

Creditworthiness using Alteryx

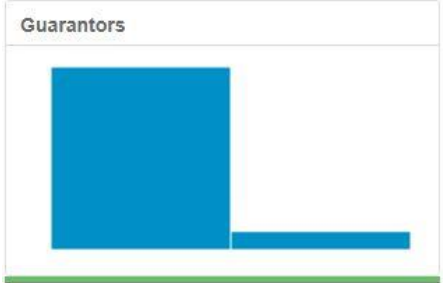
Business and Data Understanding


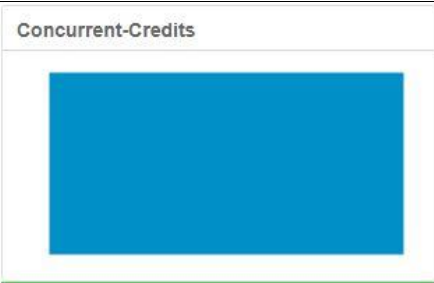



Key Decisions:

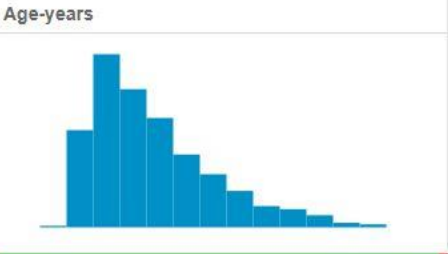
- ✓ We need to decide which of the 500 new applications will approved for a loan and provide a list of creditworthy customers to the manager.
- ✓ We need to explore the data of previous applications and build classification models using these data to build an *Estimation* and *Validation* sample (70% - 30%) to train the models. Then we will score the data of new customers with the best model to provide a list of creditworthy customers. The data we are looking for includes Account-Balance, Duration-of-Credit-Month, Age years, Payment-Status-of-Previous-Credit, Purpose and Credit-Amount for past and new loan applicants.
- ✓ Since we need to decide on the outcome whether it's creditworthy or non—creditworthy then we will use the **Binary models**.

Building the Training Set

We have to decide in the cleanup process which fields to remove or impute. And it explained with figures below:

Field	Action taken	figure
Guarantor	Removed – due to low variability	

Duration in current address	Removed – has 69% missing data	
Concurrent credits	Removed – low variability, no variation of data	
Occupation	Removed - Removed – low variability, no variation of data has a single value of 1	
No. of independents	Removed – due to low variability	
Telephone	Removed – no correlation with target variable	
Foreign worker	Removed – due to low variability	

Age years	imputed missing data by using the median value for all age values in the field.	
-----------	---	--

Training the Classification Models

models: Logistic Regression, Decision Tree, Forest Model, Boosted Model

Logistic Regression

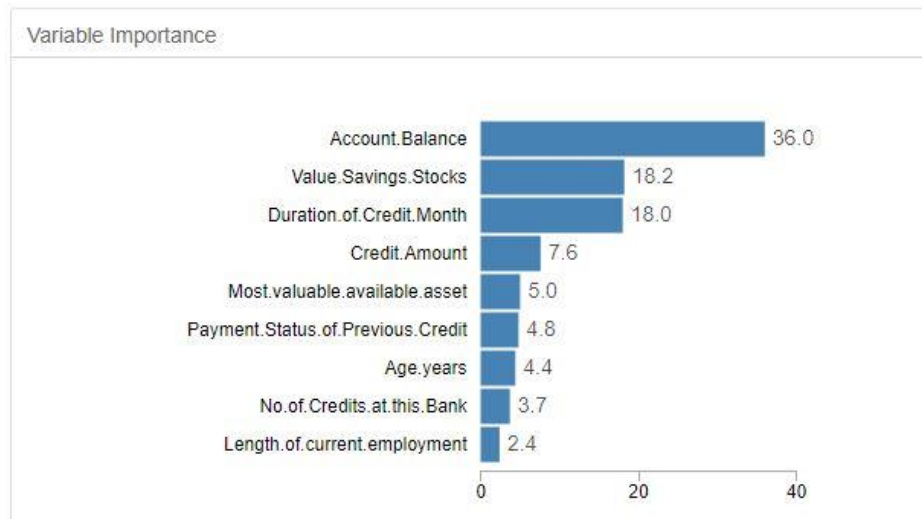
The most significant variables are as shown below after using the stepwise tool

Response: Credit.Application.Result

	LR Chi-Sq	DF	Pr(>Chi-Sq)	
Account.Balance	31.129	1	2.41e-08	***
Payment.Status.of.Previous.Credit	5.687	2	0.05823	.
Purpose	12.225	3	0.00665	**
Credit.Amount	9.882	1	0.00167	**
Length.of.current.employment	5.522	2	0.06324	.
Instalment.per.cent	5.198	1	0.02261	*
Most.valuable.available.asset	3.509	1	0.06104	.

