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?

Identify whether customers who applied for loan are creditworthy to be extended one

- What data is needed to inform those decisions?

Data on all past applications and the list of customers that need to be processed in the next few days

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

  Binary classification models such as logistics regression, decision tree and n Non-Binary such as forest model and boosted tree will be used to analyze and determine creditworthy customers

Step 2: Building the Training Set

The field Summary below show all variables , Duration in Current Address has 69% missing data and should be removed. While Age Years has 2% missing data , it is appropriate to impute the missing data with the median age. Median age is used instead of mean as the data is skewed to the left as shown below.

In addition, Concurrent credit has one value while Guarantors, Foreign Worker *and* No of Dependents show low variability where more than 80% of the data skewed towards one data. These data should be removed in order not to skew our analysis results.
Telephone field should also be removed due to its irrelevancy to the customer creditworthy.

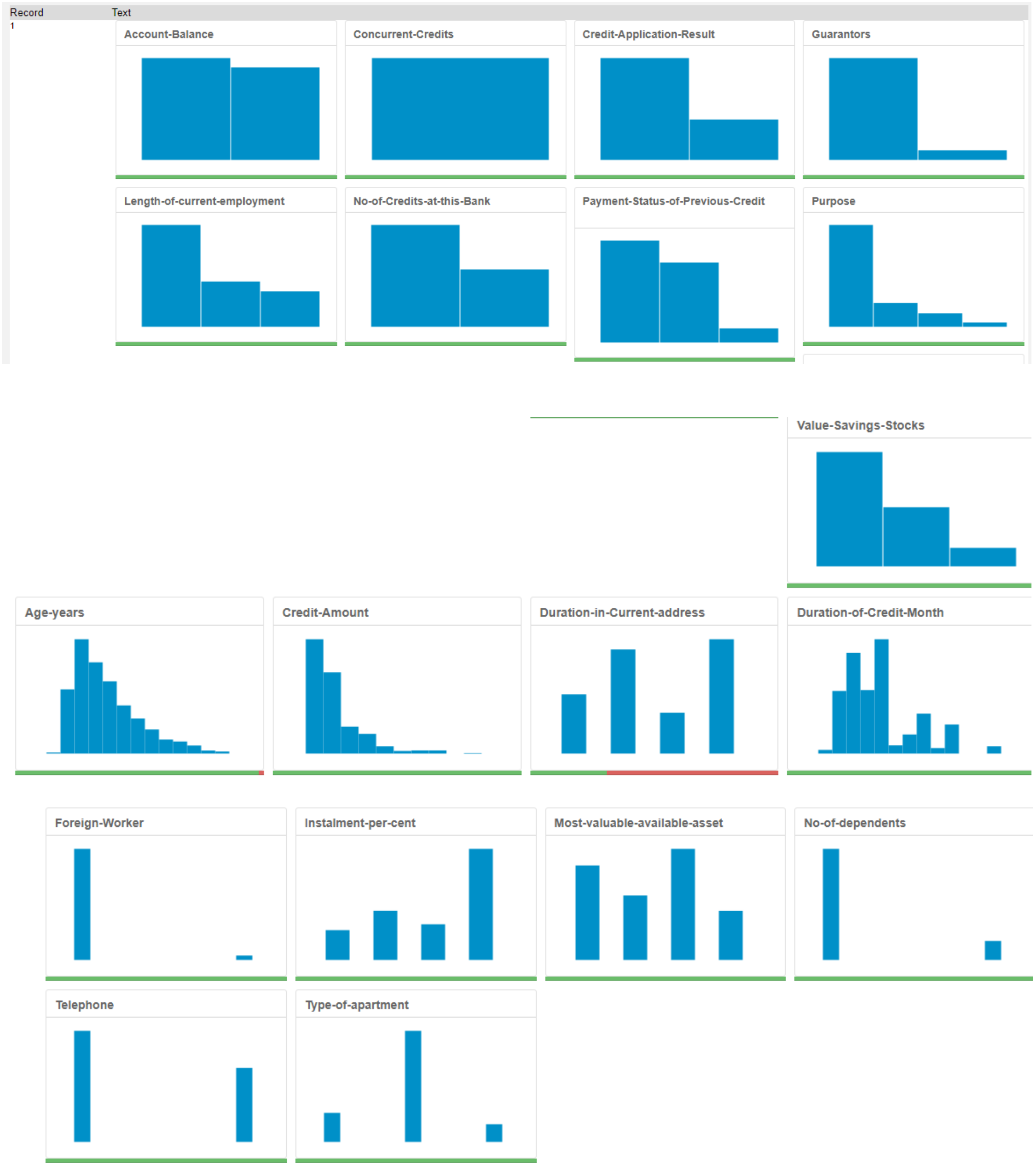
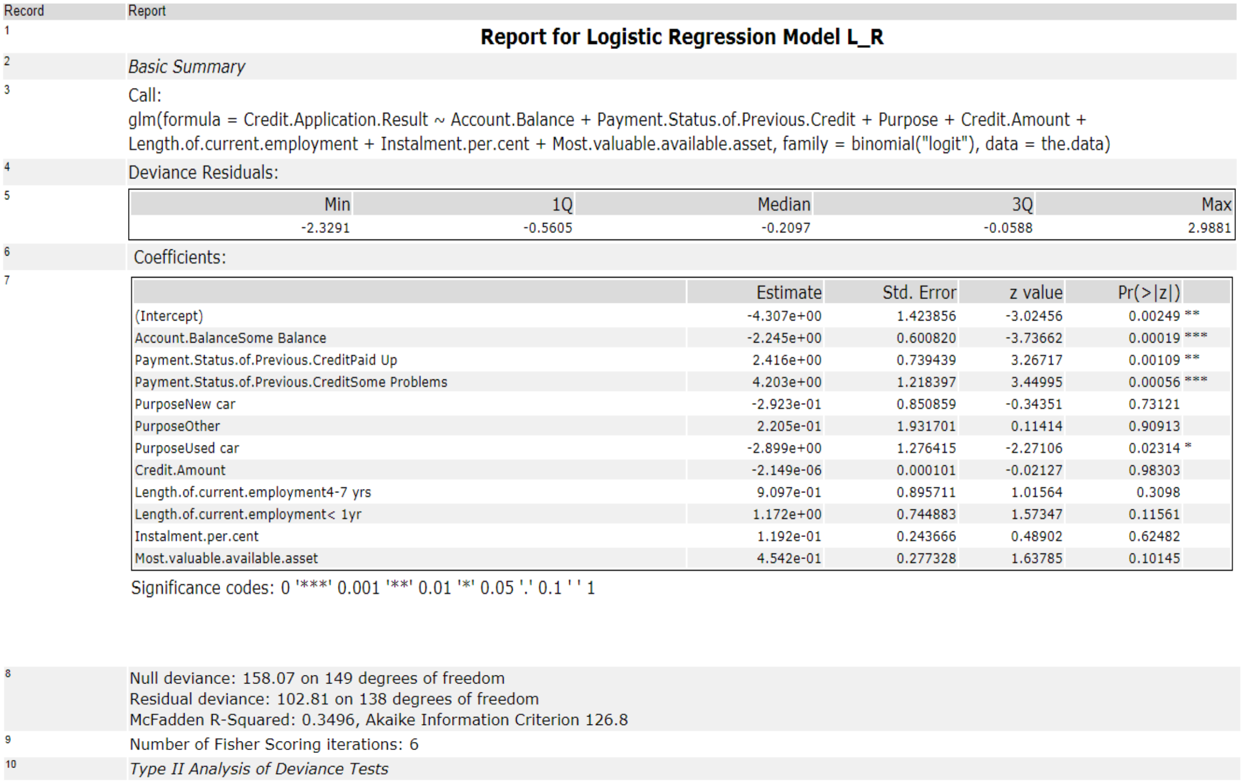


