Project: Creditworthiness

Step 1: Business and Data Understanding

Key Decisions:

Answer these questions

- **What decisions needs to be made?**
  As a loan officer, I need to predict whether a new incoming customer is creditworthy or not. Using the Problem Solving Framework, a new list on credit worthy customers has to be submitted to the manager.

- **What data is needed to inform those decisions?**
  The data needed to inform the decisions is
  1) Data on all past applications
  2) 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?**
  Since, the prediction is whether a customer is "*Creditworthy*" or "*Not-Creditworthy*", a Classification Model is implemented. The output variable in the model is Binary and the following methos will be used to decide which model will result in the highest accuracy :-
  1) Logistic Regression
  2) Decision Trees
  3) Forest Model
  4) Boosted Tree

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.***

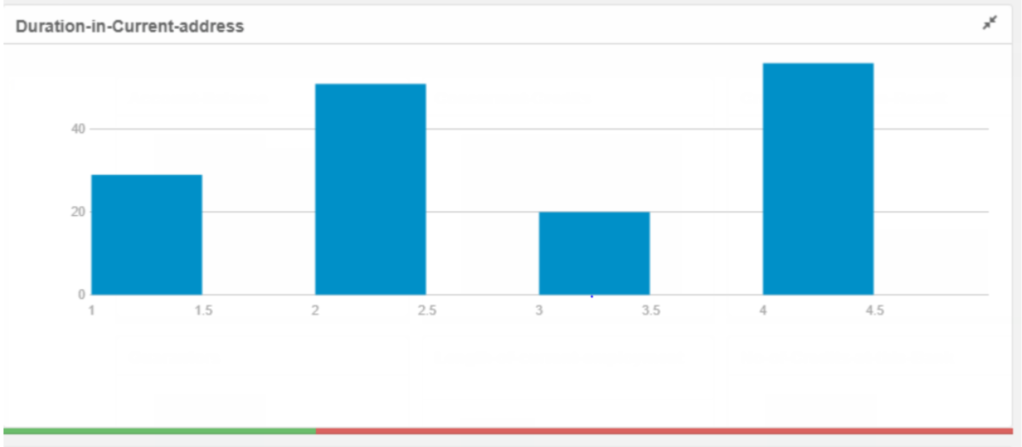
Variable	**Data Type**
Credit-Application-Result	String
Account-Balance	String
Duration-of-Credit-Month	Double
Payment-Status-of-Previous-Credit	String
Purpose	String
Credit-Amount	Double
Value-Savings-Stocks	String
Length-of-current-employment	String
Instalment-per-cent	Double
Guarantors	String
Duration-in-Current-address	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

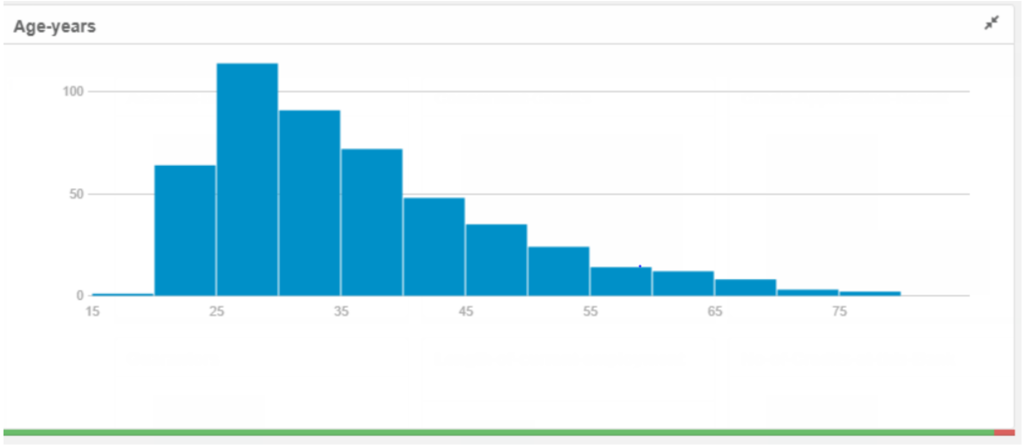
While cleaning the data, all data columns are assigned the appropriate data types using the SELECT tool. Parsing the data values, it is noted that no duplicate data and extra characters are present by aggregating the data set.

MISSING DATA

Numerical Columns

Next, we check for missing values in the numerical columns and notice that “*Age-Years*” and “*Duration-in-Current-address*” have missing values.



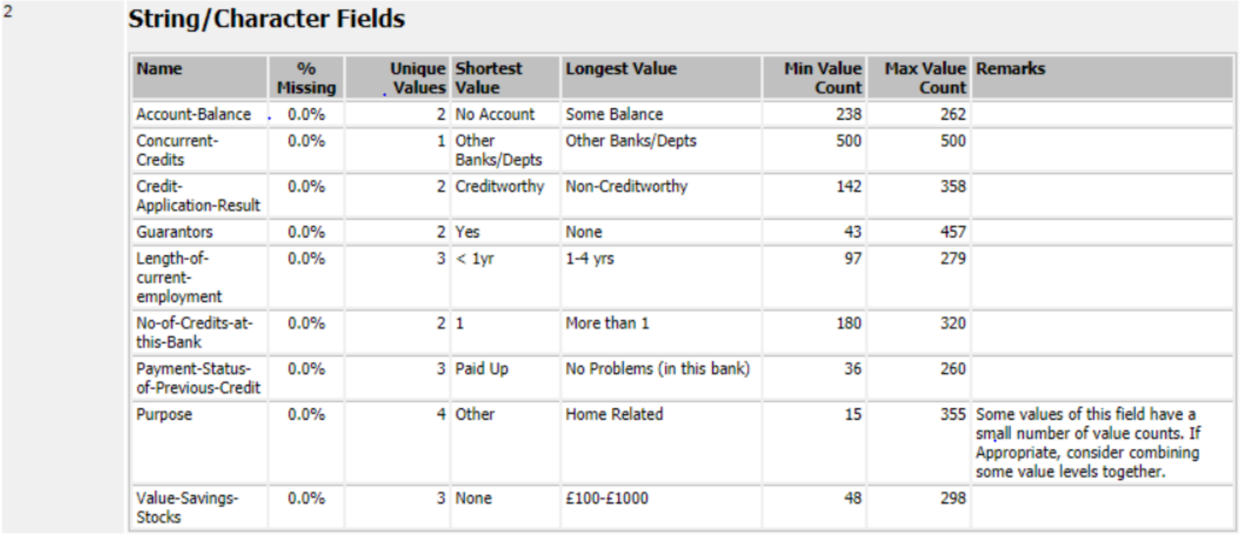


The above two figures represent the data columns that have missing values. The red line on the bottom of each figure represents the amount of missing values present in the column.

