Project #4
Predict Default Risk

Claudia Dai

19th October, 2018

1 Business and Data Understanding (125/250)

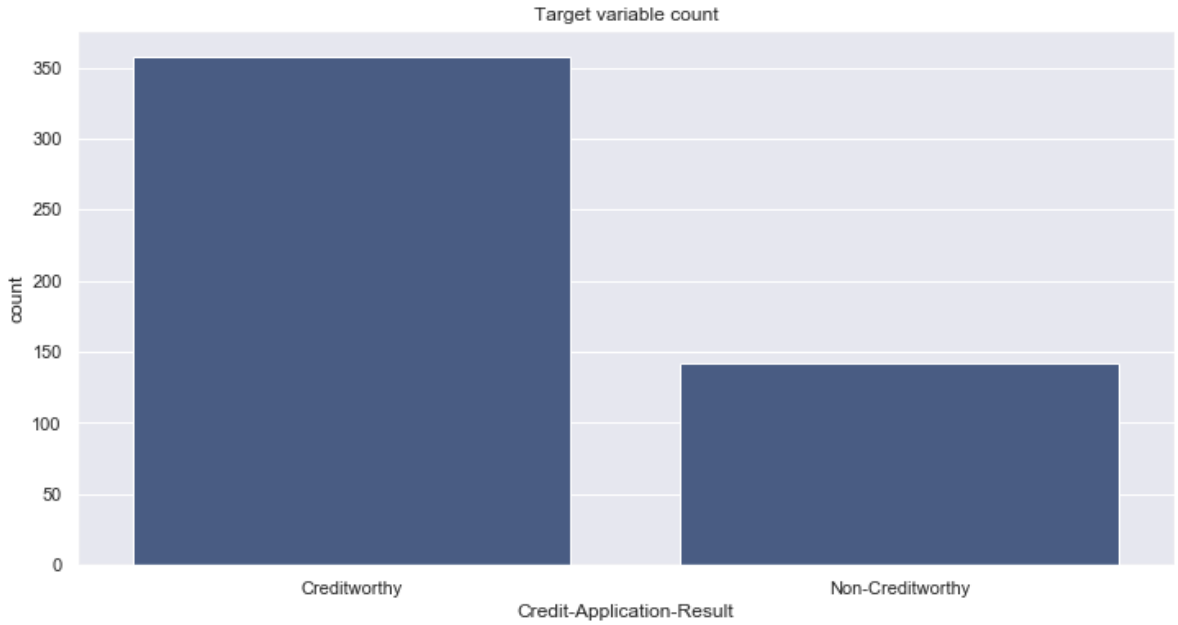
The small bank receives about 200 loan applications every week and needs to decide whether or not to approve the application. The goal is therefore to be able to identify which of these customers are creditworthy and which are not.

To inform these decisions, data on these customers' previous applications need to be available for analysis and modelling. For example, their current account balance and the amount and duration of credit they apply for could be useful indicators. It should also be considered whether they had any problems in the past of repaying the loan.

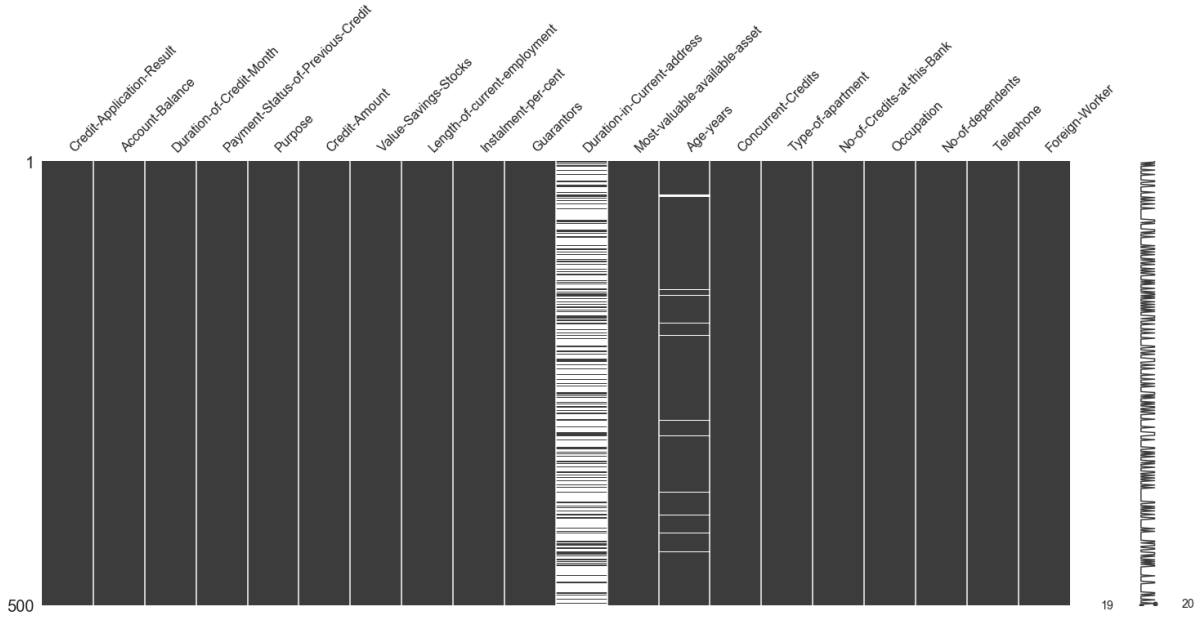
Determining the creditworthiness of a customer is a binary classification problem (i.e., a yes/no problem). Some models for binary classification are logistic regression, decision trees, random forest, and boosted models.