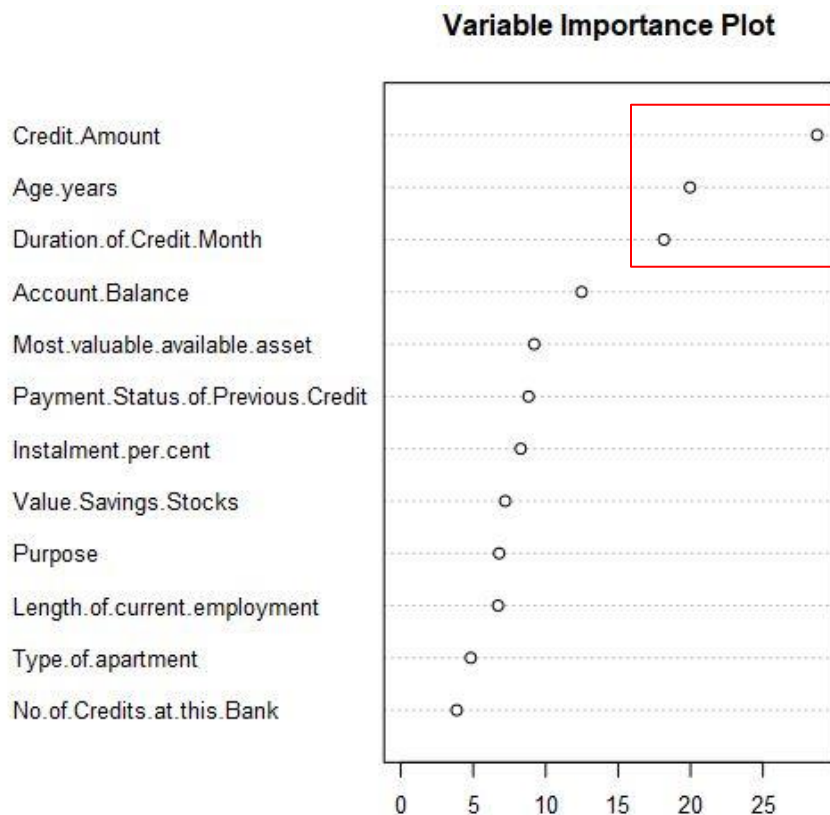
Decision Tree

The most important variables for the decision Tree model are the first 3 variables as in the figure below



Forest Model

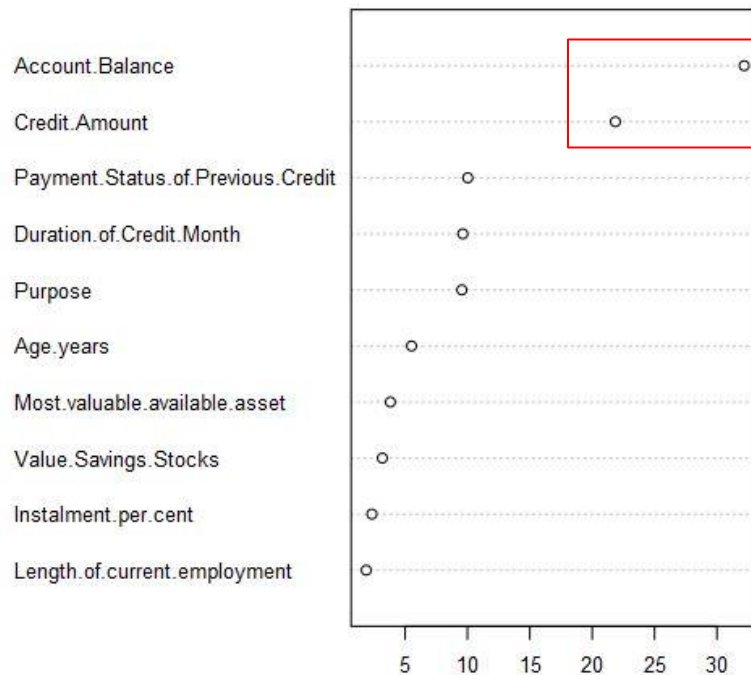
There are almost 3 important variables for the forest model



