Prediction of Risky Credit Customers

An investigation of the dataset 'German Credit Data Risk' with applied Machine Learning.

Introduction to the Dataset

The dataset used in this project is named 'German Credit Data Risk' and is collected from banks in Germany. It was published in 2016 but the data has been collected before this. The dataset contains 1000 entries with 10 categories, where one entry represent a person who has credit in a bank. Each entry is classified as having either a good or bad credit risk. The dataset was found on Kaggle.com, where it is broadly used and has been downloaded more than 10.000 times. The 10 column features that the dataset contain is listed below;

Age (numeric/int)
Sex (text/object: male, female)
Job (numeric/int: 0 - unskilled and non-resident, 1 - unskilled and resident, 2 - skilled, 3 - highly skilled,
Housing (text/object: own, rent, or free)
Saving accounts (object, describing the size of the saving account as either: little, moderate quite rich, rich)
Checking account (object, describing the size of the saving account as either: little, moderate, rich)
Credit amount (numeric/int, in DM),
Duration (numeric, in month)
Purpose (text/object: car, furniture/equipment, radio/TV, domestic appliances, repairs, education, business, vacation/others).
Risk: (text/object: Good, bad)

Problem Identification

Our aim with this project is to investigate if there is a correlation between certain features in the dataset and if there is risk of issuing a loan to certain types of customers. An example could be if there is a correlation between age and risk. We want to make a model which predicts and categorizes if a person is in risk of being a 'bad' loaner. A such solution might be in value for institutions like banks, who issue loans.

Value for banks

It is valuable for banks to have an insight in whether a person is categorized as either in good or bad risk. This knowledge can aid some of the bank employees' decision making and can have a positive effect on the banks' finances.

Not in risk of being a bad loaner

If the bank can point out that a person is not in any risk of being a bad loaner, this will ease the personal judgment when a loan has to be approved.

In risk of being a bad loaner

In cases, where a bank is in doubt of whether the loan should be approved the prediction model can help in this decision making. If the person is categorized as in risk of being a bad loaner the bank knows that they have to be careful with this person. With this knowledge, the bank can require more from these persons, e.g. being stricter with payment-deadlines. Another option is to deny the persons in risk a loan or requiring that people who wants a loan, have a bigger credit amount.

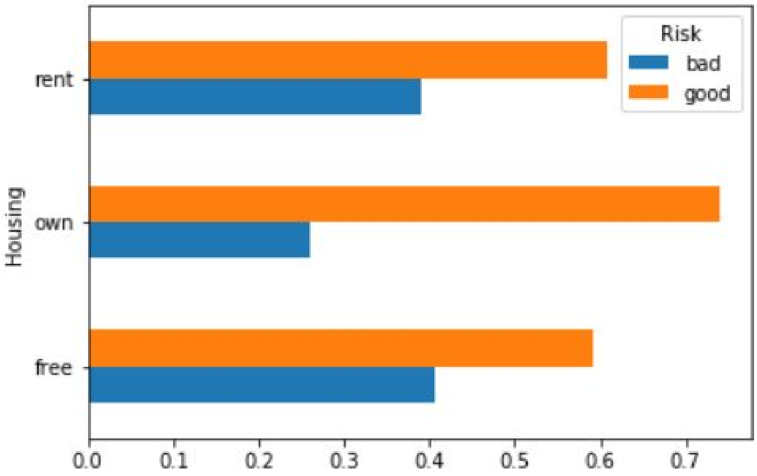
In developing this decision making tool we will apply machine learning to predict whether a loaner is in a good or bad risk. In the report we are also going to explain which factors that are relevant for predicting this and which groups of people who are more likely to be either good or bad.

Correlation between Variables

During the analysis of the dataset we search for correlations among the factors that have an effect on 'Risk'. We have chosen the factors, 'Purpose', 'Sex', 'Age', 'Housing', 'Saving accounts', Checking account', 'Credit amount' and 'Duration'.

Looking at the correlation between 'Risk' and 'Purpose' we find that this is not of high importance, since there is only a minor difference in the risk depending on the purpose. However it is possible to a slight difference, wherefor this factor has also been chosen as relevant for the prediction of 'Risk'.

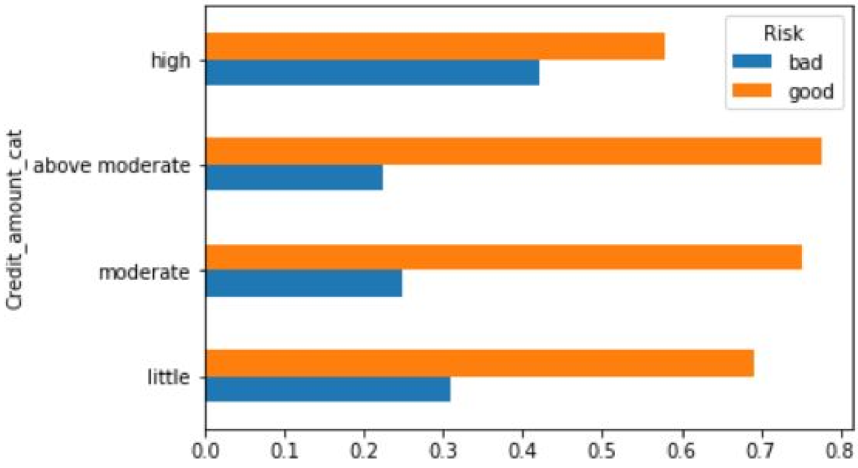
In the correlation between 'Risk' and the 'Housing' situation we find that there is a smaller probability for being bad risk for owners than for renter and people living for free. This correlation is illustrated in the graph.



