Project: Creditworthiness

Table of Contents

Step 1: Business and Data Understanding

Business Issue

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 - This file contains all credit approvals from your past loan applicants the bank has ever completed.*
2. *The list of customers that need to be processed in the next few days. This is the new set of customers that you need to score on the classification model you will create.*

Key Decisions:

Answer these questions:

- *What decisions need to be made?*
  We need to provide a list of creditworthy customers out of the 500 incoming loan applications based on a series of predictor attributes.

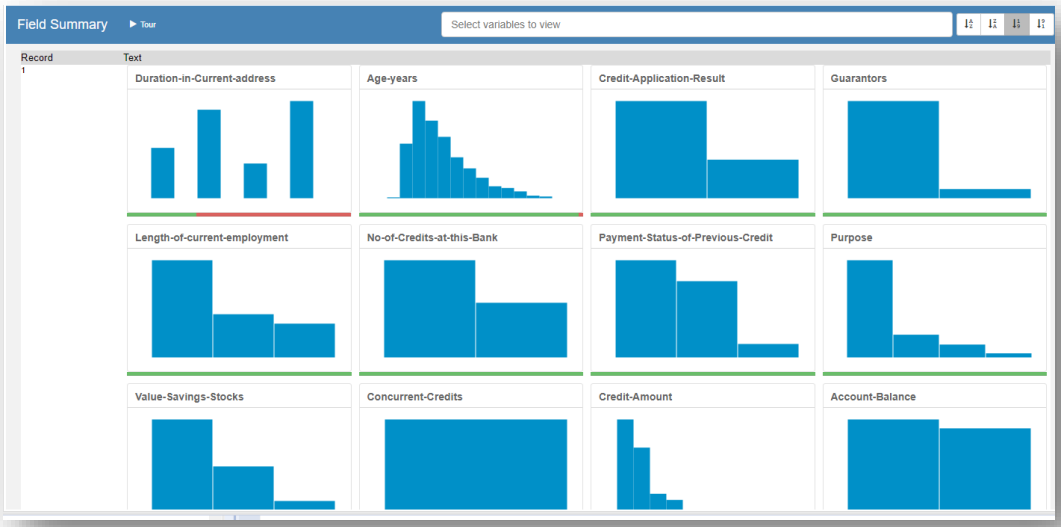
- *What data is needed to inform those decisions?*
  To predict whether a customer will be creditworthy or not, we are given a training set which includes both numeric and categorical variables such as account balance, duration of credit, payment status of previous credit, credit amount, length of current employment and several others.

- *What kind of model (Continuous, Binary, Non-Binary, Time-Series) do we need to use to help make these decisions?*
  Since our target variable is a categorical binary variable (the outcome “Creditworthiness” can be either Yes or No) and we have data related to credit approvals from past loan applicants of our bank, the Problem-Solving Framework suggests that we are in a data-rich scenario and we can go down the path of classification models.

