Project: Creditworthiness

Step 1: Identifying the business problem and the key decisions to be made.

In this case as a loan officer at a young and small bank I needed to classify new customers on whether they can be approved for a loan or not. The I provided a list of creditworthy customers to my manager.

We need the data about the previous customers and new potential customers to decide whether a new customer should be approved for a loan. The target variable is Credit Application Result and I have that information for the previous customers at hand. I also have the predictor variables which may potentially predict this result -such as Account-Balance, Payment-Status-of-Previous-Credit, Length-of-current-employment, Purpose, Credit-Amount, Value-Savings-Stocks, Age-years etc. I have those predictor variables both for our existing customers and the new potential customer at hand.

The target variable is a categorical variable and has just two categories making it a binary variable. Therefore, I used the binary models (such as logistic regression, decision tree, random forest, and boosted models) to predict the outcome variable for the new customers which is either creditworthy or not creditworthy.

Step 2: Building the Training Set

My association analysis indicated some degree of correlation between some of the predictor variables, but I did not think that they were high enough to be concerned with. The highest correlation value is .68 and I do not think that it is a major problem.

Layout

Pearson Correlation Analysis

Full Correlation Matrix

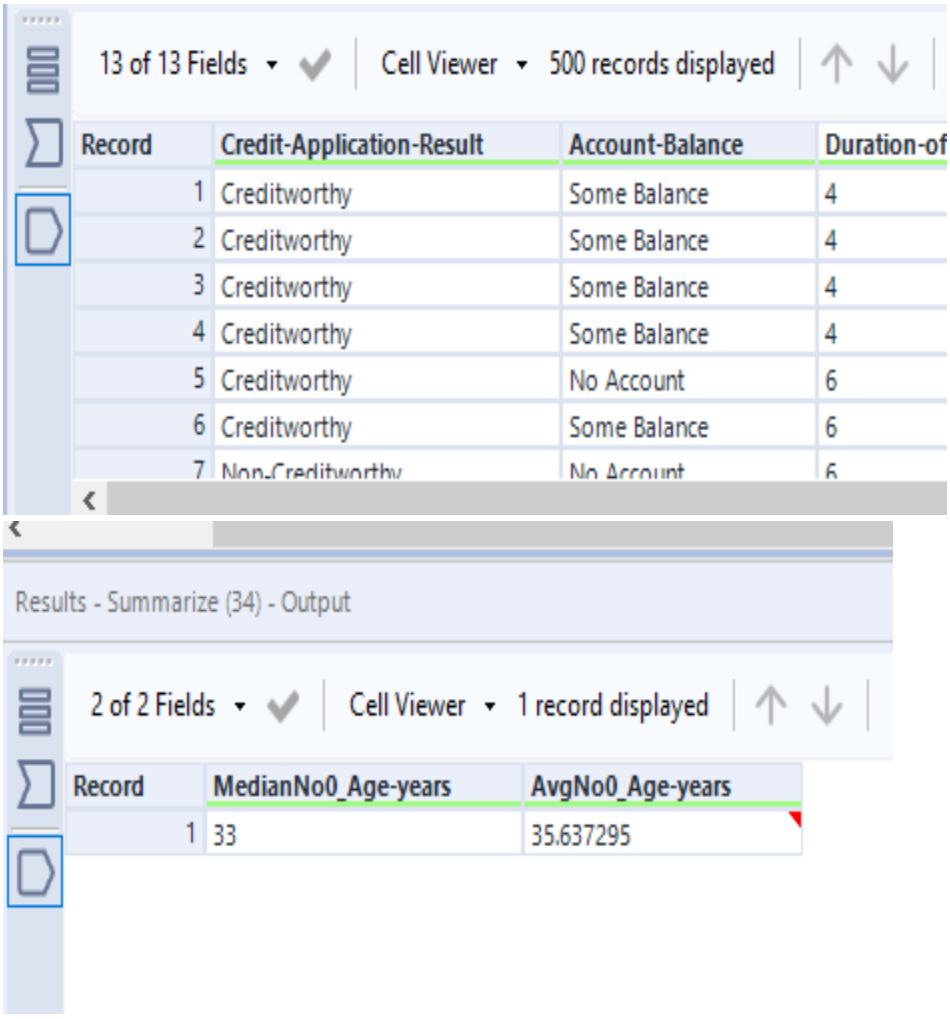
	Duration.of.Credit.Month	Credit.Amount	Instalment.per.cent	Age_Years	Type.of.apartment	Most.valuable.available.asset
Duration.of.Credit.Month	1.000000	0.573980	0.068106	-0.064197	0.152516	0.299855
Credit.Amount	0.573980	1.000000	-0.288852	0.069316	0.170071	0.325545
Instalment.per.cent	0.068106	-0.288852	1.000000	0.039270	0.074533	0.081493
Age_Years	-0.064197	0.069316	0.039270	1.000000	0.329350	0.086233
Type.of.apartment	0.152516	0.170071	0.074533	0.329350	1.000000	0.373101
Most.valuable.available.asset	0.299855	0.325545	0.081493	0.086233	0.373101	1.000000

The initial evaluation of the data indicates that: *Duration in current address* variable has variable has 68.8 % missing/Null data. It is too much missing/Null data for an analysis, and it must be removed. *Age Years* variable has just 2% missing data and it can be handled with imputation. Since mean age is 33, the null/missing *Age years* variable's values will be replaced with 33.

