Determining Creditworthiness

Business and Data Understanding

Business Problem

The bank recently gained 500 new loan applications and needs to determine each applicant's Creditworthiness. Therefore, we will take historical data from applications where the decision has been made (creditworthy or Non-Creditworthy), create a model based on that data, and then apply that model to the new applications.

Data Needed

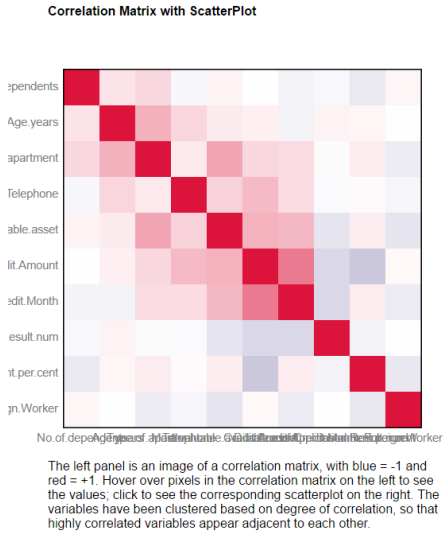
The data needed to inform these decisions is contained in the credit-data-training.xlsx Excel file and contains the following variables: Credit-Application-Result, Account-Balance, Duration-of-Credit-Month, Payment-Status-of-Previous-Credit, Purpose, Credit-Amount, Value-Savings-Stocks, Length-of-current-employment, Instalment-per-cent, Guarantors, Duration-in-Current-address, Most-valuable-available-asset, Age-years, Concurrent-Credits, Type-of-apartment, No-of-Credits-at-this-Bank, Occupation, No-of-dependents, Telephone, and Foreign-Worker. It will be determined which of these is important to solving the business problem.

Model Type

We are attempting to solve a business problem where the outcome is either Creditworthy or Non-Creditworthy. Because there are only two outcomes we will be using a binary model type, which consists of the following model types: Logistic Regression, Decision Tree, Forest Model, and Boosted Tree.

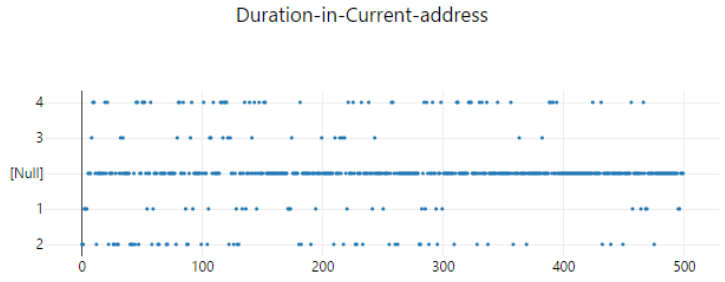
Building the Training Set

The correlation of the set does not reach the "high" (0.70) standard and therefore does not need to be addressed. However, we do need to address variables that need to be removed or fixed before analysis can be completed.