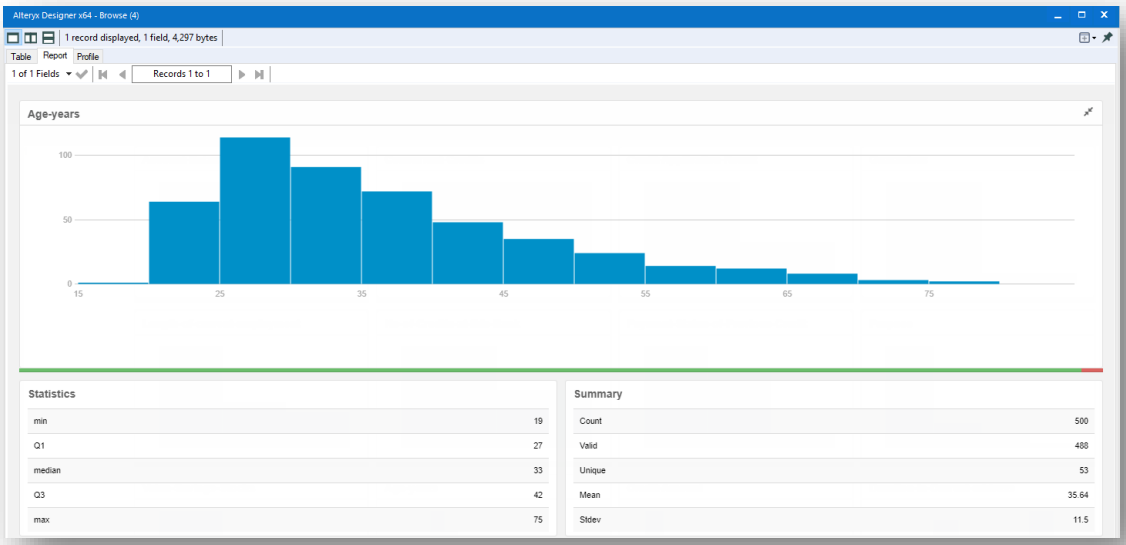
Step 2: Building the Training Set

Answer this question:

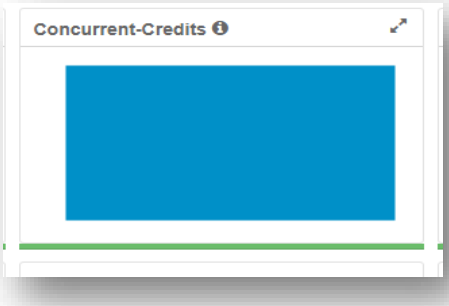
- *In your cleanup process, which fields did you remove or impute? Please justify why you removed or imputed these fields. Visualizations are encouraged.*
  By bringing in the Field Summary Tool, we can see at first glance from the red ticker bar at the bottom of each histogram which fields contain null values. It appears that the following fields have missing values: Duration-in-Current-address (69% missing) and Age-years (2%).

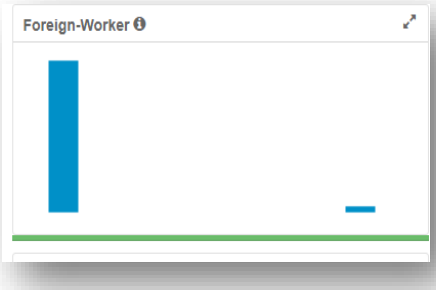


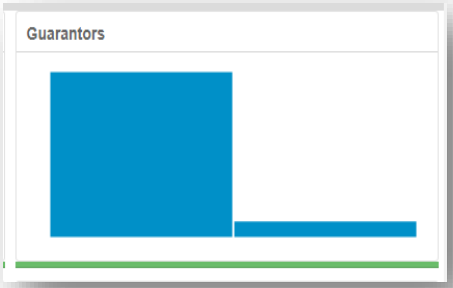
We can remove the Duration-in-Current-address subset from our analysis as the displayed gap is quite large. For the time being, we can keep the Age-years variable as it might turn out to be a valuable attribute for our model. As shown below, the distribution of Age is left or negatively skewed, hence it is recommended to impute these values with its median (33).

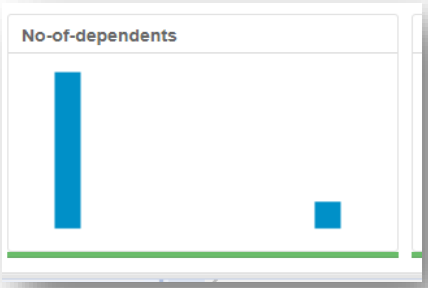


Looking at the distribution of the other variables, we can identify some fields that have zero or near zero variance such as Concurrent Credits (Value “Other Banks/Depts” for all fields), Foreign Workers, Guarantors, Occupation (value “1” for all fields), No-of dependents.

