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.

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. This transforms into a good opportunity to increase more customers to the bank. This also increases the importance of identifying a credit worthy customer accurately.

- 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.
    i. Credit-Application-Result
    ii. Account-Balance
    iii. Duration-of-Credit-Month
    iv. Payment-Status-of-Previous-Credit
    v. Purpose
    vi. Credit-Amount
    vii. Value-Savings-Stocks
    viii. Length-of-current-employment
    ix. Instalment-per-cent
    x. Guarantors
    xi. Duration-in-Current-address
    xii. Most-valuable-available-asset
    xiii. Age-years
    xiv. Concurrent-Credits
    xv. Type-of-apartment
    xvi. No-of-Credits-at-this-Bank
    xvii. Occupation
    xviii. No-of-dependents
    xix. Telephone
    xx. Foreign-Worker

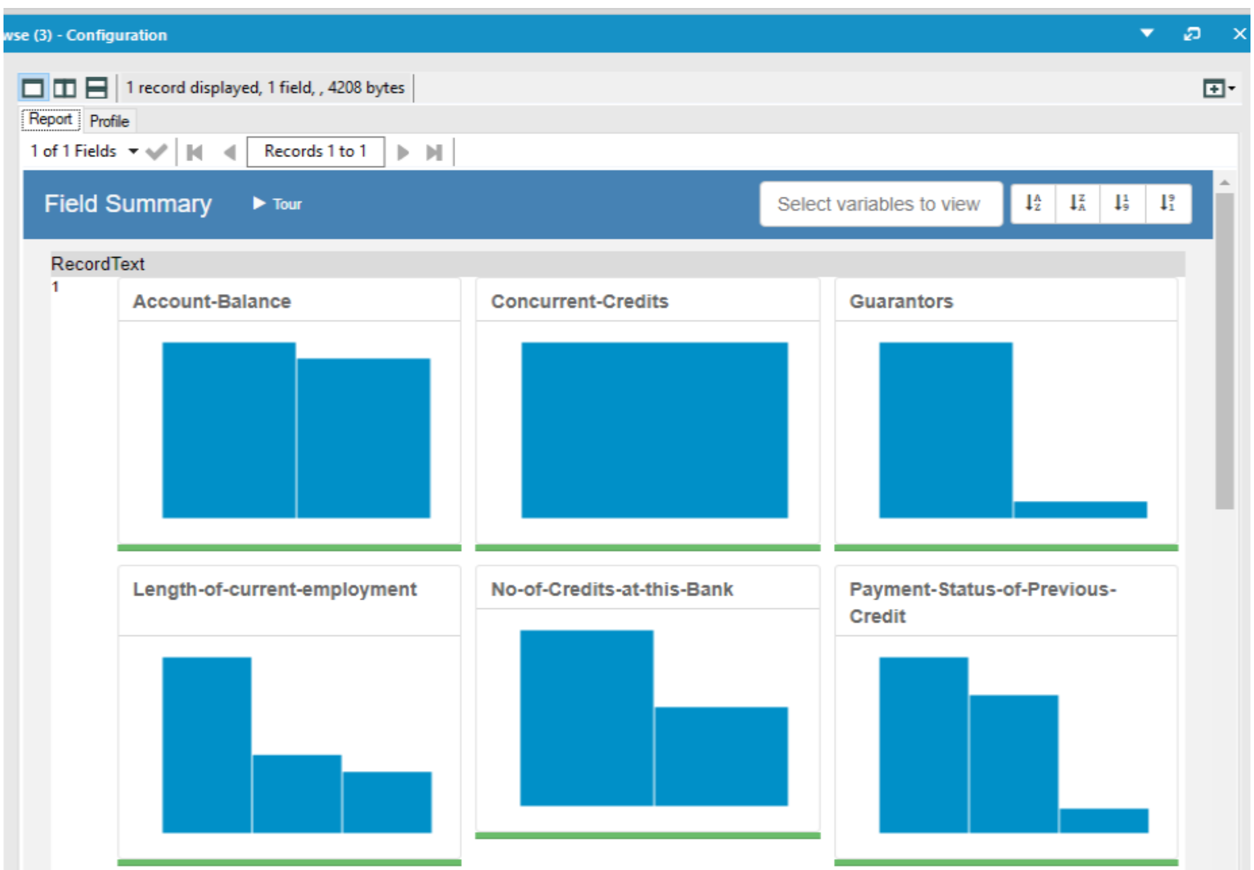
- What kind of model (Continuous, Binary, Non-Binary, Time-Series) do we need to use to help make these decisions?
  Since we are predicting whether a customer is creditworthy or not, we need a binary model.