Figure 1:Field Summary for all variables

Step 3: Train your Classification Models

1-Logistic Regression Model

Using Credit Application Result as the target variable and Account Balance ,Payment Status of previous Credit ,Purpose , Credit Amount ,Length of current employment , Instalment per cent and Most valuable available asset for predictive variables



Record	Report
1	**Report for Logistic Regression Model L_R**
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	(table below)
6	Coefficients:
7	(table below)
8	Null deviance: 158.07 on 149 degrees of freedom Residual deviance: 102.81 on 138 degrees of freedom McFadden R-Squared: 0.3496, Akaike Information Criterion 126.8
9	Number of Fisher Scoring iterations: 6
10	*Type II Analysis of Deviance Tests*

Min	1Q	Median	3Q	Max
-2.3291	-0.5605	-0.2097	-0.0588	2.9881

	Estimate	Std. Error	z value	Pr(>|z|)	
(Intercept)	-4.307e+00	1.423856	-3.02456	0.00249	**
Account.BalanceSome Balance	-2.245e+00	0.600820	-3.73662	0.00019	***
Payment.Status.of.Previous.CreditPaid Up	2.416e+00	0.739439	3.26717	0.00109	**
Payment.Status.of.Previous.CreditSome Problems	4.203e+00	1.218397	3.44995	0.00056	***
PurposeNew car	-2.923e-01	0.850859	-0.34351	0.73121	
PurposeOther	2.205e-01	1.931701	0.11414	0.90913	
PurposeUsed car	-2.899e+00	1.276415	-2.27106	0.02314	*
Credit.Amount	-2.149e-06	0.000101	-0.02127	0.98303	
Length.of.current.employment4-7 yrs	9.097e-01	0.895711	1.01564	0.3098	
Length.of.current.employment< 1yr	1.172e+00	0.744883	1.57347	0.11561	
Instalment.per.cent	1.192e-01	0.243666	0.48902	0.62482	
Most.valuable.available.asset	4.542e-01	0.277328	1.63785	0.10145	

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

Figure 2: Logistic Regression Model Report

2-Dession Tree Model

Using Credit Application Result as the target variable ,Account Balance, Payment Status and credit amount has the most top variables. The overall Accuracy is 86%.

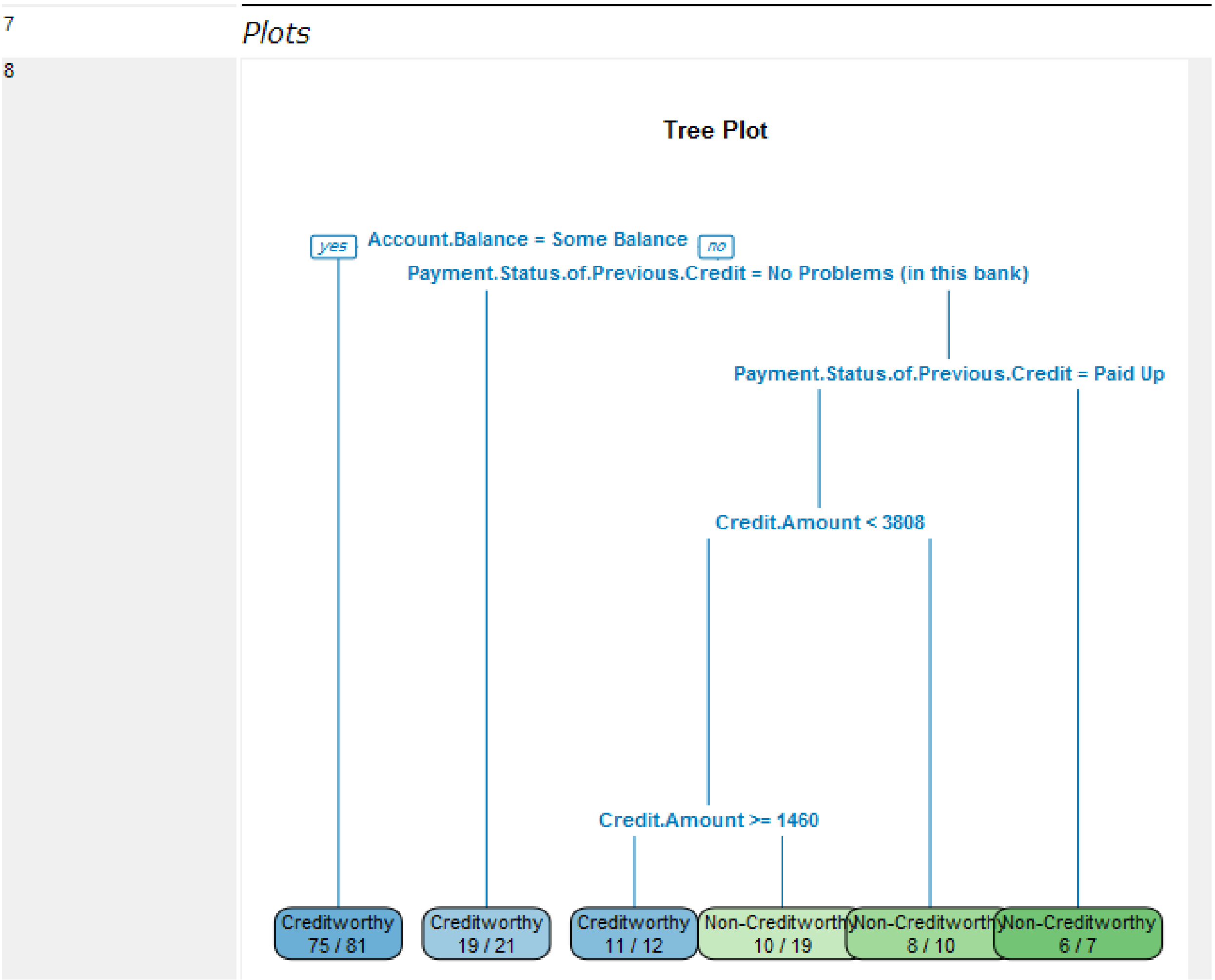


Figure 3: Decision Tree Plot

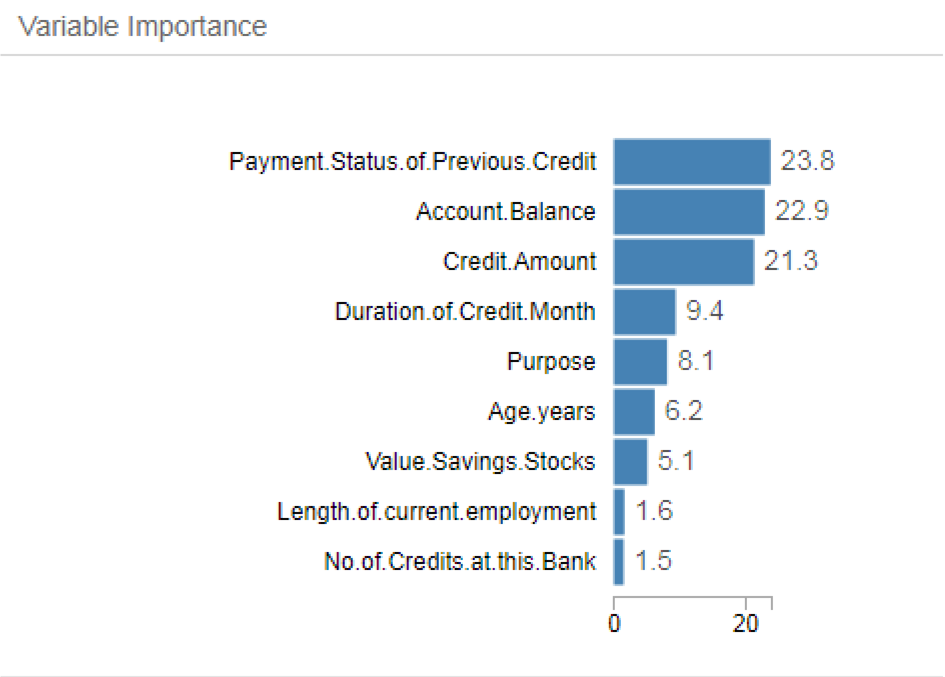
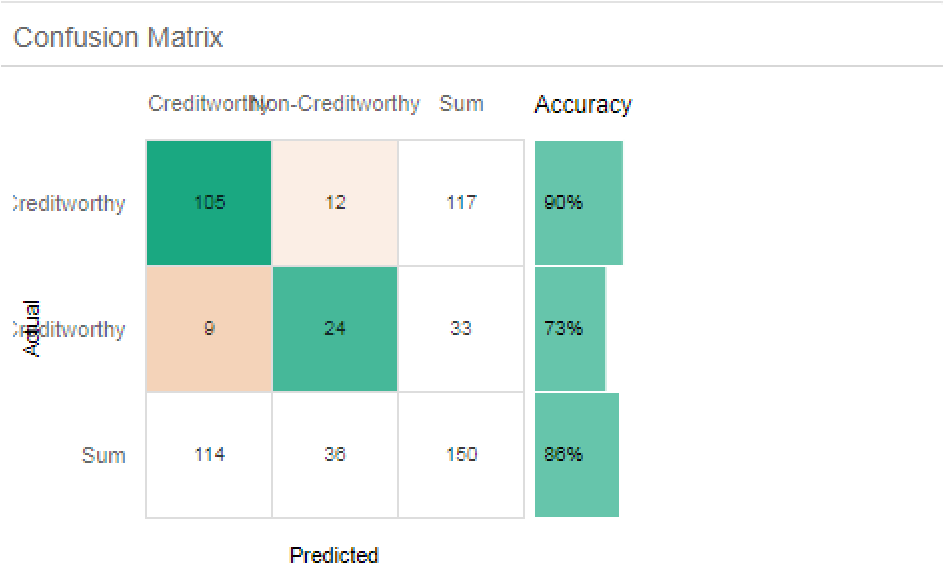


