Matt Sellman

Report: -

Common and most relevant characteristics of bad credit risk:

The most common characteristics of bad credit risk are, according to the decision tree I made: Living in a rented apartment, being a foreign worker and having no guarantor or being a co-applicant.

I found living in a rented apartment to be a very common characteristic of bad credit risk, with a large number of bad applicants living in rented accommodation, though this did not apply to all. Almost every person who counts as a bad applicant is a foreign worker, leading this to being one of the most relevant attributes of bad credit risk. Another aspect of bad credit risk was having no guarantor or being a co-applicant, this was a very common attribute of bad credit risk.

The most relevant of these were a combination of being a foreign worker and having no garuntor, which led to bad credit risk almost 90% of the time.

The process of making the models and an explanation:

The process with which I made my predictive model was by first of all, setting the role of the data, from there I then selected the appropriate attributes and used a sample (stratified) that I normalised and then split the data into two sets, one of which went through an optimise parameters model and the other set went past that and straight into the apply model and performance operators.

Inside optimise parameters I implemented a cross validation operator and inside that I put metacost, apply model and performance operators. From there I used the decision tree operator. Cross validation also split my data into a training set which contained 60% of the data and a test set which contained 40% of the data.

The second model I made was similar to the first, I imported the first process I made and from there I retrieved the dataset, then I used set role operator to select the attributes, I then used sample stratified in order to increase the accuracy before putting the process through cross validation. Inside the cross validation I imported the model and from there applied and then tested the performance of the model.

Confusion matrix and performance:

The confusion matrix I made was using the performance measures for bad creditability, including guarantors, accommodation type, number of loans and whether or not they are a foreign worker.

accuracy: 78.87%

	true bad	true good	class precision
pred. bad	30	17	63.83%
pred. good	13	82	86.32%
class recall	69.77%	82.83%	

And the overall performance measures were the accuracy and root mean squared error. These are relevant to my application because I thought they represented the consistency of my decision tree.

The criteria I used to judge the quality of my best predictive model were the highest accuracy and the class precision. I chose accuracy first because it was a good indicator of how effective my decision tree was at predicting good or bad credit risk customers. The second was precision, I

wanted to know how close the predictions were to one another, or whether they were scattered all over the place. The precision, as you can see above, is higher for the predicted good than it is for the predicted bad this is because of the higher number of true good examples (82) compared to the true bad examples (30).

The performance of the second rapdiminer process is shown below:

accuracy: 77.42% +/- 3.13% (mikro: 77.46%)

	true bad	true good	class precision
pred. bad	18	13	58.06%
pred. good	3	37	92.50%
class recall	85.71%	74.00%	

As you can see, the accuracy of this piece is slightly lower though the precision is higher. The pre-processing that occurred from the model has had a marked effect on the outcome and changed the accuracy and the overall performance of the model.