Project: Predicting Default Risk

Step 1: Business and Data Understanding

The Business Problem
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

1. What decisions need to be made?

The objective is to identify whether an applicant is creditworthy or eligible for loan approval.

2. What data is needed to inform those decisions?

Data on past applications **(Credit Data Training file)** and list of customers that need to be processed in the next few days **(Customers to Score file)** such as *Account Balance*, *Credit Amount*, *Purpose*, *Duration of Credit Month*, *Value/Savings/Stocks*, *Type of Apartment*, *Length of Current Employment*, *Age_Years*, *Payment Status of Previous Credit* etc, are needed to build a predictive model to analyze and categorize customers into creditworthy and non-creditworthy.

3. 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 logistic regression, decision tree, forest model and boosted tree are needed for the analysis of creditworthy customers.

Step 2: Building the Training Set

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

Association Analysis

An association analysis shows there are no numerical variables that are highly correlated with each other (higher than 0.70).

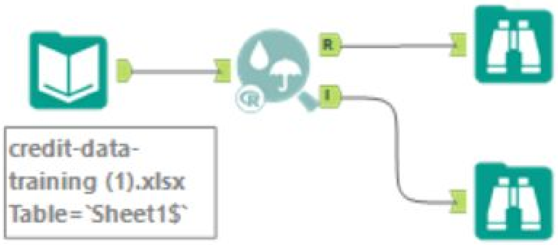


Figure 1: Alteryx workflow for Pearson correlation analysis

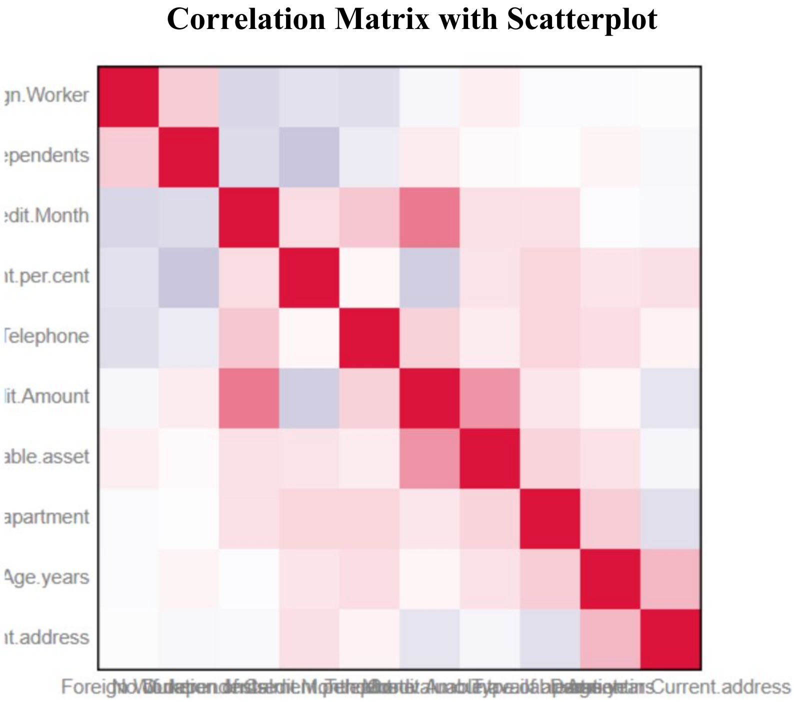


Figure 2: Correlation Matrix of variables - Pearson (I output)