Concurrent-credits variable has only one value and no variability. Therefore, it needs to be removed from the analysis. In the same way, variable *foreign worker* does not seem to have enough variability for an analysis. *Occupation* has only one value with no variability and must be removed from the analysis. *Guarantor, Telephone* and *Number of Dependents* also seem to have low variability.

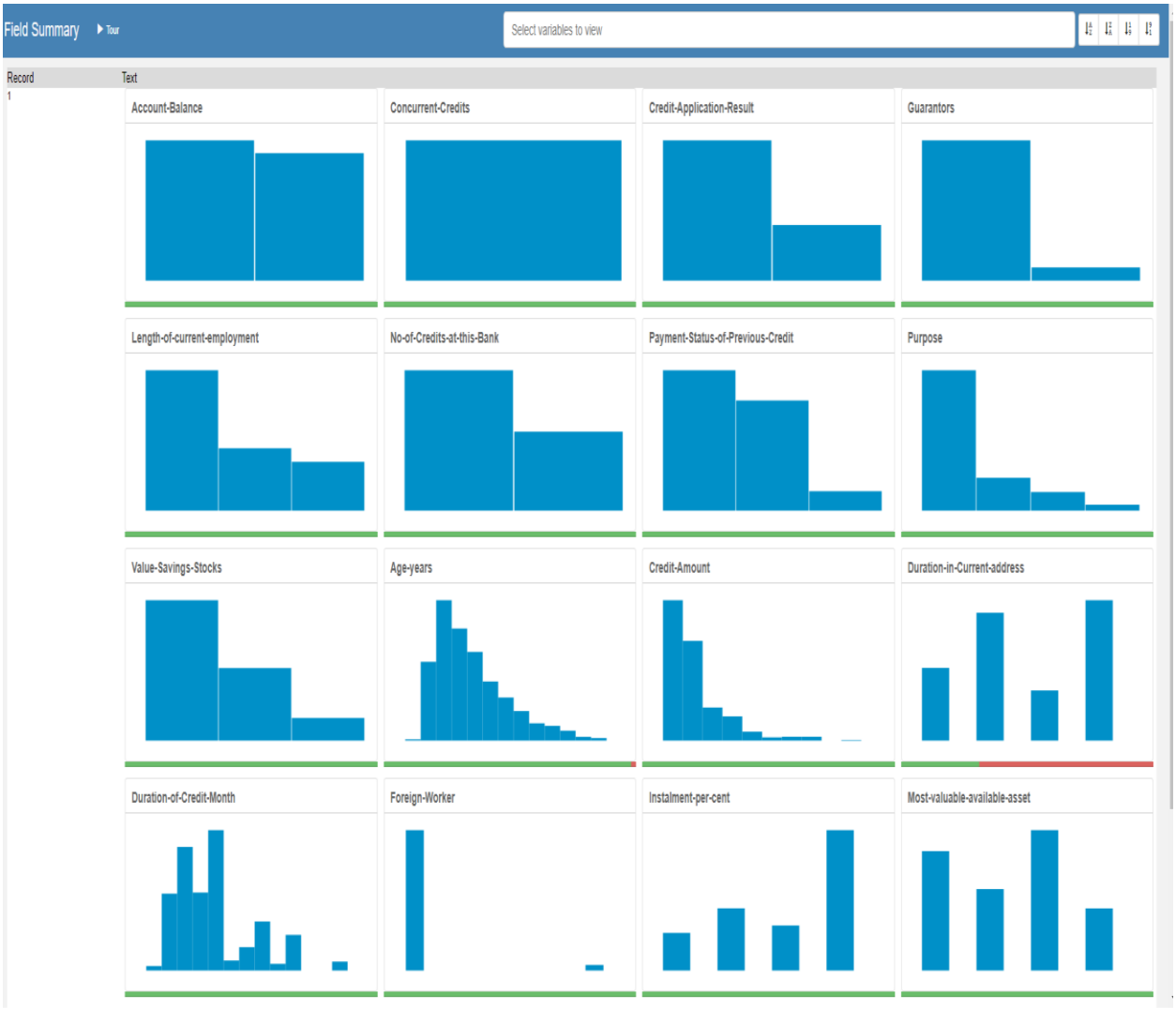
Therefore, I decided to remove them from the analysis, as well. After doing all this data cleaning work I have 13 remaining variables for the analyses.

As shown below; my cleaned data has 13 fields and 500 records, and the mean age is 36. However, I used the median value of age for the imputation, which is 33.

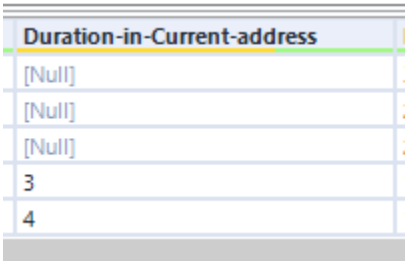


I removed some variables as they had too much missing/Null data (*Duration in Current Address, Age Years)* some had low or no variability (*Concurrent-credits, foreign worker, Occupation, Guarantor* and *Number of Dependents, Telephone).*

Please refer to the graphs below for variables with missing data and low or no variability.



Yellow color on the results window indicates Null values which is 68.8% for the shown variable.