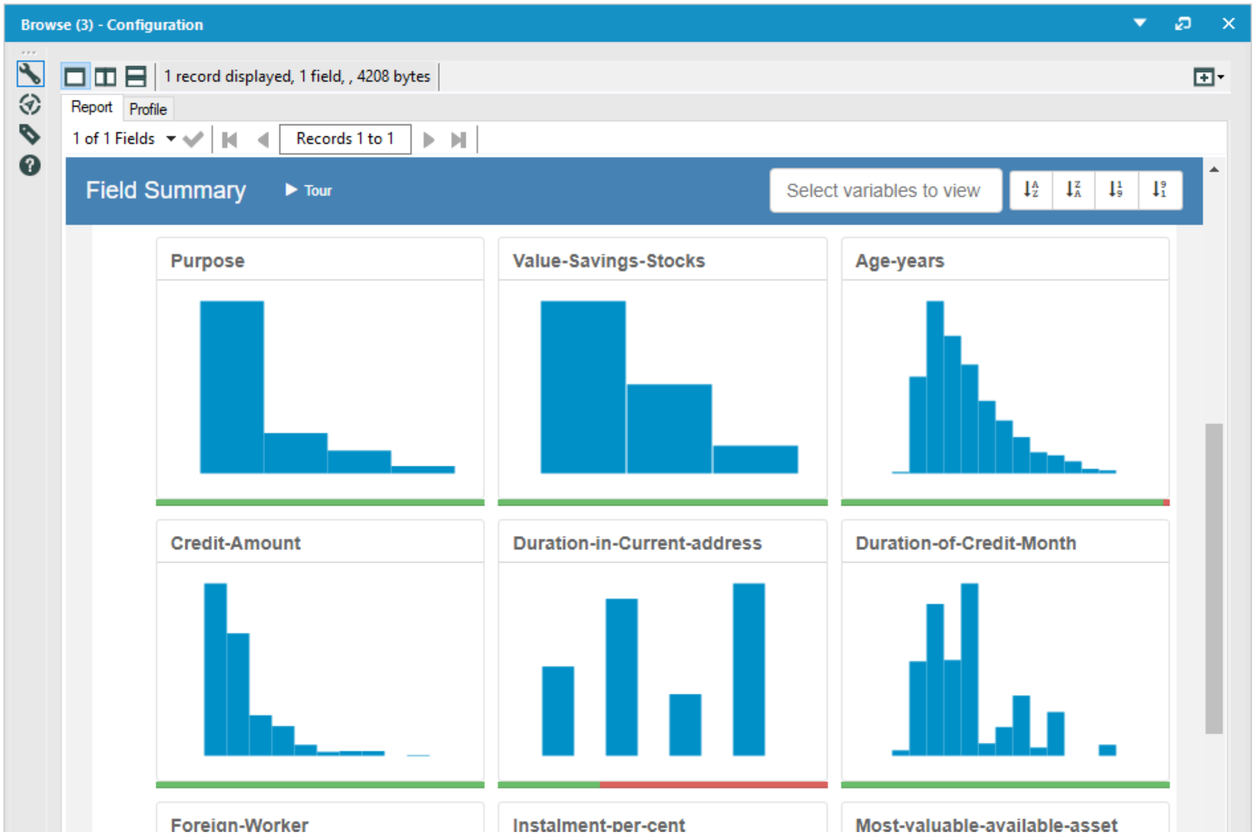
Step 2: Building the Training Set

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.

All the fields were analyzed using the Alteryx Field Summary tool. Below are the results of the analysis.




Browse (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
Purpose
Value-Savings-Stocks
Age-years
Credit-Amount
Duration-in-Current-address
Duration-of-Credit-Month
Foreign-Worker
Instalment-per-cent
Most-valuable-available-asset


Foreign-Worker
Instalment-per-cent
Most-valuable-available-asset
No-of-dependents
Telephone
Type-of-apartment

Following variables were removed either due to low variability (having only one value) or missing data.

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 did not seem to have enough values to impact the target variable.
6. Removed installment-per-cent 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

*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 model took account balance to be the strong predictor variables. It had an error rate of 17.3 % as indicated by the R-squared value. This implies it classified 82.7 % correctly.

Browse (79) - Configuration

13 records displayed, 2 fields, , 90 KB

Report | Profile

1 of 1 Fields | Records 1 to 10

1 **Report for Logistic Regression Model LR_CreditWorthy**

2 *Basic Summary*

3 Call:
glm(formula = Credit.Application.Result ~ Account.Balance + Duration.of.Credit.Month + Purpose + Credit.Amount + Duration.in.Current.address + Most.valuable.available.asset + Type.of.apartment + No.of.Credits.at.this.Bank + Age_years + PaymentStatus + ValueInStocks + LengthOfEmployment, family = binomial(logit), data = the.data)