Here we notice, that “*Duration-in-Current-address*” has several missing values. These missing values can lead to inaccuracy and bias. So we decide to impute this column with the median for training the model.

“*Age-Years*” have less amount of data points (2.4%) and we can either impute or remove the records containing missing values. In this scenario, where we have **500** total records and the number of records with null is **12**. Here we replace the null values with the median 33.

Categorical columns



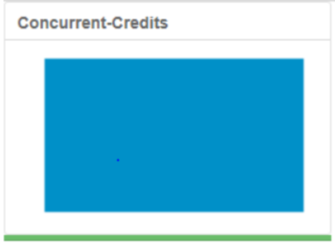
2

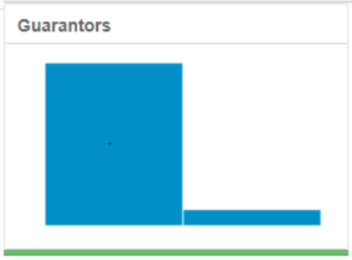
String/Character Fields

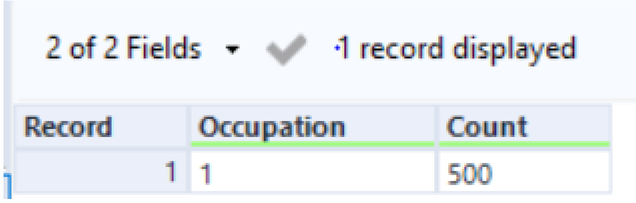
Name	% Missing	Unique Values	Shortest Value	Longest Value	Min Value Count	Max Value Count	Remarks
Account-Balance	0.0%	2	No Account	Some Balance	238	262	
Concurrent-Credits	0.0%	1	Other Banks/Depts	Other Banks/Depts	500	500	
Credit-Application-Result	0.0%	2	Creditworthy	Non-Creditworthy	142	358	
Guarantors	0.0%	2	Yes	None	43	457	
Length-of-current-employment	0.0%	3	< 1yr	1-4 yrs	97	279	
No-of-Credits-at-this-Bank	0.0%	2	1	More than 1	180	320	
Payment-Status-of-Previous-Credit	0.0%	3	Paid Up	No Problems (in this bank)	36	260	
Purpose	0.0%	4	Other	Home Related	15	355	Some values of this field have a small number of value counts. If Appropriate, consider combining some value levels together.
Value-Savings-Stocks	0.0%	3	None	£100-£1000	48	298	

The above figure justifies that string fields have 0% missing values.

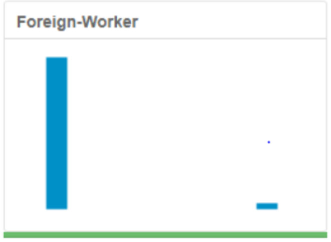
Further, we now need to deal with the low variability of our categorical variables, where a data field(within a categorical column) can heavily skew towards one type of data.

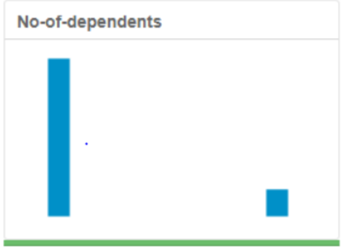






Record	Occupation	Count
1	1	500

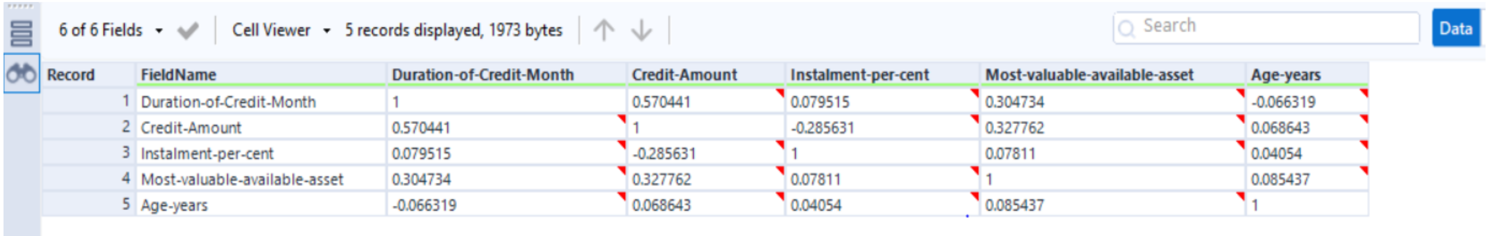




Therefore, we eliminate -
Concurrent Credits – all values have a uniform distribution.
Guarantors – majority of data is skewed towards one side.
Foreign-Worker – majority of data is skewed towards one side.
No-of-dependents - majority of data is skewed towards one side.
Telephone – for the consideration of this model.
Occupation – all values have uniform distribution

CORRELATION

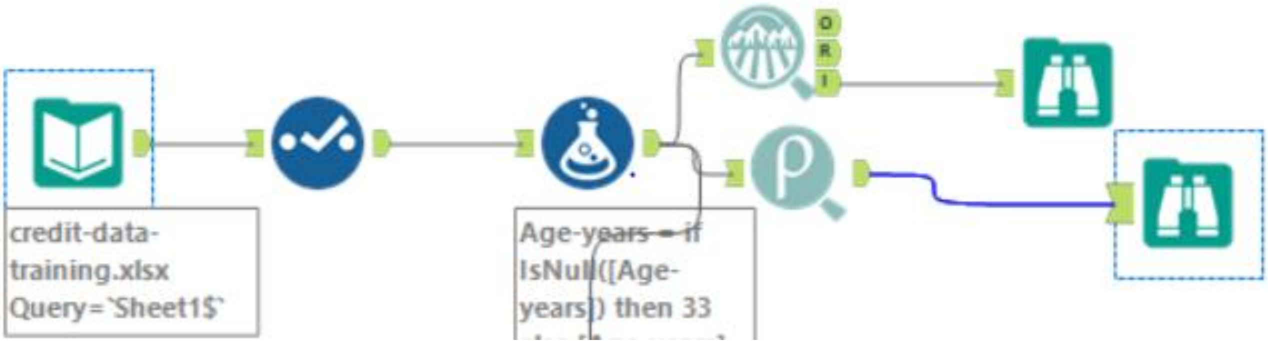
Before training the model with the cleaned dataset, it is necessary to check for correlation between the numerical variables in the dataset. Using the *Pearson-Correlation tool* and the browse tool to display results, we get the correlation values between 0 and 1 between all variables. A correlation of more than +/- 0.7 is highly-correlated and hence, we need to deal with these fields as required.