Boosted Model

For the Boosted model there are 2 most important variables

Variable Importance Plot



Validating the models against the Validation set

The figure below shows the **overall accuracy** for each model, with a 79% of accuracy for both Boosted model & Forest Model being the highest. And the lowest accuracy percentage is for the Decision Tree model with a value of 75%

Model Comparison Report					
Fit and error measures					
Model	Accuracy	F1	AUC	Accuracy_Creditworthy	Accuracy_Non-Creditworthy
Logistic_Reg	0.7800	0.8520	0.7314	0.9048	0.4889
Decision_Tree	0.7467	0.8273	0.7054	0.8667	0.4667
forest_Model	0.7933	0.8681	0.7368	0.9714	0.3778
Boosted_model	0.7867	0.8632	0.7524	0.9619	0.3778

For the **Confusion Matrix**, both Boosted model & Forest Model have performed well in predicting the creditworthy customer with values of 101(96%) and 102 (97%)

Confusion matrix of Boosted_model		
	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 Logistic_Reg		
	Actual_Creditworthy	Actual_Non-Creditworthy
Predicted_Creditworthy	95	23
Predicted_Non-Creditworthy	10	22

Confusion matrix of forest_Model		
	Actual_Creditworthy	Actual_Non-Creditworthy
Predicted_Creditworthy	102	28
Predicted_Non-Creditworthy	3	17

Evaluating Bias in model prediction

We will use the prediction accuracy for both outcomes in the confusion matrix above to evaluate the bias:

Model	Calculation	Evaluation
Boosted Model	PPV= $101 / (101+28) = 0.78$ NPV= $17 / (17+4) = 0.81$	No bias in the model prediction
Decision Tree	PPV= $91 / (91+24) = 0.79$ NPV= $21 / (21+14) = 0.60$	Biased to predicting Creditworthy
Logistic Regression	PPV= $95 / (95+23) = .80$ NPV= $22 / (22+10) = .68$	Biased to predicting Creditworthy
Forest Model	PPV= $102 / (102+28) = 0.78$ NPV= $17 / (17+3) = 0.85$	No bias in the model prediction

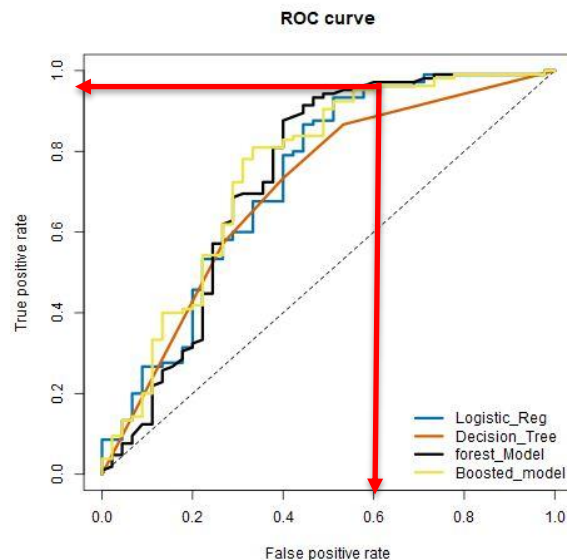
conclusion

After training our classification model using all of the 4 models (Logistic Regression, Decision Tree, Forest Model, Boosted Tree)

The confusion matrix shows that the **Forest Model** is the best fit for our data with overall accuracy of **79.33%** and **97%** accuracy in predicting creditworthy and **38%** Non-creditworthy against the validation set. Furthermore, there is no bias in the model's prediction toward any of the outcomes based on the confusion matrix results:

($PPV = 102 / (102 + 28) = 0.78$ And $NPV = 17 / (17 + 3) = 0.85$).

Another reason on why the **Forest Model** is the best fit is that by looking at the ROC curve we can see that even if we have a false positive rate of 0.6 the model will still have a True positive rate of 0.9 or higher that means it can give a higher number of True positive prediction which makes the curve of the model higher than all other 3 models.



Next, we will score the new data set of 500 customers with the **Forest Model** to predict which customers are creditworthy.

The result of scoring the model with the new customers data shows that **408** out of 500 new applications are predicted to be creditworthy.

Record #	CountNonNull_Creditworthy_list	CountNull_Creditworthy_list
1	408	92

Alteryx workflow