4 Deviance Residuals:

5

Min	1Q	Median	3Q	Max
-1.632	-0.655	-0.440	-0.166	2.282

6 Coefficients:

7

	Estimate	Std. Error	z value	Pr(>|z|)	
(Intercept)	-2.475e+00	2.490e+00	-0.993726	0.32036	
Account.BalanceSome Balance	-1.437e+00	6.698e-01	-2.145089	0.03195	*
Duration.of.Credit.Month	5.384e-02	3.522e-02	1.528810	0.12631	
PurposeNew car	-1.553e+01	1.898e+03	-0.008183	0.99347	
PurposeOther	-3.538e-01	1.339e+00	-0.264119	0.79169	
PurposeUsed car	-5.783e-01	8.963e-01	-0.645131	0.51884	
Credit.Amount	-1.982e-04	1.791e-04	-1.106450	0.26853	
Duration.in.Current.address	-1.317e-01	2.707e-01	-0.486656	0.6265	
Most.valuable.available.asset	8.543e-01	7.138e-01	1.196844	0.23137	
Type.of.apartment	2.304e-01	6.471e-01	0.356002	0.72184	
No.of.Credits.at.this.BankMore than 1	2.367e-01	9.573e-01	0.247295	0.80468	
Age_years	-1.169e-02	3.237e-02	-0.361276	0.71789	
PaymentStatusGood	3.560e-02	1.042e+00	0.034172	0.97274	
PaymentStatusPoor	1.413e+00	1.172e+00	1.206121	0.22777	

Browse (79) - Configuration

13 records displayed, 2 fields, , 90 KB

Report | Profile

1 of 1 Fields | Records 1 to 10

6 Coefficients:

7

	Estimate	Std. Error	z value	Pr(>|z|)	
(Intercept)	-2.475e+00	2.490e+00	-0.993726	0.32036	
Account.BalanceSome Balance	-1.437e+00	6.698e-01	-2.145089	0.03195	*
Duration.of.Credit.Month	5.384e-02	3.522e-02	1.528810	0.12631	
PurposeNew car	-1.553e+01	1.898e+03	-0.008183	0.99347	
PurposeOther	-3.538e-01	1.339e+00	-0.264119	0.79169	
PurposeUsed car	-5.783e-01	8.963e-01	-0.645131	0.51884	
Credit.Amount	-1.982e-04	1.791e-04	-1.106450	0.26853	
Duration.in.Current.address	-1.317e-01	2.707e-01	-0.486656	0.6265	
Most.valuable.available.asset	8.543e-01	7.138e-01	1.196844	0.23137	
Type.of.apartment	2.304e-01	6.471e-01	0.356002	0.72184	
No.of.Credits.at.this.BankMore than 1	2.367e-01	9.573e-01	0.247295	0.80468	
Age_years	-1.169e-02	3.237e-02	-0.361276	0.71789	
PaymentStatusGood	3.560e-02	1.042e+00	0.034172	0.97274	
PaymentStatusPoor	1.413e+00	1.172e+00	1.206121	0.22777	
ValueInStocksLow	5.937e-01	7.363e-01	0.806349	0.42004	
ValueInStocksMedium	1.028e+00	1.430e+00	0.719084	0.47209	
LengthOfEmploymentLow	1.918e-01	8.527e-01	0.224929	0.82203	
LengthOfEmploymentMedium	-1.274e+00	1.179e+00	-1.080898	0.27974	