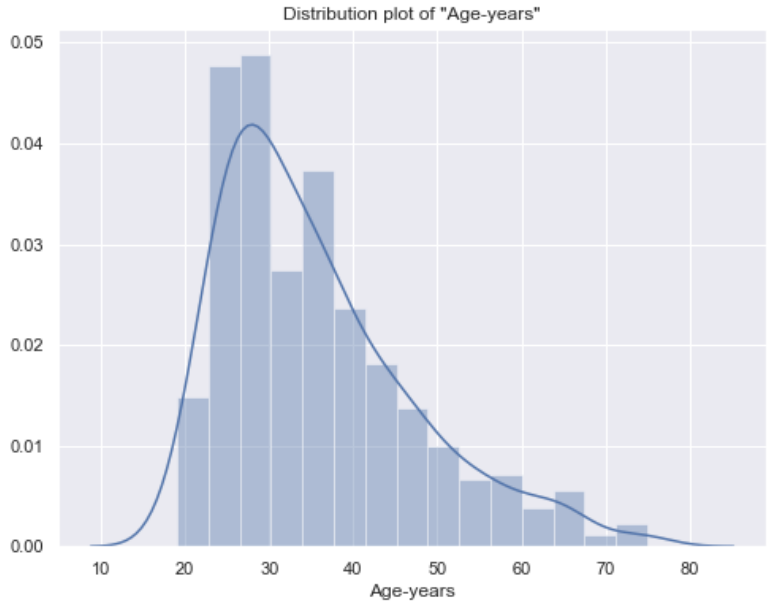
2 Building the Training Set (200/100)

We have an imbalanced-class problem at hand. There are far more creditworthy entries than non-creditworthy ones.