Record	FieldName	Duration-of-Credit-Month	Credit-Amount	Instalment-per-cent	Most-valuable-available-asset	Age-years
1	Duration-of-Credit-Month	1	0.570441	0.079515	0.304734	-0.066319
2	Credit-Amount	0.570441	1	-0.285631	0.327762	0.068643
3	Instalment-per-cent	0.079515	-0.285631	1	0.07811	0.04054
4	Most-valuable-available-asset	0.304734	0.327762	0.07811	1	0.085437
5	Age-years	-0.066319	0.068643	0.04054	0.085437	1

From the above figure, we notice that none of the values have a correlation greater or lesser than +/- 0.7. None of the values are correlated with each other. Hence, we can move ahead to use this dataset for training in our classification models.

Alteryx Workflow


credit-data-training.xlsx
Query=`Sheet1$`
Age-years = if IsNull([Age-years]) then 33
O
R
I

Final List of Predictor Variables

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
Most-valuable-available-asset	Double

Age-years	Double
No-of-Credits-at-this-Bank	String
Type – of – Apartment	Double

Step 3: Train your Classification Models

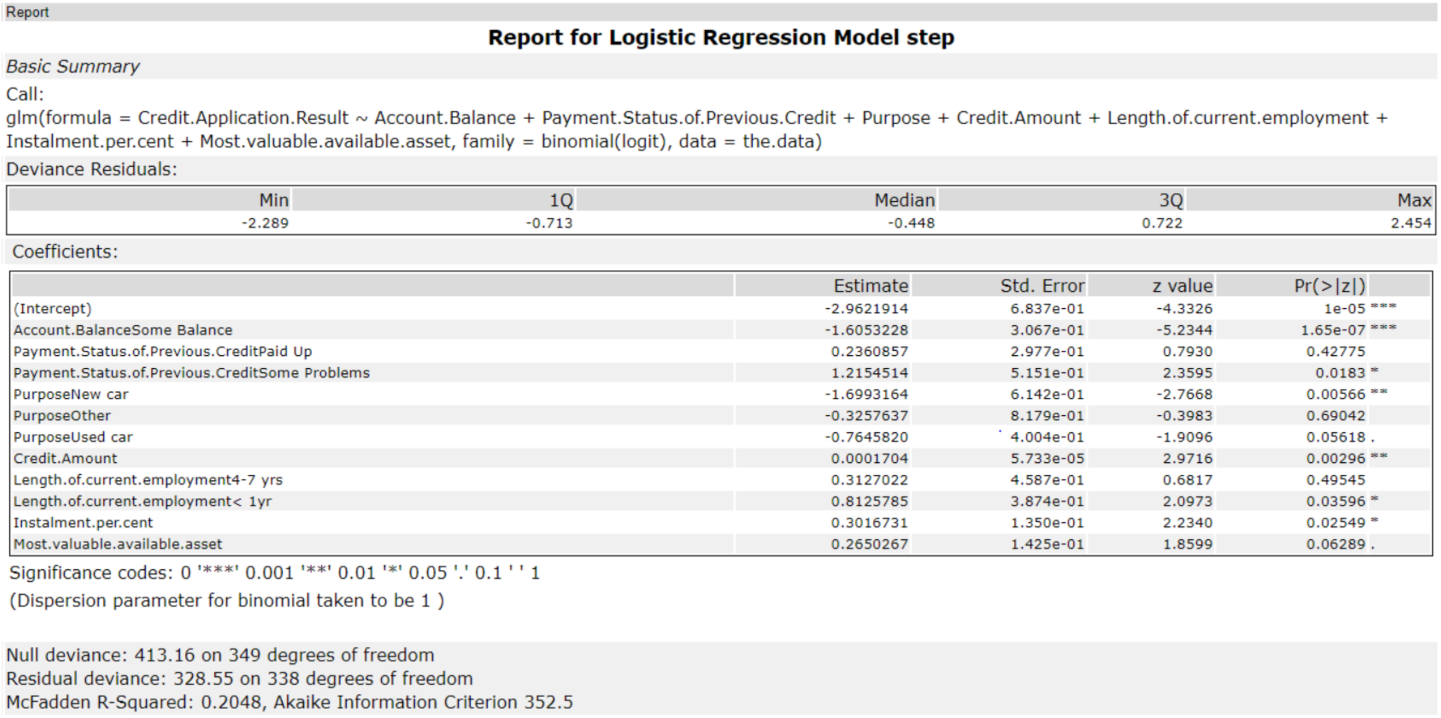
First, create your Estimation and Validation samples where 70% of your dataset should go to Estimation and 30% of your entire dataset should be reserved for Validation. Set the Random Seed to 1.

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

Logistic Regression

Using this classification model, we test the feasibility of a model by predicting a binary output variable. Logistic distribution tool is suitable for making a binary prediction. Along with the Logistic Tool we have used the stepwise tool. The stepwise tool is useful for selecting statistically significant predictor variables from the logistic distribution model. In other words, it simply selects variables that are more useful in making a prediction in the model. We may obtain a lower R-square value while we use stepwise. But in this model, the R-Square value for stepwise model is only less by around 2.
For the stepwise tool, we have used Adjusted Fit measure as AIC and used a Backward and Forward Search Direction.

Results



Report

Report for Logistic Regression Model step

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	.

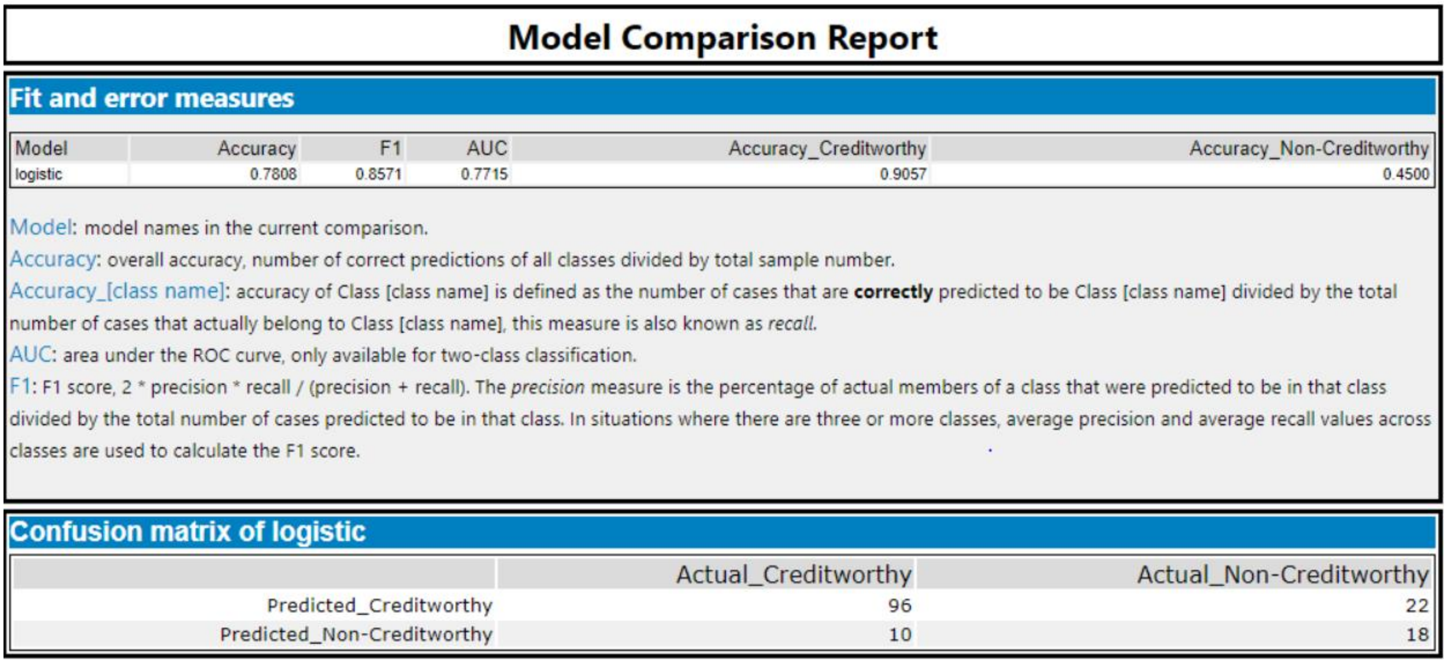
Significance codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
(Dispersion parameter for binomial taken to be 1)

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

In the above report, most of the predictor variables are statistically significant (p-value < 0.5). We obtain a R-square value of 0.2048 which explains that the predictor variables account for 20.5% variability for the “*Credit-*

Application Result" output. (Note : Some predictor variables are not significant but fall in a field of a statistically significant variable. In this case, we consider the entire field.)

Validation



Model Comparison Report

Fit and error measures

Model	Accuracy	F1	AUC	Accuracy_Creditworthy	Accuracy_Non-Creditworthy
logistic	0.7808	0.8571	0.7715	0.9057	0.4500

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 logistic

	Actual_Creditworthy	Actual_Non-Creditworthy
Predicted_Creditworthy	96	22
Predicted_Non-Creditworthy	10	18

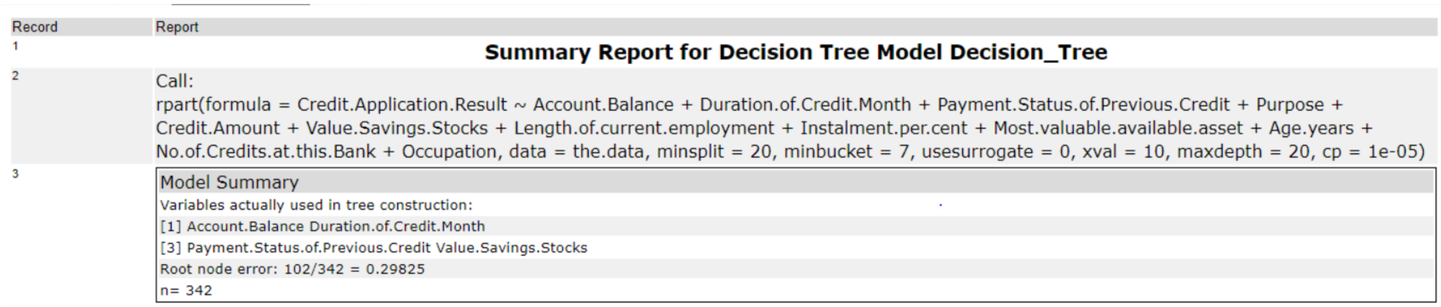
The accuracy for this model on the validation dataset is 78%.
The accuracy for predicting Creditworthy customers (90.5%) is more than that of Non-Creditworthy customers(45%).
This should not be an issue for the model since we are trying to predict the customers who have credit scores worthy for a loan. Therefore, a low value for Non-creditworthy customers is feasible.

The confusion matrix tells us that the model predicts 96 customers as trustworthy and 22 customers as non-creditworthy. Therefore,
Precision – 96/(94+22) = 0.827
Recall – 96/(94+10) = 0.923
Error Rate – (22+10)/146 = 0.219

Decision Tree

The decision tree tool is used in the on the cleaned and sampled training dataset.

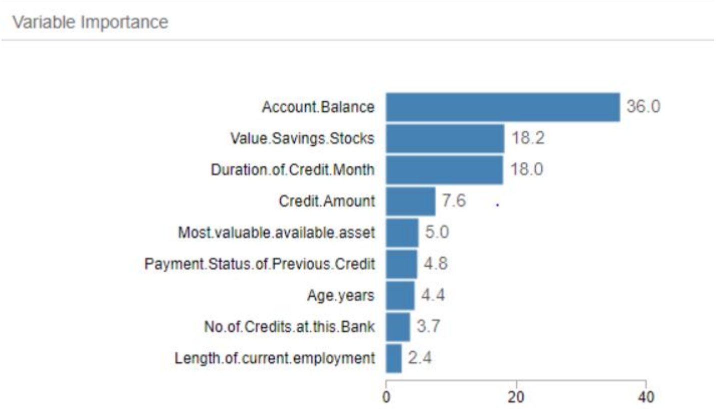
Results

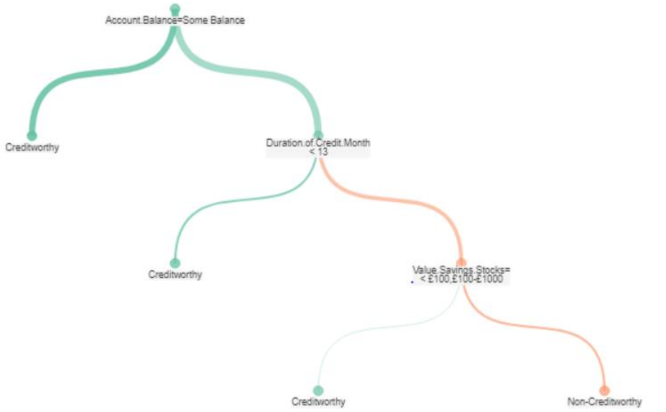


Record	Report
1	**Summary Report for Decision Tree Model Decision_Tree**
2	Call: rpart(formula = Credit.Application.Result ~ Account.Balance + Duration.of.Credit.Month + Payment.Status.of.Previous.Credit + Purpose + Credit.Amount + Value.Savings.Stocks + Length.of.current.employment + Instalment.per.cent + Most.valuable.available.asset + Age.years + No.of.Credits.at.this.Bank + Occupation, data = the.data, minsplit = 20, minbucket = 7, usesurrogate = 0, xval = 10, maxdepth = 20, cp = 1e-05)
3	Model Summary Variables actually used in tree construction: [1] Account.Balance Duration.of.Credit.Month [3] Payment.Status.of.Previous.Credit Value.Savings.Stocks Root node error: 102/342 = 0.29825 n= 342

The predictor variables used to develop nodes in the decision tree are shown in the above figure. The root node error is 0.2985, which means 29.85% of data points went to the incorrect terminal node (predicted incorrectly) when all of the data points were validated against themselves within the entire training set (the Estimation dataset).