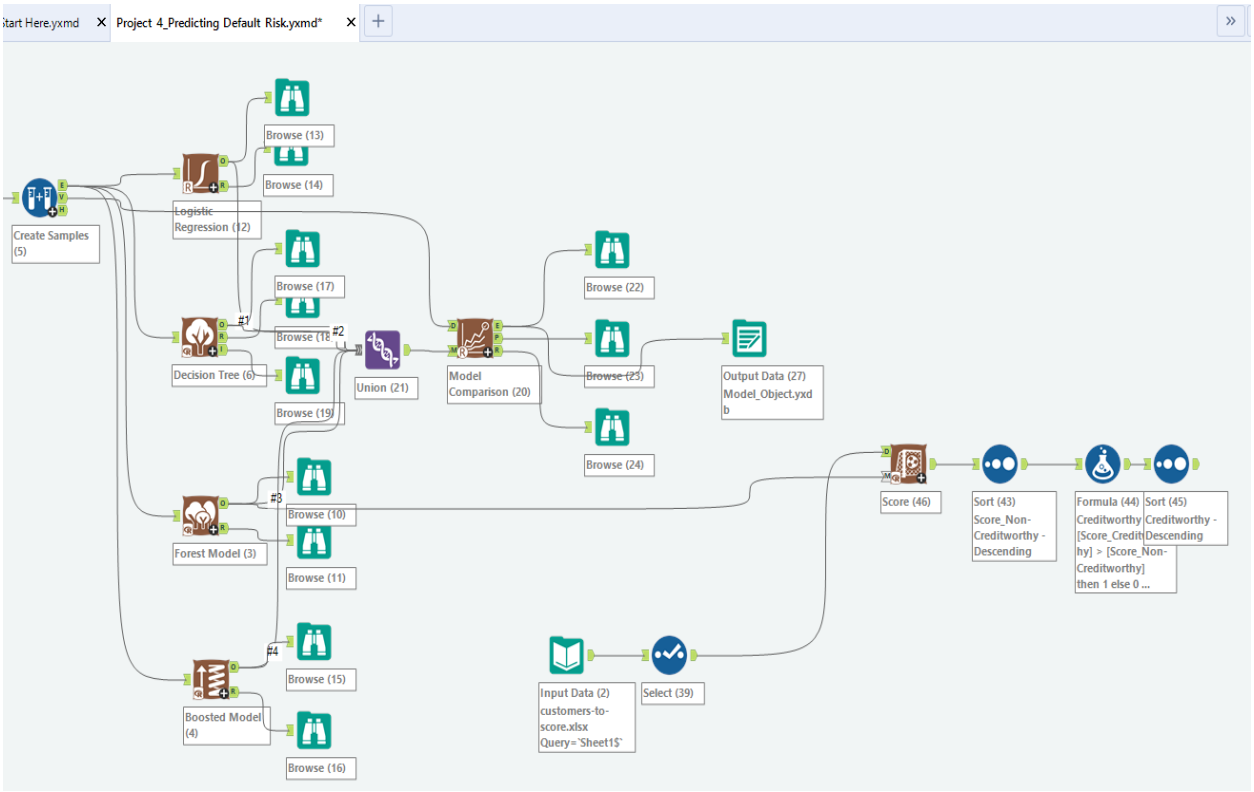


Duration-in-Current-address
[Null]
[Null]
[Null]
3
4

Step 3: Training the Classification Models

First, I created the Estimation and Validation samples where 70% of the dataset went to Estimation and 30% of the entire dataset were reserved for Validation. I set the Random Seed to 1.

I created the models such as Logistic Regression, Decision Tree, Forest Model, Boosted Model to pick the best model among them. Please refer to the Alteryx workflow below for all those steps.



As can be seen below, four different models in general provide similar results but there are also some slight differences in terms of the significance/importance of the variables.

However, variables such as Account Balance, Credit Amount, Value Savings Stocks, Duration of Credit Month, Payment Status of Previous Credit seem to be the most important significant variables.

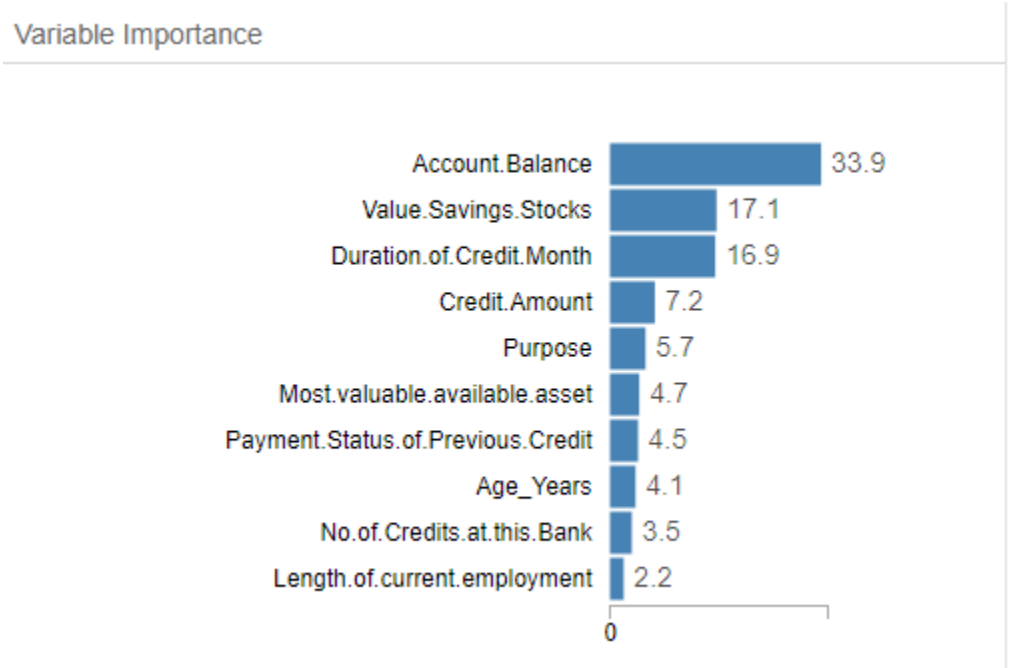
There are also other significant variables which are relatively less important than the variables above.

