Project: Creditworthiness

By: Alhanoof Alyabes
Date: 17/4/2018

Step 1: Business and Data Understanding

Key Decisions:

- What decisions needs to be made?

The main decision is to make an analysis and predict if the customers are creditworthy or not to give them a loan.

- What data is needed to inform those decisions?

Different data are required to complete the analysis such as: account balance, credit amount, age years, duration of credit month, payment statues, purpose and value saving stocks

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

Binary mode used to analysis and make the results, such models are: Logistic Regression, Decision Tree, Forest Model, Boosted Model

Step 2: Building the Training Set

By using filed summary of all variables, we can notice that **duration of current address** has about 69% missing data, which more than half of the data, so the decision here is to remove this columns for better analysis. on the other hand, the age column has 2% missing data, which is low, and it is good to keep it and use a median age here.

Furthermore, **(no-of-dependents, foreign workers, guarantors)** are removed, since they have more than 80% at one of their categories (Low Variability), which may skew the result. while **concurrent credit and occupation** are removed, because they have one value. **telephone** column removed because not related to the creditworthy

figure (1) field summary for all variables

while when using the associations analysis for numeric variables, there are no strong relationship (correlation should be at least .70). the highest value .57 between credit amount and duration of credit month.

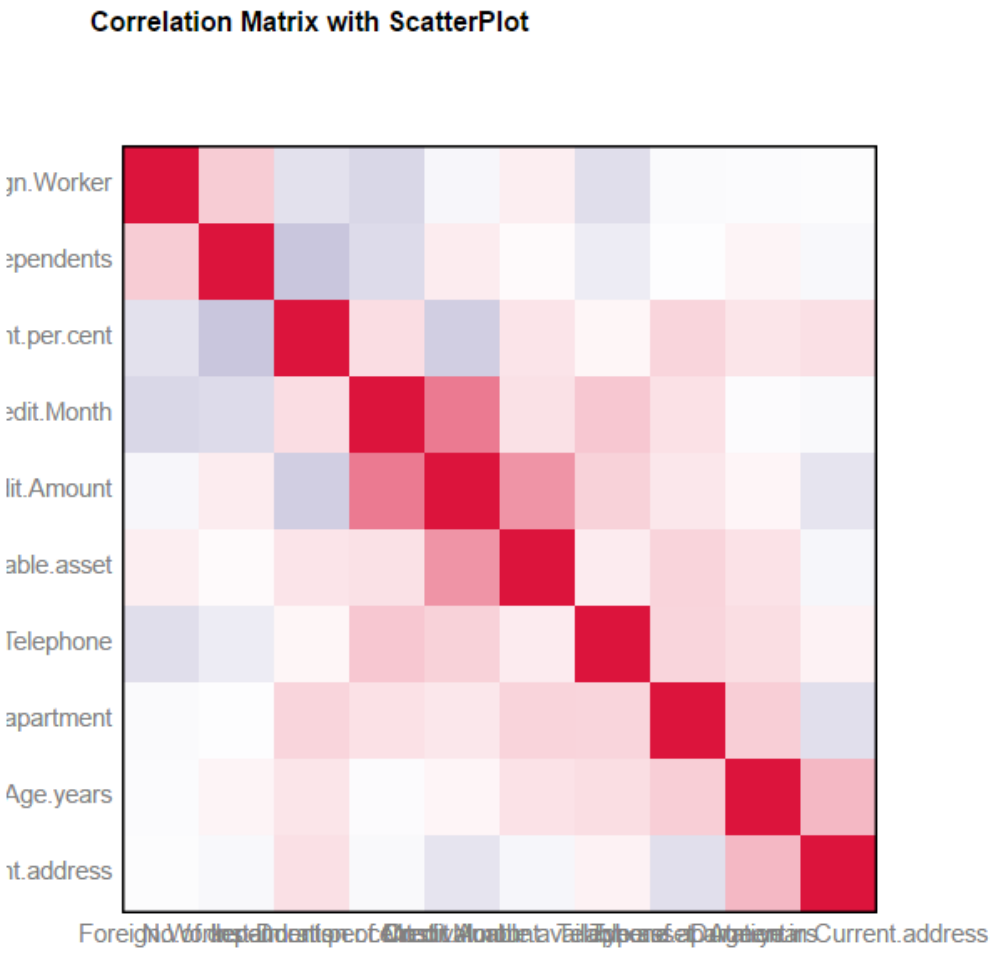


figure (2) correlation matrix for numeric variables

Step 3: Train your Classification Models

1. **Logistic Regression**

Top three predicted variables for logistic model are account balance, purposeNew car and credit amount with p-value of less than .05

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, AIC: 352.5

Number of Fisher Scoring iterations: 5

Type II Analysis of Deviance Tests

figure (3) report for logistic regression model

the accuracy for logistic model is 76%, accuracy creditworthy 80% and accuracy noncredit 62%. Biased toward non-creditworthy.

Model Comparison Report

Fit and error measures

Model	Accuracy	F1	AUC	Accuracy_Creditworthy	Accuracy_Non-Creditworthy
step_Creditworthy	.7600	.8364	.7306	.8000	.6286

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], number of samples that are **correctly** predicted to be Class [class name] divided by number of samples predited to be Class [class name]
AUC: area under the ROC curve, only available for two-class classification.
F1: F1 score, precision * recall / (precision + recall)

Confusion matrix of step_Creditworthy

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

figure (4) model comparison for stepwise logistic model

2. Decision Tree

The top three predicted variables for decision tree are: account balance, value saving stock and duration of credit moth with over all accuracy 74%. Accuracy credit worthy 97% and non-credit 60%. Biased toward predicting customers non-creditworthy.

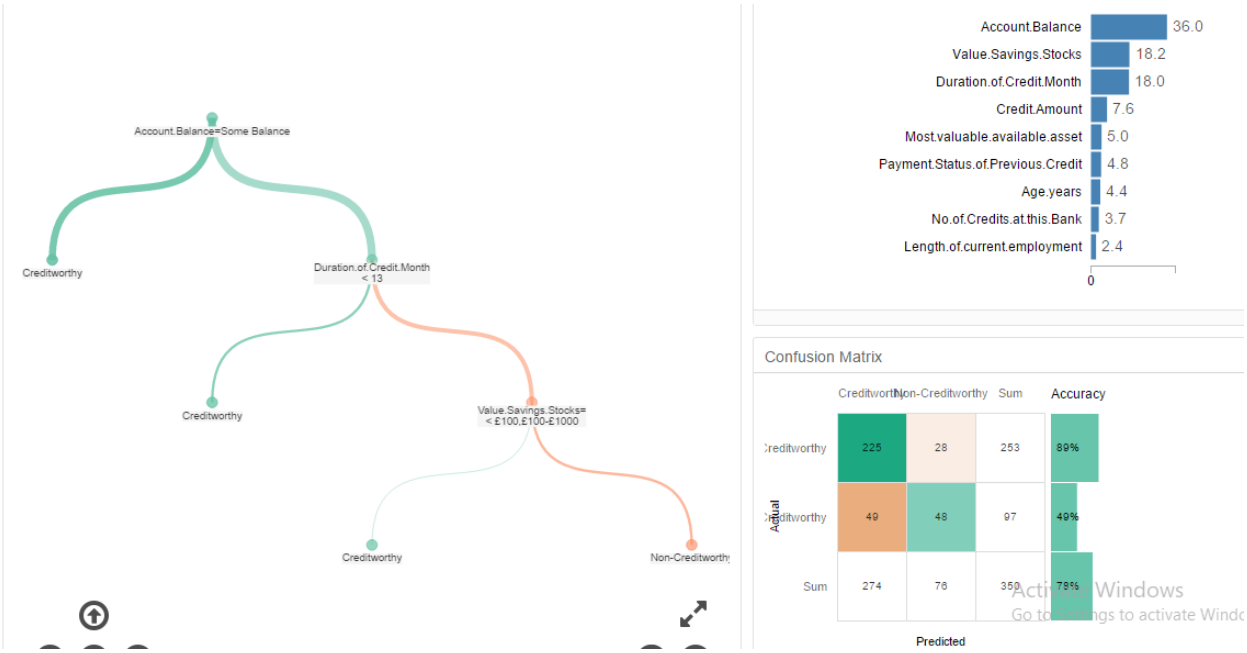


figure (5) decision tree, confusion matrix and variables importance

Model Comparison Report

Fit and error measures

Model	Accuracy	F1	AUC	Accuracy_Creditworthy	Accuracy_Non-Creditworthy
tree_Creditworthy	.7467	.8273	.7054	.7913	.6000

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], number of samples that are **correctly** predicted to be Class [class name] divided by number of samples predited to be Class [class name]
AUC: area under the ROC curve, only available for two-class classification.
F1: F1 score, precision * recall / (precision + recall)