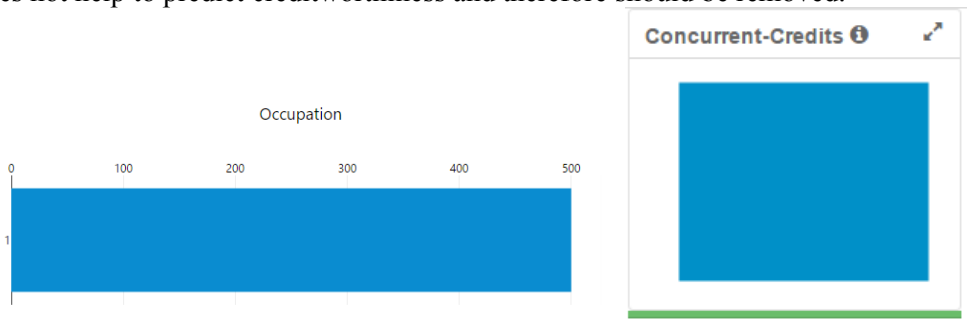


Removed Variables

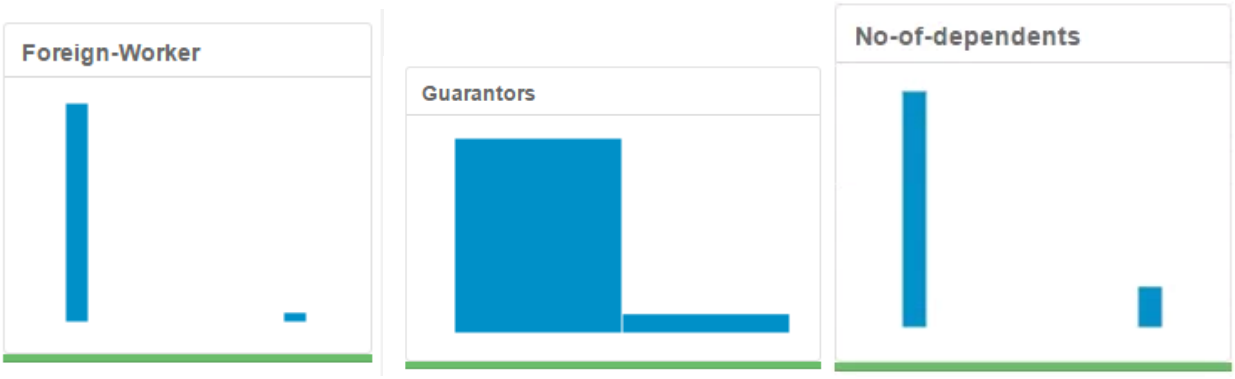
Duration-in-Current-address was removed because 69% of the data is not available. Imputation of this high percent should not happen because the data could be skewed.



Occupation and **Concurrent-Credits** - removed because all the records are the same. The records being the same does not help to predict creditworthiness and therefore should be removed.



Foreign-Worker, **Guarantors and Numer-of-Dependants** - have values that are over 99% of one value. Because of the low amount of variance, these variables do not provide enough information to make predictions, therefore they were removed.

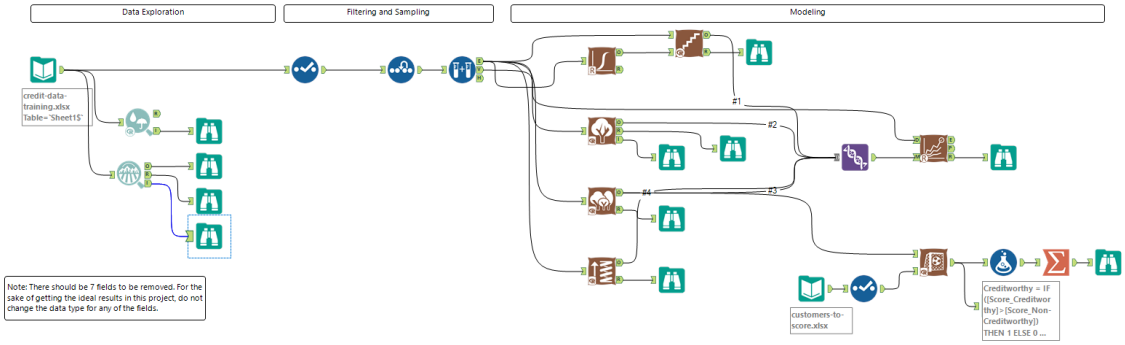


Telephone - was removed from the predictor variable because there is no logical reason for including them as they do not relate to outcome.

Imputed Data

Age-Years - needed to be imputed because there were 2% of there values missing. These values needed to align with the others in the data set and needed to remain whole numbers. Using imputation, these values were created using median. Median provides a value that will not negatively adjust the data set as it is in the middle, and will provide a whole number.

Training Classification Models



First, I created Estimation and Validation samples where 70% of the dataset goes to Estimation and 30% of the dataset is reserved for Validation. Next, these were tested against four model types: Logistic Regression, Decision Tree, Forest Model, and Boosted Tree. The following chart is the model comparison report.

Model	Accuracy	F1	AUC	Accuracy_Creditworthy	Accuracy_Non-Creditworthy
DecTree	0.7467	0.8273	0.7054	0.7913	0.6000
Forest	0.8133	0.8783	0.7350	0.8080	0.8400
Boosted	0.7867	0.8632	0.7524	0.7829	0.8095
LogRegStep	0.7600	0.8364	0.7306	0.8000	0.6286

Logistic Regression

The Logistic Regression model was 76% accurate. The variables that are significant are:

Account-Balance (Some Balance) – pvalue=1.65e-07
Credit-Amount - pvalue=0.00296
Instalment-per-cent - pvalue=0.02549
Length of current employment (< 1yr) - pvalue=0.03596
Payment-Status-of-Previous-Credit (Some Problems) - pvalue=0.0183
Purpose (New car) - pvalue=0.00566

Any p-value under .05 is considered to statistically significant.