Field Summary Analysis

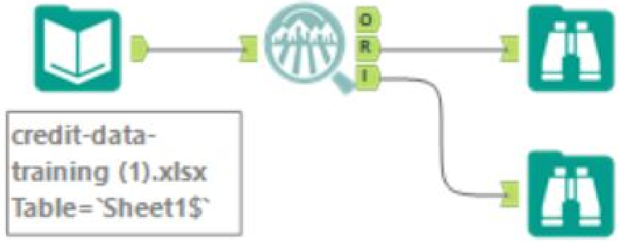


Figure 3: Alteryx Workflow for Field Summary analysis

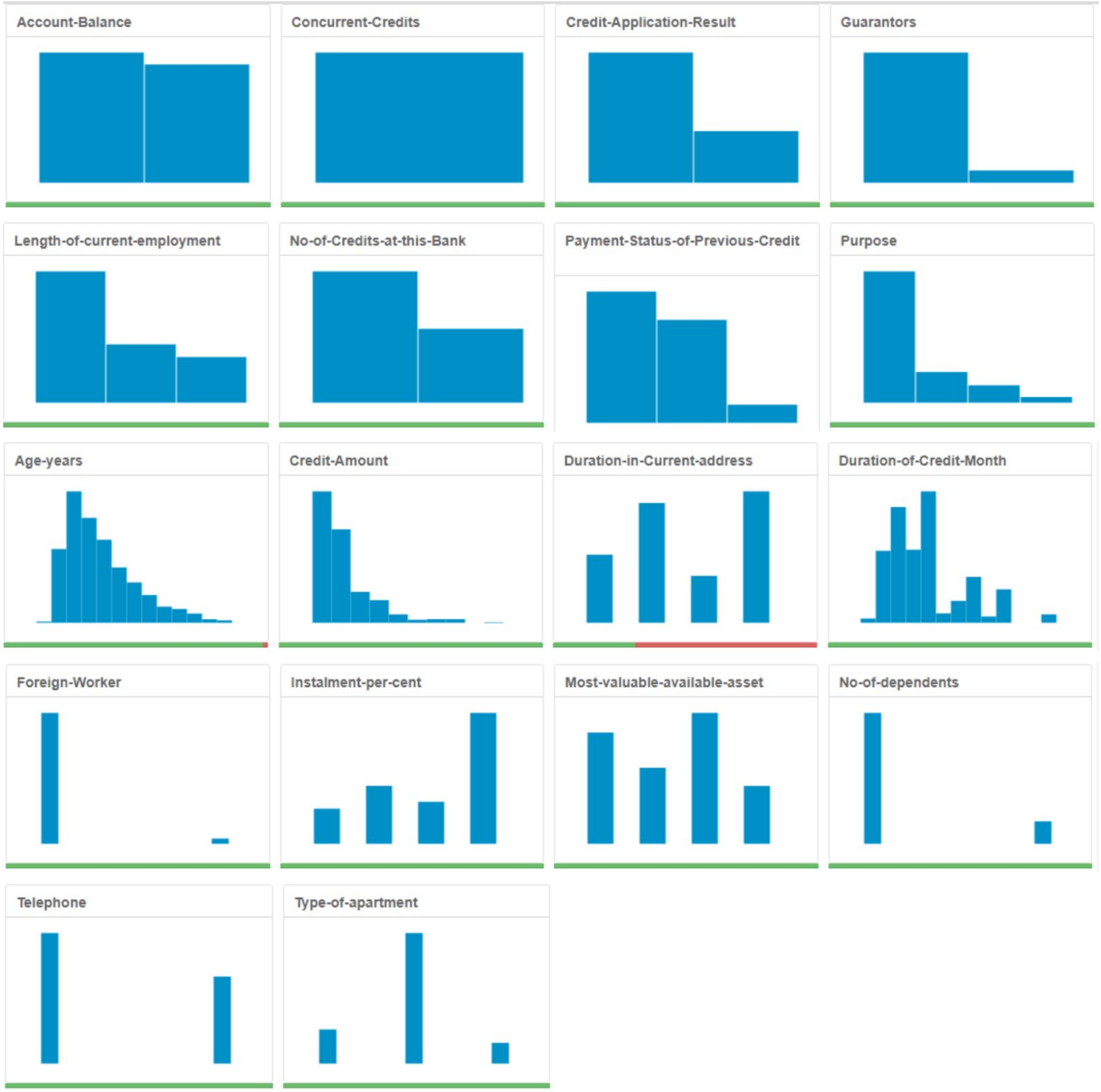
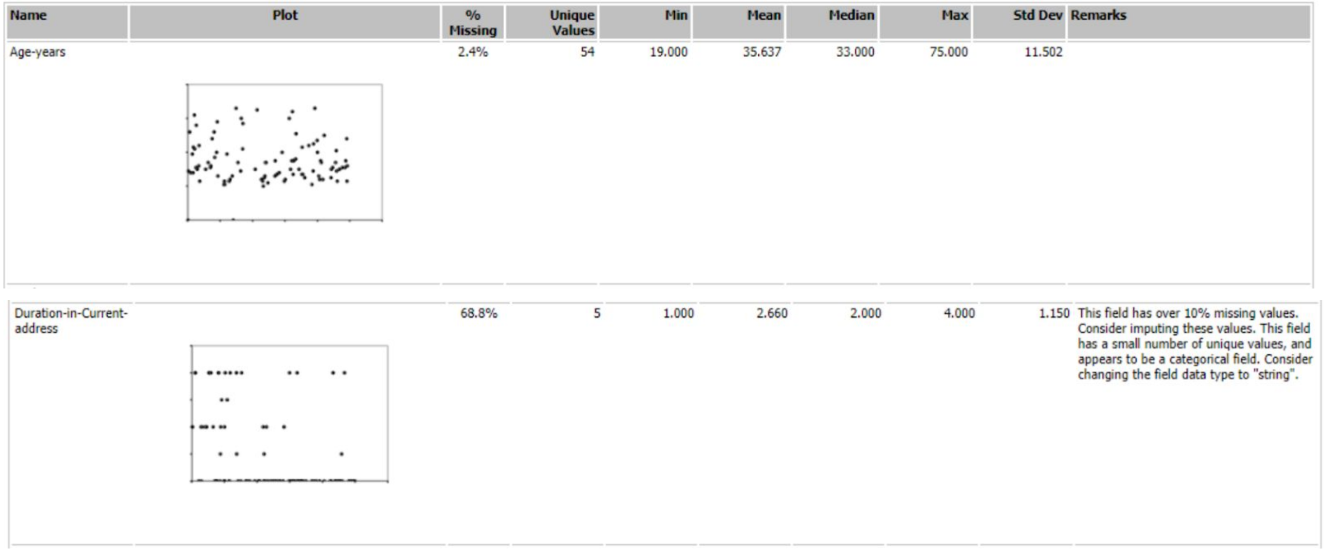


Figure 4: Field Summary of all variables (I output)



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	
Duration-in-Current-address		68.8%	5	1.000	2.660	2.000	4.000	1.150	This field has over 10% missing values. Consider imputing these values. This field has a small number of unique values, and appears to be a categorical field. Consider changing the field data type to "string".

Figure 5: Field Summary (R output) showing missing data

Removed field	Reason
Duration in Current Address	69% missing data
Concurrent Credit	Only one type of data for entire field
Occupation	Only one type of data for entire field
Telephone	Irrelevant data for determining the creditworthiness of customers
No of Dependents	Low variability; heavily skewed
Guarantors	Low variability; heavily skewed
Foreign Workers	Low variability; heavily skewed

Age Years has 2.4% missing data so it is appropriate to impute the missing data with the median age of 33. Median age is considered because data for age is skewed to the left.

Step 3: Train you 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?

1. Logistic Regression (stepwise)

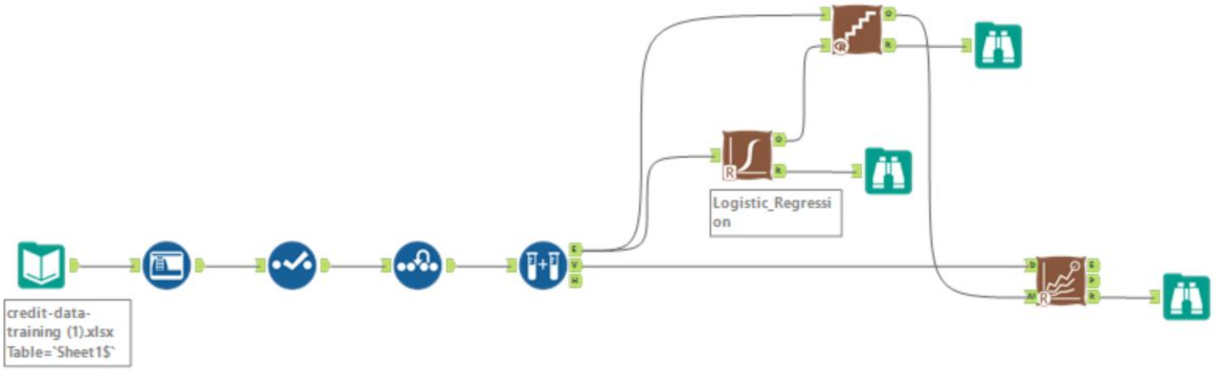
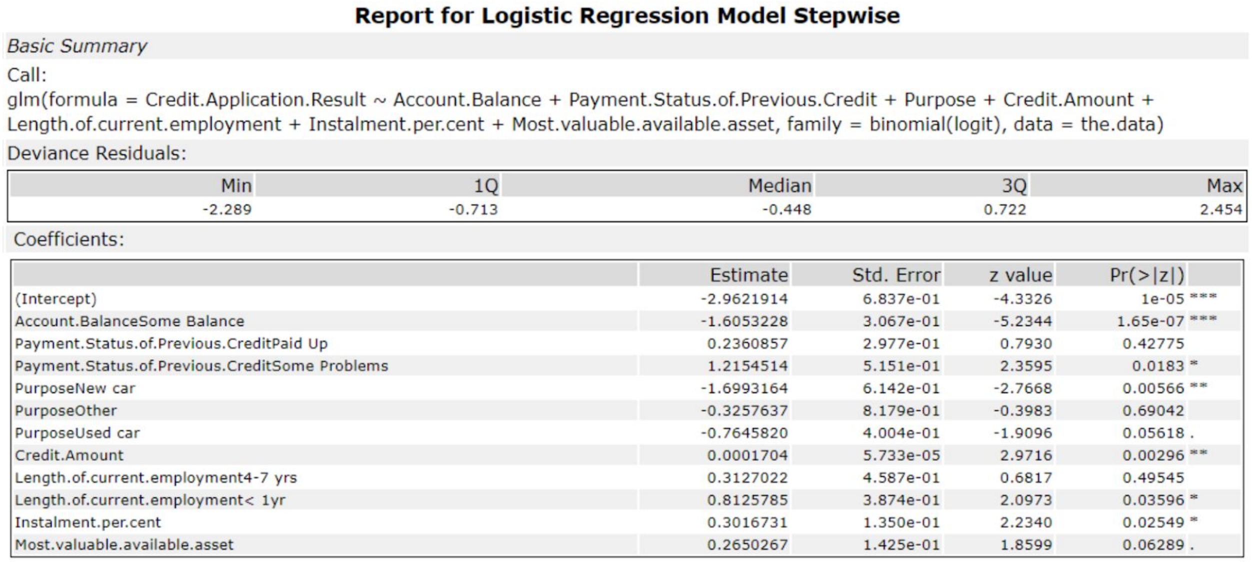