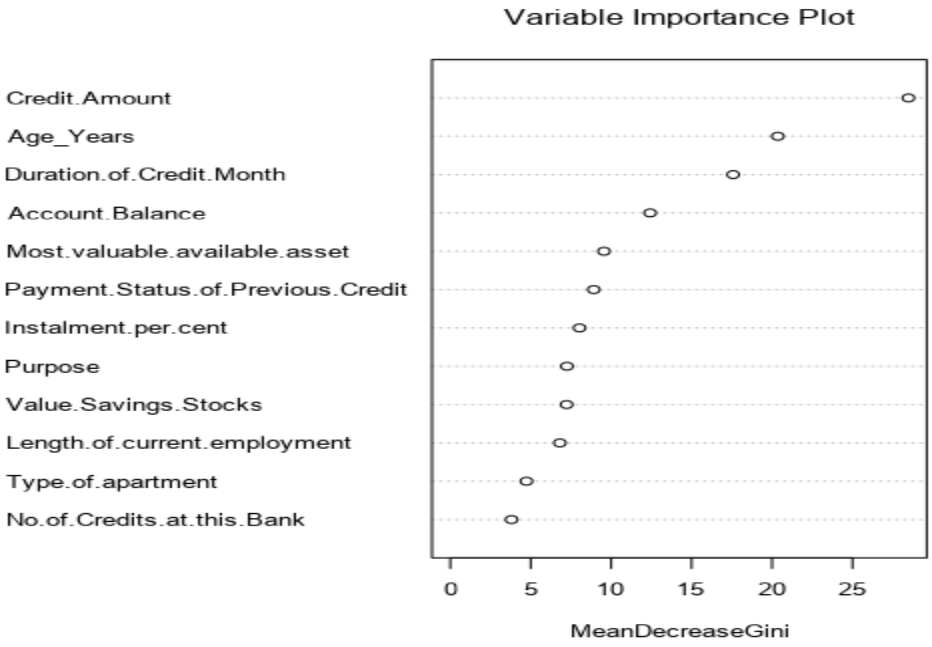
Coefficients:

	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	*
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	
Most.valuable.available.asset	0.3258706	1.556e-01	2.0945	0.03621	*

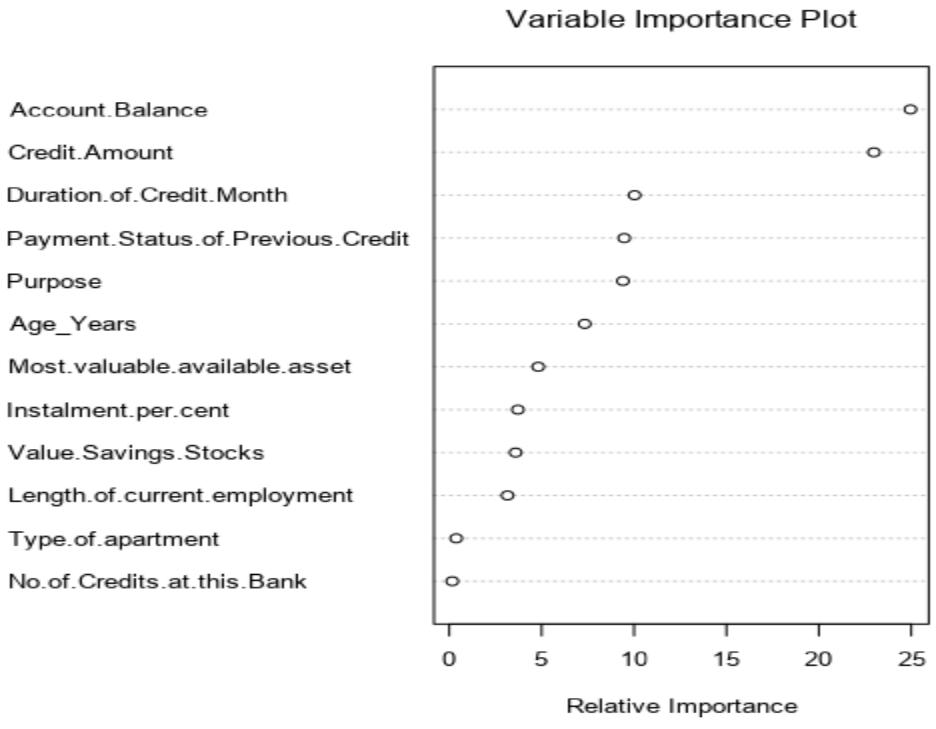
Significance codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial taken to be 1)




Variable Importance Plot
Credit.Amount
Age_Years
Duration.of.Credit.Month
Account.Balance
Most.valuable.available.asset
Payment.Status.of.Previous.Credit
Instalment.per.cent
Purpose
Value.Savings.Stocks
Length.of.current.employment
Type.of.apartment
No.of.Credits.at.this.Bank
0
5
10
15
20
25
MeanDecreaseGini