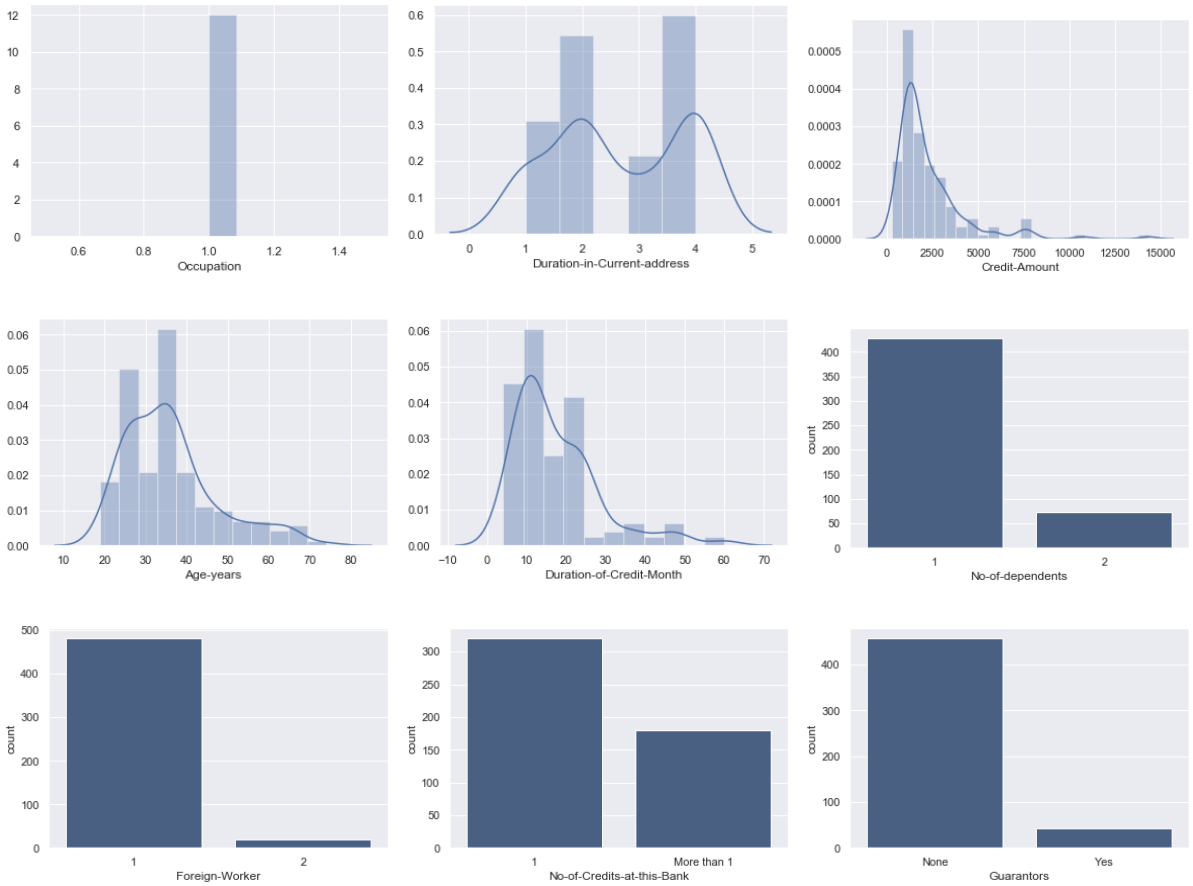


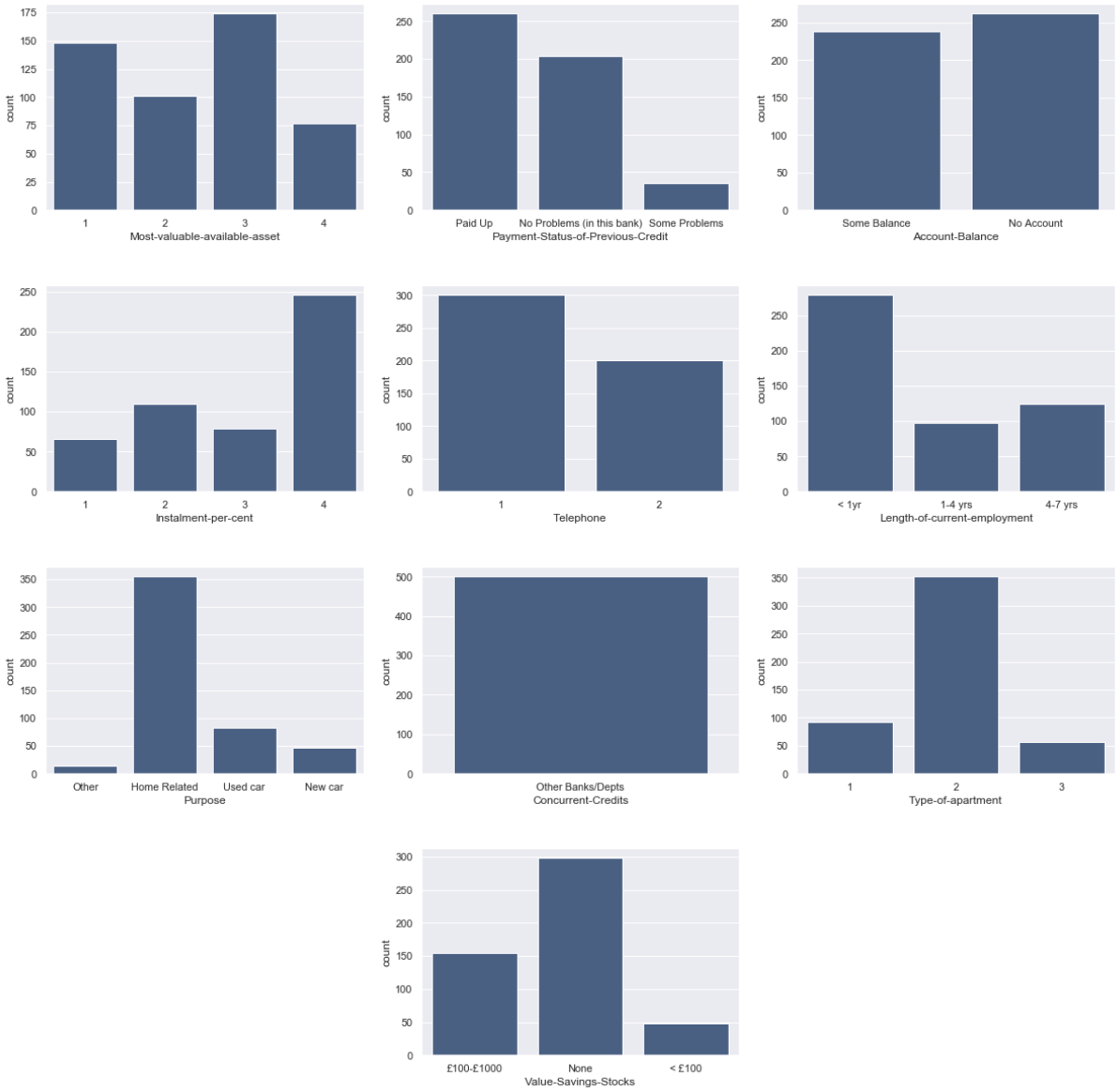
According to the paper "The Relationship between Precision-Recall and ROC Curves" (Davis & Goadrich, 2006), the Precision-Recall AUC (PR AUC) should be used instead of the ROC AUC when it comes to imbalanced-class problems. We will use the ROC curve regardless, because this is given in the assignment.