Figure 6: Alteryx workflow for Logistic-Stepwise Regression



Report for Logistic Regression Model Stepwise

Basic Summary

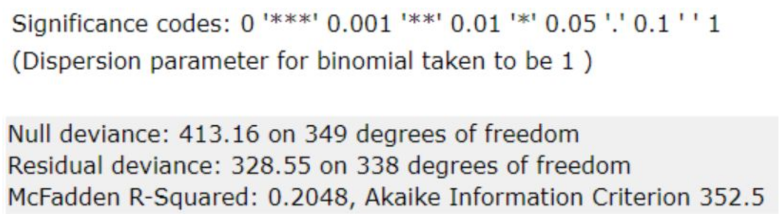
Call:
glm(formula = Credit.Application.Result ~ Account.Balance + Payment.Status.of.Previous.Credit + Purpose + Credit.Amount + Length.of.current.employment + Instalment.per.cent + Most.valuable.available.asset, family = binomial(logit), data = the.data)

Deviance Residuals:

Min	1Q	Median	3Q	Max
-2.289	-0.713	-0.448	0.722	2.454

Coefficients:

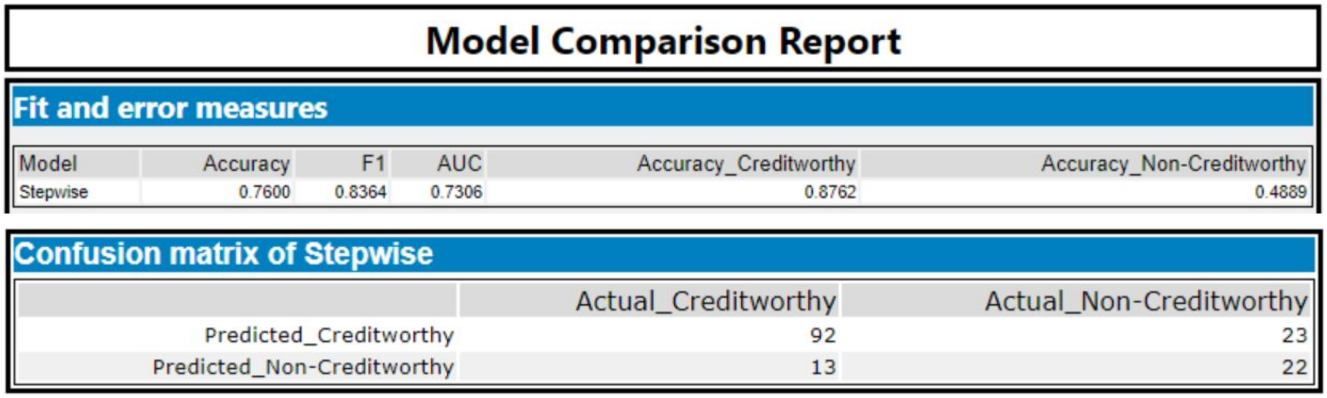
	Estimate	Std. Error	z value	Pr(>|z|)	
(Intercept)	-2.9621914	6.837e-01	-4.3326	1e-05	***
Account.BalanceSome Balance	-1.6053228	3.067e-01	-5.2344	1.65e-07	***
Payment.Status.of.Previous.CreditPaid Up	0.2360857	2.977e-01	0.7930	0.42775	
Payment.Status.of.Previous.CreditSome Problems	1.2154514	5.151e-01	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 )