	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	***
Credit.Amount	0.0001704	5.733e-05	2.9716	0.00296	**
Instalment.per.cent	0.3016731	1.350e-01	2.2340	0.02549	*
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	*
Most.valuable.available.asset	0.2650267	1.425e-01	1.8599	0.06289	.
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	.

Within the confusion matrix, the Logistic Regression model was 80% accurate with regard to Creditworthy, and 63% accurate with regard to Non-Creditworthy. There is negative bias towards Non-Creditworthy.

Confusion matrix of LogRegStep

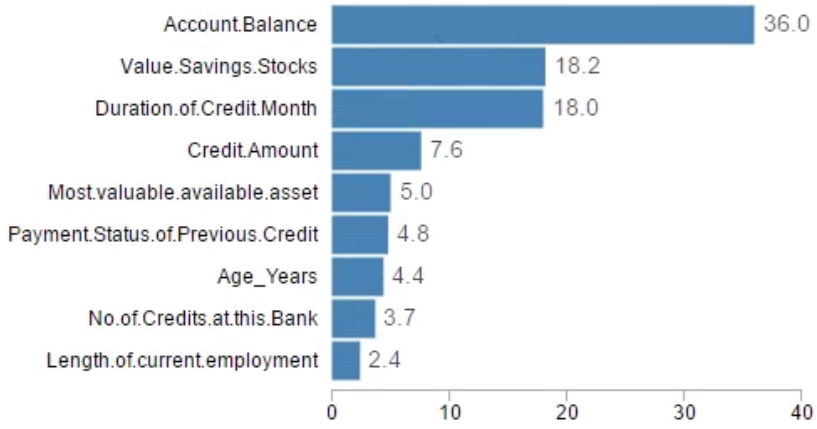
	Actual_Creditworthy	Actual_Non-Creditworthy
Predicted_Creditworthy	92	23
Predicted_Non-Creditworthy	13	22

Decision Tree

The Decision Tree model was 75% accurate.

The variables that are significant are:
Primary (these appear farthest to the right in the Variable Importance Plot):
Account-Balance

Secondary (These appear in the middle of the Variable Importance Plot):
Value-Saving-Stocks
Duration-of-Credit-Month



Within the confusion matrix, the Decision Tree model was 79% accurate with regard to Creditworthy, and 60% accurate with regard to Non-Creditworthy. There is negative bias towards Non-Creditworthy.

Confusion matrix of DecTree

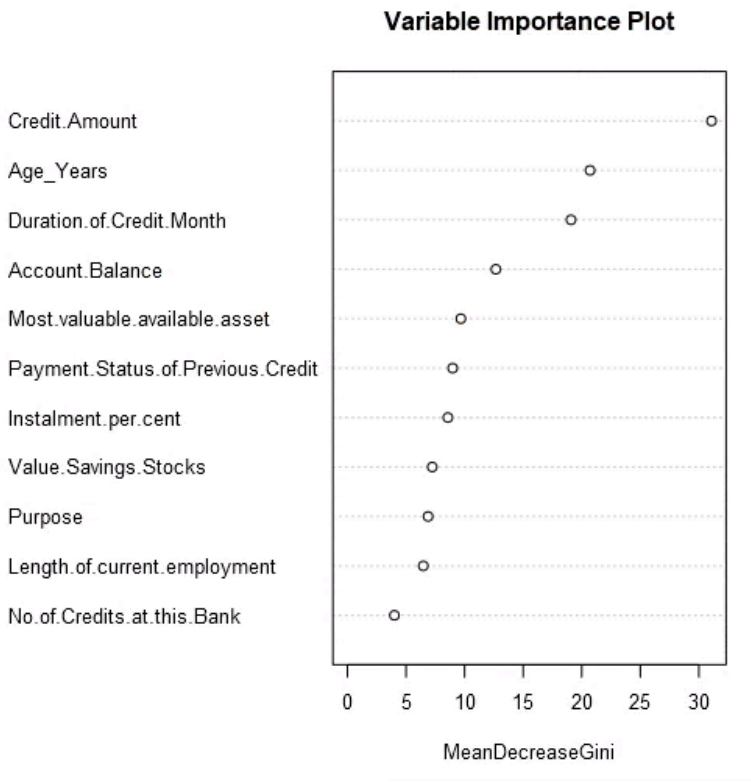
	Actual_Creditworthy	Actual_Non-Creditworthy
Predicted_Creditworthy	91	24
Predicted_Non-Creditworthy	14	21

Forest model

The Forest model was 81% accurate.

