Project: Creditworthiness

Step 1: Business and Data Understanding

Key Decisions:

Answer these questions

- What decisions needs to be made?

There is a group of 500 people that are interested in the loan, and we need to decide who should receive credit and who we should reject.

- What data is needed to inform those decisions?

- Past loan applicant's data – personal details about customer such as age and how long they are employed in the current job.

- Individuals financial history.

- What is the purpose of the loan and how big it is.

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

The model is a binary model – as this is yes or no answer.

Step 2: Building the Training Set

Build your training set given the data provided to you. The data has been cleaned up for you already so you shouldn't ***need to convert any data fields to the appropriate data types.***

To achieve consistent results reviewers expect.

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.

This are the positions that I was sure about my logic.

Only one position. It would skew data so I will remove it from the model.

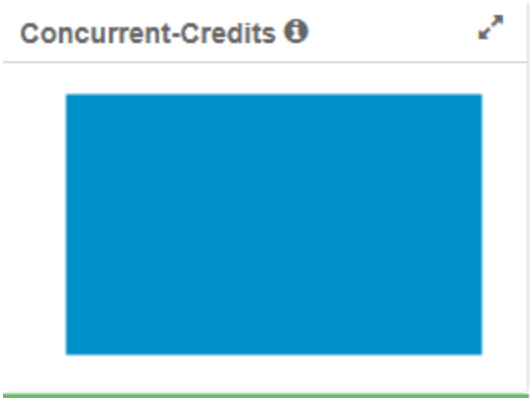


Figure 1. Field Summary – Concurrent - Credits

There are only two options available and one is majority. It will skew the data - I will remove it from the model. Also it has very small impact on the model.

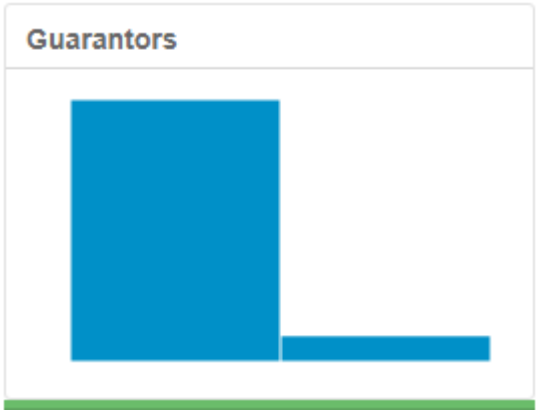


Figure 2. Field Summary – Guarantors

Too many null values - I will remove it from the model.

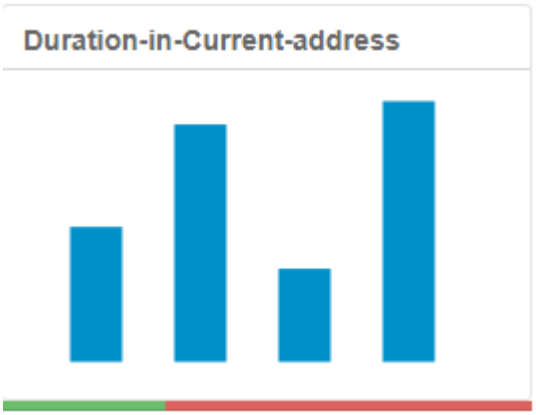


Figure 3. Field Summary – Duration in Current Address

Imputed median to replace nulls - but I will keep it in the model. I have decided to choose median as thanks to that we are mitigating the risk of skewing the data.

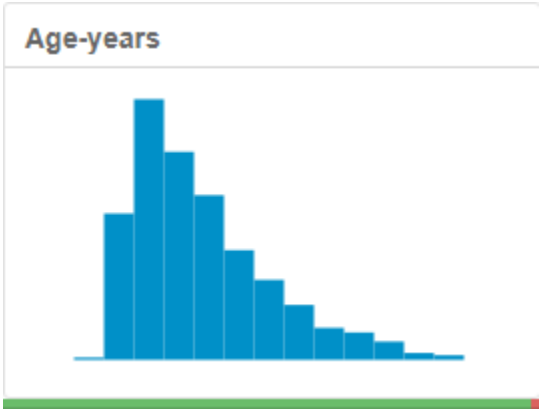


Figure 4. Field Summary – Age Years

Foreign worker. One category has the majority of the positions. It would skew the data - I will remove it from the model.

