Project: Creditworthiness

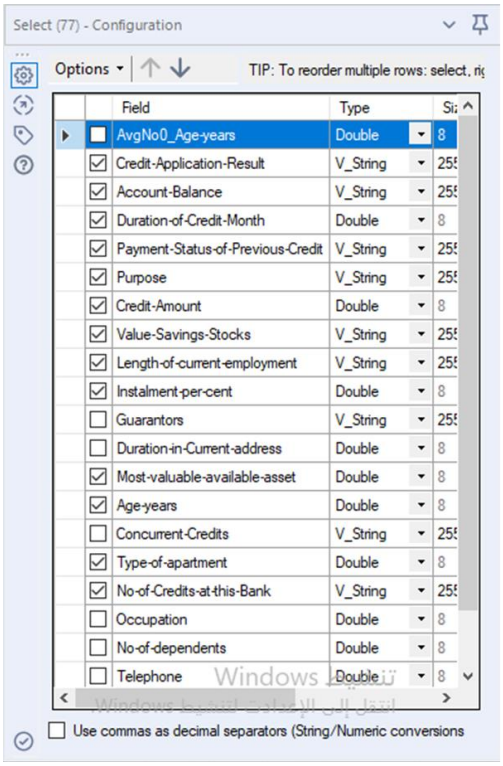
Step 1: Business and Data Understanding

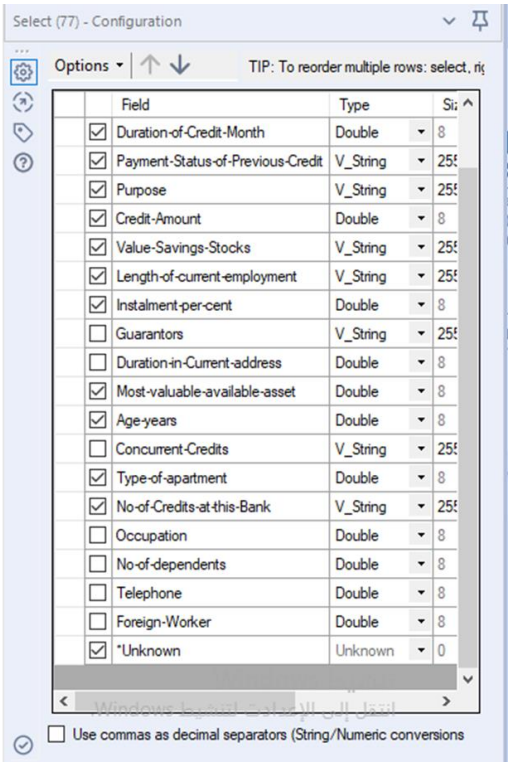
Key Decisions:

- What decisions needs to be made?
  predict\decide if each new customer is credit worthy or not.
- What data is needed to inform those decisions?
  Two datasets: old customers (credit-data-training) and new customers data (customers-to-score).
- What kind of model (Continuous, Binary, Non-Binary, Time-Series) do we need to use to help make these decisions?
  Binary model, cause the decision we'll take has two options only (credit worthy or not).

Step 2: Building the Training Set

- In your cleanup process, which fields did you remove or impute? Please justify why you removed or imputed these fields. Visualizations are encouraged.
  Removed Guarantors, Concurrent credits, Occupation, No of dependents, Duration in current address, Telephone and Foreign worker as their data has no relation with predicting if a customer is credit worthy or not, especially Duration in current address as its majority are Null.
  Imputed Age_years with average age (in old customers data only).





Step 3: Train your Classification Models

- Which predictor variables are significant or the most important? Please show the p-values or variable importance charts for all of your predictor variables.

Logistic Regression:

Report for Logistic Regression Model CreditWorthyLogistic

Basic Summary

Call:
glm(formula = Credit.Application.Result ~ Account.Balance + Duration.of.Credit.Month, family = binomial(logit), data = the.data)

Deviance Residuals:

Min	1Q	Median	3Q	Max
-1.541	-0.889	-0.520	0.990	2.224

Coefficients:

	Estimate	Std. Error	z value	Pr(>|z|)	
(Intercept)	-0.99695	0.26931	-3.702	0.00021	***
Account.BalanceSome Balance	-1.66212	0.28495	-5.833	5.44e-09	***
Duration.of.Credit.Month	0.03033	0.01003	3.022	0.00251	**

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

(Dispersion parameter for binomial taken to be 1)

Decision Tree:

Summary Report for Decision Tree Model CreditWorthyTree

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 + Type.of.apartment + No.of.Credits.at.this.Bank, data = the.data, minsplit = 20, minbucket = 7, usesurrogate = 1, xval = 10, maxdepth = 20, cp = 1e-05)

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

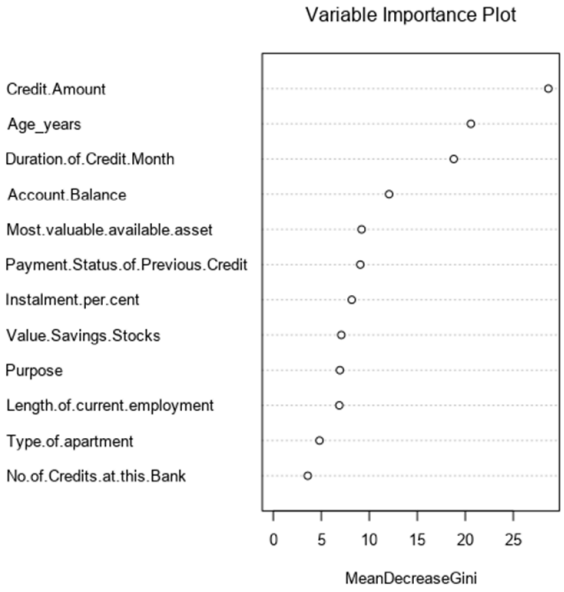
Pruning Table

Level	CP	Num Splits	Rel Error	X Error	X Std Dev
1	0.068729	0	1.00000	1.00000	0.086326
2	0.041237	3	0.79381	0.92784	0.084295