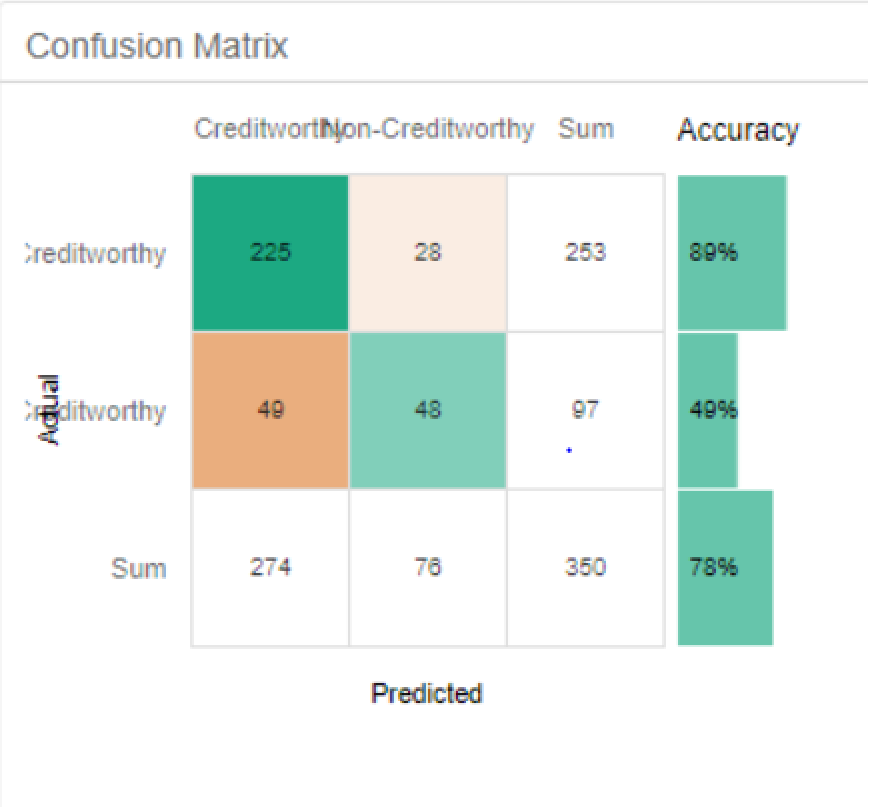
Variability Importance Chart





Duration of Credit Month, Value Savings Stocks and Account Balance are the highest three values and are present in the start of the decision trees. This also means that these fields are the basis for all decisions in the decision tree i.e. as splitting increases

Confusion Matrix



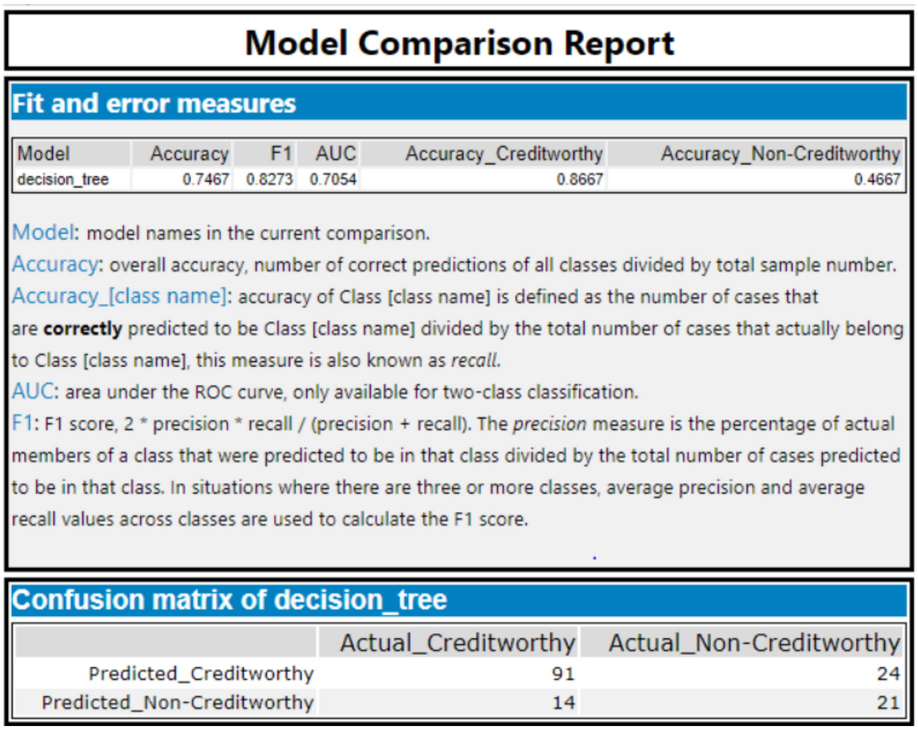
Confusion Matrix

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

Predicted

The accuracy for predicting creditworthy customers is 89% and the overall accuracy is 78%. Therefore, the model is feasible. This confusion matrix is for the estimation dataset. A high accuracy in this model doesn't assure accuracy in a novel dataset.