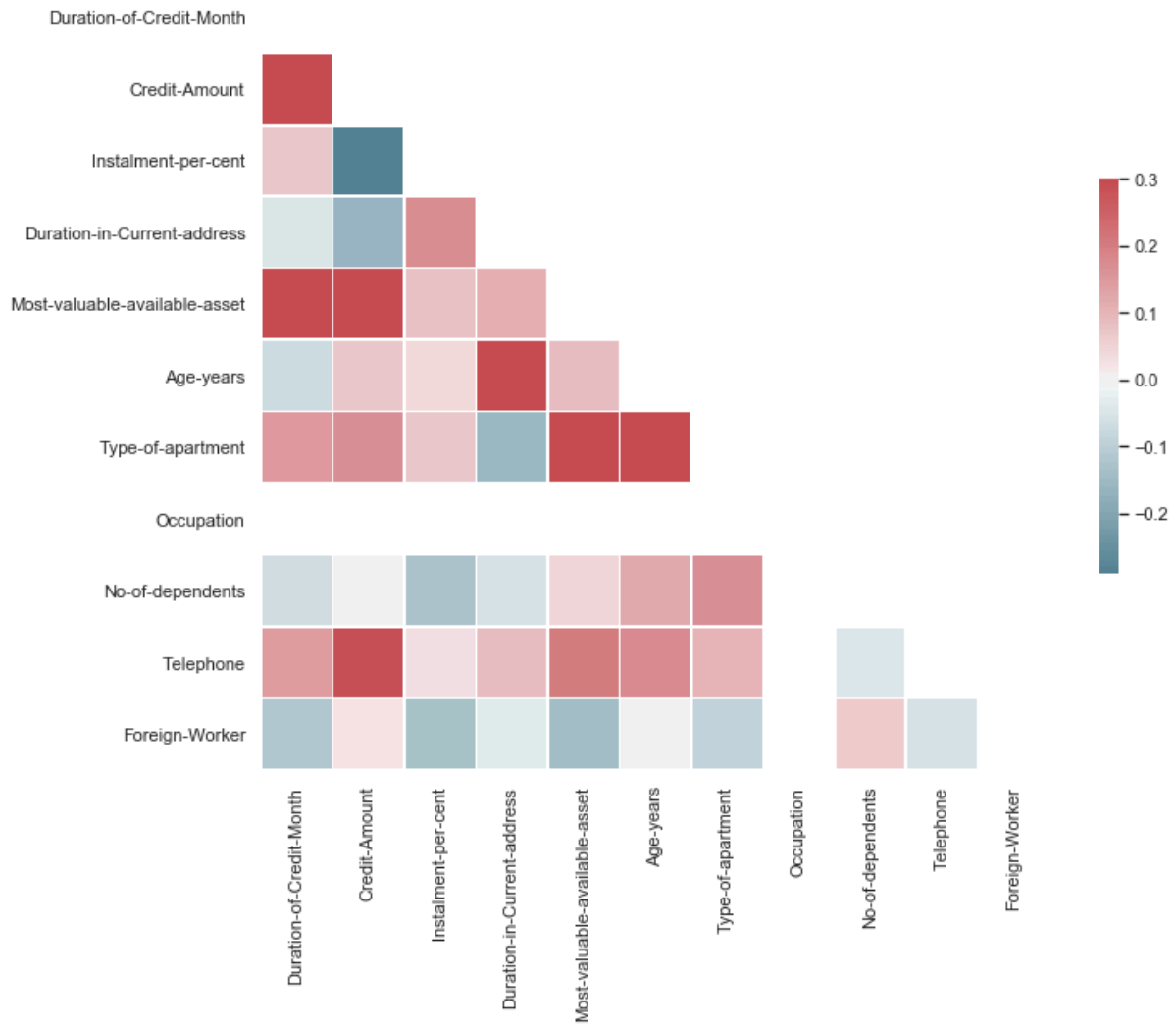


The column **`Age-years`** is missing 12 values, accounting for 2.4% of the whole column. The data for `Age-years` is skewed, and we will, thus, impute the median = 33. The column **`Duration-in-Current-address`** is missing 344 values, accounting for a whopping 68.8% of the whole column! It is decided to drop this column.