Plots:


Variable Importance Plot
Account.Balance
Credit.Amount
Duration.of.Credit.Month
Payment.Status.of.Previous.Credit
Purpose
Age_Years
Most.valuable.available.asset
Instalment.per.cent
Value.Savings.Stocks
Length.of.current.employment
Type.of.apartment
No.of.Credits.at.this.Bank
0
5
10
15
20
25
Relative Importance

I validated the model against the Validation set as follows. After creating both training (70%) and validation data (30%) I ran four different binary models, then joined the Model Objects with Union Tool and then compared those models in terms of their success/accuracy via Model Comparison tool.
Below is the accuracy of each model. According to Model Comparison Report it seems that Random Forest Model (81%) has the highest overall accuracy rate and then comes Boosted Model (78%).
For Random Forest Model:
PPV= true positives \ (true positives + false positives) = 101 / (101+24) =.81.
NPV= true negatives \ (true negatives + false negatives) = 21/ (21+4) = .84.
According to the confusion matrix there does not seem to be bias in the model's prediction.

To choose the best model, I also looked at the other results like the accuracy of the predicted categories, Confusion Matrixes, Lift Curves, Gain Charts, and ROC Curves.
ROC Curve indicates that Random Forest Models arrives to the highest point and then gets stable. When we look at the Random Forest Model's performance to predict creditworthy people it is the same with Boosted Model; however, it is way better then Boosted Model in predicting noncreditworthy people. Taken altogether, Random Forest Model which has the highest overall accuracy score seems to be the best model.

Model Comparison Report

Fit and error measures

Model	Accuracy	F1	AUC	Accuracy_Creditworthy	Accuracy_Non-Creditworthy
Decision_Tree_Default_Risk	0.7467	0.8304	0.7035	0.8857	0.4222
Random_Forest_Model_Default_Risk	0.8133	0.8783	0.7385	0.9619	0.4667
Boosted_Model_Default_Risk	0.7867	0.8632	0.7515	0.9619	0.3778
LG_Reg_Model	0.7800	0.8520	0.7314	0.9048	0.4889

Confusion matrix of Boosted_Model_Default_Risk

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

Confusion matrix of Decision_Tree_Default_Risk

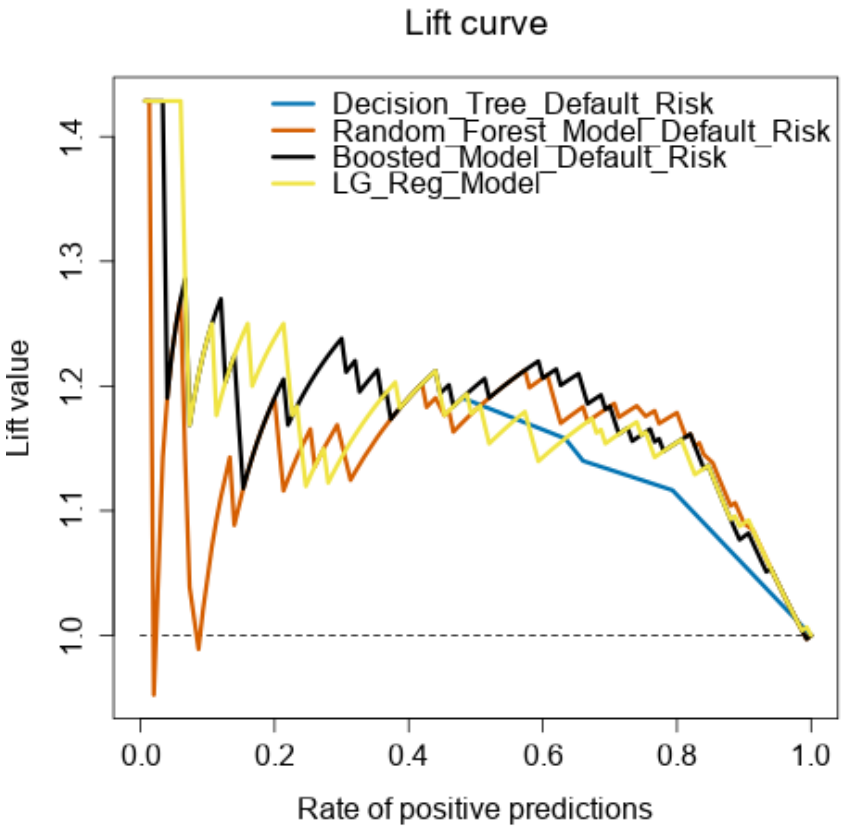
	Actual_Creditworthy	Actual_Non-Creditworthy
Predicted_Creditworthy	93	26
Predicted_Non-Creditworthy	12	19