The variables that are significant are:
Primary (these appear farthest to the right in the Variable Importance Plot):
Credit Amount
Age-Years
Duration-of-Credit-Months

Secondary (These appear in the middle of the Variable Importance Plot):
Account Balance
Most-valuable-available-asset
Payment-Status-of-Previous-Credit
Instalment-per-cent



Within the confusion matrix, the Forest model was 80% accurate with regard to Creditworthy, and 84% accurate with regard to Non-Creditworthy. The bias was close to equal between the two variables.

Confusion matrix of Forest

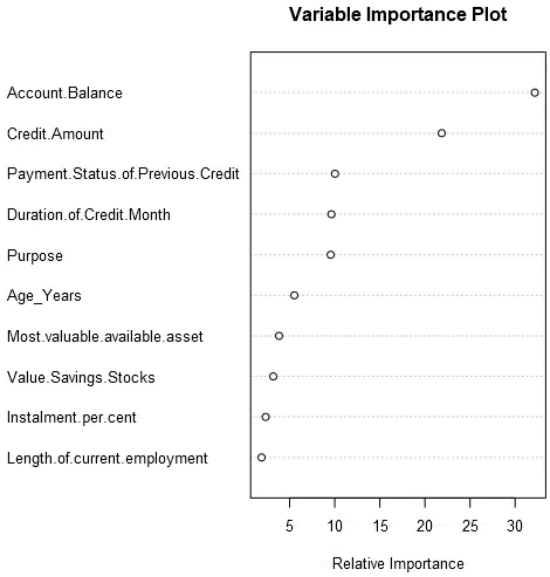
	Actual_Creditworthy	Actual_Non-Creditworthy
Predicted_Creditworthy	101	24
Predicted_Non-Creditworthy	4	21

Boosted model

The Boosted model was 79% accurate.

The variables that are significant are:
Primary (these appear farthest to the right in the Variable Importance Plot):
Account Balance
Credit Amount

Secondary (These appear in the middle of the Variable Importance Plot):
Payment-Status-of-Previous-Credit
Duration-of-Credit-Month
Purpose



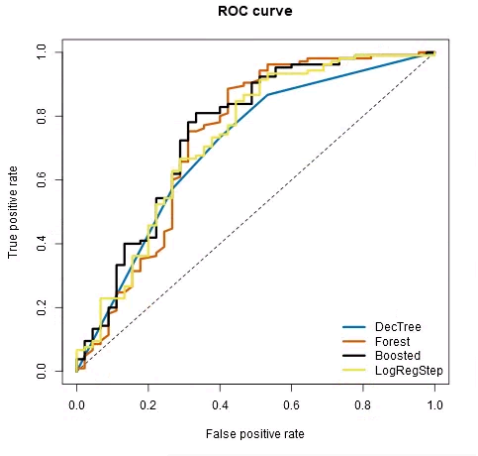
Within the confusion matrix, the Boosted model was 78% accurate with regard to Creditworthy, and 81% accurate with regard to Non-Creditworthy. The bias was close to equal between the two variables.

Confusion matrix of Boosted

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

Writeup

By comparing the models, it can be seen that the forest model had the highest accuracy among the four models when tested against the validation set. the Forest model was 80% accurate with regard to Creditworthy, and 84% accurate with regard to Non-Creditworthy, where both accuracies were higher than any other model. The Forest model also had the highest percentage of accuracy and lowest bias of the four models in the confusion chart. And finally, on the ROC chart, the forest model reached stayed above the others more often and reached the top faster than the others.



For these reasons, the Forest model was one chosen to check for Creditworthiness in the 500-applications data set. After adjusting the likelihood of an application being Creditworthy, based on the previous criteria, 406 applications were seen as creditworthy.