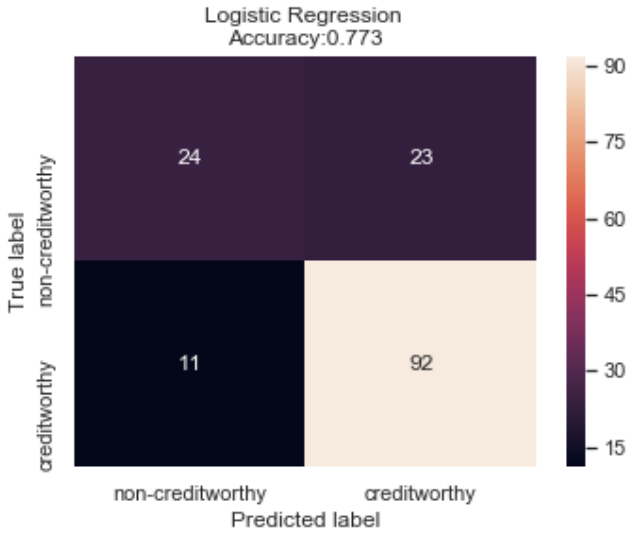
`Occupation` and `Concurrent-Credits` take in only one value, so we will drop these columns. `Guarantors`, `Foreign-Worker` and `No-of-Dependents` are removed – these seem to have low variability, since more than 80% of the data is skewed towards a value. `Telephone` might be useful to collect debts to make a customer creditworthier, but otherwise there is no logical reason for including it in our model. We will hence drop it.



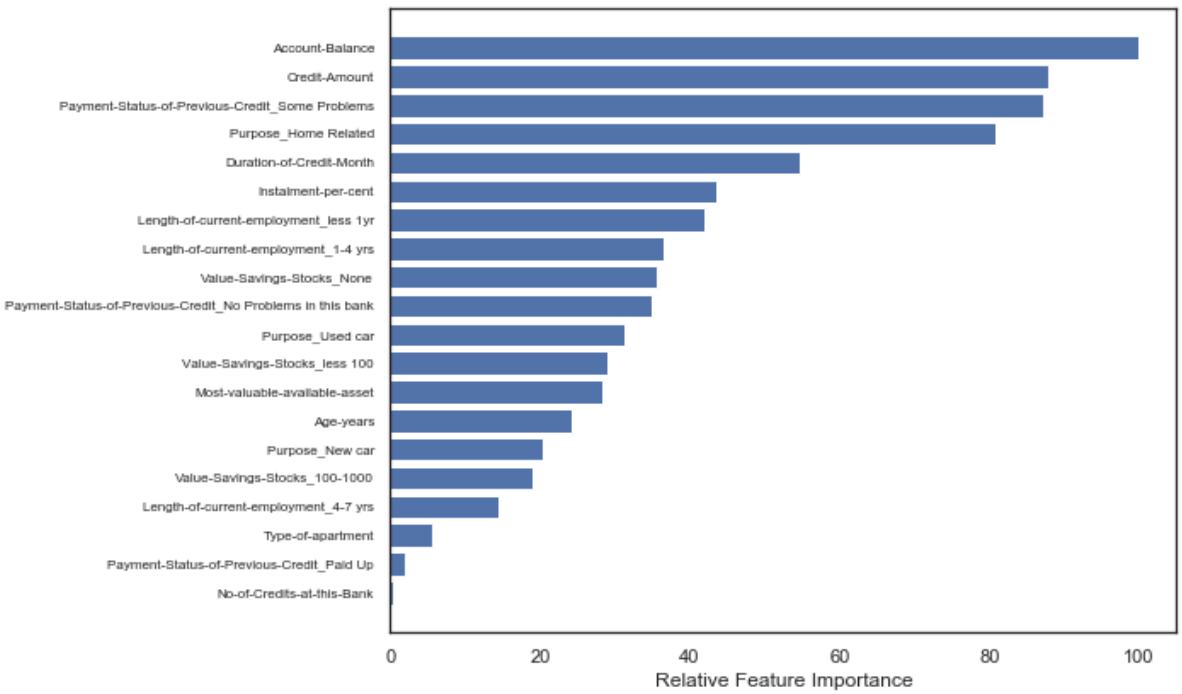
There are no fields that correlate more than 0.3 or less than -0.3 with one another.

3 Train your Classification Models (/500)

Logistic regression yielded 0.77 accuracy. The model is biased towards non-creditworthy customers, since precision and recall are higher for creditworthy predictions than non-creditworthy predictions. The most important features were `Account-Balance`, `Credit-Amount`, and `Payment-Status-of-Previous-Credit_Some Problems`.