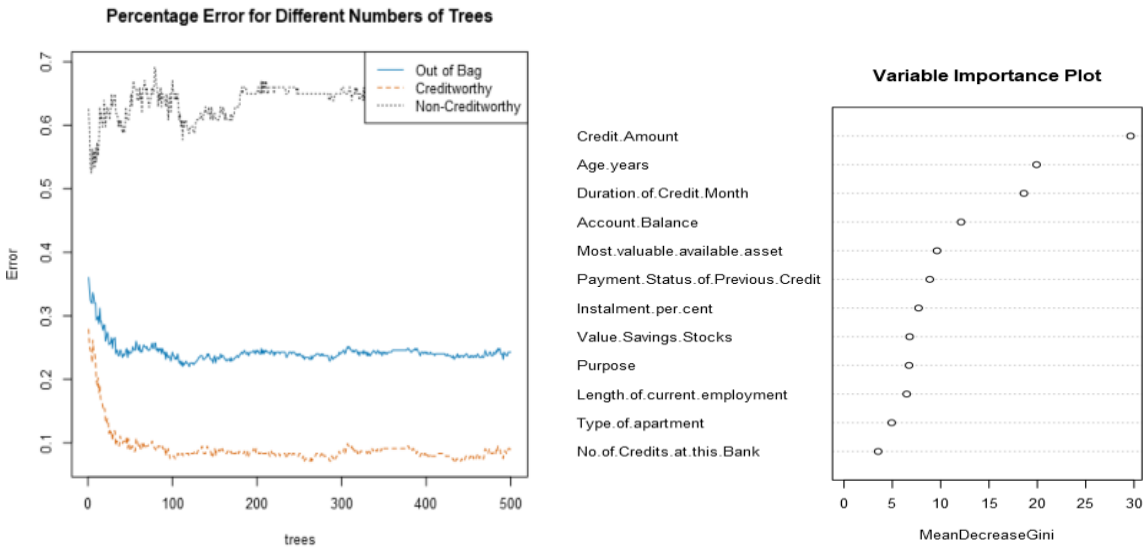
Confusion matrix of tree_Creditworthy

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

figure (6) model comparison for decision tree

3. Forest Model

The top three predicted variables for forest model are: credit amount, age years and duration of credit month. the accuracy 80%, credit worthy accuracy 79% while noncredit 86%. The model is not biased toward any accuracies, because accuracies for both are similar.



figure(7)percentage error for different numbers of tree and variable importance plot

Model Comparison Report

Fit and error measures

Model	Accuracy	F1	AUC	Accuracy_Creditworthy	Accuracy_Non-Creditworthy
forest_Creditworthy	.8067	.8755	.7392	.7969	.8636

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], number of samples that are **correctly** predicted to be Class [class name] divided by number of samples predited to be Class [class name]
AUC: area under the ROC curve, only available for two-class classification.
F1: F1 score, precision * recall / (precision + recall)

Confusion matrix of forest_Creditworthy

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

figure(8)model comparison for forest model

4. Boosted Model

Account balance and credit amount are top predicted variables for boosted model. The overall accuracy 78%, creditworthy accuracy 78% and non-creditworthy 80%.the model is not biased because it is similar accuracies

