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.

Machine generated alternative text:
Concurrent-credits O 

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.

Machine generated alternative text:
Guarantors 

Figure 2. Field Summary – Guarantors

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

Machine generated alternative text:
Duration-in-current-address 

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.

Machine generated alternative text:
Age-years 

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.

Machine generated alternative text:
Foreign 
0.0% 
0.191 
This field has a 
small number of 
unique values, and 
appears to be a 
categorical field. 
Consider changing 
the field data type 
to "string 

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.

Machine generated alternative text:
(Intercept) 
Account.BalanceSome Balance 
Duration. of. Credit.Month 
Payment.Status.of. Previous.CreditPaid Up 
Payment. Status.of. Previous.CreditSome Problems 
PurposeNew car 
PurposeOther 
PurposeUsed car 
Credit.Amount 
Value.Savings.StocksNone 
value.savings.stocksE100-E1000 
Length.of.current.employment4-7 yrs 
Length.of.current.employmentc lyr 
Instalment.per.cent 
Most.valuable.available.asset 
Age.years 
Type. of. apartment 
No.of.Credits.at.this.aankMore than 
Estimate 
-3.0136120 
1.5433699 
0.0064973 
0.4054309 
1.2607175 
1.7541034 
-0.3191177 
-o. 7839554 
0.0001764 
0.6074082 
0.1694433 
0.5224158 
o. 777g4g2 
0.3109833 
0.3258706 
-0.0141206 
-0.2603038 
0.3619545 
Std. Error 
1.013e+00 
3.232e-01 
1.371e-02 
3.841e-01 
5.335e-01 
6.276e-01 
8.342e-01 
4.124e-01 
6.838e-05 
5.100e-01 
5.64ge-01 
4.930e-01 
3.956e-01 
1.3gge-01 
1.556e-01 
1.535e-02 
2.956e-01 
3.815e-01 
z value 
-2.g760 
-4.7752 
O 38 
1.0554 
2.3632 
-2.7951 
-0.3825 
l.goos 
2.57gs 
1.1911 
0.3000 
I.osgs 
1.g664 
2.2232 
2.0945 
-0.9202 
-0.8805 
0. g487 
Pre Izl) 
0.002g2 * 
1.79e-06 
0.63565 
0.29124 
0.01812 * 
0.00519 * 
o. 70206 
0.05733 . 
o.oogsg * 
0.23361 
0.7642 
0.28934 
0.04925 * 
0.0262 * 
0.03621 * 
O. 35747 
0.3786 
0.34275 

 Figure 6. Report for Logistic Regression Model

Boosted Model – Credit amount + Account balance

Machine generated alternative text:
Credit A mount 
A ccount. Balance 
Duration. of. Credit. Month 
Purpose 
Payment. Status. of. Previous. Credit 
Age. years 
Most. valuable. available. asset 
Value. Savings. Stocks 
Instalment. per. cent 
Length. of. current. employment 
Type. of. apartment 
No. of Credits. at. this. Bank 
o 
Variable Importance Plot 
25 
5 
10 
15 
20 

Figure 7. Variable Importance Plot – Boosted Model

Forest model

Machine generated alternative text:
Credit A mount 
Age. years 
Duration. of. Credit. Month 
AGGount Balance 
Most. valuable. available. asset 
Payment. Status. of. Previous. Credit 
Instalment. per. cent 
Value. Savings. Stocks 
Purpose 
Length. of. current. employment 
Type. of. apartment 
No. of Credits. at. this. Bank 
o 
Variable Importance Plot 
5 
10 
15 
20 
25 

Figure 8. Variable Importance Plot – Forest Model

Tree model – Account balance

Machine generated alternative text:
Variable Importance 
Account Balance 
Duration_of.Credit Month 
Credit Amount 
Value_Savings_Stocks 
Ageyears 
Purpose 
Length ffcurrent employment 
Most valuable available.asset 
No ofCredits.at this Bank 
Payment.Status of Previous_Credit 
16.4 
11.8 
9.0 
7.9 
• 39 
5.7 

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.