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
```

Figure 7: Summary Report for Stepwise Logistic Regression Model (R output)



Model Comparison Report

Fit and error measures

Model	Accuracy	F1	AUC	Accuracy_Creditworthy	Accuracy_Non-Creditworthy
Stepwise	0.7600	0.8364	0.7306	0.8762	0.4889

Confusion matrix of Stepwise

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

Figure 8: Model Comparison Report for Stepwise Logistic Regression Model (R output)

The Logistic Regression-stepwise summary report shows that *Credit Application Result* is the target variable and *Account Balance*, *Payment Status of Previous Credit*, *Purpose*, *Credit Amount*, *Length of current employment*, and *Instalment percent* are the 4 most significant predictor variables with p-value of less than 0.05. The R-Squared value sounds not good at all, with a value of 0.2048

The model comparison report for the Logistics-Stepwise Regression shows that this model has an overall accuracy of 76.0%. Though the accuracy for creditworthiness is high at 87.2%, the accuracy for non-creditworthiness is low at 48.9%. Thus, the model is biased towards predicting customers as creditworthy.

b. Decision Tree

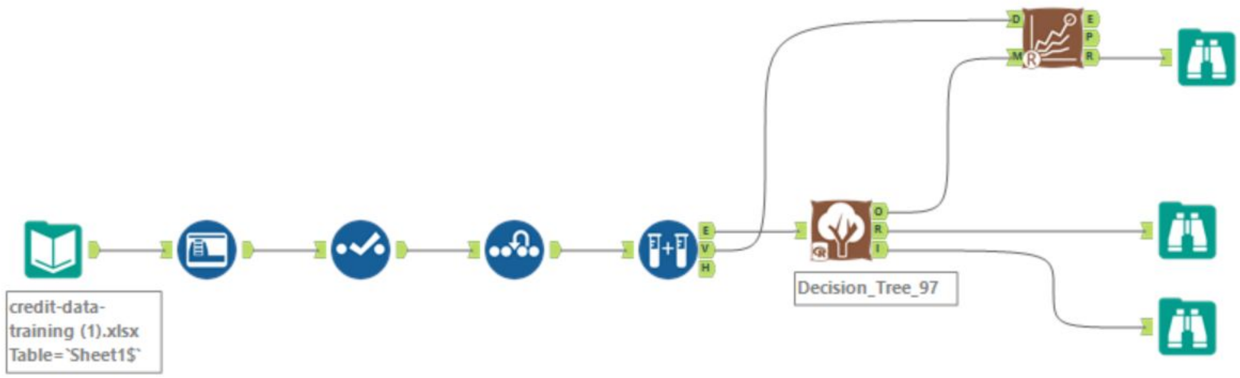
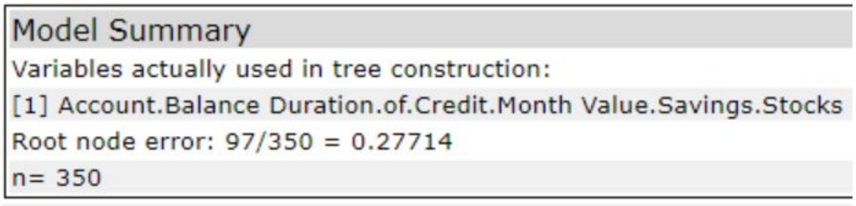
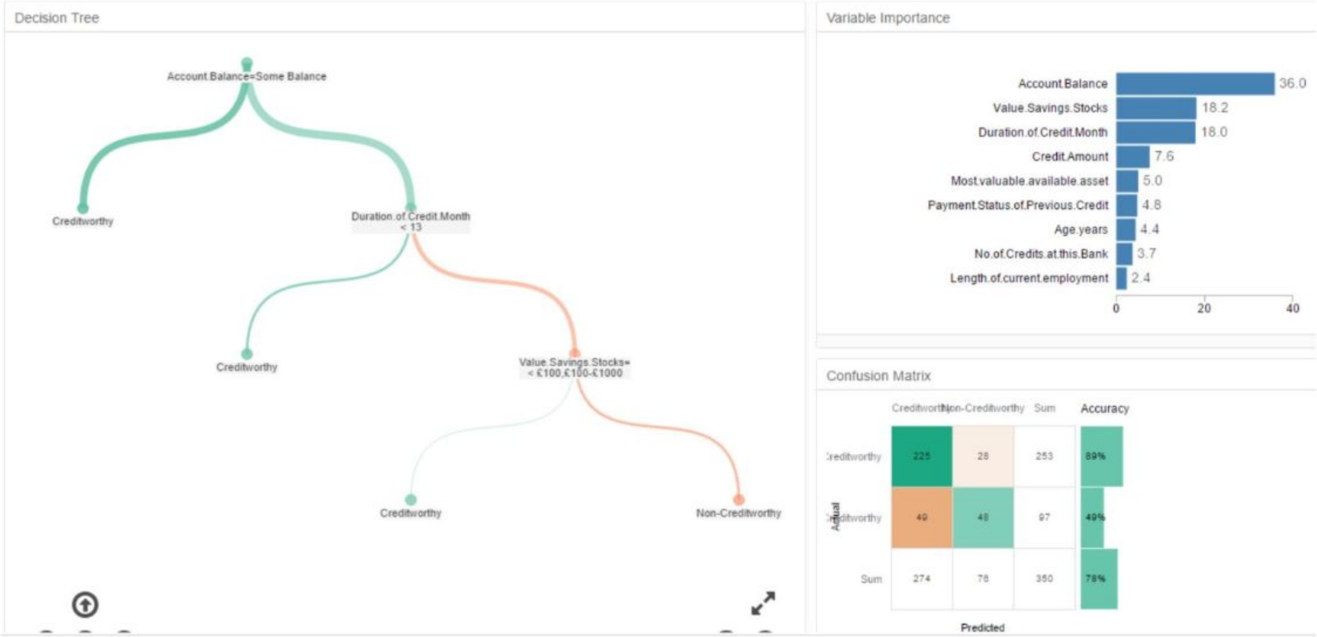


Figure 9: Alteryx workflow for Decision-Tree model



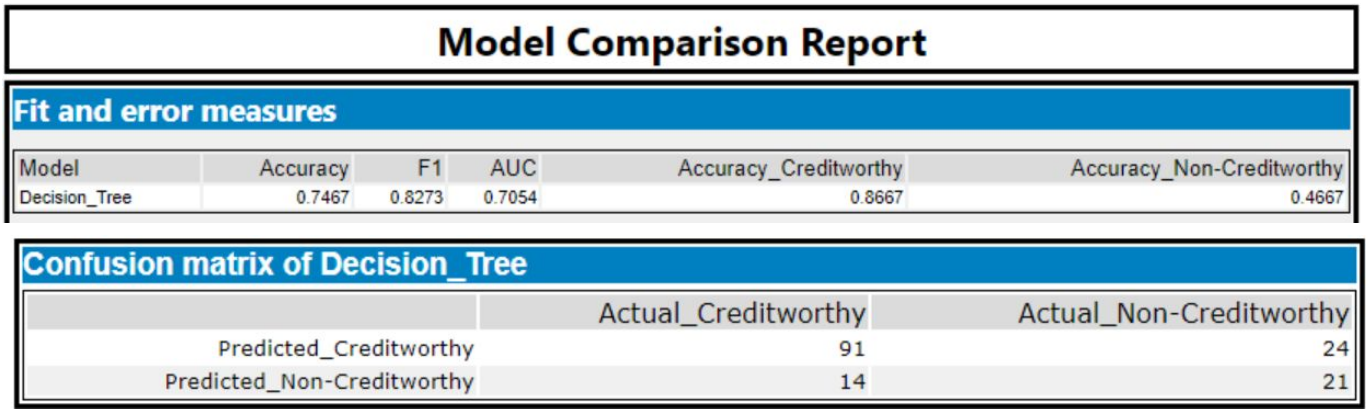
```
Model Summary