Figure 4:Variable Importance



Confusion Matrix

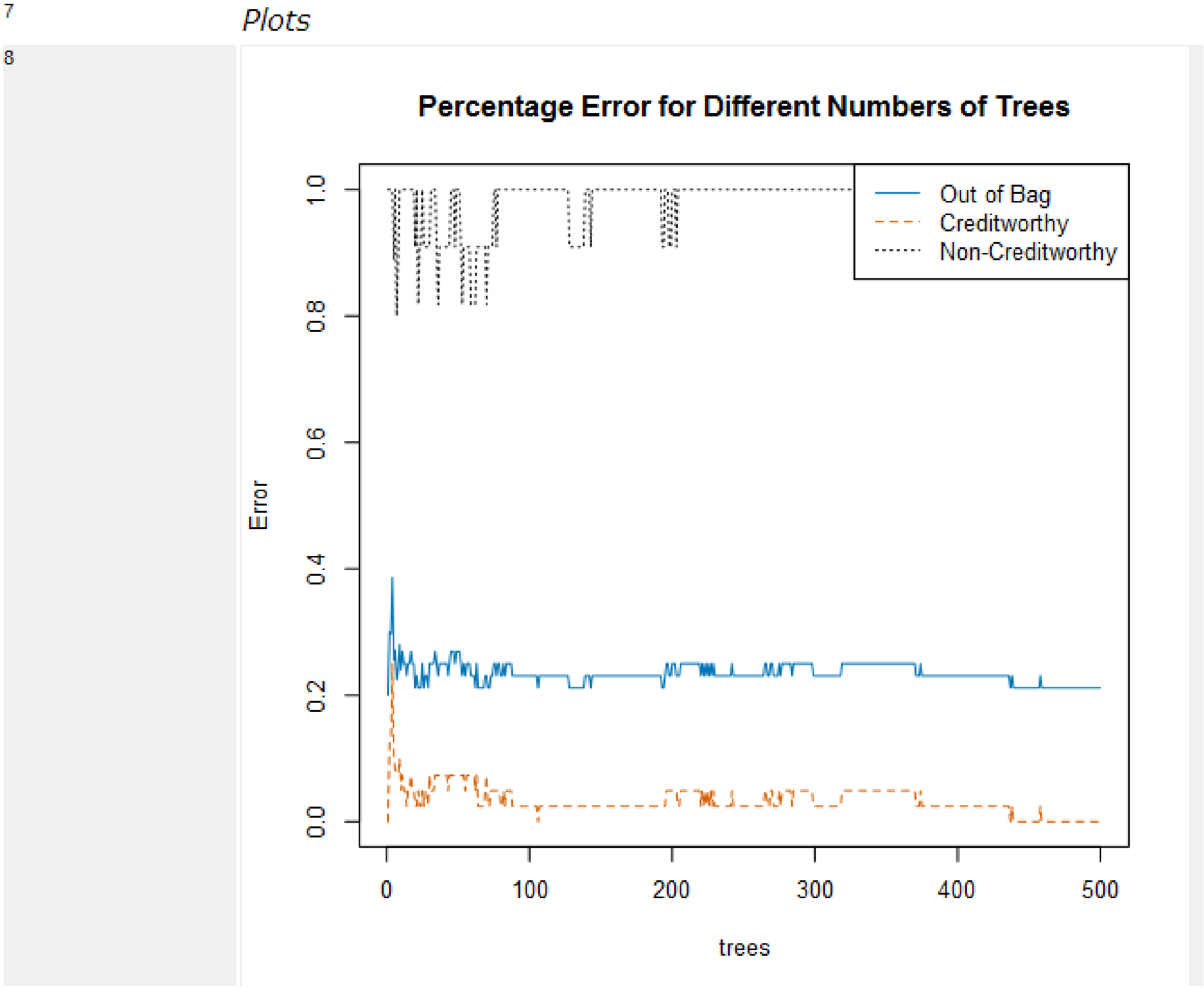
Actual	Creditworthy	Non-Creditworthy	Sum	Accuracy
:reditworthy	105	12	117	90%
:reditworthy	9	24	33	73%
Sum	114	36	150	86%

Predicted

Figure 5:Confusion Matrix

3- Forest Model

