Predicting Default Risk

Business and Data Understanding

Key Decisions:

What decisions needs to be made?

We need to determine which new loan applicants are creditworthy to give a loan to.

What data is needed to inform those decisions?

Our *credit-data-training.xlsx* file contains data on previous loan applications including:

Credit-Application-Result, Account-Balance, Duration-of-Credit-Month, Payment-Status-of-Previous-Credit, Purpose, Credit-Amount, Value-Savings-Stocks, Length-of-current-employment, Installment-per-cent, Most-valuable-available-asset, Age-years, Type-of-apartment, and No-of-Credits-at-this-Bank.

Our *customers-to-score.xlsx* file contains data for our 500 new loan applications.

What kind of model (Continuous, Binary, Non-Binary, Time-Series) do we need to use to help make these decisions?

We need to use a Binary classification model to help us decide which loan applicants are deemed either creditworthy or non-creditworthy.

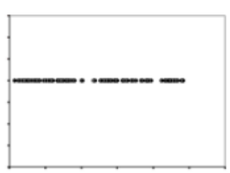
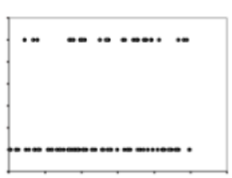
Building the Training Set

Using the Summarize tool to count the number of null values for each field, I chose to remove **Duration-In-Current-address** because it contained too many null values. I chose to impute **Age-years** because only 12 values were null so replacing those missing values with the median average age in years would be appropriate.

CountNull_Duration-in-Current-address	CountNull_Most-valuable-available-asset	CountNull_Age-years
344	0	12

Looking at the interactive output of the Field Summary tool, I chose to remove **Concurrent-Credit, Foreign-Worker, No-of-dependents** and **Guarantors** because they showed low variability in their bar graphs.

The reports output of the Field Summary tool showed 0 variance for **Occupation** field, so I removed this field. Similarly, **Telephone** showed low variability with a standard deviation of .490 and logically there is no reason to include this variable.

Occupation		.0%	1	1.000	1.000	1.000	1.000	.000	This field has a small number of unique values, and appears to be a categorical field. Consider changing the field data type to "string".
Telephone		.0%	2	1.000	1.400	1.000	2.000	.490	This field has a small number of unique values, and appears to be a categorical field. Consider changing the field data type to "string".

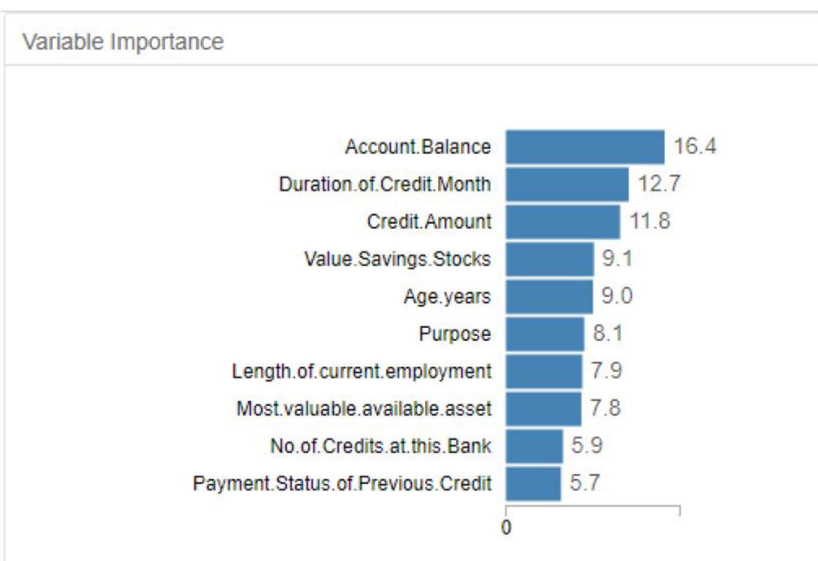
Train your Classification Models

The Logistic Regression model report identified **Account-Balance, Purpose,** and **Credit-Amount** as the most significant variables with p-values less than .01.

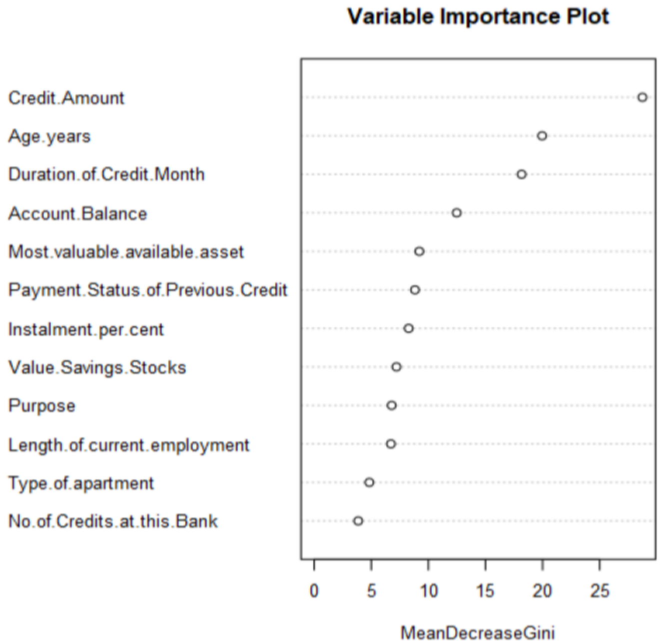
Coefficients:

	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	.