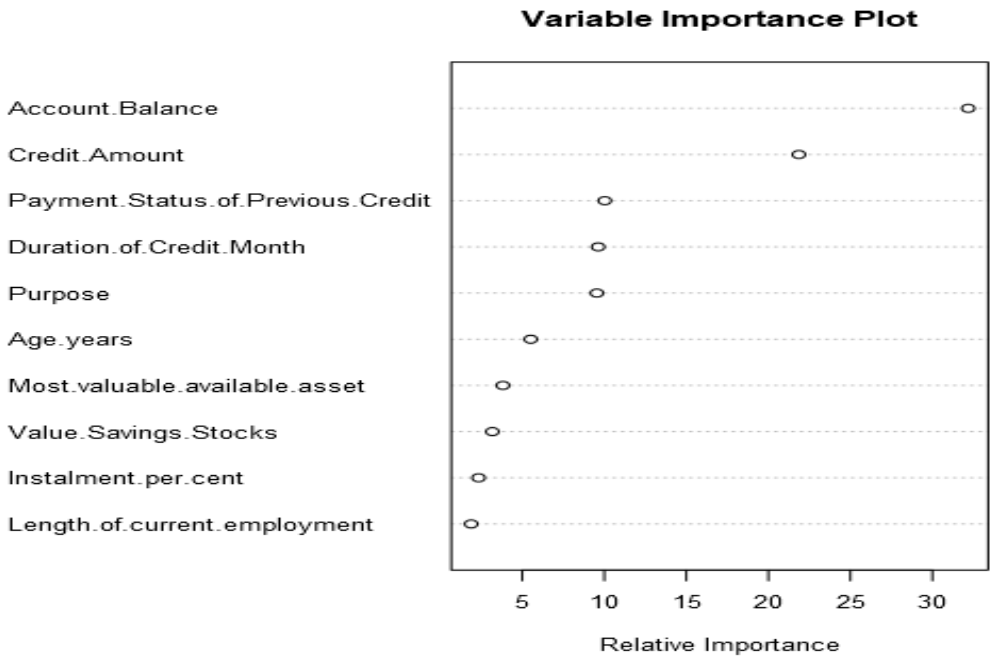


figure (9) variable importance plot

Model Comparison Report

Fit and error measures

Model	Accuracy	F1	AUC	Accuracy_Creditworthy	Accuracy_Non-Creditworthy
bo_Creditworthy	.7867	.8632	.7524	.7829	.8095

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], number of samples that are **correctly** predicted to be Class [class name] divided by number of samples predited to be Class [class name]
AUC: area under the ROC curve, only available for two-class classification.
F1: F1 score, precision * recall / (precision + recall)

Confusion matrix of bo_Creditworthy

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

figure (10) model comparison for boosted model

Step 4: Writeup

The best model is forest model with the highest accuracy 80% against validation set. the accuracy of creditworthy and non-creditworthy are the highest among other models.

we can see from the ROC curve that the forest is the fast model for true positive. the accuracy between creditworthy and non-creditworthy are similar which means have the least bias toward decision
Total number of creditworthy at the forest model are **408**

Model Comparison Report

Fit and error measures

Model	Accuracy	F1	AUC	Accuracy_Creditworthy	Accuracy_Non-Creditworthy
tree_Creditworthy	.7467	.8273	.7054	.7913	.6000
forest_Creditworthy	.8067	.8755	.7392	.7969	.8636
bo_Creditworthy	.7867	.8632	.7524	.7829	.8095
step_Creditworthy	.7600	.8364	.7306	.8000	.6286