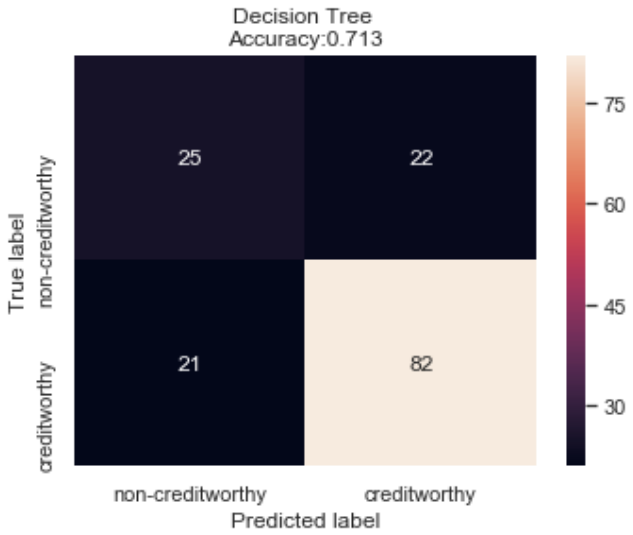


Logistic Regression	Precision	Recall	F1-score	Support
Non-creditworthy	0.54	0.53	0.54	47
Creditworthy	0.79	0.80	0.79	103
Avg/Total	0.71	0.71	0.71	150

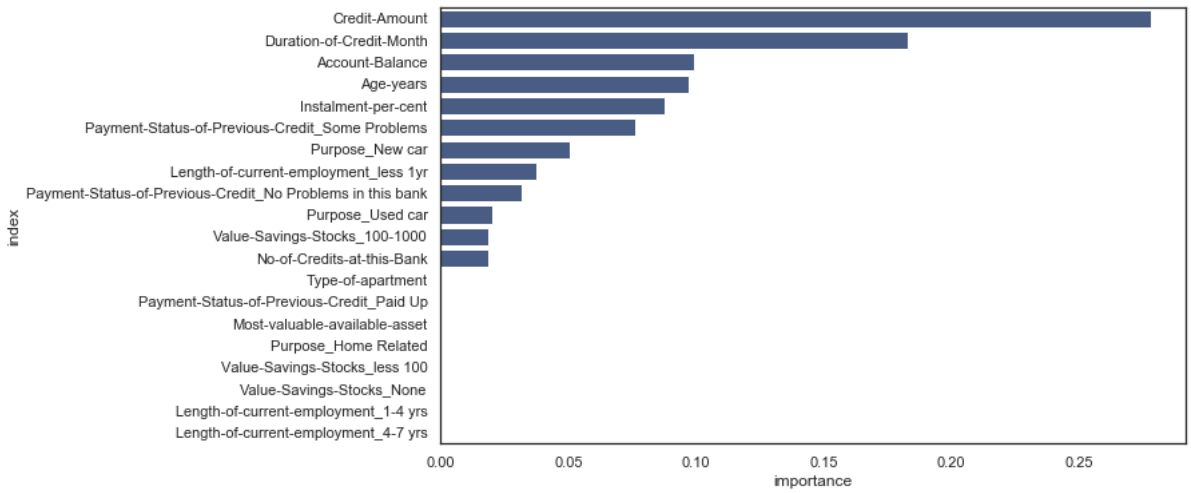


Decision tree yielded 0.71 accuracy. The model is biased towards non-creditworthy customers, since precision and recall are higher for creditworthy predictions than non-creditworthy

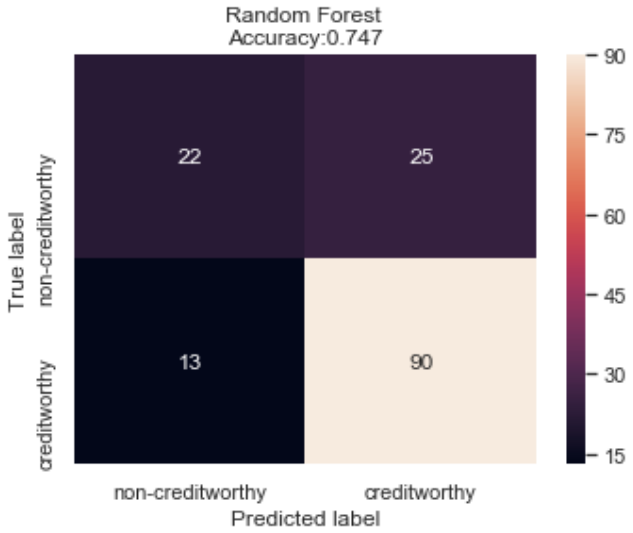
predictions. The most important features were `Credit-Amount`, `Duration-of-Credit-Month`, and `Account-Balance`.