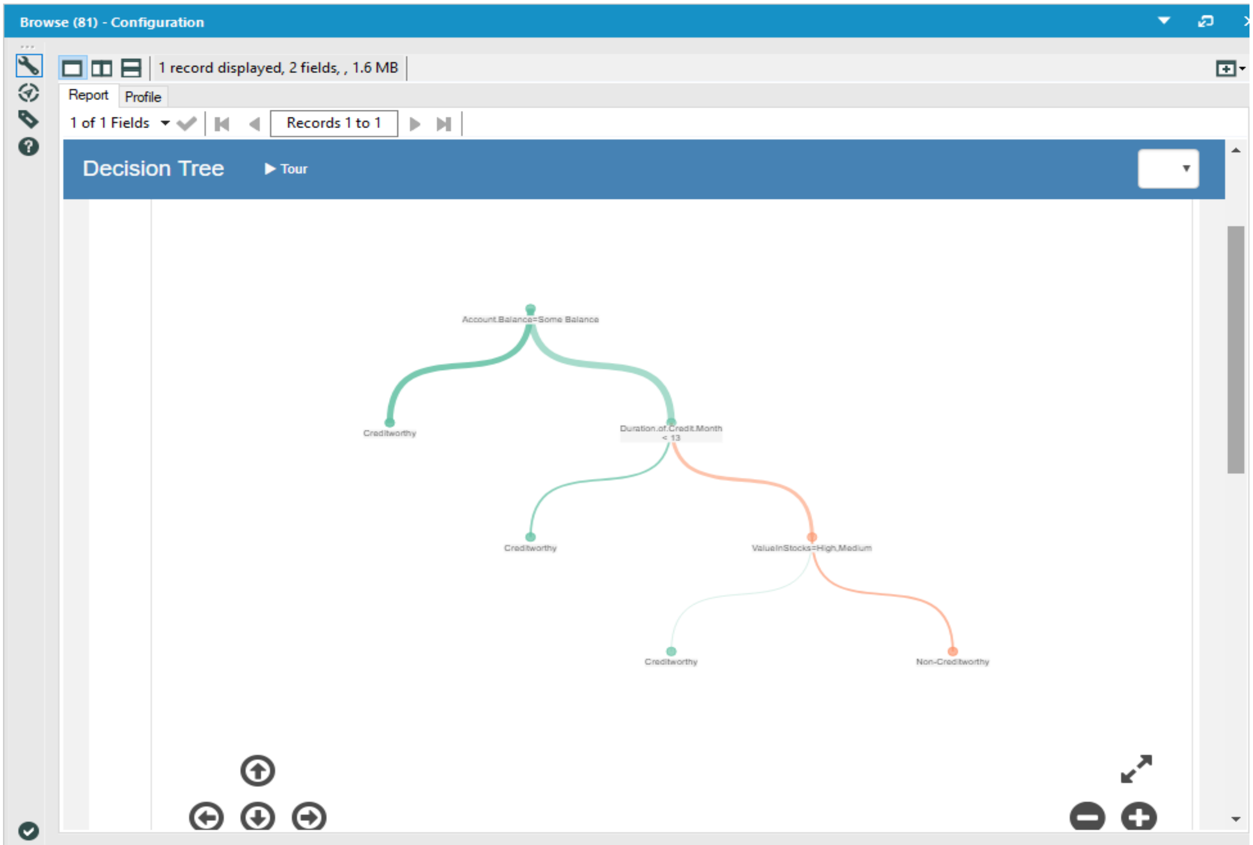
Significance codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

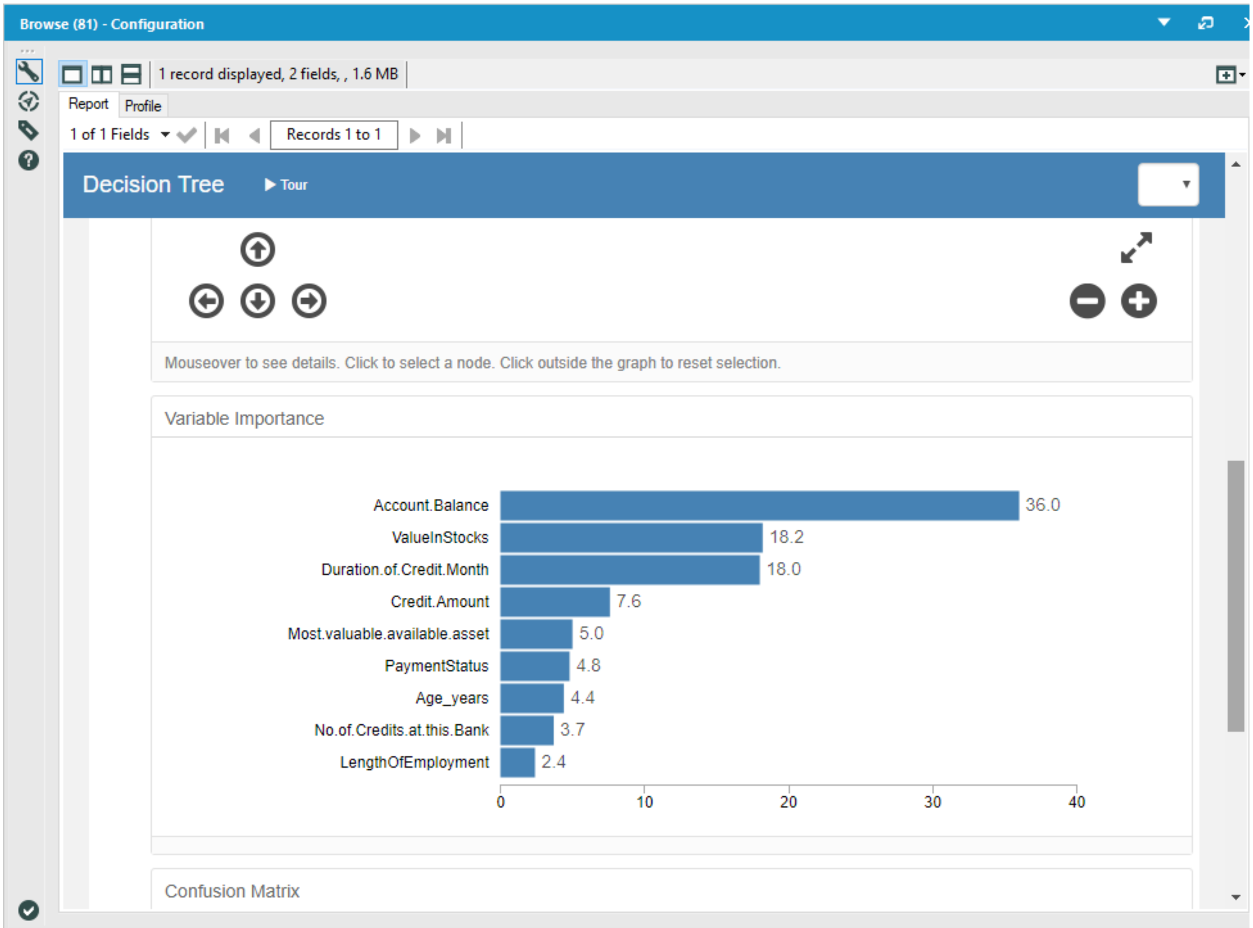
(Dispersion parameter for binomial taken to be 1)