Using Credit Application Result as the target variables, Credit Amount, Age Years and Account Balance are the 3 most important variables.



7 Plots

8

Figure6 :Percentage Error for Different Numbers of Trees

9

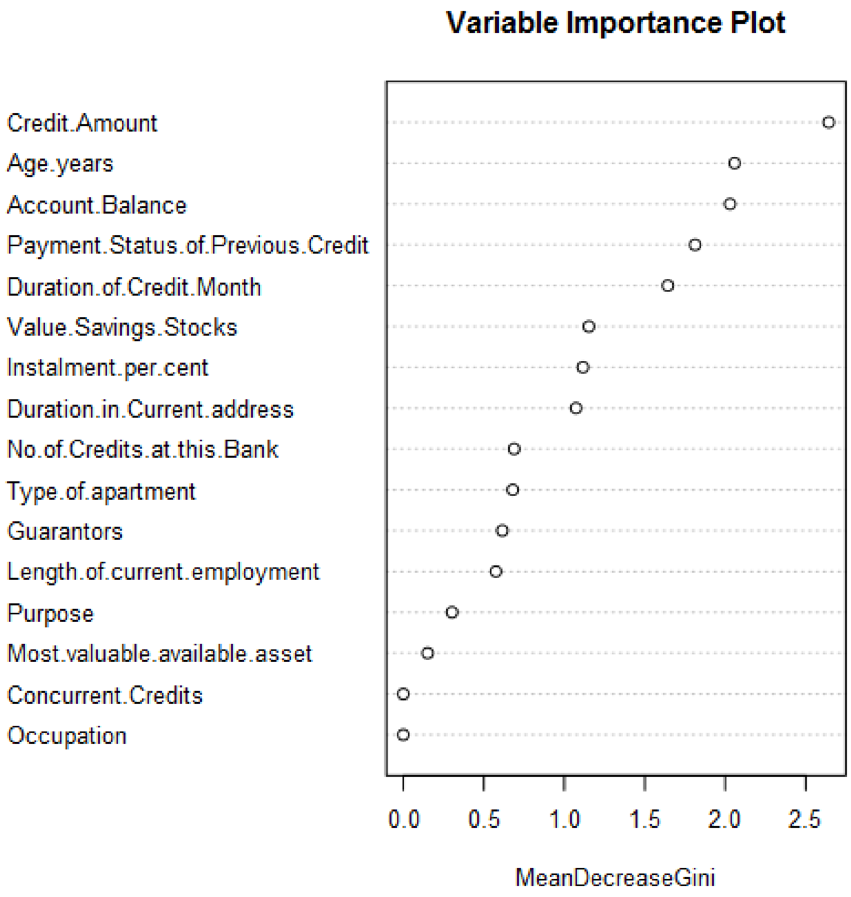
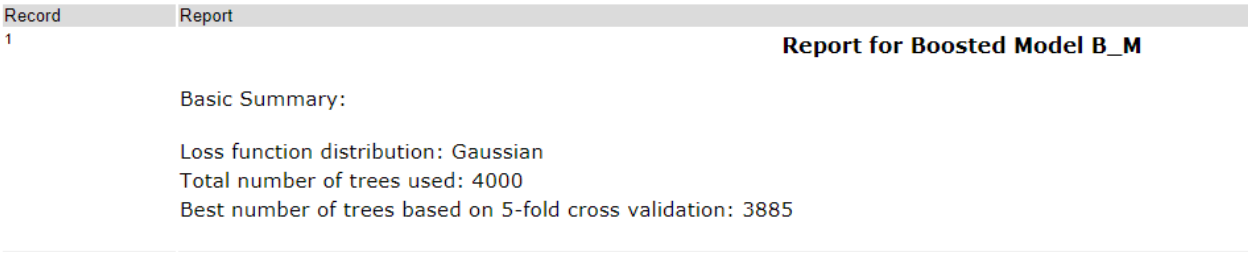