Validation



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

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 decision_tree

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

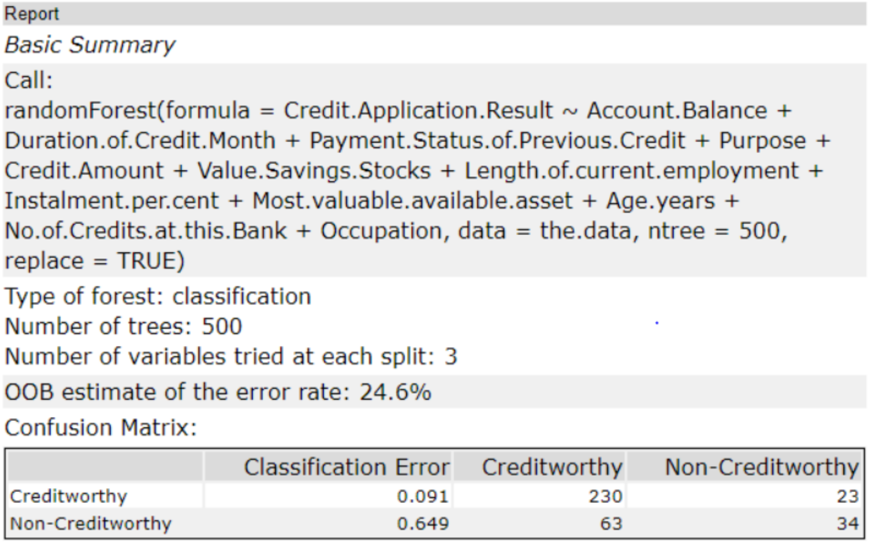
The accuracy for this model on the validation dataset is 74.67%.
The accuracy for predicting Creditworthy customers (86%) is more than that of Non-Creditworthy customers(46%).
This should not be an issue for the model since we are trying to predict the customers who have credit scores worthy for a loan. Therefore, a low value for Non-creditworthy customers is feasible.

The confusion matrix tells us that the model predicts 94 customers as trustworthy and 19 customers as non-creditworthy. Therefore,
Precision – 91/(91+24) = 0.791
Recall – 91/(91+14) = 0.866
Error Rate – (24+14)/146 = 0.2602

Forest Model

A random forest model uses an ensemble of decision trees. This technique is used to eliminate the overfitting caused in decision trees. The error is averaged out by taking into account the average error of all trees.

Results



Report

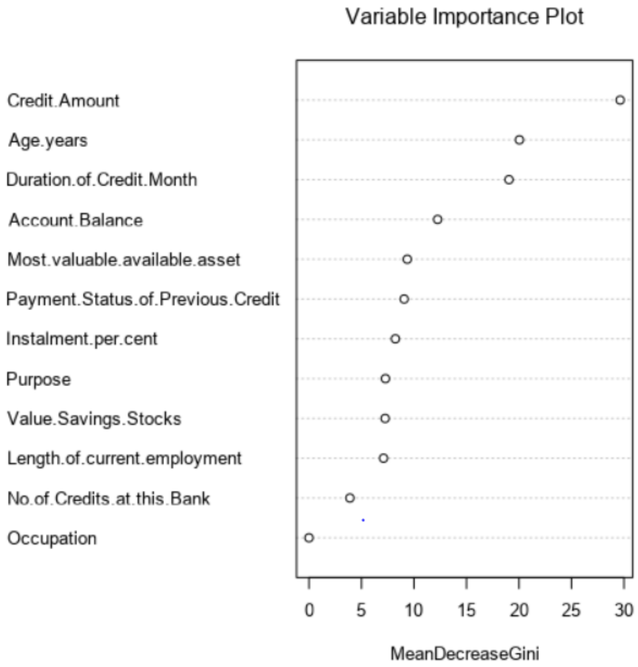
Basic Summary

```
Call:
randomForest(formula = 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 + Most.valuable.available.asset + Age.years +
No.of.Credits.at.this.Bank + Occupation, data = the.data, ntree = 500,
replace = TRUE)
Type of forest: classification
Number of trees: 500
Number of variables tried at each split: 3
OOB estimate of the error rate: 24.6%
Confusion Matrix:
```

	Classification Error	Creditworthy	Non-Creditworthy
Creditworthy	0.091	230	23
Non-Creditworthy	0.649	63	34

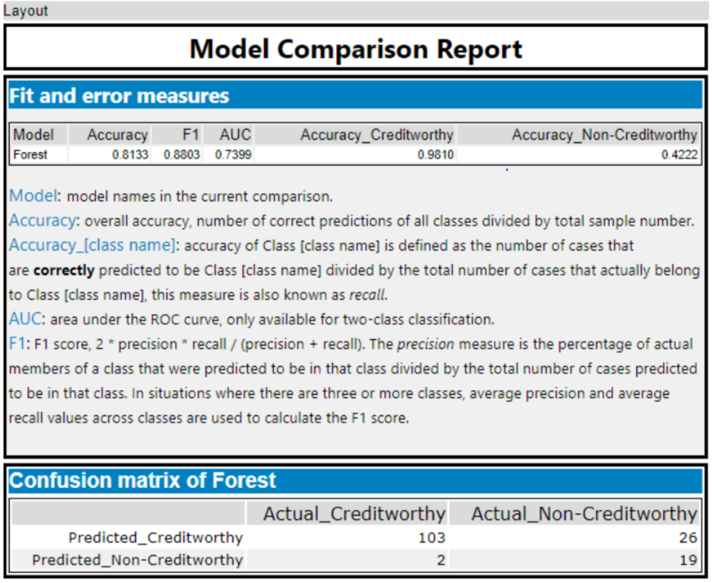
The OOB estimate error is 24.6%, which is the error when the estimation model creates its own validation set within the model this is the error rate. This can also be denoted as the R-Square value.
The confusion matrix shows that Non-creditworthy customers again have a high classification error due to less number of records in the training set.

Variability Importance Chart



The fields with the highest mean decrease Gini are the variables that contributed more towards the model. Therefore, we can use the top contributing fields for further analysis such as PCA.
In this case Credit amount, age and duration of credit month seems like the top contributing factors in creating a random forest model.

Validation



Layout

Model Comparison Report

Fit and error measures

Model	Accuracy	F1	AUC	Accuracy_Creditworthy	Accuracy_Non-Creditworthy
Forest	0.8133	0.8803	0.7399	0.9810	0.4222

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.

