Project: Creditworthiness

Valencia Becerra Gerardo

Step 1: Business and Data Understanding

1. What decisions needs to be made?

   As employees from a small bank, we want to create a model to automatically classify, as accurately as possible using data from our previous customers, if 500 new loan applicants are creditworthy or not.

2. What data is needed to inform those decisions?

   We want to make a decision based on the same features of creditworthiness that past applicants (approved by hand) satisfy. We are provided with the following datasets:

   - Data on all past applications (*credit-data-training.xlsx*)
   - Customers to be processed in the next few days (*customers-to-score.xlsx*)

   The variables on these files are:

Variable	**Data Type**
Credit-Application-Result	String
Account-Balance	String
Duration-of-Credit-Month	Double
Payment-Status-of-Previous-Credit	String
Purpose	String
Credit-Amount	Double
Value-Savings-Stocks	String
Length-of-current-employment	String
Instalment-per-cent	Double
Guarantors	String
Duration-in-Current-address	Double
Most-valuable-available-asset	Double
Age-years	Double
Concurrent-Credits	String
Type-of-apartment	Double
No-of-Credits-at-this-Bank	String
Occupation	Double
No-of-dependents	Double
Telephone	Double
Foreign-Worker	Double

However, we must choose which features are the best to build the model instead of using all the data.

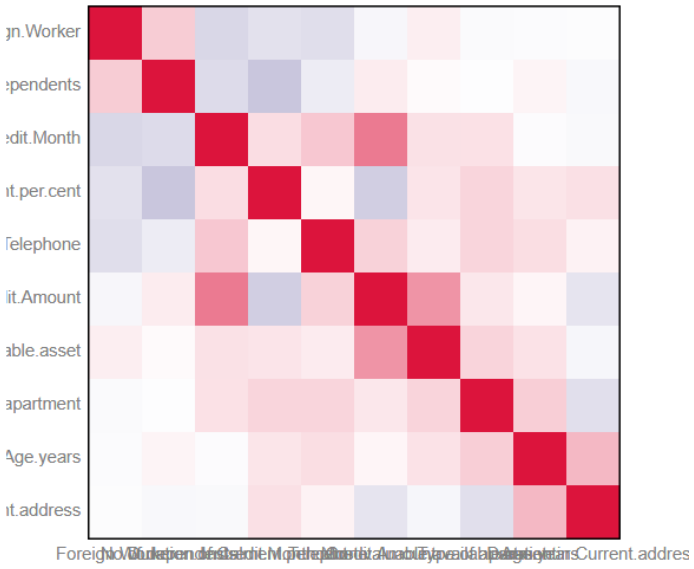
3. What kind of model (Continuous, Binary, Non-Binary, Time-Series) do we need to use to help make these decisions?

   We are trying to predict if a loan applicant is creditworthy or not, so the model should be Binary.

Step 2: Building the Training Set

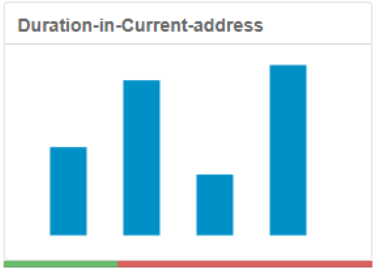
- In your cleanup process, which fields did you remove or impute? Please justify why you removed or imputed these fields. Visualizations are encouraged.

  All the numerical data fields (except for Occupation, as it has only one value) don't have a high positive/negative correlation as they aren't at least 0.70: *Duration of Credit Month* and *Credit Amount* have the strongest positive correlation (0.57) and *Instalment per Cent* and *No. Of dependents* the strongest negative correlation (-0.29).