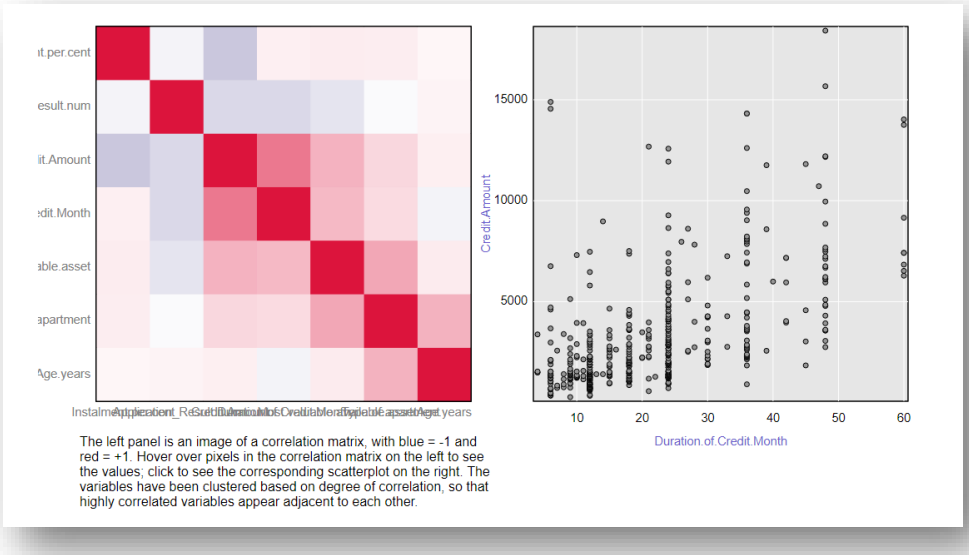








Furthermore, the phone number field should be removed as it is not a useful information in predicting creditworthiness. Eventually, the Association tool shows that numeric variables have little to no inter-correlation. The plot below illustrates the highest correlation value between credit amount and duration of credit amount being around 0.57. It appears that there are not redundant data that we need to remove.



Step 3: Train your Classification Models

*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.*

In a first step, we will examine each model results against the estimation sample. In a second step, we are going to perform cross-validation against the validation set.

Starting with the **Logistic Regression Model**, below is a result overview of the report which detects the following significant predictor variables:

Variable Importance	Account.Balance, Payment.Status.of.Previous..Credit, Purpose, Credit.Amount, Length.of.current.employment, Instalment.per.cent and Most.valuable.available.asset are statistically signicant as shown by their p values and significance codes.
	Account.Balance has the lowest p value and hence the hightest probability that a relationship exists between this predictor and the target variable.
R squared	The R squared is low at 0.22 and this might suggest that the explanatory power of this model is not very strong.

Coefficients:

	Estimate	Std. Error	z value	Pr(>|z|)	
(Intercept)	-3.0136120	1.013e+00	-2.9760	0.00292	**
Account.BalanceSome Balance	-1.5433699	3.232e-01	-4.7752	1.79e-06	***
Duration.of.Credit.Month	0.0064973	1.371e-02	0.4738	0.63565	
Payment.Status.of.Previous.CreditPaid Up	0.4054309	3.841e-01	1.0554	0.29124	
Payment.Status.of.Previous.CreditSome Problems	1.2607175	5.335e-01	2.3632	0.01812	*
PurposeNew car	-1.7541034	6.276e-01	-2.7951	0.00519	**
PurposeOther	-0.3191177	8.342e-01	-0.3825	0.70206	

PurposeUsed car	-0.7839554	4.124e-01	-1.9008	0.05733	.
Credit.Amount	0.0001764	6.838e-05	2.5798	0.00989	**
Value.Savings.StocksNone	0.6074082	5.100e-01	1.1911	0.23361	
Value.Savings.Stocks£100-£1000	0.1694433	5.649e-01	0.3000	0.7642	
Length.of.current.employment4-7 yrs	0.5224158	4.930e-01	1.0596	0.28934	
Length.of.current.employment< 1yr	0.7779492	3.956e-01	1.9664	0.04925	*
Instalment.per.cent	0.3109833	1.399e-01	2.2232	0.0262	*
Most.valuable.available.asset	0.3258706	1.556e-01	2.0945	0.03621	*
Age.years	-0.0141206	1.535e-02	-0.9202	0.35747	
Type.of.apartment	-0.2603038	2.956e-01	-0.8805	0.3786	
No.of.Credits.at.this.BankMore than 1	0.3619545	3.815e-01	0.9487	0.34275	