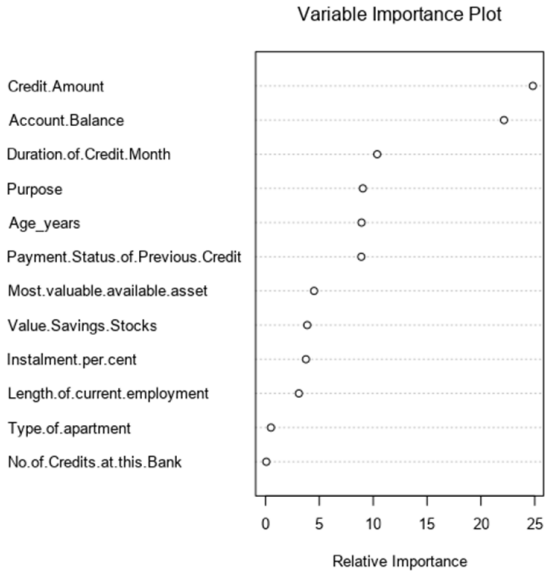
Leaf Summary

node), split, n, loss, yval, (yprob)
* denotes terminal node
1) root 350 97 Creditworthy (0.7228571 0.2771429)
2) Account.Balance=Some Balance 166 20 Creditworthy (0.8795181 0.1204819) *
3) Account.Balance=No Account 184 77 Creditworthy (0.5815217 0.4184783)
6) Duration.of.Credit.Month< 13 74 18 Creditworthy (0.7567568 0.2432432) *
7) Duration.of.Credit.Month>=13 110 51 Non-Creditworthy (0.4636364 0.5363636)
14) Value.Savings.Stocks=< £100,£100-£1000 34 11 Creditworthy (0.6764706 0.3235294) *
15) Value.Savings.Stocks=None 76 28 Non-Creditworthy (0.3684211 0.6315789) *

Forest Model:



Boosted Model:



- 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?

Overall Accuracy:

Model Comparison Report

Fit and error measures

Model	Accuracy	F1	AUC	Accuracy_Creditworthy	Accuracy_Non-Creditworthy
CreditWorthyLogistic	0.7333	0.8291	0.7108	0.9238	0.2889
CreditWorthyTree	0.7467	0.8273	0.7054	0.8667	0.4667
CreditWorthyForest	0.7867	0.8644	0.7389	0.9714	0.3556
CreditWorthyBoosted	0.7867	0.8621	0.7526	0.9524	0.4000

Confusion Matrix:

Confusion matrix of CreditWorthyBoosted

	Actual_Creditworthy	Actual_Non-Creditworthy
Predicted_Creditworthy	100	27
Predicted_Non-Creditworthy	5	18

Confusion matrix of CreditWorthyForest

	Actual_Creditworthy	Actual_Non-Creditworthy
Predicted_Creditworthy	102	29
Predicted_Non-Creditworthy	3	16

Confusion matrix of CreditWorthyLogistic

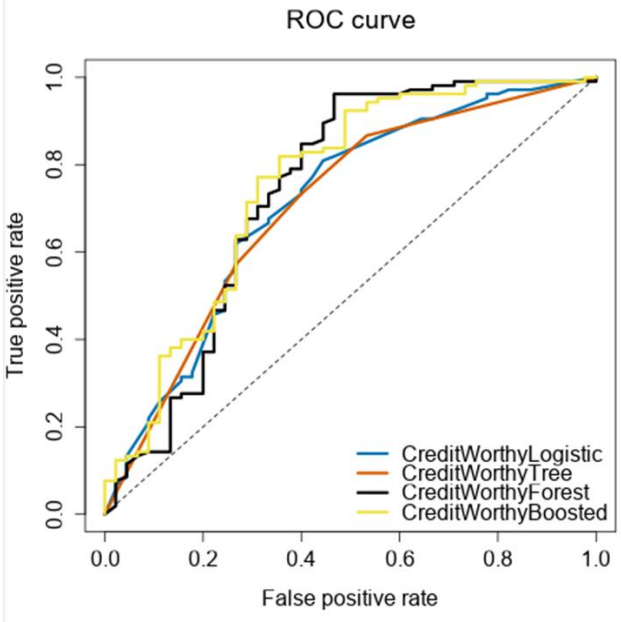
	Actual_Creditworthy	Actual_Non-Creditworthy
Predicted_Creditworthy	97	32
Predicted_Non-Creditworthy	8	13

Confusion matrix of CreditWorthyTree

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

Step 4: Writeup

- 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:
  Boosted Model is the best model to use as its:
  - Overall Accuracy = 0.7867.
  - Almost highest accuracies: Creditworthy = 0.95 and Non creditworthy = 0.4.
  - ROC graph illustrates how this model has the highest AUC:



  - No bias in this model (PPV = 0.78 and NPV = 0.78).
- How many individuals are creditworthy?
  412 customers.