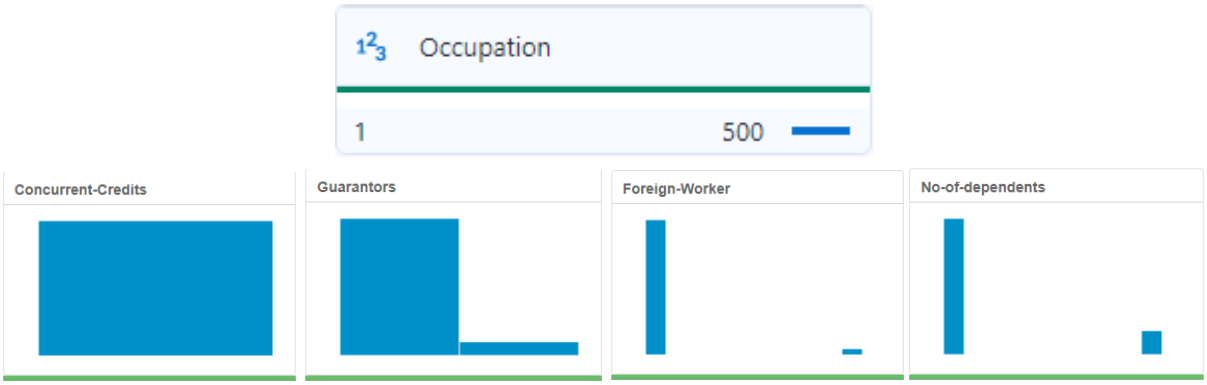


The following fields were removed:

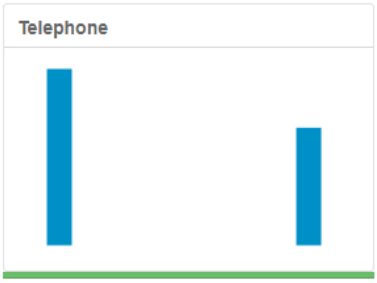
- Missing Data: 69% of the values were missing



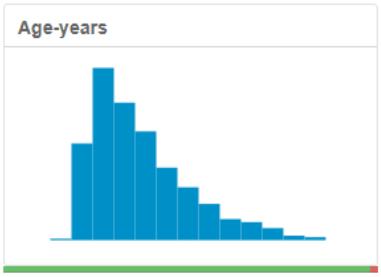
- Low Variability: Very few values in a subset of the data field



- Not relevant: Hinders the model's accuracy.



The following field was imputed:



Missing values were imputed using the dataset's median.

Step 3: Train your Classification Models

- Which predictor variables are significant or the most important? Please show the p-values or variable importance charts for all of your predictor variables.

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

Logistic Regression 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	

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

The most important predictor variables are:

Account.BalanceSome Balance (***)
Payment.Status.of.Previous.CreditSome Problems (*)
PurposeNew car (**)
PurposeUsed car (.)
Credit.Amount (**)
Length.of.current.employment< 1yr (*)
Instalment.per.cent (*)
Most.valuable.available.asset (*)

Fit and error measures

Model	Accuracy	F1	AUC	Accuracy_Creditworthy	Accuracy_Non-Creditworthy
Creditworthiness_Log	0.7800	0.8520	0.7314	0.9048	0.4889
Creditworthiness_DT	0.7467	0.8304	0.7035	0.8857	0.4222
Creditworthiness_RF	0.7933	0.8681	0.7368	0.9714	0.3778
Creditworthiness_Boosted	0.7867	0.8632	0.7515	0.9619	0.3778