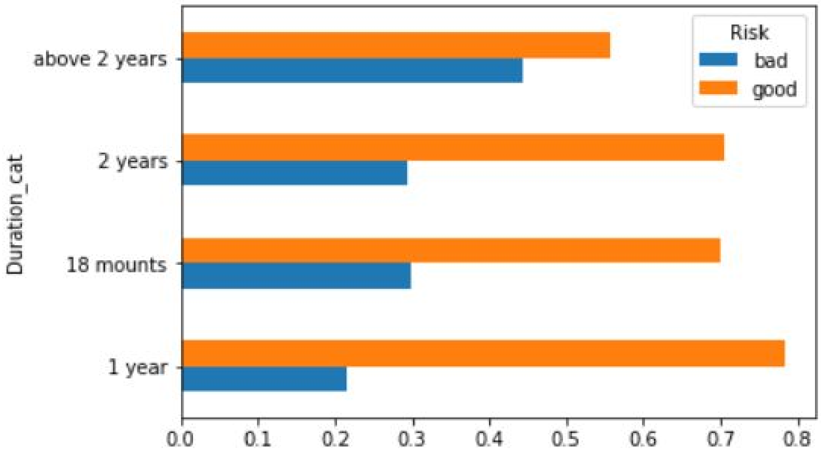
This leads on to finance. Here we find that for both the savings account and the checking account the probability for being in bad risk decreased the more money you have.

Saving accounts	Good Risk	Bad Risk	Checking Accounts	Good Risk	Bad Risk
Little	64 %	36 %	Little	50,7 %	49,3 %
Moderate	67 %	33 %	moderate	61 %	39 %
quite rich	82,5 %	17,5 %			
rich	87,5 %	12,5 %	rich	77,8%	22,2 %

For the credit amount we find that the probability of being bad risk decreases when going from little to moderate and again from moderate to above moderate, but then there it increases when having a high credit amount. Although we do see a correlation why this factor is also included in the analysis.



The last correlation we find is between risk and the duration of the loan. Here we see that the fewer months the lower probability of being in bad risk.



Between 'Risk' and the 'Job' situation we find no correlation, therefore this feature will not be a part of the further analysis.

Summarizing the factors which seem to be correlated in having an effect on risk level are sex, age, purpose, housing situation, duration of the loan and saving and checking accounts.

Whereas we find the purpose seems to be of less importance than the others. We find, based on the data that the probability for being in bad risk is higher for women than men. Furthermore, people younger than 30 have a higher probability of being in bad risk than people older than 30

Based on these correlations we find it interesting to investigate whether the loaners cluster together in groups who look alike on the above mentioned factors, and what you can say about the 'Risk' in such groups.

Clusters of loaners and prediction models

Based on the selected factors we have chosen to cluster the data in three clusters.

The first cluster is characterized as women and the youngest loaners, with an average age on 32,6 years. Financially they are placed in the middle, although only decimals separate this cluster from the third cluster, who has the best economic.

The second cluster is older than the first one with an average age on 36,2 years and with a higher probability of being a man than a woman. This cluster has a little less money on their saving than the first cluster. However the credit amount for this cluster is more than double the size of the two other clusters and in relation to this the duration of the loan is also much higher for this cluster.
This cluster represents the loaners with the highest probability of being in bad risk.

The third and last cluster is characterized as men and are a bit older but almost the same age as the second cluster. Financially they look a lot like the first cluster, but they have a slightly higher saving and checking account, also the credit amount and duration of the loan is equal to the first cluster. This cluster represents the loaners with the lowest probability of being in bad risk.

Because of the factors that are used to represents the loaners in clusters, the same factors can be used to predict if they are in bad risk. To make a prediction we are going to use different types of models. All of these models are based on the ‘Risk’ factor. The analysis in Supervised shows us that it is easier to predict the risk of a good loaner than a bad loaner. When it comes to using the column ‘Risk’ it is relevant to say that we have developed a distribution table which illustrates that 70% represents a good risk and 30% represents a bad risk. This will of course have some sort of influence on the result we get, therefore we tried to convert this distribution into 50/50 in the train set.

Conclusion

The most representative prediction model we recommend is based on XGBoost. The model have a 64 % chance to predict loaners with a bad ‘Risk’. Since the prediction rate is 64 %, there is certainly space for improvements, wherefore we have our doubts towards whether this prediction model creates enough value for a bank. On the other hand, it could be the first steps in creating a decent prediction model on evaluating future loaners. To change the prediction model and make it more credible, it would be necessary to include more data which gives a more equal distribution, so it becomes a more accurate prediction. Furthermore the cluster could be a supplement to evaluating if a loaners have a good or a bad risk.