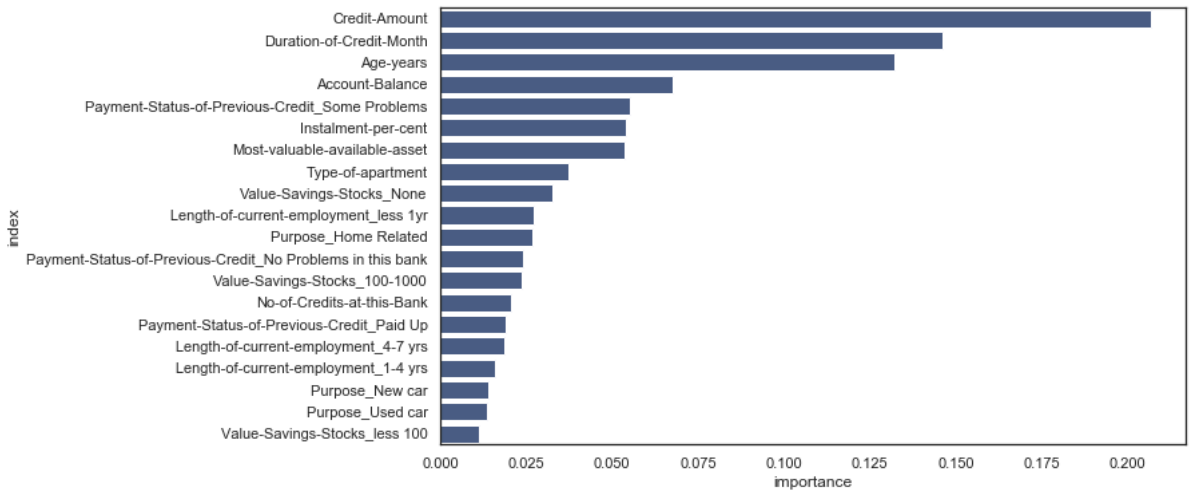
Decision Tree	Precision	Recall	F1-score	Support
Non-creditworthy	0.54	0.53	0.54	47
Creditworthy	0.79	0.80	0.79	103
Avg/Total	0.71	0.71	0.71	150



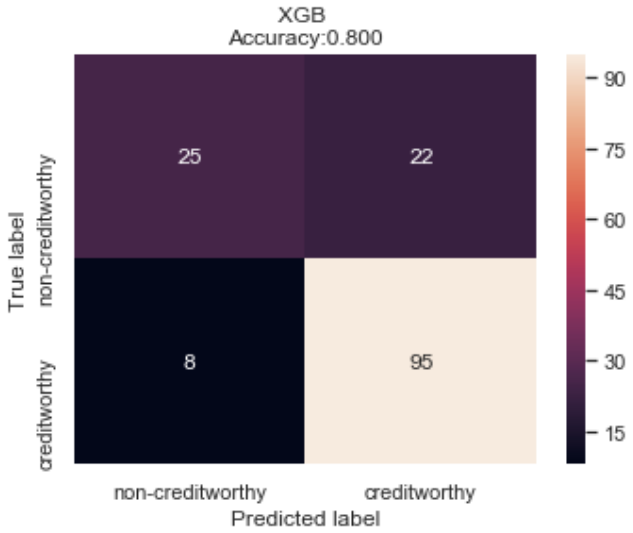
Random forest yielded 0.75. The model is biased towards non-creditworthy customers, since precision and recall are higher for creditworthy predictions than non-creditworthy predictions. The most important features were `Credit-Amount`, `Duration-of-Credit-Month`, and `Age-years`.



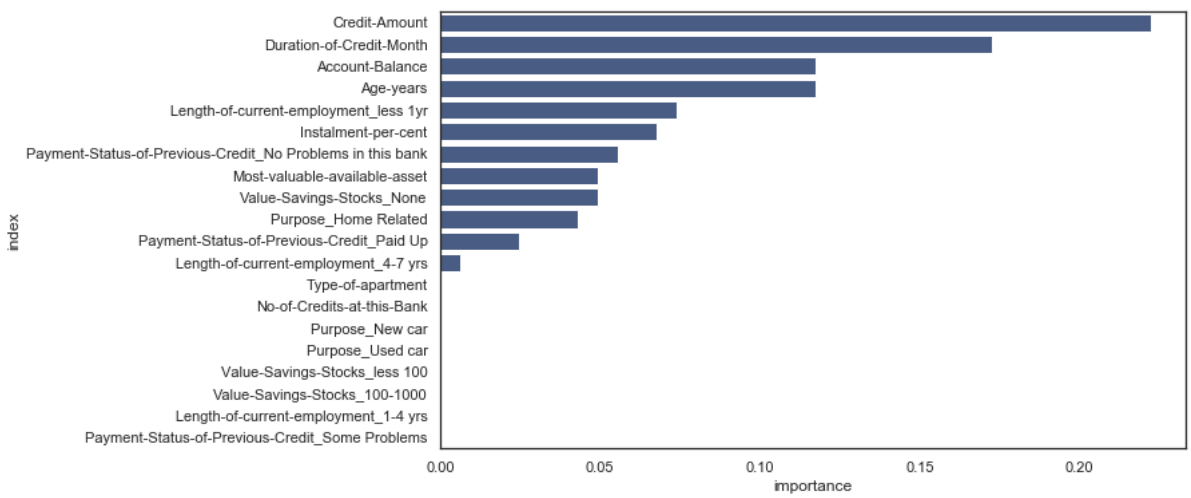
Random Forest	Precision	Recall	F1-score	Support
Non-creditworthy	0.63	0.47	0.54	47
Creditworthy	0.78	0.87	0.83	103
Avg/Total	0.73	0.75	0.74	150