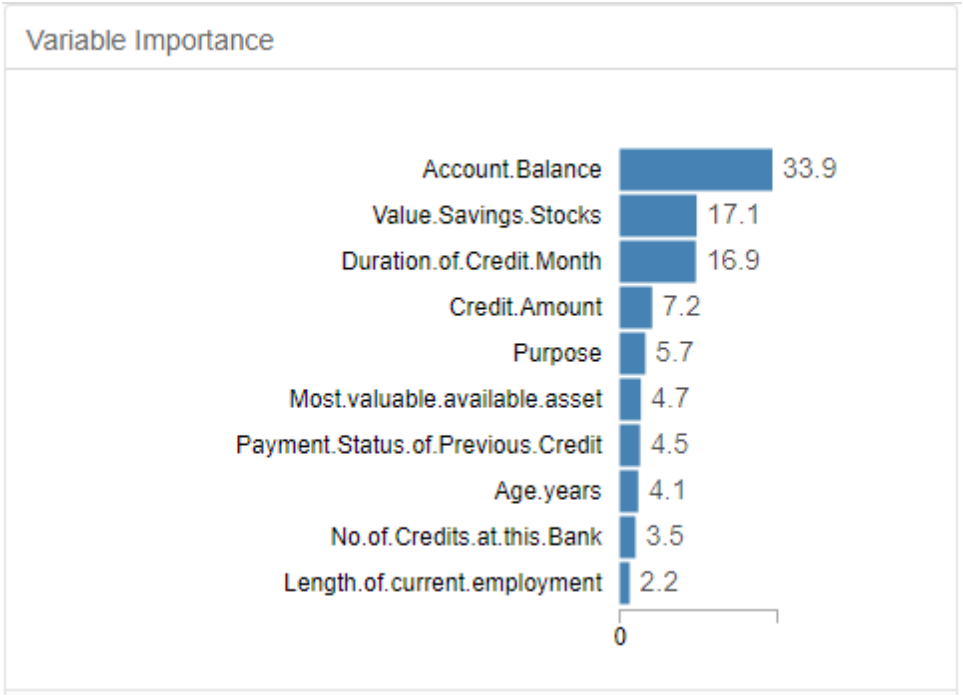
Accuraccy: 0.7800

Confusion matrix of Creditworthiness_Log

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

The accuracy for Creditworthy is 0.9048, for Non-Creditworthy it's 0.4889. The Positive Predictive Value (PPV) is 0.80 and the Negative Predictive Value (NPV) is 0.6875, so the model is biased towards correctly predicting Creditworthy individuals.

Decision Tree Model



The most important predictor variables, in order, are:

1. Account.Balance
2. Value.Savings.Stocks
3. Duration.of.Credit.Month
4. Credit.Amount
5. Purpose
6. Most.valuable.available.asset
7. Payment.Status.of.Previous.Credit
8. Age.years
9. No.of.Credits.at.this.Bank
10. Length.of.current.employment

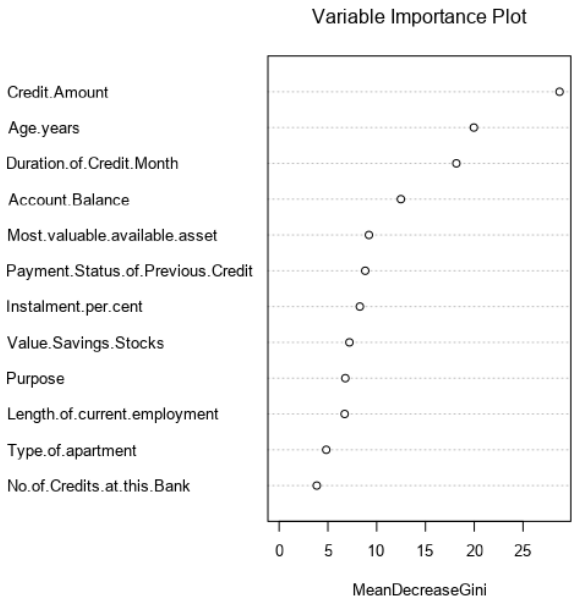
Accuraccy: 0.7467

Confusion matrix of Creditworthiness_DT

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

The accuracy for Creditworthy is 0.8857, for Non-Creditworthy it's 0.4222. The PPV is 0.7815 and the NPV is 0.6129, so the model is biased towards correctly predicting Creditworthy individuals.