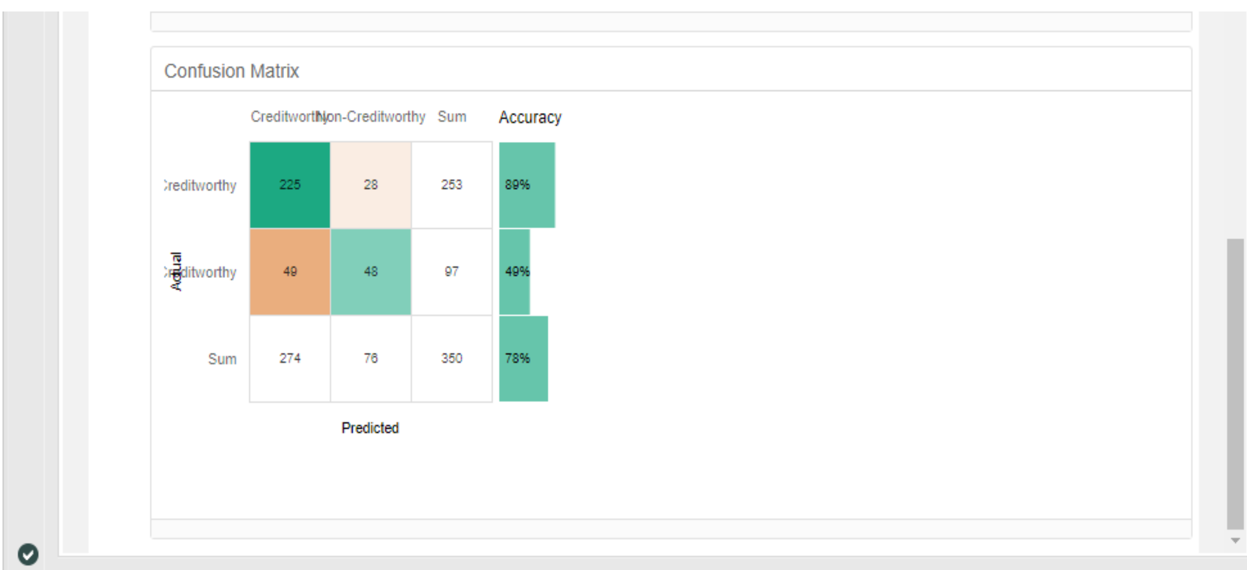
8 Null deviance: 114.64 on 113 degrees of freedom
Residual deviance: 94.827 on 96 degrees of freedom
McFadden R-Squared: 0.1729, Akaike Information Criterion 130.8