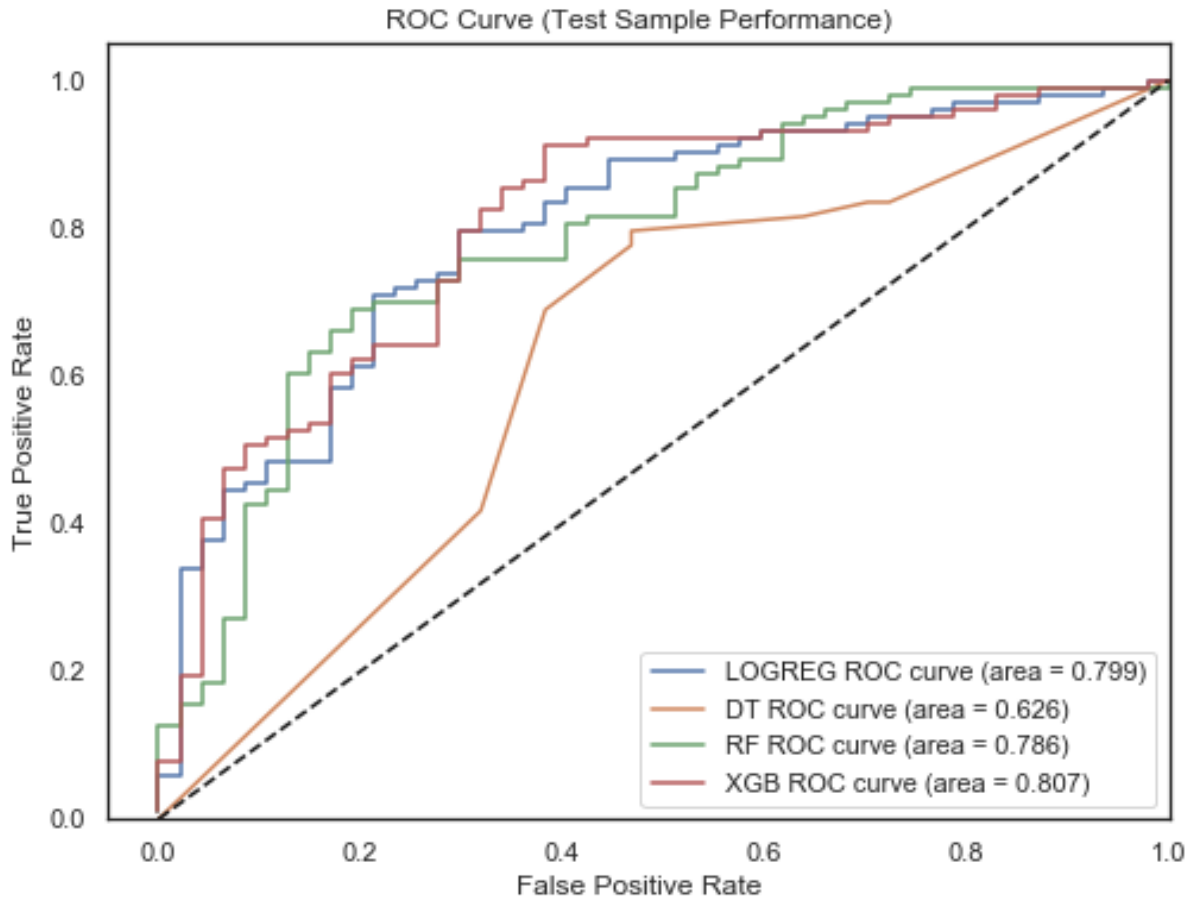
XGBoost yielded 0.80. The model is biased towards non-creditworthy customers, since precision and recall are higher for creditworthy predictions than non-creditworthy predictions. The most important features were `Credit-Amount`, `Duration-of-Credit-Month`, and `Account-Balance`.



XGBoost	Precision	Recall	F1-score	Support
Non-creditworthy	0.76	0.53	0.62	47
Creditworthy	0.81	0.92	0.86	103
Avg/Total	0.79	0.80	0.79	150



4 Write-Up



XGBoost is chosen with the highest accuracy of 0.80 against the validation set. This model also has the highest precision and recall scores and a ROC AUC of 0.807. It has the lowest false positives and false negatives. For the bank, it is important to avoid approving loans to customers with high default risk, while at the same time making sure that creditworthy customers get their loans in order to ensure profitability from these loan applications.

There are 425 creditworthy customers according to the model's prediction.

5 References

Davis, J., & Goadrich, M. (2006). The relationship between Precision-Recall and ROC curves. In *Proceedings of the 23rd international conference on Machine learning* (pp. 233-240). ACM.