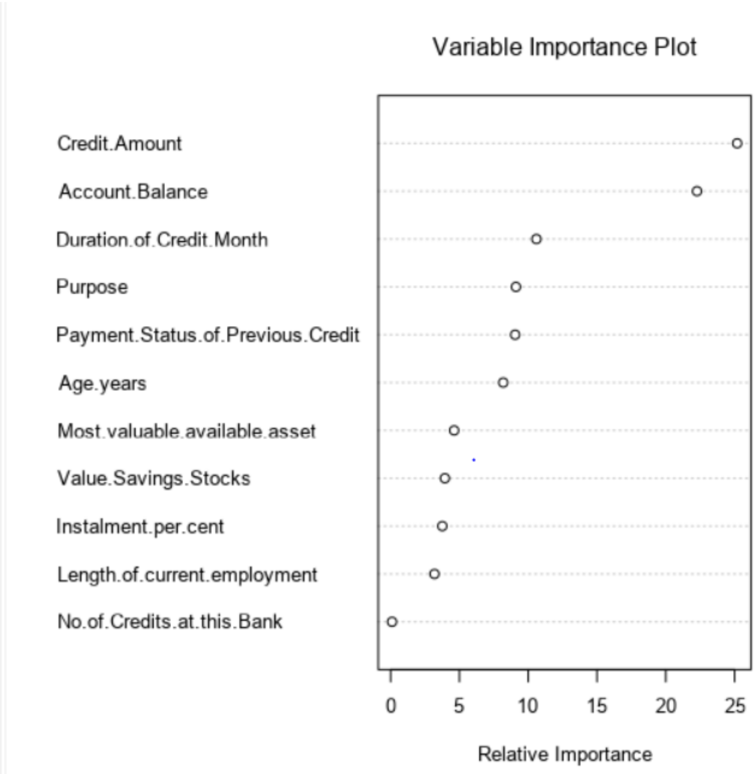
Confusion matrix of Forest

	Actual_Creditworthy	Actual_Non-Creditworthy
Predicted_Creditworthy	103	26
Predicted_Non-Creditworthy	2	19

The accuracy for this model is 81.33%, which is higher than the previous two models (Decision Trees and Logistic Regression).Also, according to our business requirements, we need to predict the creditworthy customers and the model does a great job in predicting the creditworthy customers (98%). The confusion matrix has least number of false negatives as compared to other models

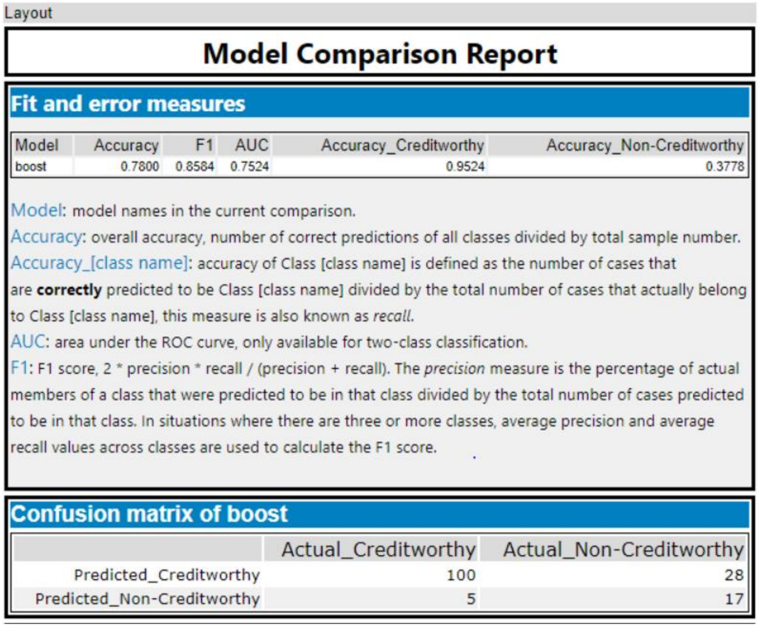
Boosted Model

Variability importance chart



The fields used in the model are mapped against their relative importance (X-axis). We can see the top two predictor variables were Account Balance and Credit-Amount. Hence, these two factors were most important for the prediction in the model

Validation



Model Comparison Report

Fit and error measures

Model	Accuracy	F1	AUC	Accuracy_Creditworthy	Accuracy_Non-Creditworthy
boost	0.7800	0.8584	0.7524	0.9524	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 boost

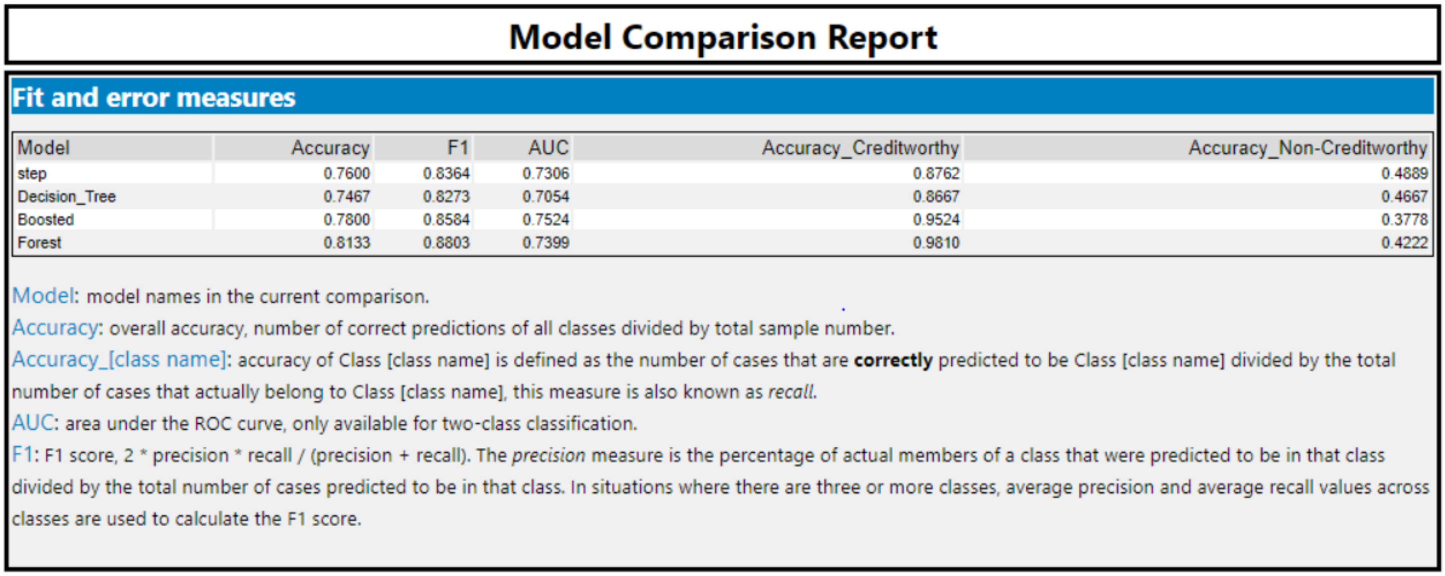
	Actual_Creditworthy	Actual_Non-Creditworthy
Predicted_Creditworthy	100	28
Predicted_Non-Creditworthy	5	17

The accuracy for the Boosted Model is the highest at 78%. Also, the accuracy of predicting creditworthy customers is 95%. This model has the highest accuracy as compared to previous models.
The confusion matrix is almost the same.