9 Number of Fisher Scoring iterations: 16

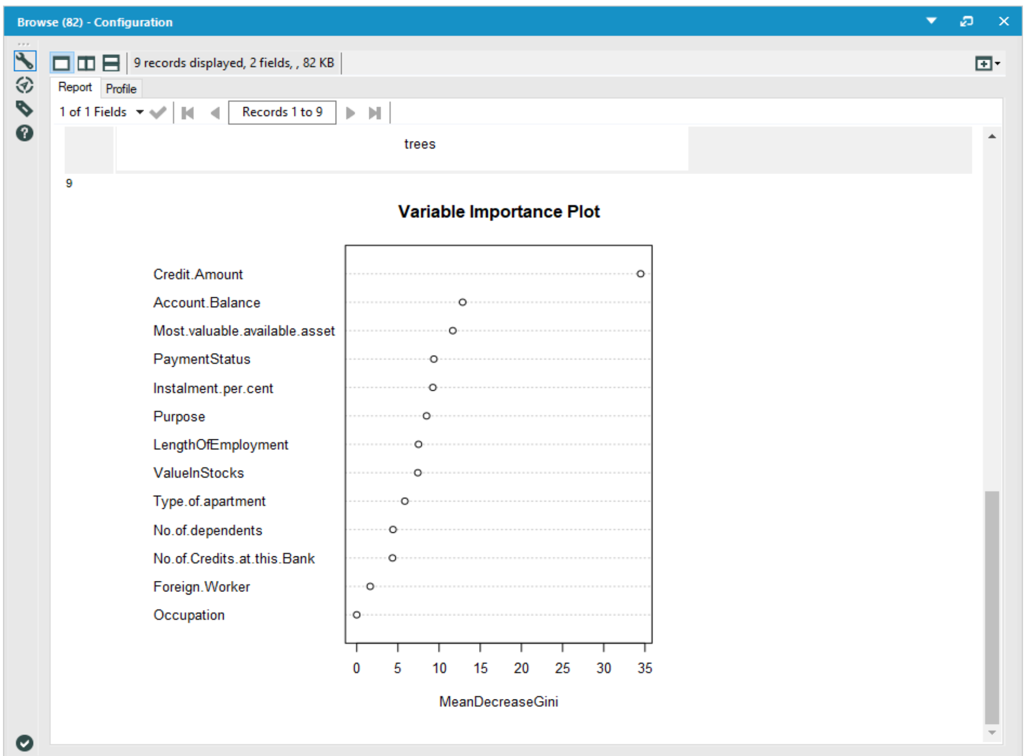
2. **Decision Tree**: This model took account balance, value in stocks and duration of credit month as the top predictor variables. It classified 89% correctly as worthy while incorrectly classified 49% who were not credit worthy as worthy. It got an overall 78% accuracy for classifying the data.

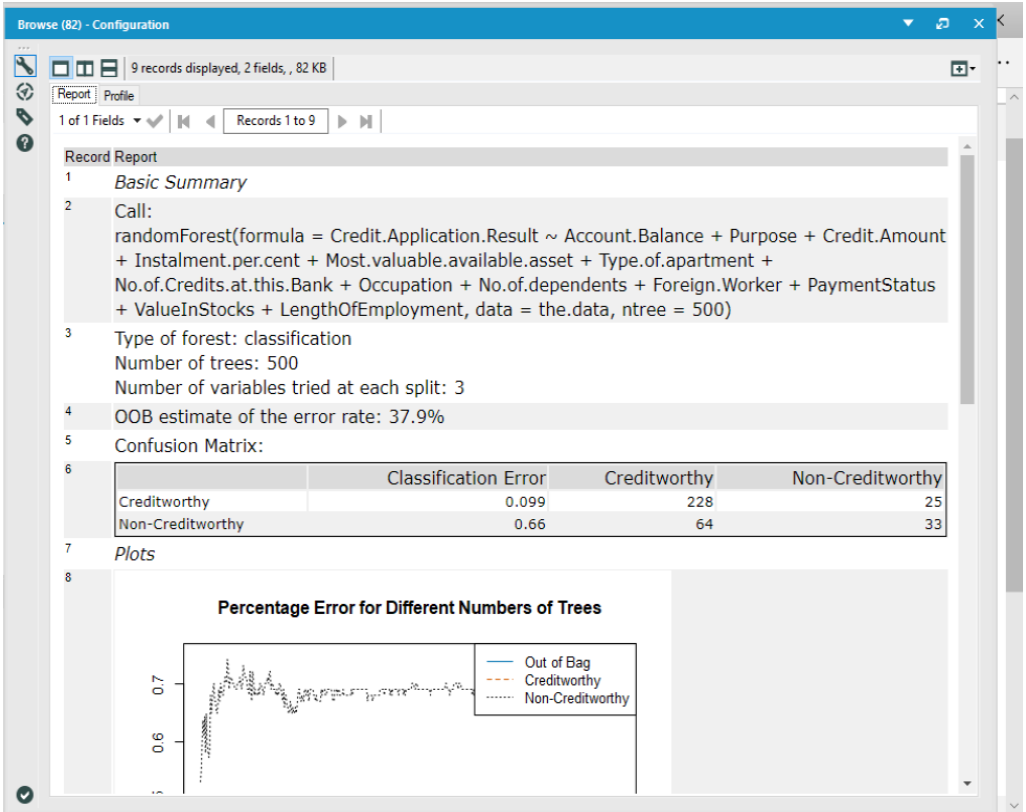



Browse (81) - Configuration
1 record displayed, 2 fields, , 1.6 MB
Report
Profile
1 of 1 Fields
Records 1 to 1
Decision Tree
Tour
Mouseover to see details. Click to select a node. Click outside the graph to reset selection.
Variable Importance
Account.Balance 36.0
ValueInStocks 18.2
Duration.of.Credit.Month 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
Confusion Matrix