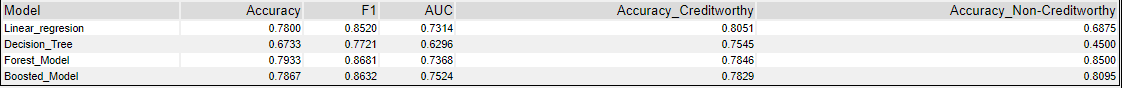


Figure 10. Accuracy of all tested models

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.

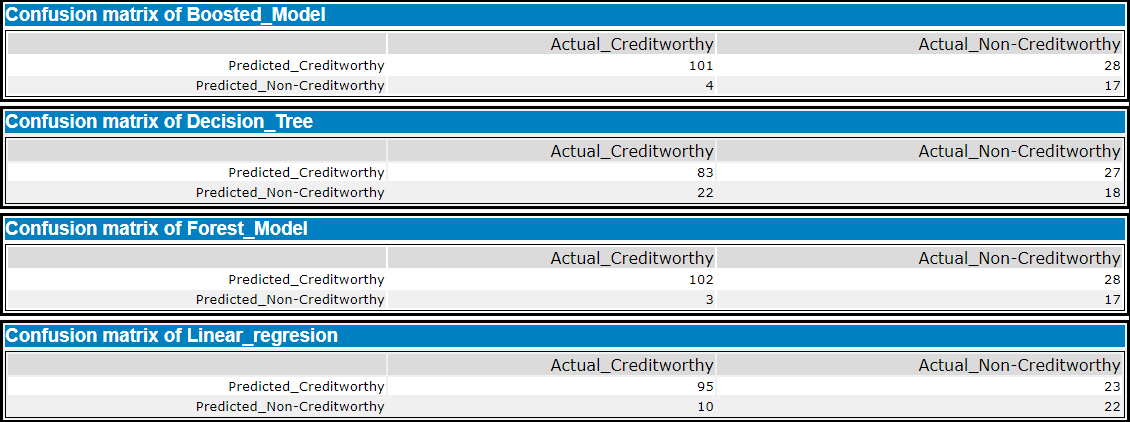


Figure 11. Confusion Matrix of Each Model

# 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:

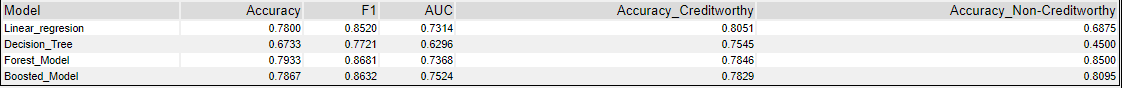


Figure 12. Accuracy of all tested models

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.

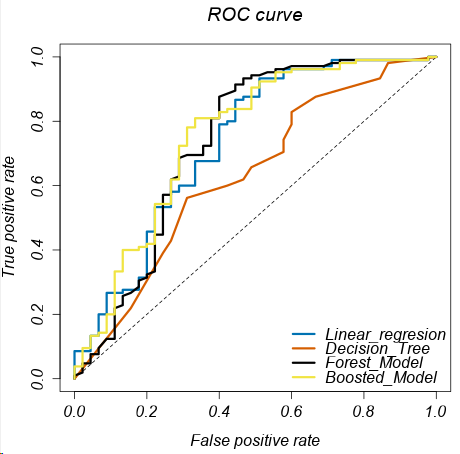


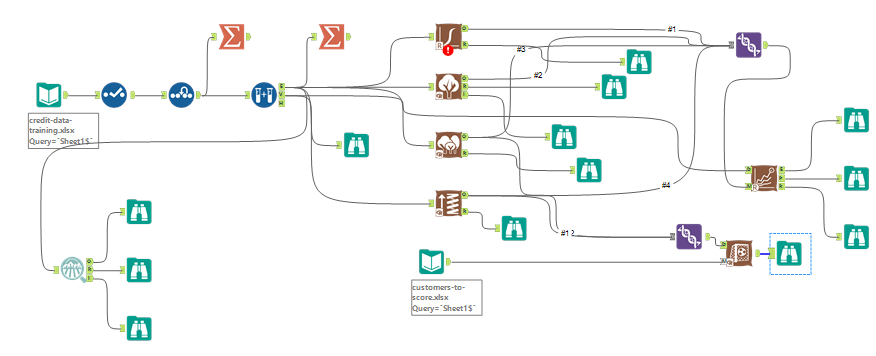
Figure 13. Roc Curve

**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