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: 322.31 on 332 degrees of freedom
McFadden R-Squared: 0.2199, Akaike Information Criterion 358.3

By applying the **Stepwise Regression** tool, the model does not improve considerably with an R squared that is still around 0.20 while the variable Most.valuable.available.asset turns out to be not statistically significant:

Variable importance	Account.Balance, Payment.Status.of.Previous..Credit, Purpose, Credit.Amount, Length.of.current.employment, Instalment.per.cent are statistically signicant as shown by their p values and significance codes.
	Account.Balance has still the lowest p value and hence the hightest probability that a relationship exists between this predictor and the target variable.
R squared	The R squared is low at 0.20 and this might suggest that the explanatory power of this model is not very strong.

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

The **Decision Tree Model** overview illustrates the following results taken from the report below:

Variable importance	Account.Balance, Value.Savings.Stocks and Duration.of.Credit.Month are considered important variables.
Root node error	The root node error shows that about 27% of the data went into the incorrect terminal node.
Confusion Matrix	The confusion matrix illustrates how the model is not accurate at detecting non-creditworthiness as opposed to creditworthiness: about 53% of false positives (non-creditworthy segments that have been incorrectly classified as creditworthy segments). Since the accuracy of creditworthiness is far greater than the accuracy for non-creditworthiness, we may conclude that this model is biased towards creditworthiness. This scenario is common when the number of observations belonging to one class (number of actual creditworthy individuals in the dataset) is significantly higher than those belonging to the other classes (non-creditworthy segments). The overall model accuracy is 79%.

Model Summary

Variables actually used in tree construction:

[1] Account.Balance Duration.of.Credit.Month Purpose

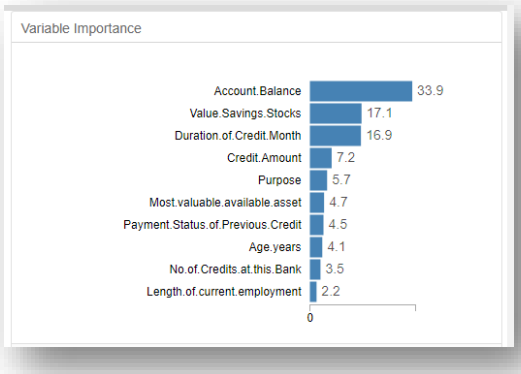
[4] Value.Savings.Stocks

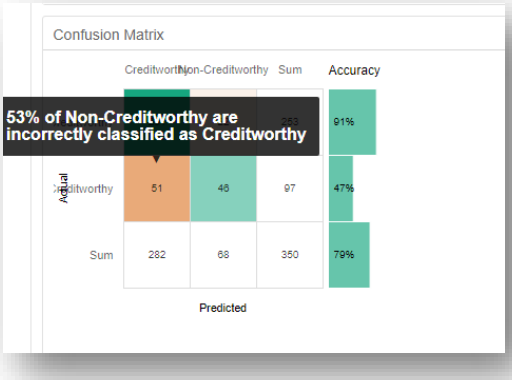
Root node error: 97/350 = 0.27714

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	0.94845	0.084898
3	0.025773	4	0.75258	0.88660	0.083032





The **Forest Model** overview reveals the following results:

Variable importance	Credit.amount, Age.years, DurationofCredit.month and Account.Balance are considered important variables according to the MeanDecreaseGini.
OOB	With 500 trees the out of the bag error lies within the acceptable value of 23.1%, so the missclassification error is lower than the one appearing in the Decision Tree Model.
Confusion Matrix	The confusion matrix illustrates how non-creditworthiness as opposed to creditworthiness is still the toughtest variable to predict as result of class imbalances. This is reflected by the higher classification error (0.649).