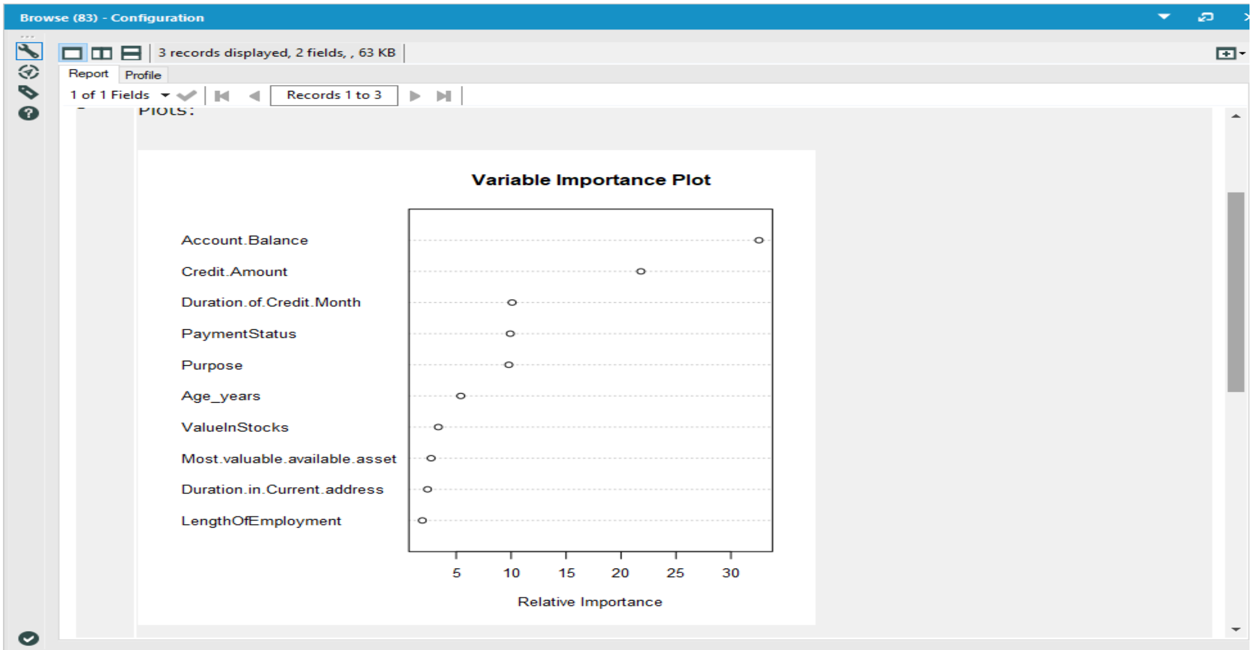

Confusion Matrix
Creditworthy
Non-Creditworthy
Sum
Accuracy
225
28
253
89%
49
48
97
49%
Sum
274
76
350
78%
Actual
Predicted

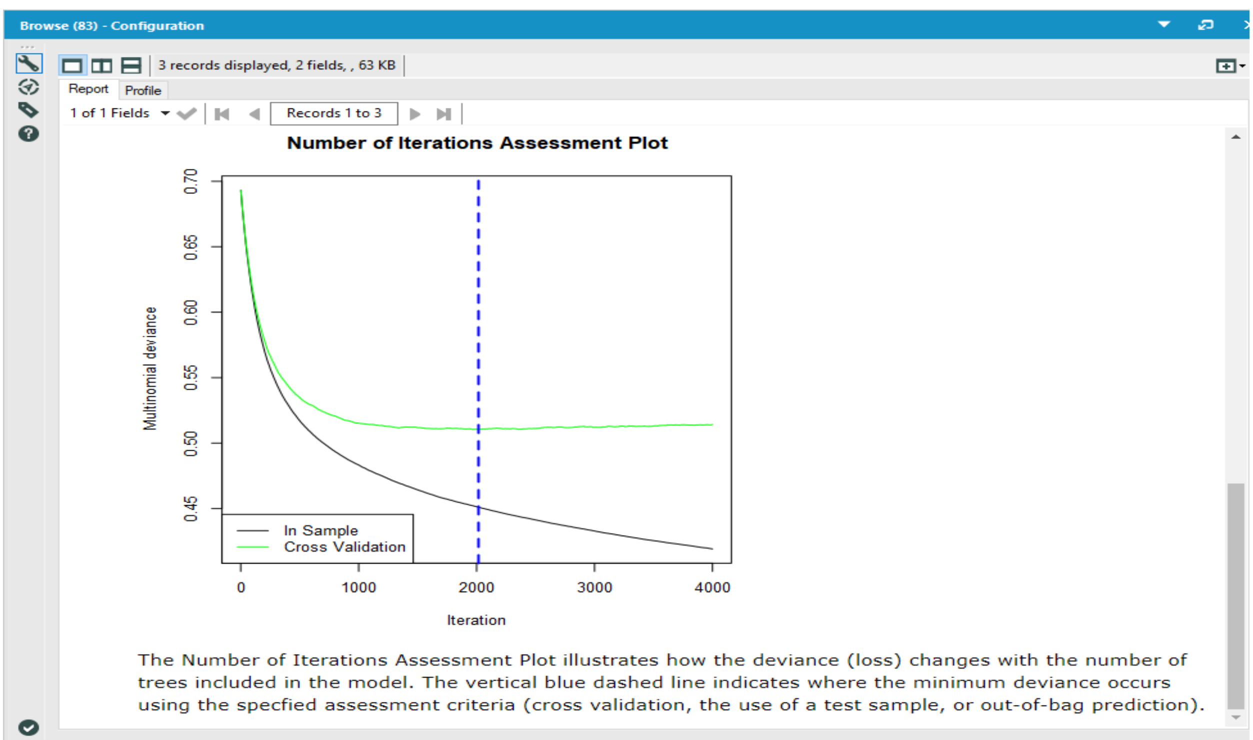
Random Forrest: This model took credit_amount, account balance, most valuable available asset as top 3 predictors. It classified 9.9% creditworthy customers incorrectly and classified 66% noncreditworthy customers incorrectly. It had a out of the bag error rate of 37.9%, which implies only 62.1 % are classified properly.