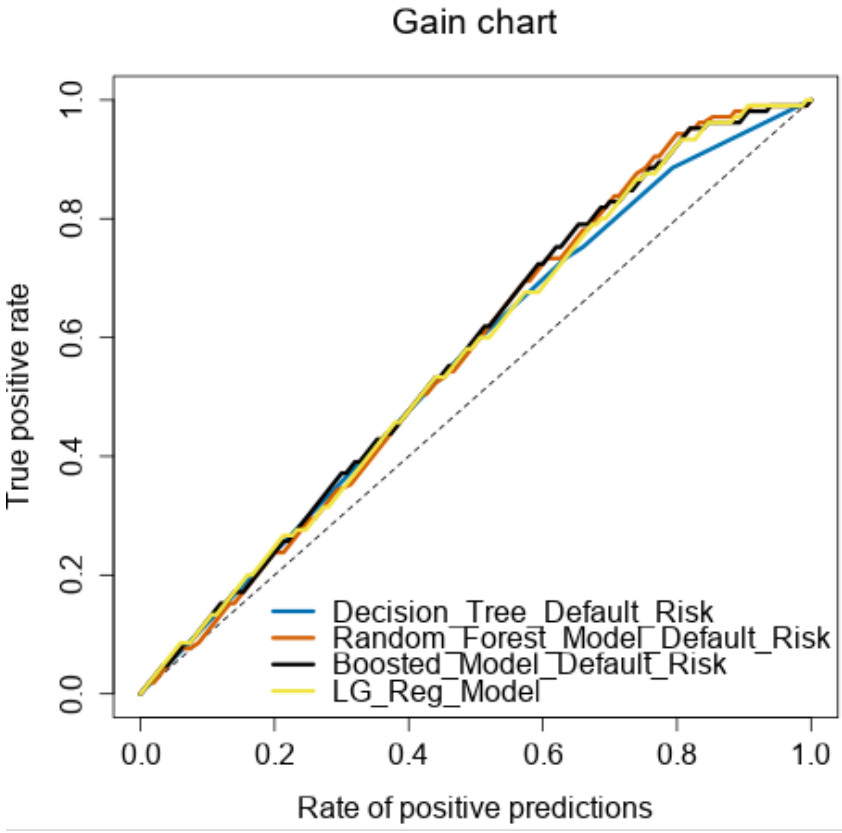
Confusion matrix of LG_Reg_Model

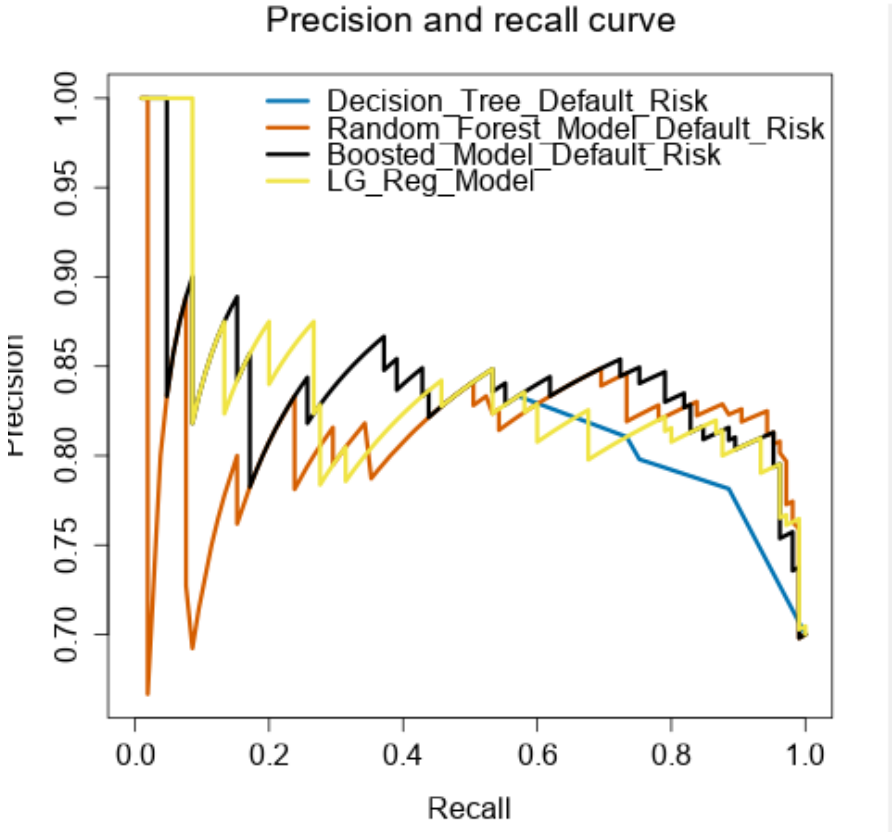
	Actual_Creditworthy	Actual_Non-Creditworthy
Predicted_Creditworthy	95	23
Predicted_Non-Creditworthy	10	22