Type of forest: classification

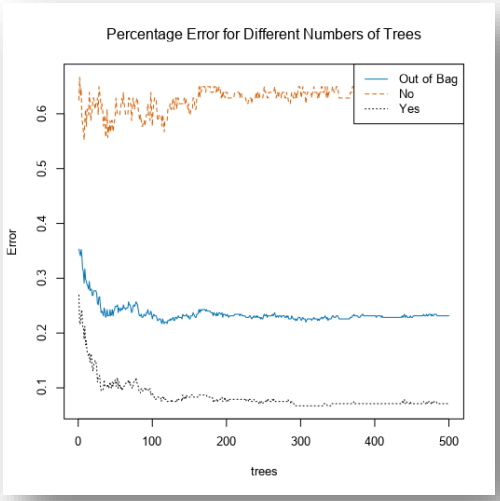
Number of trees: 500

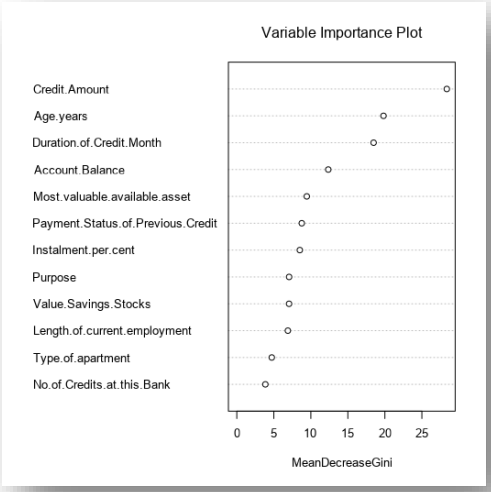
Number of variables tried at each split: 3

OOB estimate of the error rate: 23.1%

Confusion Matrix:

	No	Yes	Classification Error
No	34	63	0.649
Yes	18	235	0.071

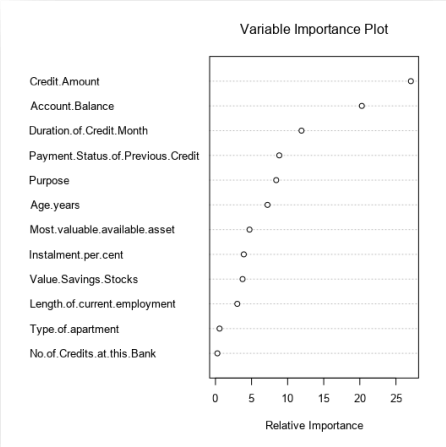


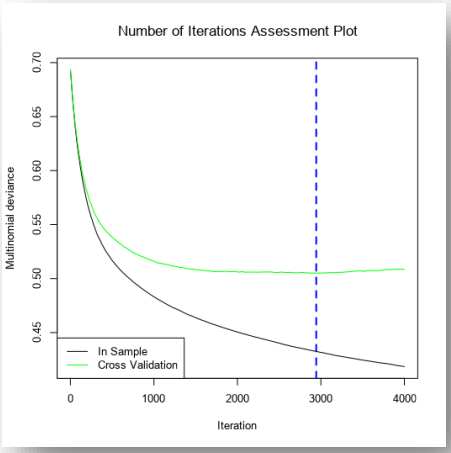


The **Boosted Model** overview has some commonalities with the Forest Model in terms of variable importance:

Variable importance	As shown by the Iterations Assessment Plot, the optimal number of trees needed to to detect important variables such as Credit.amount, Account.Balance, DurationofCredit.Month and Payment.Status.of.Previous.Credit is 2944.

Basic Summary:
Loss function distribution: Bernoulli
Total number of trees used: 4000
Best number of trees based on 5-fold cross validation: 2944

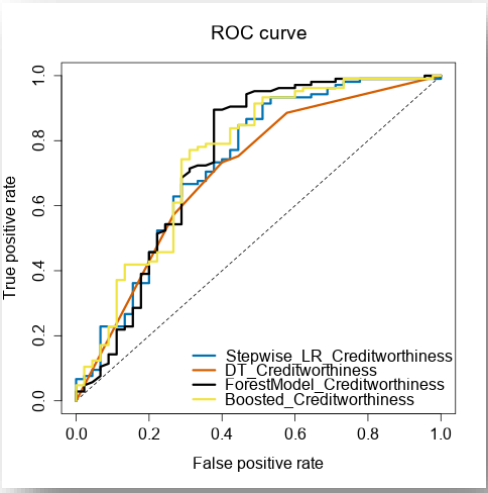


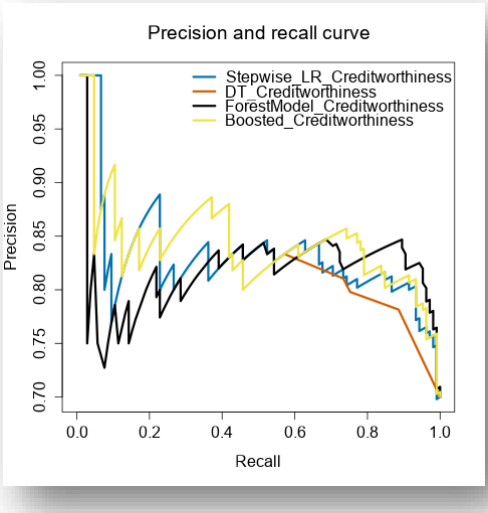


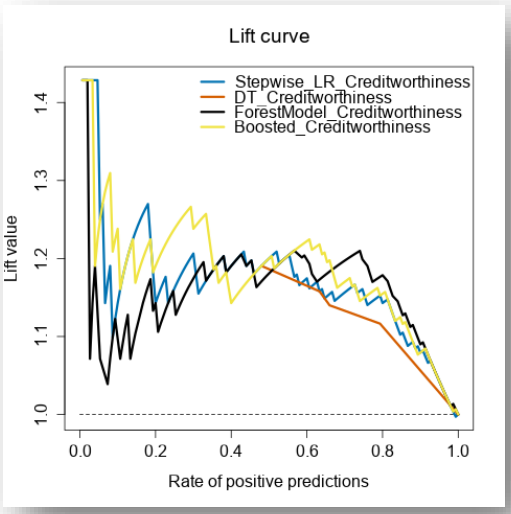
- *Validate your model against the Validation set. What was the overall percent accuracy? Show the confusion matrix. Are there any bias seen in the model's predictions?*

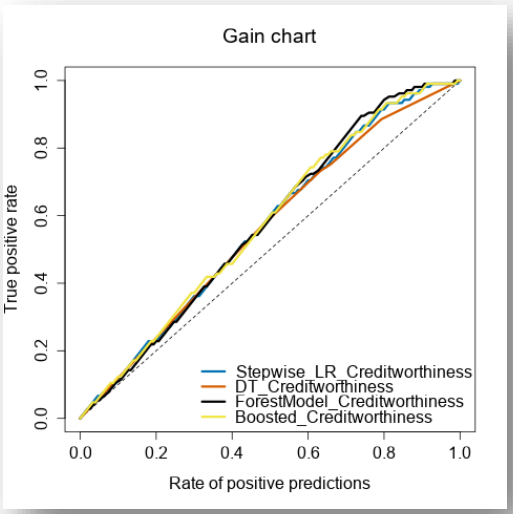
When validating all models against the validation set, we obtain the following results:

Model	Accuracy	Accuracy_No	Accuracy_Yes	F1	AUC
Stepwise_LR_Creditworthiness	0.7600	0.4889	0.8762	0.5500	0.7306
DecisionTree_Creditworthiness	0.7467	0.4222	0.8857	0.5000	0.7035
ForestModel_Creditworthiness	0.8000	0.4000	0.9714	0.5455	0.7460
Boosted_Creditworthiness	0.7867	0.4000	0.9524	0.5294	0.7503