Browse (82) - Configuration
9 records displayed, 2 fields, , 82 KB
Report
Profile
1 of 1 Fields
Records 1 to 9
Record Report
1 Basic Summary
2 Call:
randomForest(formula = Credit.Application.Result ~ Account.Balance + Purpose + Credit.Amount + Instalment.per.cent + Most.valuable.available.asset + Type.of.apartment + No.of.Credits.at.this.Bank + Occupation + No.of.dependents + Foreign.Worker + PaymentStatus + ValueInStocks + LengthOfEmployment, data = the.data, ntree = 500)
3 Type of forest: classification
Number of trees: 500
Number of variables tried at each split: 3
4 OOB estimate of the error rate: 37.9%
5 Confusion Matrix:
6 Classification Error
Creditworthy
Non-Creditworthy
Creditworthy 0.099 228 25
Non-Creditworthy 0.66 64 33
7 Plots
8 Percentage Error for Different Numbers of Trees
0.7
0.6
Out of Bag
Creditworthy
Non-Creditworthy

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





Based on the ROC and error rates, I used Boosted model to classify the data.

- 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?

	Predicted CreditWorthy	Predicted Non Creditworthy
Actual Creditworthy	135	15
Actual NonCreditworthy	22	128

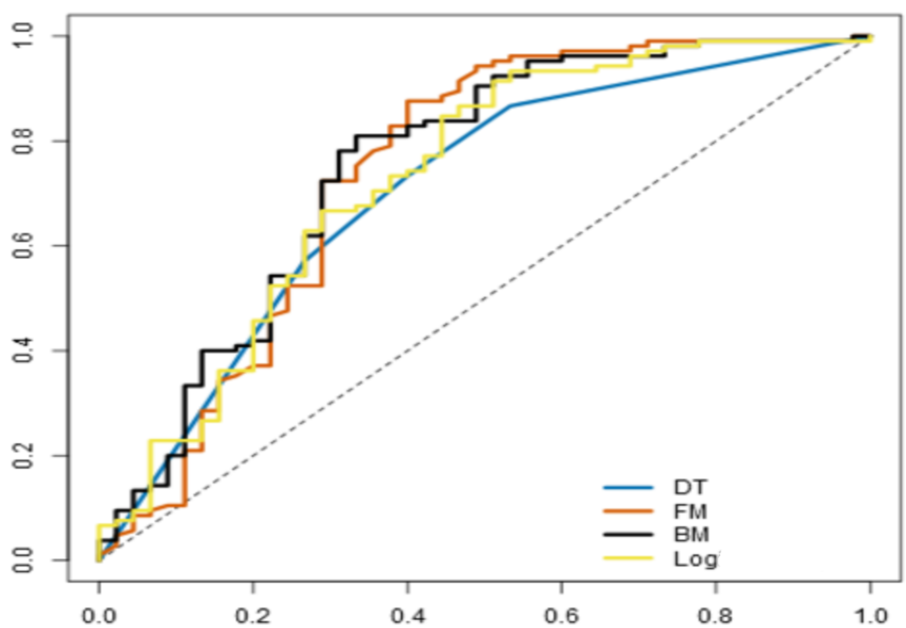
It classified 91% correctly as worthy while incorrectly classified 15% who were not credit worthy as worthy. It got an overall 90% accuracy for classifying the data.

Step 4: Writeup

Answer these questions:

- Which model did you choose to use?

It classified 9.9% creditworthy customers incorrectly and classified 66% noncreditworthy customers incorrectly. It had a out of the bag error rate of 37.9%, which implies only 62.1 % are classified properly.



- How many individuals are creditworthy?
  I used boosted model to find out the individuals who are creditworthy. I found 454 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.