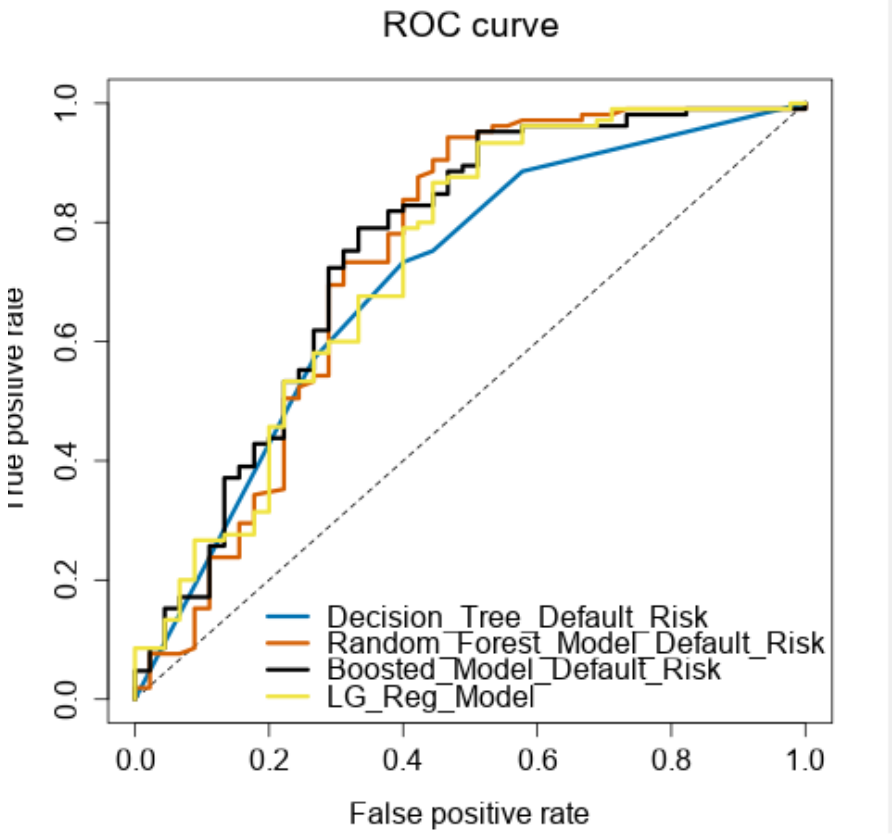
Confusion matrix of Random_Forest_Model_Default_Risk

	Actual_Creditworthy	Actual_Non-Creditworthy
Predicted_Creditworthy	101	24
Predicted_Non-Creditworthy	4	21


Lift curve
Decision_Tree_Default_Risk
Random_Forest_Model_Default_Risk
Boosted_Model_Default_Risk
LG_Reg_Model
Lift value
1.0
1.1
1.2
1.3
1.4
0.0
0.2
0.4
0.6
0.8
1.0
Rate of positive predictions

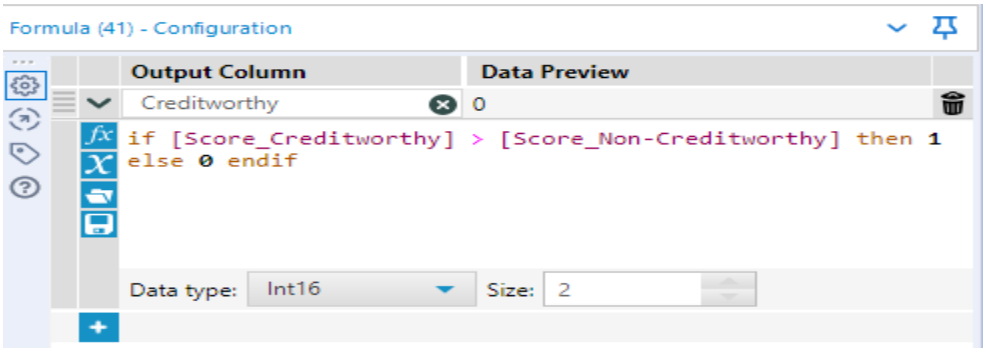

Gain chart
True positive rate
0.0
0.2
0.4
0.6
0.8
1.0
Decision_Tree_Default_Risk
Random_Forest_Model_Default_Risk
Boosted_Model_Default_Risk
LG_Reg_Model
0.0
0.2
0.4
0.6
0.8
1.0
Rate of positive predictions


Precision and recall curve
Decision_Tree_Default_Risk
Random_Forest_Model_Default_Risk
Boosted_Model_Default_Risk
LG_Reg_Model
Precision
1.00
0.95
0.90
0.85
0.80
0.75
0.70
0.0
0.2
0.4
0.6
0.8
1.0
Recall