Confusion matrix of Boosted_Creditworthiness

	Actual_No	Actual_Yes
Predicted_No	18	5
Predicted_Yes	27	100

Confusion matrix of DecisionTree_Creditworthiness

	Actual_No	Actual_Yes
Predicted_No	19	12
Predicted_Yes	26	93

Confusion matrix of ForestModel_Creditworthiness

	Actual_No	Actual_Yes
Predicted_No	18	3
Predicted_Yes	27	102

Confusion matrix of Stepwise_LR_Creditworthiness

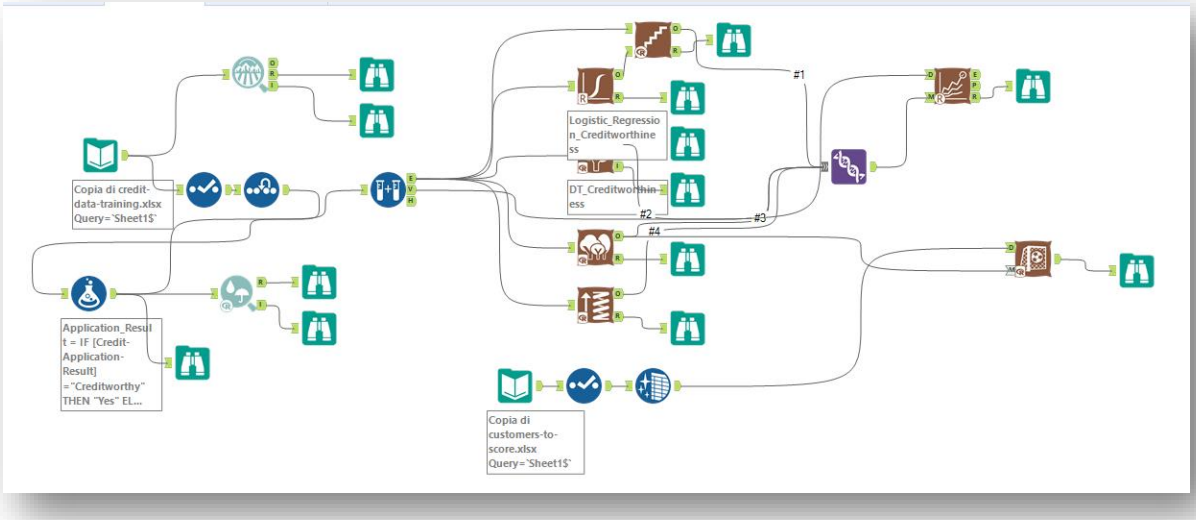
	Actual_No	Actual_Yes
Predicted_No	22	13
Predicted_Yes	23	92

Step 4: Writeup

Decide on the best model and score your new customers.
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.
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*

The workflow is illustrated below:



When validating the models against the validation set, we may conclude that the Forest Model displays the best fit for the following reasons:

Accuracy	The Forest Model displays the highest overall accuracy that is estimated at 80%. In addition to this, it can best predict the creditworthy segments (around 97%) which represent our target variable and lie at the core of our business problem. The confusion matrix of the model reflects this observation. The Boosted Model also does a good job in predicting actual creditworthy applicants. Conversely, the Decision Tree and the Logistic Regression models are biased towards Creditworthiness as they incorrectly classify non-creditworthy segments as creditworthy ones to a much larger extent.
AUC	The area under the curve of the Forest Model - which is the amount of space underneath the ROC curve - is among the largest displayed in the model set. In fact, the ROC plot shows a curve that reaches the upper-left part of the graph at a faster rate, thus originating a larger AUC.
Gain Chart	The gain chart of the Forest Model shows the true response rate on the y axis as compared to the rate of positive predictions on the x axis: the Forest Model reaches the top of true positives quickly and is overall the highest.
Lift Chart	For the Forest Model, the area between the lift curve and the baseline is also large enough and points to an accurate model.

- *How many individuals are creditworthy?*

When scoring the model on the new loan applicants, we predict that the bank will be able to process 405 creditworthy customers.