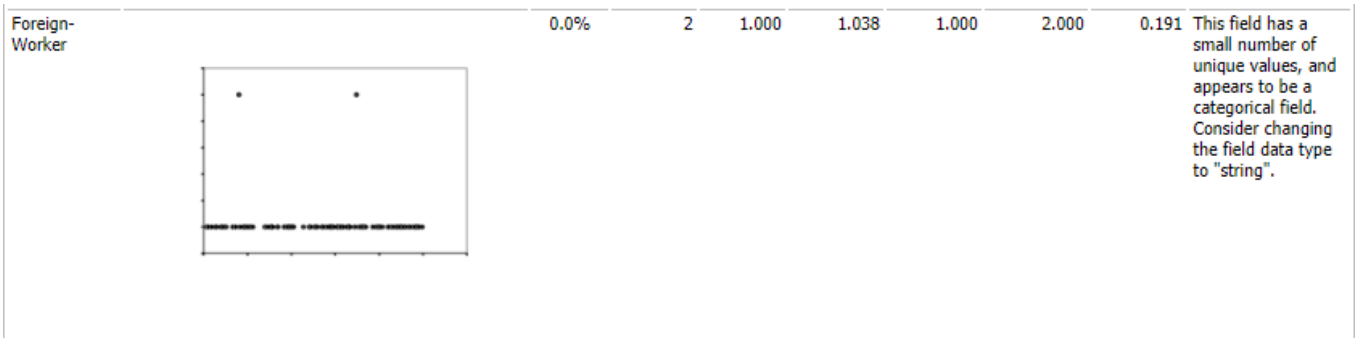


Figure 5. Field Summary – Foreign Worker

The last that I have removed is no of dependents. In the majority of the model it was the least useful information regarding the model quality so I decided to remove it.

Step 3: Train your Classification Models

First, create your Estimation and Validation samples where 70% of your dataset should go to Estimation and 30% of your entire dataset should be reserved for Validation. Set the Random Seed to 1.

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.

Please see the screens below. I will also write the most important predictor variables next to the name of the model.

Logistic regresion – Account.Balance. We can observe 3 stars at the end of the table that is a information that this is really good predictor variable for a model.