Confusion matrix of bo_Creditworthy

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

Confusion matrix of forest_Creditworthy

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

Confusion matrix of step_Creditworthy

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

Confusion matrix of tree_Creditworthy

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

figure (11) model comparison for all models

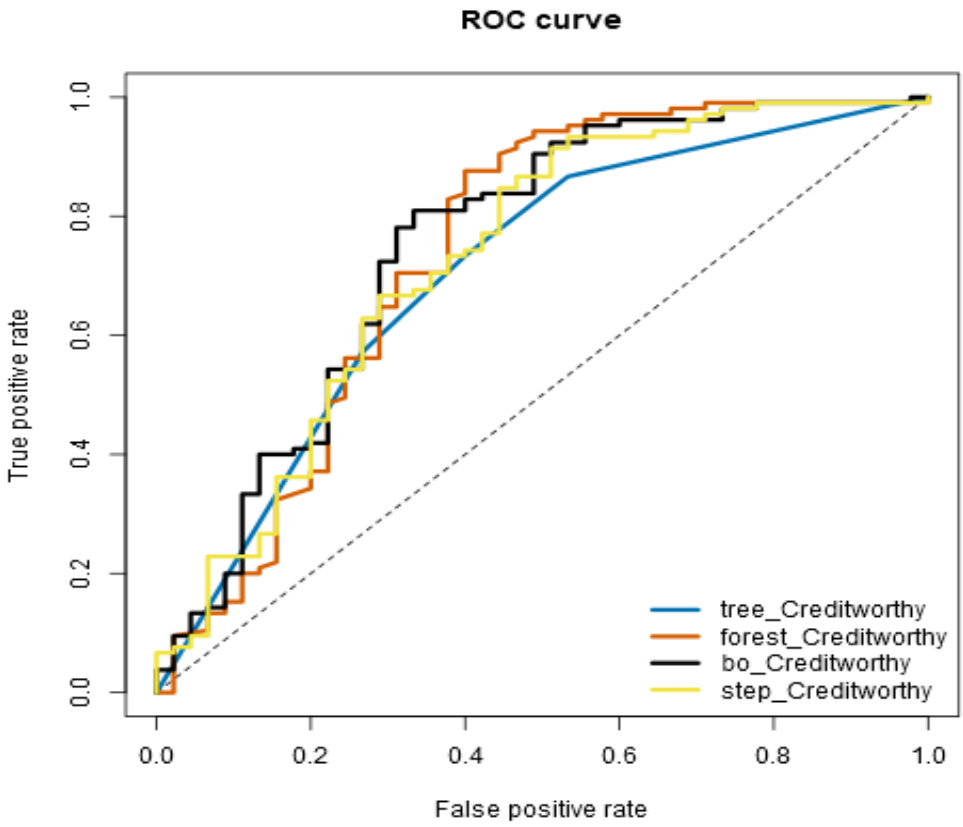


figure (12) ROC curve for 4 models

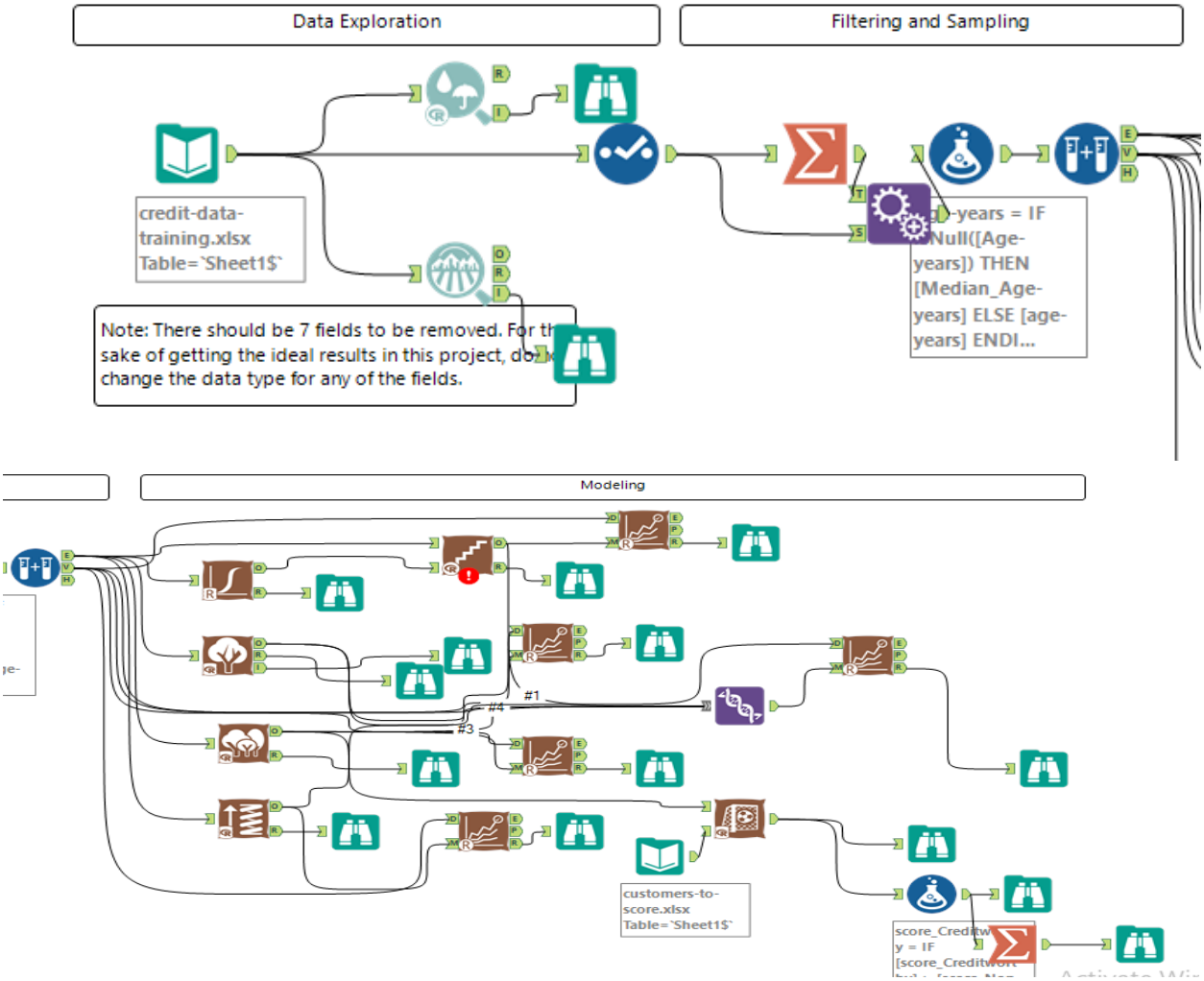


figure (13) alteryx workflow