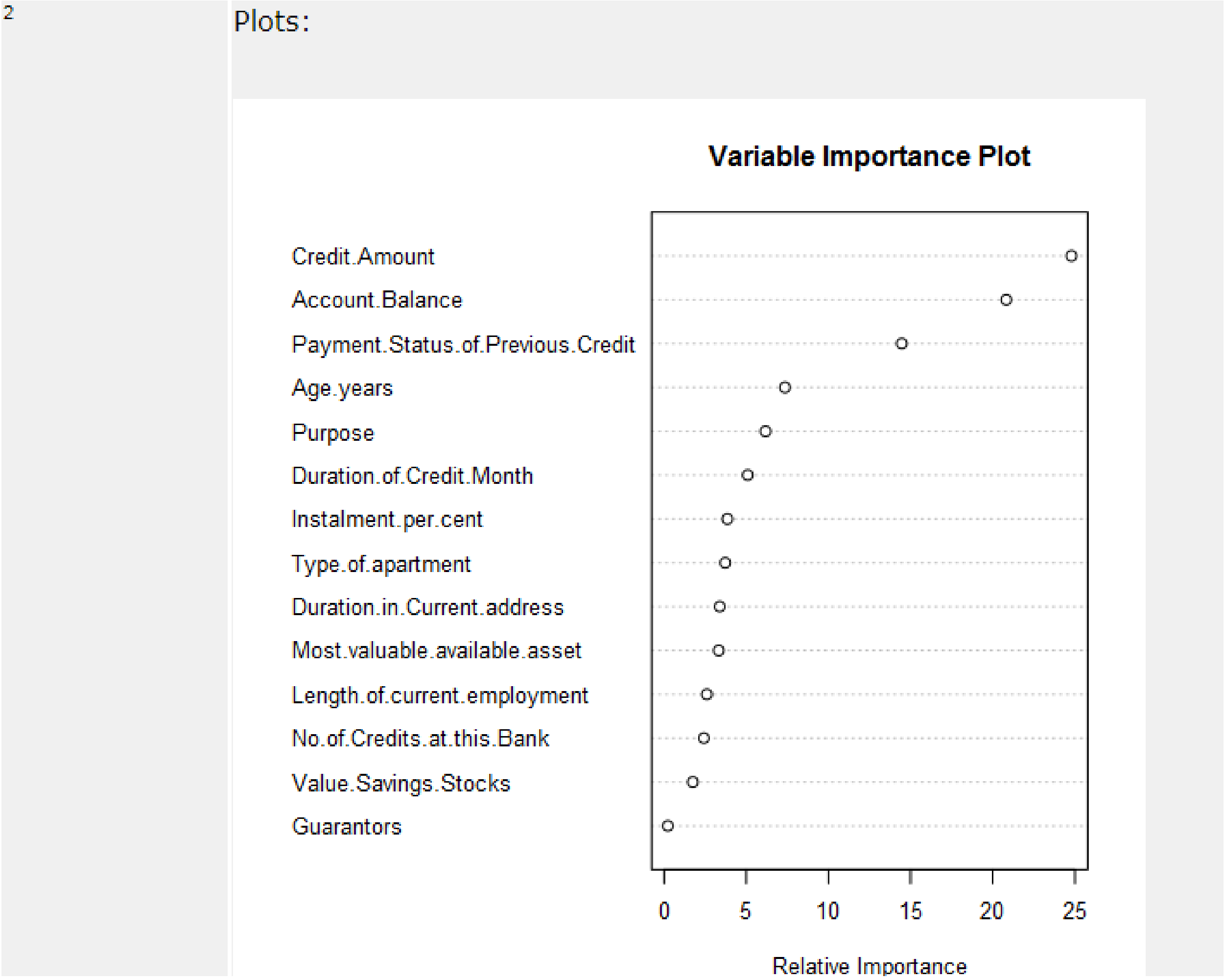
Figure7:Variable Impotence Plot for Forest Model