	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	

Figure 6. Report for Logistic Regression Model

Boosted Model – Credit amount + Account balance

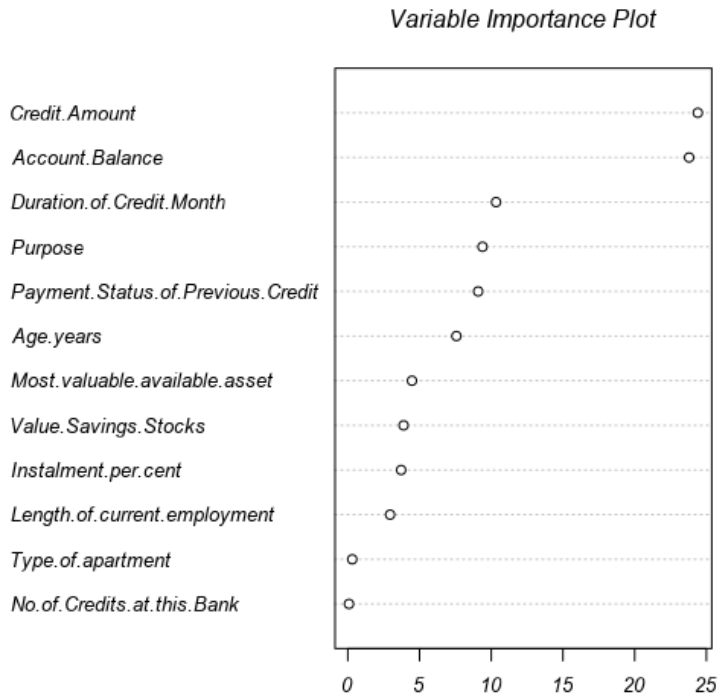


Figure 7. Variable Importance Plot – Boosted Model

Forest model

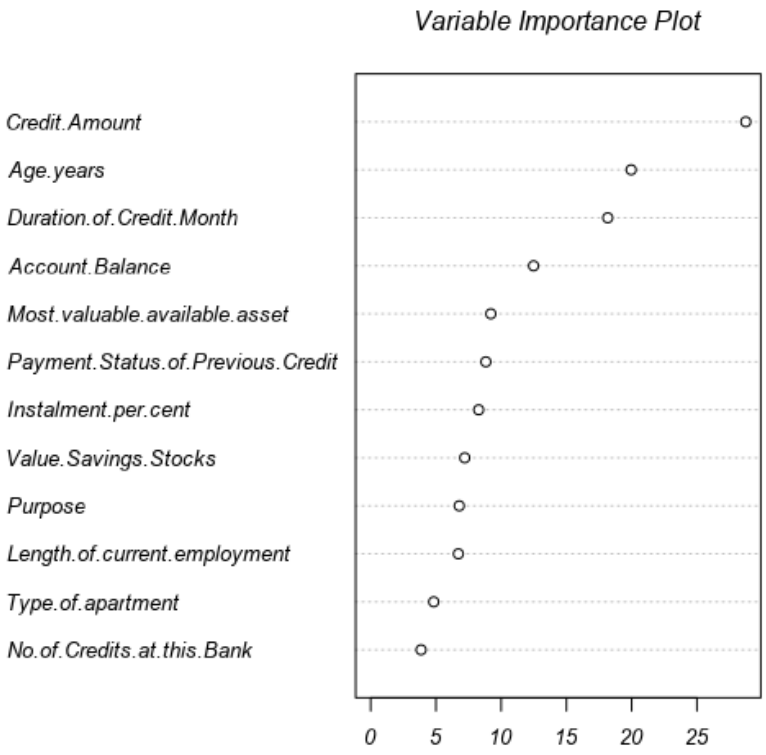


Figure 8. Variable Importance Plot – Forest Model

Tree model – Account balance

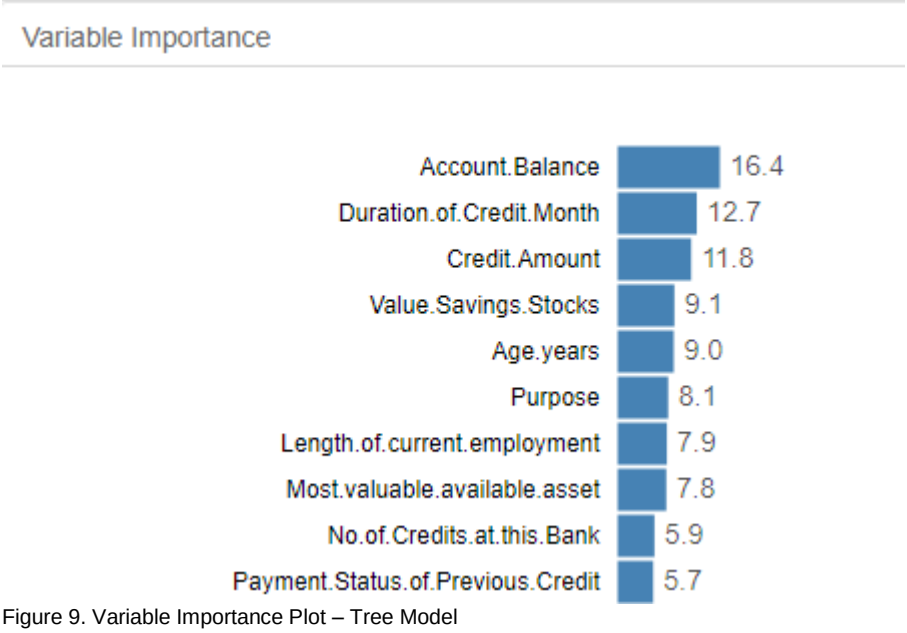


Figure 9. Variable Importance Plot – Tree Model

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

Using the table below, I identified in which model we can observe biases and why.

In the first table below we can see that one model is less accurate than the rest(decision tree). Rest of the models have close accuracy.

Model	Accuracy	F1	AUC	Accuracy_Creditworthy	Accuracy_Non-Creditworthy
Linear_regresion	0.7800	0.8520	0.7314	0.8051	0.6875
Decision_Tree	0.6733	0.7721	0.6296	0.7545	0.4500
Forest_Model	0.7933	0.8681	0.7368	0.7846	0.8500
Boosted_Model	0.7867	0.8632	0.7524	0.7829	0.8095

Figure 10. Accuracy of all tested models

Below table show model names and their biases if they have any.

Model name	*Bias*
Linear Regresion	*The model prediction for the noncredit worthy people is lower than the rate for the credit worthy part.*
Decision Three	*The model prediction for the noncredit worthy people is lower than the rate for the credit worthy part.*
Forest Model	*No bias*
Boostem Model	*No bias*

The previous table was build using confusion, matrixes that I have pasted below.