Variables actually used in tree construction:
[1] Account.Balance Duration.of.Credit.Month Value.Savings.Stocks
Root node error: 97/350 = 0.27714
n= 350
```

Figure 10: Summary Report for Decision Tree (R output)



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

Figure 11: Decision Tree, Variable Importance and Confusion Matrix (I output)



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

Confusion matrix of Decision_Tree

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

Figure 12: Model Comparison Report for Decision Tree (R output)

Summary report: testing the Decision Tree model into our dataset we could see that even though the Root node error is quite high it still under 30%, which is considered as an acceptable error.

Variable Importance Plot: *Credit Application Result* is the target variable and *Account Balance*, *Value Savings Stocks*, and *Duration of Credit Month* are the 3 most significant variables --- according to the interactive report of the DT tool.

Confusion Matrix Plot: When we are validating our model against itself with the Confusion Matrix, we can see that the Sum of Accuracy is 78%, classifying it as a reliable model.

The model comparison report for the Decision-Tree model shows that this model has an overall accuracy of 74.7%. The accuracy for creditworthiness is 86.7% whereas the accuracy for non-creditworthiness is 46.7%. This particular model has bias towards correctly predicting creditworthy individuals because its accuracy in this segment is way higher than in the ot

c. Forest Model

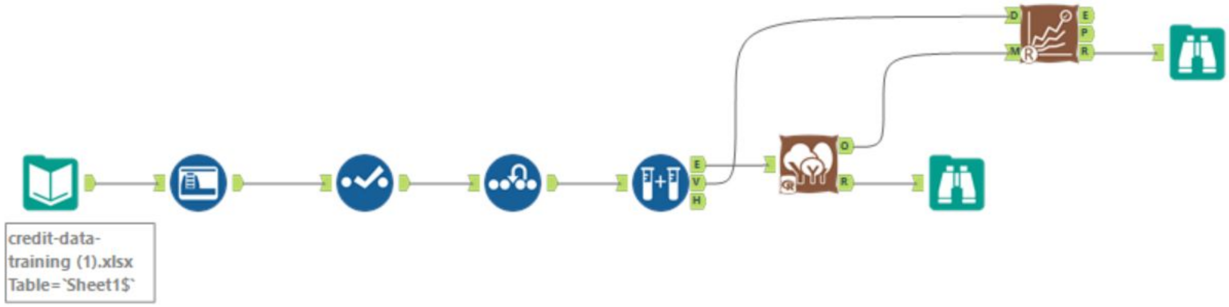
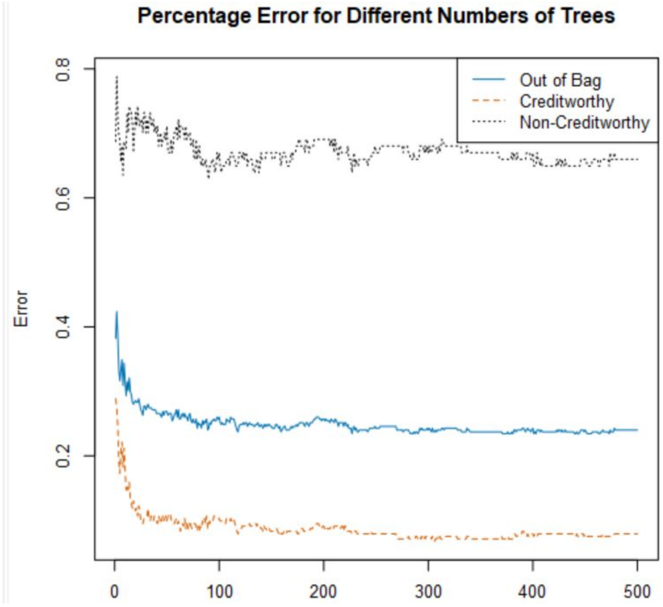


Figure 13: Alteryx Workflow for Forest-tree Model



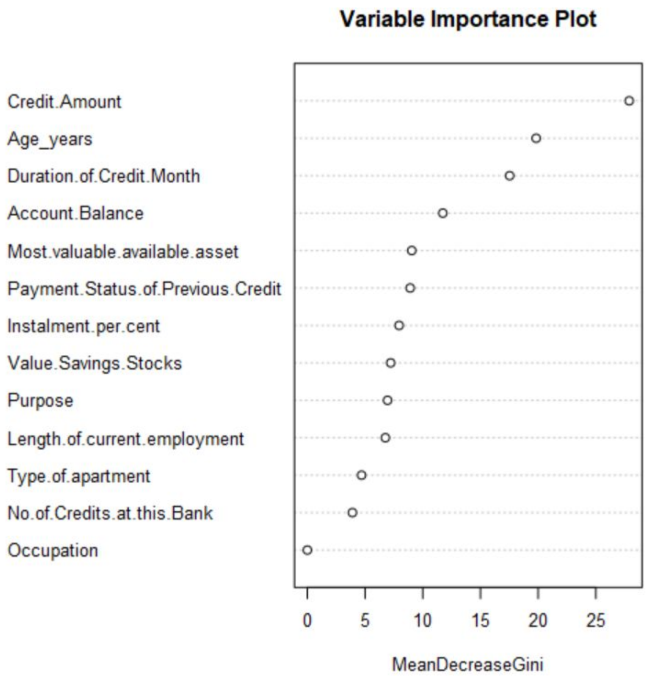
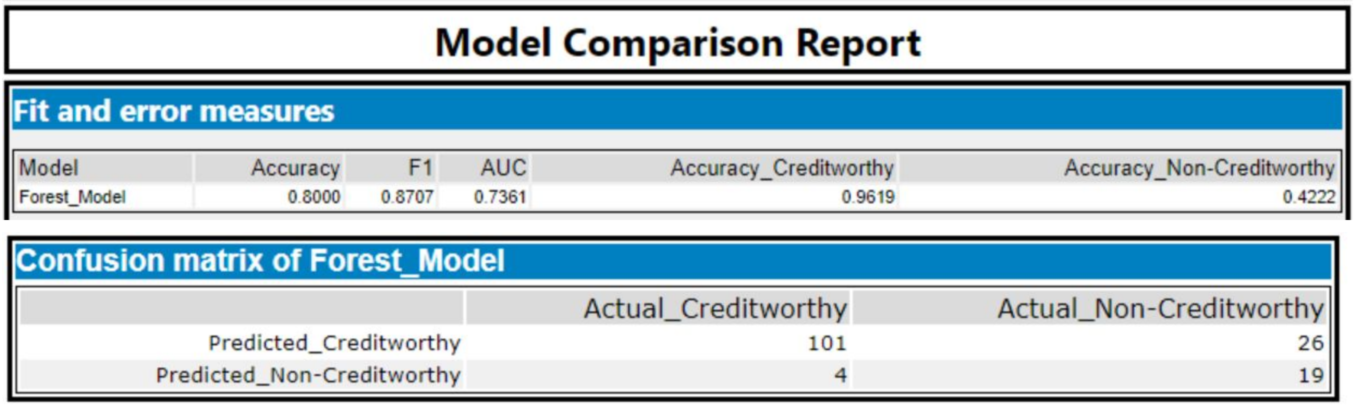


Figure 14: Percentage Error for Different Number of Trees and Variable Importance Plot (FM - R output)



Model Comparison Report

Fit and error measures

Model	Accuracy	F1	AUC	Accuracy_Creditworthy	Accuracy_Non-Creditworthy
Forest_Model	0.8000	0.8707	0.7361	0.9619	0.4222

Confusion matrix of Forest_Model

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

Figure 15: Model Comparison Report for Forest Model (R output)

Variable Importance Plot: *Credit Application Result* is the target variable and *Credit Amount*, *Age Years*, and *Duration of Credit Month* are the 3 most significant variables.

The model comparison report for the Forest Model shows that this model has an overall accuracy of 80.0%. The accuracy for creditworthiness is 79.5% whereas the accuracy for non-creditworthiness is 82.6%. The accuracies are comparable and the model is not substantially biased in its predictions for creditworthiness of customers.

d. Boosted Model

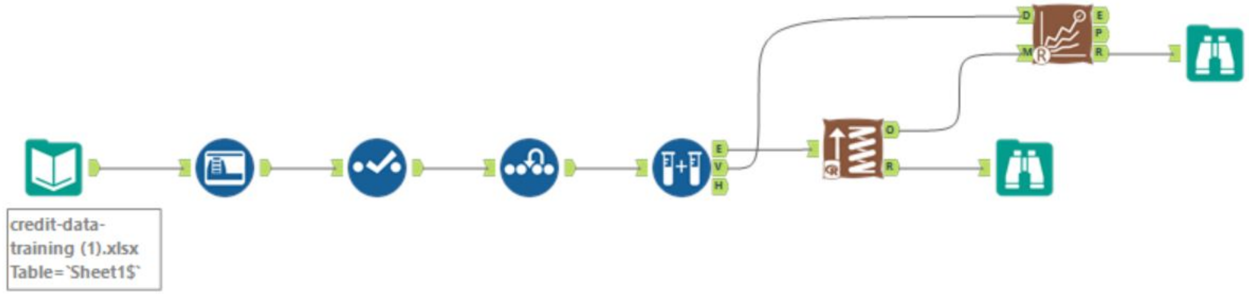
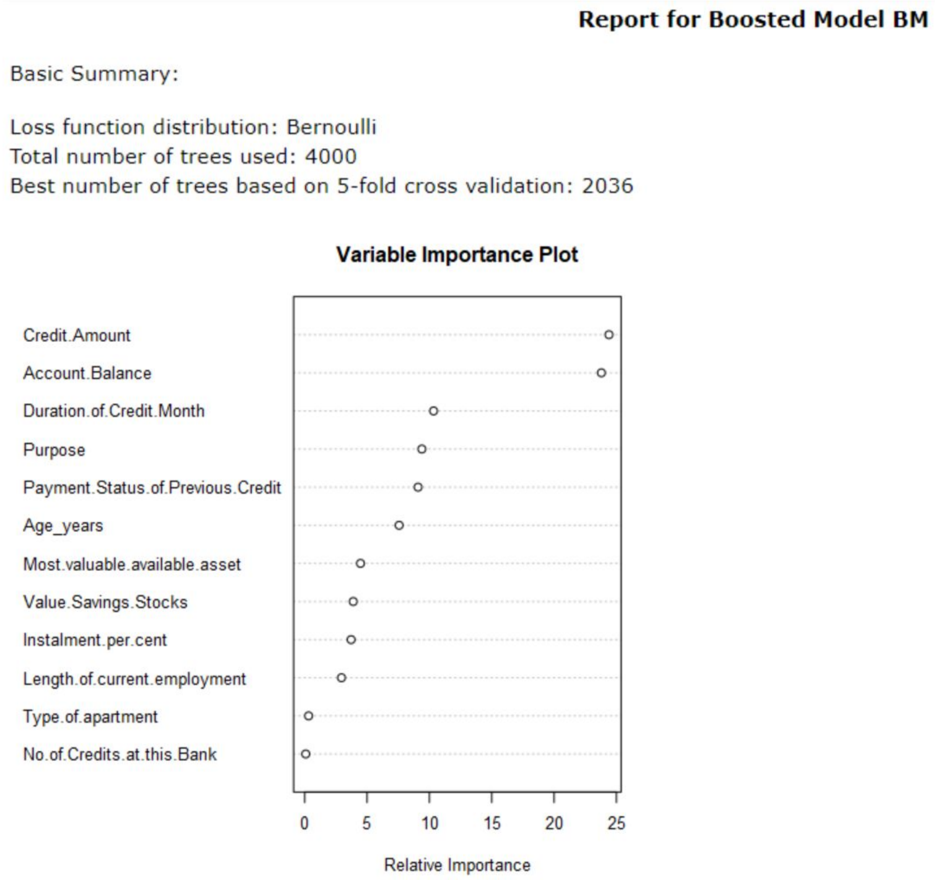


Figure 16: Alteryx workflow for Boosted Model



Report for Boosted Model BM

Basic Summary:

```
Loss function distribution: Bernoulli
Total number of trees used: 4000
Best number of trees based on 5-fold cross validation: 2036
```

Figure 17: Variable Importance Plot for Boosted Model (R output)

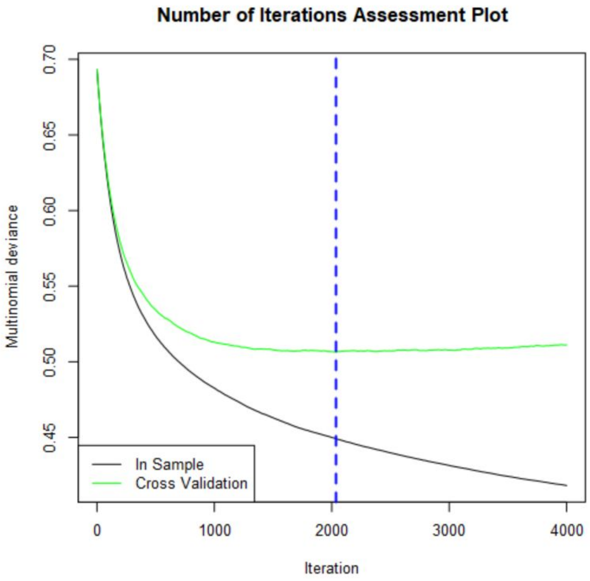
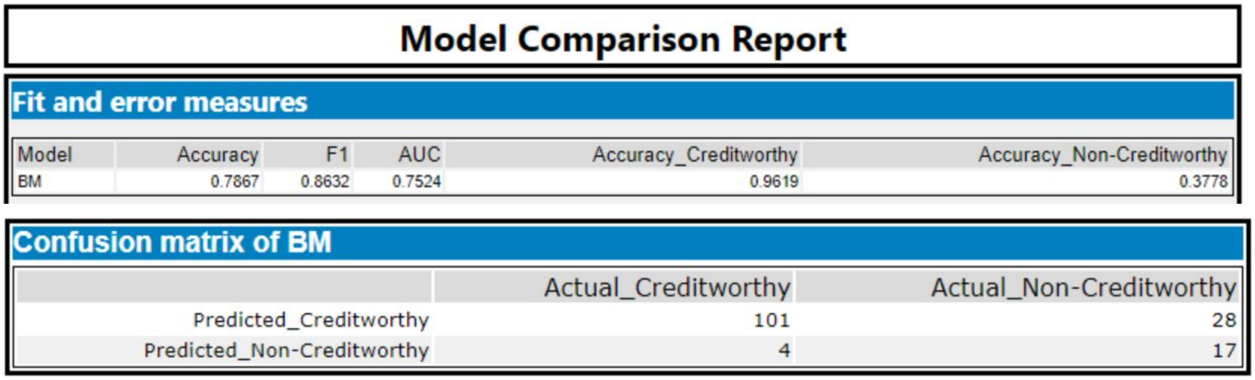


Figure 18: Number of Iteration Assessment Plot (Boosted Model - R output)



Model Comparison Report

Fit and error measures

Model	Accuracy	F1	AUC	Accuracy_Creditworthy	Accuracy_Non-Creditworthy
BM	0.7867	0.8632	0.7524	0.9619	0.3778

Confusion matrix of BM

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

Figure 19: Model Comparison Report for Boosted Model (R output)

The two most significant predictor variables are *Account Balance* and *Credit Amount*. The model comparison reports displays that the overall accuracy is 78.7%. This time, the accuracies for creditworthiness and non-creditworthiness are 78.3% and 82.0% respectively and they both share close percentages—indicating a lack of bias in predicting credit eligibility.

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.

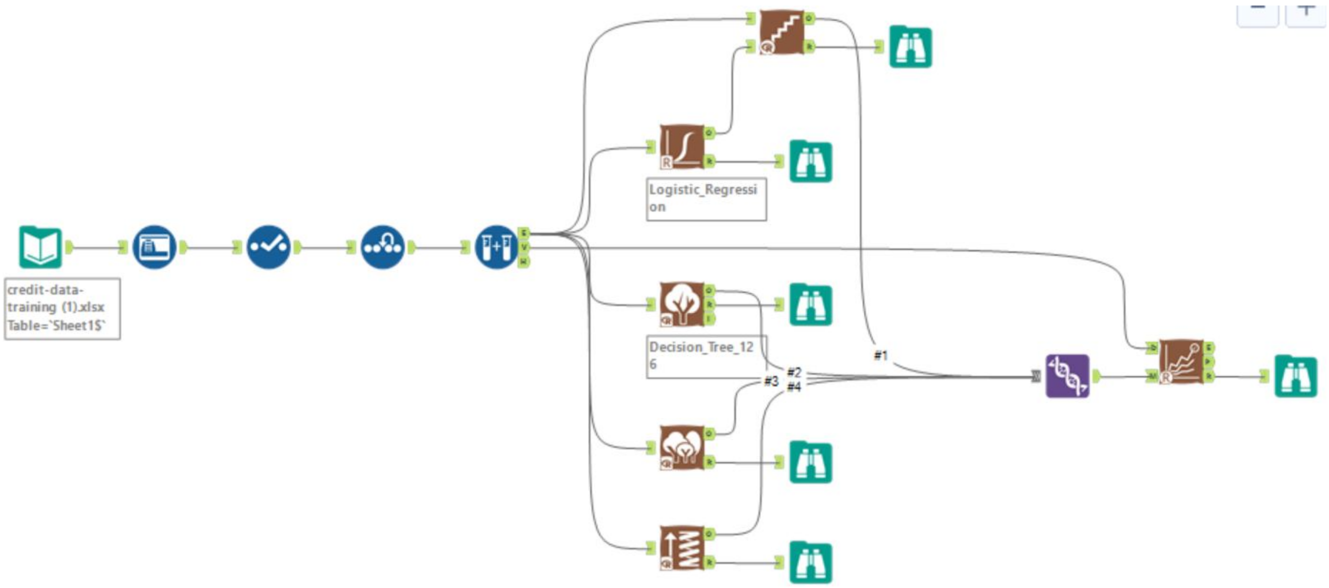
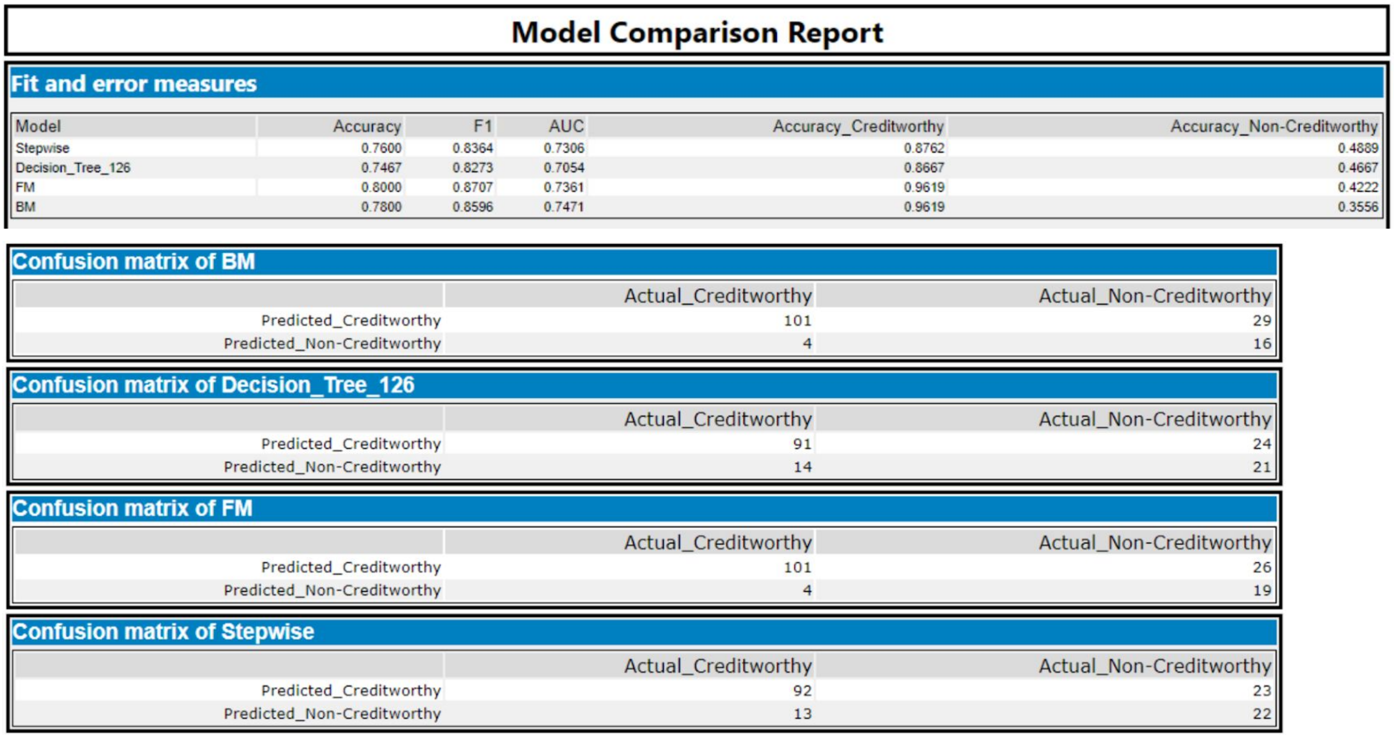


Figure 20: Side-by-side model comparison



Model Comparison Report

Fit and error measures

Model	Accuracy	F1	AUC	Accuracy_Creditworthy	Accuracy_Non-Creditworthy
Stepwise	0.7600	0.8364	0.7306	0.8762	0.4889
Decision_Tree_126	0.7467	0.8273	0.7054	0.8667	0.4667
FM	0.8000	0.8707	0.7361	0.9619	0.4222
BM	0.7800	0.8596	0.7471	0.9619	0.3556

Confusion matrix of BM

	Actual_Creditworthy	Actual_Non-Creditworthy