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"

From the validation results of all the models, it is clear that the Forest Model is a better model for predicting whether a new incoming customer for a loan is creditworthy or not. To justify our answer, we take the following four levels under consideration :-

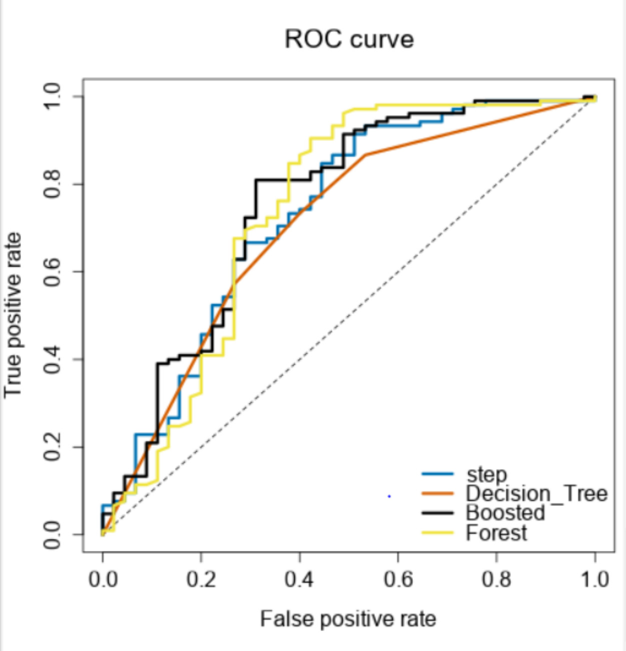


Model Comparison Report

Fit and error measures

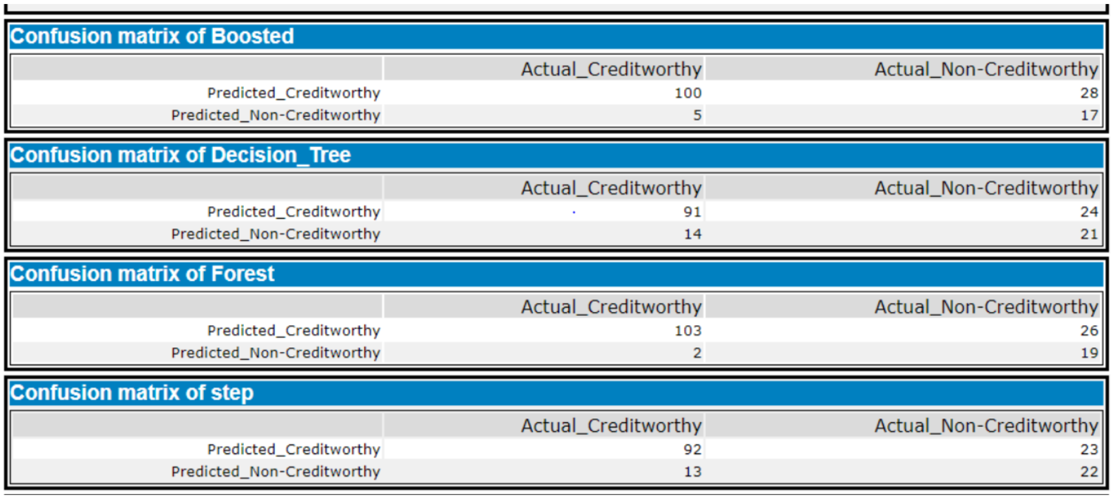
Model	Accuracy	F1	AUC	Accuracy_Creditworthy	Accuracy_Non-Creditworthy
step	0.7600	0.8364	0.7306	0.8762	0.4889
Decision_Tree	0.7467	0.8273	0.7054	0.8667	0.4667
Boosted	0.7800	0.8584	0.7524	0.9524	0.3778
Forest	0.8133	0.8803	0.7399	0.9810	0.4222

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.



1) From the above figure, it is noted that the accuracy for the **Forest Model** is highest at 0.81(81%) and has the lowest error rate (Since, error rate = 1-accuracy)

2) Also, the accuracy for predicting creditworthy customers for the loan is the highest at 0.98(98%). The accuracy for predicting non-creditworthy customers is less. But it doesn't matter in this scenario since the focus is on generating most number of potential customers!
3) The ROC Curve Graph indicates that the Area Under Curve(AUC) for Boosted Model is the highest at 75%. The second highest is the Forest Model 74%. We can consider Forest Model, since the difference is minute. Also, the lines curving closer to the diagonal have less AUC



Confusion matrix of Boosted

	Actual_Creditworthy	Actual_Non-Creditworthy
Predicted_Creditworthy	100	28
Predicted_Non-Creditworthy	5	17

Confusion matrix of Decision_Tree

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

Confusion matrix of Forest

	Actual_Creditworthy	Actual_Non-Creditworthy
Predicted_Creditworthy	103	26
Predicted_Non-Creditworthy	2	19

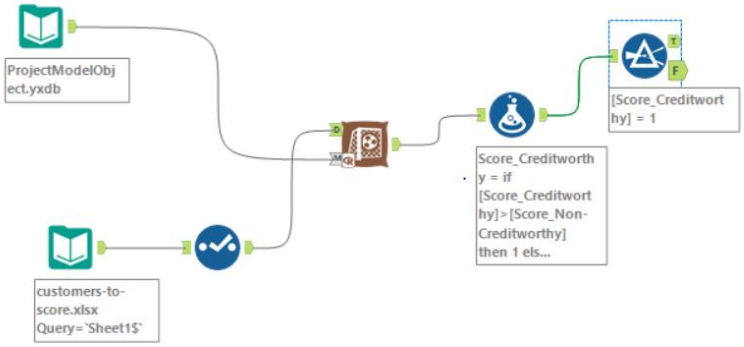
Confusion matrix of step

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

4) In the above figure, as compared to all the models we can see that the Forest model has the highest accuracy for predicting creditworthy customers(103).
The Sensitivity and Accuracy of **Forest Model** is the highest.

Conclusion

Therefore, keeping all the above factors in mind. We can consider ***Forest model*** as an appropriate model. Since, it has the highest overall accuracy (0.8133) and the highest precision for predicting creditworthy customers (0.98). Also the AUC for the model is the second highest at 75% and the Recall is also second highest.
From this Forest model, we conclude the number of creditworthy customers as **411** and Non-Creditworthy as **89**. Using this workflow and the Score Tool in Alteryx we concluded the creditworthiness for the new incoming customers next week. Therefore, 411 customers are targeted as the final set of potential customers.


ProjectModelObject.yxdb
customers-to-score.xlsx Query=`Sheet1$`
D
M
Score_Creditworthy = if [Score_Creditworthy]>[Score_Non-Creditworthy] then 1 els...
T
F
[Score_Creditworthy] = 1