Random Forests Model



The order of importance of the predictor variables is:

1. Credit.Amount
2. Age.years
3. Duration.of.Credit.Month
4. Account.Balance
5. Most.valuable.available.asset
6. Payment.Status.of.Previous.Credit
7. Instalment.per.cent
8. Value.Savings.Stocks
9. Purpose
10. Length.of.current.employment
11. Type.of.apartment
12. No.of.Credits.at.this.Bank

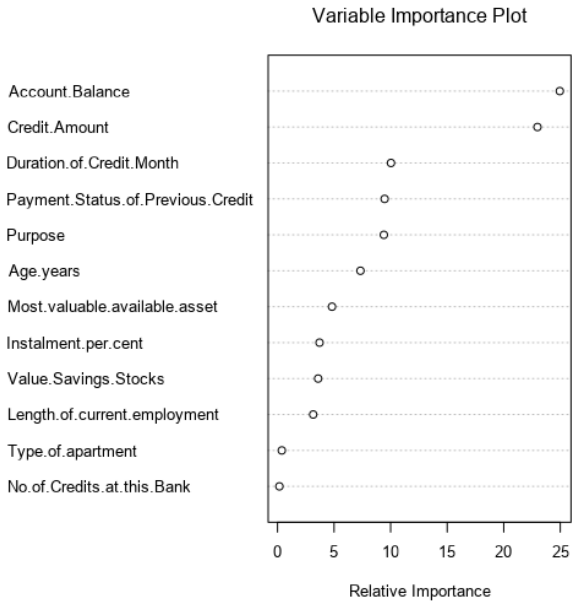
Accuraccy: 0.7933

Confusion matrix of Creditworthiness_RF

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

The accuracy for Creditworthy is 0.9714, for Non-Creditworthy it is 0.3778. The PPV is 0.7846 and the NPV is 0.85, so the model has little to no bias in its predictions.

Boosted Model



The order of importance of the predictor variables is:

1. Account.Balance
2. Credit.Amount
3. Duration.of.Credit.Month
4. Payment.Status.of.Previous.Credit
5. Purpose
6. Age.years
7. Most.valuable.available.asset
8. Instalment.per.cent
9. Value.Savings.Stocks
10. Length.of.current.employment
11. Type.of.apartment
12. No.of.Credits.at.this.Bank

Accuracy: 0.7867

Confusion matrix of Creditworthiness_Boosted

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

The accuracy for Creditworthy is 0.9619, for Non-Creditworthy it is 0.3778. The PPV is 0.7829 and the NPV is 0.8095, so the model has little to no bias.

Step 4: Writeup

- Which model did you choose to use?

  I chose the Random Forest Model.

- Please justify your decision using **all** of the following techniques. Please only use these techniques to justify your decision:

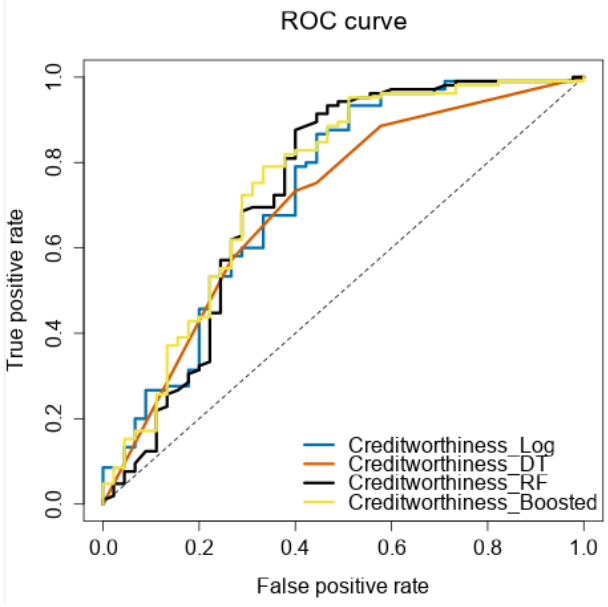
  - Overall Accuracy against your Validation set

    Its overall accuracy is 0.7933 (The highest of all 4 models)

  - Accuracies within “Creditworthy” and “Non-Creditworthy” segments

    Its Creditworthy accuracy is 0.9714 (The highest of all 4 models)
    Its Non-Creditworthy accuracy is 0.3778

  - ROC graph



    All 4 ROC curves show similar behaviours, but the logistic regression and Random Forests models reach the top faster.

  - Bias in the Confusion Matrices

    The PPV is 0.7846 and the NPV is 0.85, so the model has little to no bias in its predictions.

- How many individuals are creditworthy? 408 individuals are creditworthy.