4- Boosted Tree Model

Account Balance and Credit Amount are the most significant variables



Record	Report
1	**Report for Boosted Model B_M** Basic Summary: Loss function distribution: Gaussian Total number of trees used: 4000 Best number of trees based on 5-fold cross validation: 3885



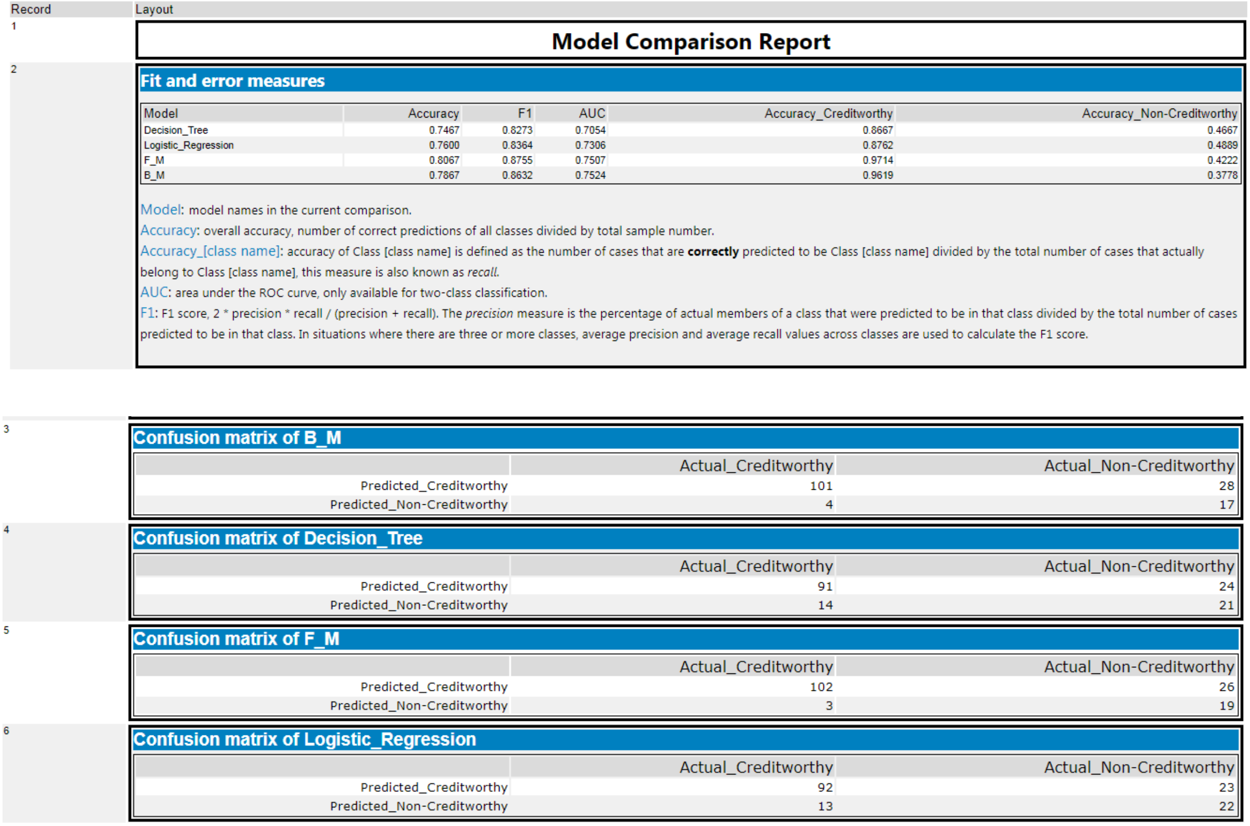
2

Plots:

Figure 8: Variable Impotence Plot for Boosted Tree Model

Step 4: Writeup

Forest model is chosen as it offers the highest accuracy at 80% against validation set.
Its accuracies for creditworthy and non-creditworthy are among the highest of all.
The accuracy difference between creditworthy and non-creditworthy are also comparable which makes it least bias towards any decisions.



Model Comparison Report

Fit and error measures

Model	Accuracy	F1	AUC	Accuracy_Creditworthy	Accuracy_Non-Creditworthy
Decision_Tree	0.7467	0.8273	0.7054	0.8667	0.4667
Logistic_Regression	0.7600	0.8364	0.7306	0.8762	0.4889
F_M	0.8067	0.8755	0.7507	0.9714	0.4222
B_M	0.7867	0.8632	0.7524	0.9619	0.3778

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 B_M

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

Confusion matrix of Decision_Tree

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

Confusion matrix of F_M

	Actual_Creditworthy	Actual_Non-Creditworthy
Predicted_Creditworthy	102	26
Predicted_Non-Creditworthy	3	19