Predicted_Creditworthy	101	29
Predicted_Non-Creditworthy	4	16

Confusion matrix of Decision_Tree_126

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

Confusion matrix of FM

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

Confusion matrix of Stepwise

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

Figure 21: Side-by-side comparison of models - confusion matrix

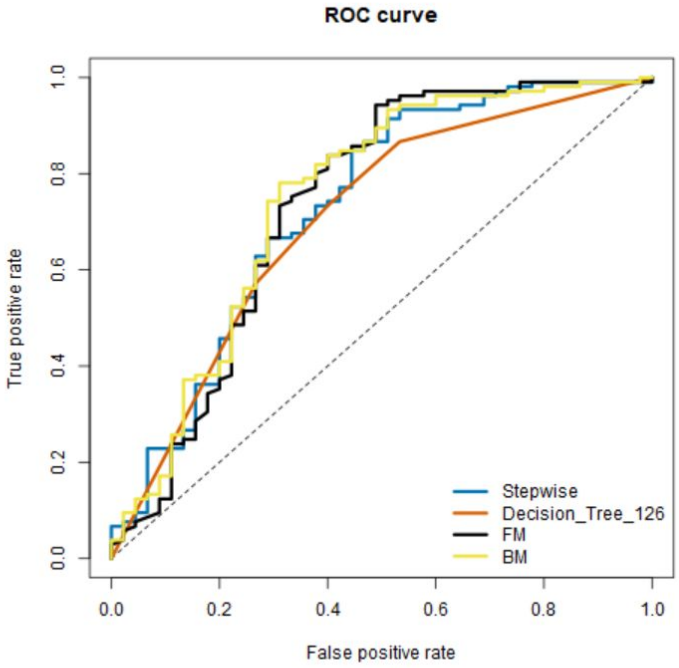
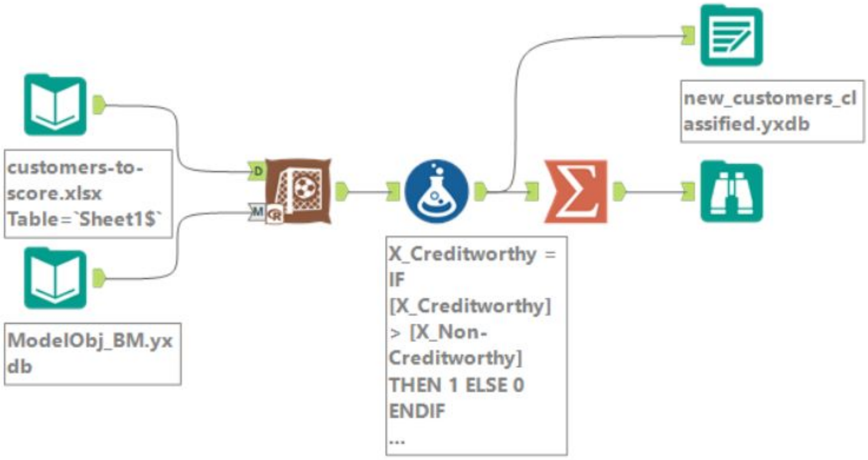


Figure 22: ROC curve for all 4 classification models

Forest model is the best choice because it provides the highest accuracy at 80% against validation set. At 96% and 42% for creditworthy and non-creditworthy respectively, the accuracies for these two groups are the highest compared to the other models. The model is not biased towards a group as the accuracy difference between creditworthy and non-creditworthy are very minimal. Based on our ROC curve, the forest model reaches the "positive rate" at the fastest rate or hugs the most positive side of the graph. These values regarding bias and accuracies are important for a lender and the loan customer since they equalize the opportunities for loan acceptance or denial based on each customer's individual ability to responsibly utilize the loan.


customers-to-score.xlsx Table=`Sheet1$`
ModelObj_BM.yxdb
D
M
R
new_customers_classified.yxdb

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

Figure 23: Scoring the model to predict creditworthiness of new customers

There are **416** creditworthy customers using the forest model to score our new customer dataset.