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 Linear_regresion

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

Figure 11. Confusion Matrix of Each Model

Step 4: Writeup

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

The answer for the point 1 and 2 can be found in the table below:

Model	Accuracy	F1	AUC	Accuracy_Creditworthy	Accuracy_Non-Creditworthy
Linear_regresion	0.7800	0.8520	0.7314	0.8051	0.6875
Decision_Tree	0.6733	0.7721	0.6296	0.7545	0.4500
Forest_Model	0.7933	0.8681	0.7368	0.7846	0.8500
Boosted_Model	0.7867	0.8632	0.7524	0.7829	0.8095

Figure 12. Accuracy of all tested models

Ideal ROC curve hugs the top left corner, indicating a high true positive rate and a low false-positive rate. Looking at the chart below we can see that the boosted model raises the fastest and is highest for most of the graph. Forest model and linear regression are also very high, but we not high enough. The decision tree has the lowest performance as it is lower than lines for other models.

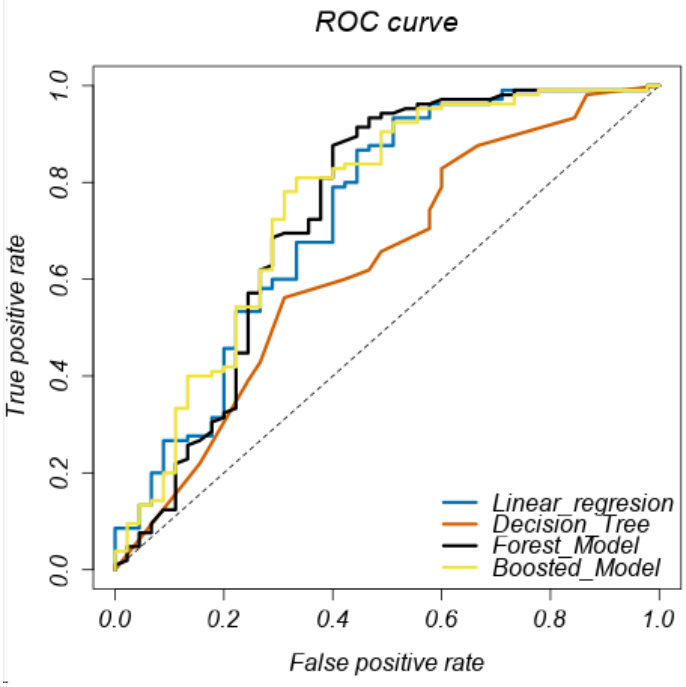


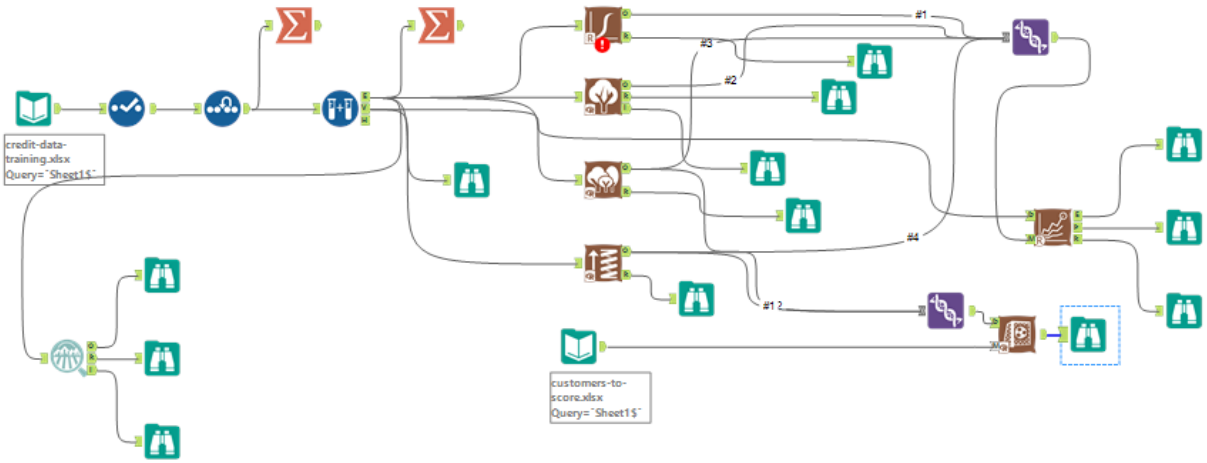
Figure 13. Roc Curve

Note: Remember that your boss only cares about prediction accuracy for Creditworthy and Non-Creditworthy segments.

- How many individuals are creditworthy?

Four hundred ten individuals are creditworthy. To provide this result, I have combined the outcome of the two most accurate models.

Alteryx Workflows



Workflow 1. Comparing the Prediction Models – Providing Recommendation