Confusion matrix of Logistic_Regression

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

Figure 9: Model Comparison Report for all 4 classification models

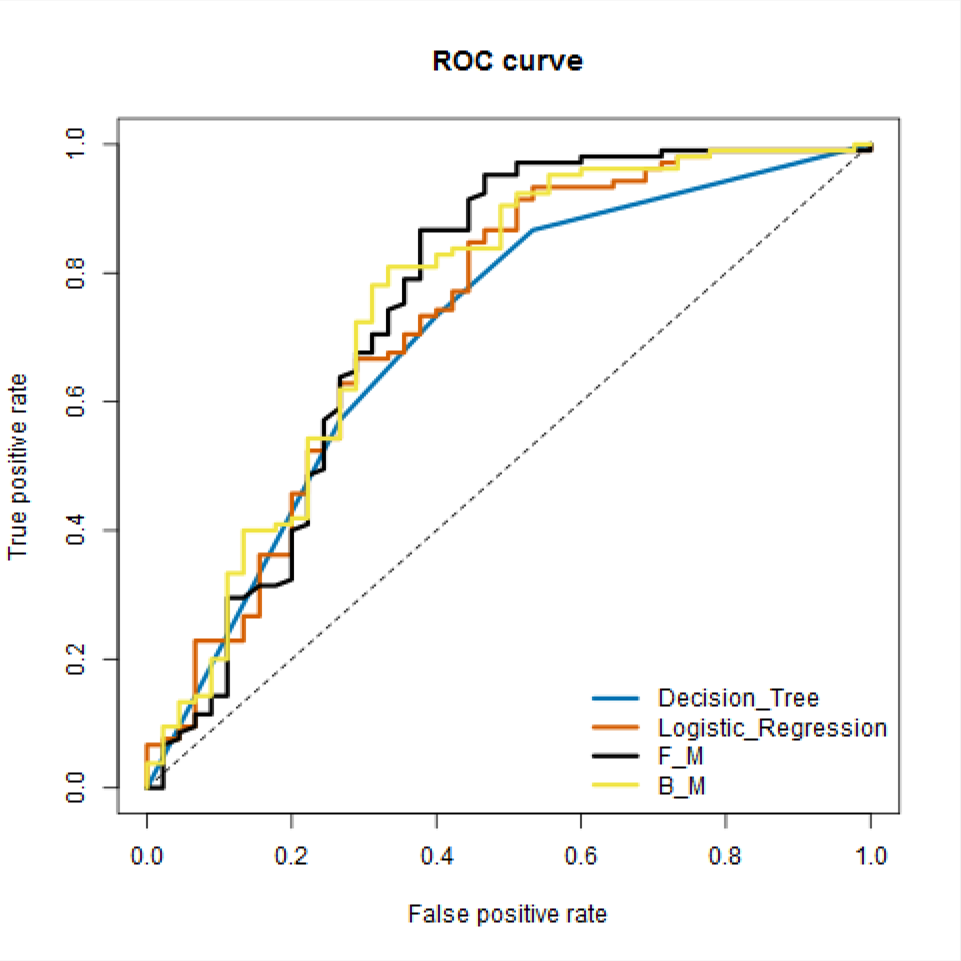


Figure10 :ROC Curve for all 4 classification models

Alteryx Workflow:

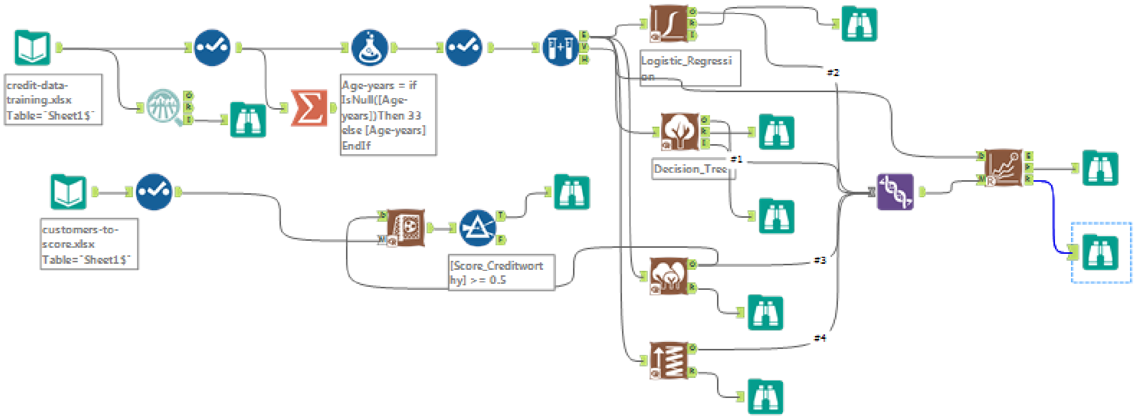

credit-data-training.xlsx Table=`Sheet1$`
Age-years = if IsNull([Age-years])Then 33 else [Age-years] EndIf
Logistic_Regression
Decision_Tree
#1
#2
#3
#4
customers-to-score.xlsx Table=`Sheet1$`
[Score_Creditworthy] >= 0.5

Figure 11: Alteryx Workflow