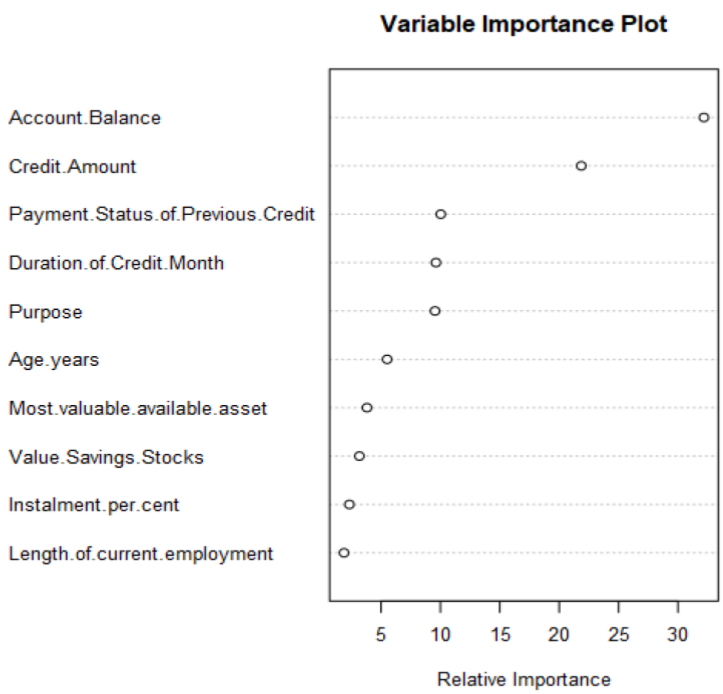
For the *Decision Tree* model, **Account-Balance, Duration-of-credit-month, Credit Amount,** and **Value-savings-stocks** were among the most important predictor variables.



For the *Forest Model* **Credit Amount, Age-Years,** and **Duration-of-credit-month** were among the most important predictor variables.



For the *Boosted Model*, **Account Balance** and **Credit Amount** were the 2 most important predictor variables.



The Model Comparison tool compared the performance of each our respective predictive models against the validation set with the accuracy and confusion matrices shown below:

Model	Accuracy	F1	AUC	Accuracy_Creditworthy	Accuracy_Non-Creditworthy
Decision_Tree	.6733	.7721	.6296	.7905	.4000
Forest_Model	.7933	.8681	.7368	.9714	.3778
Boosted_Model	.7867	.8632	.7524	.9619	.3778
Logistic_Regression_Stepwise	.7600	.8364	.7306	.8762	.4889

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	83	27
Predicted_Non-Creditworthy	22	18

Confusion matrix of Forest_Model

	Actual_Creditworthy	Actual_Non-Creditworthy
Predicted_Creditworthy	102	28
Predicted_Non-Creditworthy	3	17

Confusion matrix of Logistic_Regression_Stepwise

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

The Logistic Regression model had a strong overall percent accuracy of 76%.

PPV = true positives / (true positives + false positives) = 92 / (92+23) = .8

NPV = true negatives / (true negatives + false negatives) = 22 / (22+13) = .62

Checking the confusion matrix, there is bias seen in the model's prediction to Creditworthy.

The Decision Tree model had a good overall percent accuracy of 67.33%.

PPV = 83 / (83 + 27) = .75

NPV = 18 / (18 + 22) = .45

Checking the confusion matrix, there is bias towards Creditworthy.

The Forest Model had a strong overall percent accuracy of 79.33%.

PPV = 102 / (102 + 28) = .78

NPV = 17 / (17 + 3) = .85

Checking the confusion matrix, there is no bias seen in the model's prediction.

The Boosted model had a strong overall percent accuracy of 78.67%.

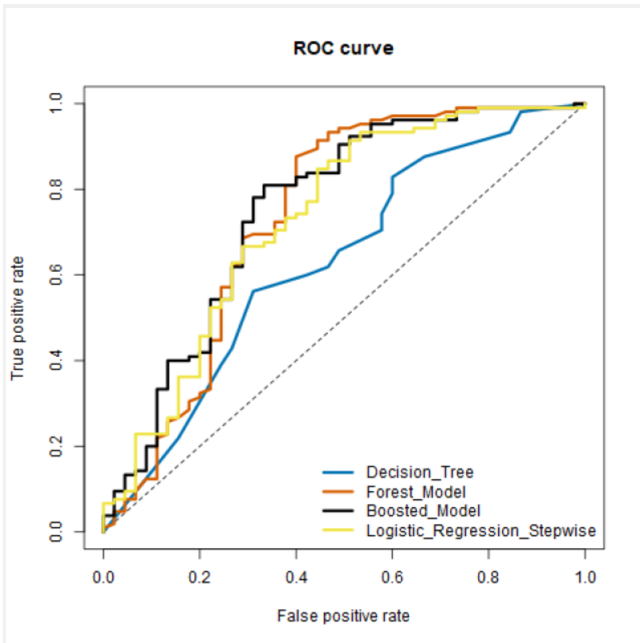
PPV = 101 / (101 + 28) = .78

NPV = 17 / (17 + 4) = .81

Checking the confusion matrix, there is no bias seen in the model's prediction.

Writeup

I decided to use the Forest model because it yielded the highest overall percent accuracy for identifying Creditworthy customers against the Validation set at 79.33%. It had the strongest accuracy rate of 97.14% within the “Creditworthy” segment and comparable accuracy rates to other models predicting the “Non-creditworthy” segment at 37.78%.



The Forest model has one of the highest ROC curves displayed in red above. It has one of the greatest AUC’s at .7368, where a higher AUC is associated with a better model. Furthermore, the model is not biased as the confusion matrix reveals a positive predicted value of .78 and negative predicted value of .85. Because the difference between the values is very small, it is safe to assume that there is a negligible effect of bias.

Among the 500 new customers, 410 loan applicants were Creditworthy.

Alteryx Workflows: