 101 28
Predicted_Non-Creditworthy 4 17
4
Confusion matrix of DT_CreditWorthy
Actual_Creditworthy
Actual_Non-Creditworthy
Predicted_Creditworthy 91 24
Predicted_Non-Creditworthy 14 21
5
Confusion matrix of LR_CreditWorthy
Actual_Creditworthy
Actual_Non-Creditworthy
Predicted_Creditworthy 97 24
Predicted_Non-Creditworthy 8 21
6
Confusion matrix of RF_CreditWorthy
Actual_Creditworthy
Actual_Non-Creditworthy
Predicted_Creditworthy 96 27
Predicted_Non-Creditworthy 9 18
7
Performance Diagnostic Plots
8
Lift curve



Based on the area under the curve in the above ROC, both Boosted model and Random Forrest seem to be good models to chose from. However, Random Forrest has highest overall and credit worthy accuracies. Hence chose Random Forrest to classify the data for credit worthiness.

- How many individuals are creditworthy?
  I used Random Forrest model to find out the individuals who are creditworthy. I found 402 out of 500 to be 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.