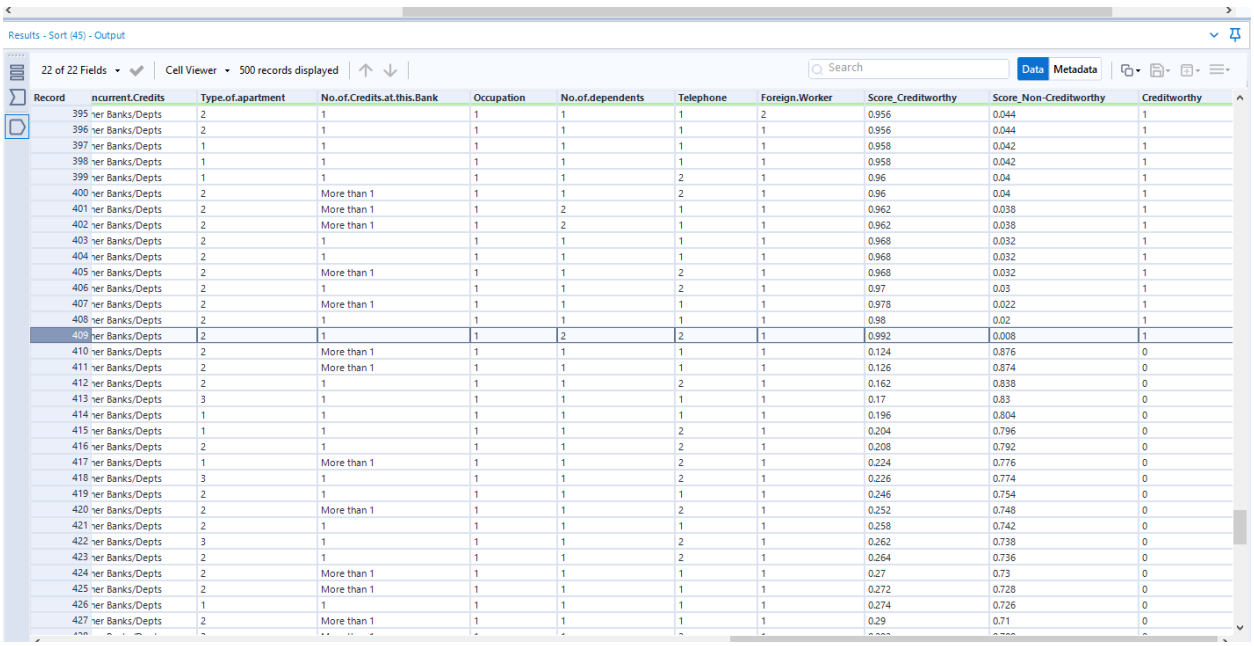

ROC curve
True positive rate
1.0
0.8
0.6
0.4
0.2
0.0
Decision_Tree_Default_Risk
Random_Forest_Model_Default_Risk
Boosted_Model_Default_Risk
LG_Reg_Model
0.0
0.2
0.4
0.6
0.8
1.0
False positive rate

Step 4: Reporting the Results

For the reasons I explained above in detail; I chose the Random Forest Model to best predict the creditworthiness and then applied the Score Tool to identify the customers' creditworthy and non-creditworthy score. Then, I categorized them into two groups. Customers whose creditworthy score is higher than the non-creditworthy score were coded with the number 1 signifying the creditworthiness.



Then I sorted this (Creditworthy) variable and the results indicated that there were 409 people who were creditworthy.



Record	ncurrent.Credits	Type.of.apartment	No.of.Credits.at.this.Bank	Occupation	No.of.dependents	Telephone	Foreign.Worker	Score_Creditworthy	Score_Non-Creditworthy	Creditworthy
395	her Banks/Depts	2	1	1	1	1	2	0.956	0.044	1
396	her Banks/Depts	2	1	1	1	1	1	0.956	0.044	1
397	her Banks/Depts	1	1	1	1	1	1	0.958	0.042	1
398	her Banks/Depts	1	1	1	1	1	1	0.958	0.042	1
399	her Banks/Depts	1	1	1	1	2	1	0.96	0.04	1
400	her Banks/Depts	2	More than 1	1	1	2	1	0.96	0.04	1
401	her Banks/Depts	2	More than 1	1	2	1	1	0.962	0.038	1
402	her Banks/Depts	2	More than 1	1	2	1	1	0.962	0.038	1
403	her Banks/Depts	2	1	1	1	1	1	0.968	0.032	1
404	her Banks/Depts	2	1	1	1	1	1	0.968	0.032	1
405	her Banks/Depts	2	More than 1	1	1	2	1	0.968	0.032	1
406	her Banks/Depts	2	1	1	1	2	1	0.97	0.03	1
407	her Banks/Depts	2	More than 1	1	1	1	1	0.978	0.022	1
408	her Banks/Depts	2	1	1	1	1	1	0.98	0.02	1
409	her Banks/Depts	2	1	1	2	2	1	0.992	0.008	1
410	her Banks/Depts	2	More than 1	1	1	1	1	0.124	0.876	0
411	her Banks/Depts	2	More than 1	1	1	1	1	0.126	0.874	0
412	her Banks/Depts	2	1	1	1	2	1	0.162	0.838	0
413	her Banks/Depts	3	1	1	1	1	1	0.17	0.83	0
414	her Banks/Depts	1	1	1	1	1	1	0.196	0.804	0
415	her Banks/Depts	1	1	1	1	2	1	0.204	0.796	0
416	her Banks/Depts	2	1	1	1	2	1	0.208	0.792	0
417	her Banks/Depts	1	More than 1	1	1	2	1	0.224	0.776	0
418	her Banks/Depts	3	1	1	1	2	1	0.226	0.774	0
419	her Banks/Depts	2	1	1	1	1	1	0.246	0.754	0
420	her Banks/Depts	2	More than 1	1	1	2	1	0.252	0.748	0
421	her Banks/Depts	2	1	1	1	1	1	0.258	0.742	0
422	her Banks/Depts	3	1	1	1	2	1	0.262	0.738	0
423	her Banks/Depts	2	1	1	1	2	1	0.264	0.736	0
424	her Banks/Depts	2	More than 1	1	1	1	1	0.27	0.73	0
425	her Banks/Depts	2	More than 1	1	1	1	1	0.272	0.728	0
426	her Banks/Depts	1	1	1	1	1	1	0.274	0.726	0
427	her Banks/Depts	2	More than 1	1	1	1	1	0.29	0.71	0

In conclusion the Random Forest Model has the highest overall accuracy which is 81 %. It predicts both the creditworthy and non-creditworthy people better than the other models according to the confusion matrixes. ROC Curve indicates that Random Forest Models arrives to the highest point and then gets stable signifying a better performance. There does not seem to be bias in the Random Forest Model as there was not much difference in its success in accurately predicting the creditworthy and non-creditworthy people.

According to the results of my analysis, **409 people are creditworthy**, and **91 people are non-creditworthy**.

PREPARED BY: KADIR AKYUZ