German credit analisys assignment

Practical assignment: The German Credit Case Introduction This report summarizes our findings on German Credit case. The task for this first assignment is to develop a credit scoring rule such that, we will be able to classify the new applicants as good credit risk or bad credit risk with the best accuracy and considering the cost/gain of applying that rule. In order to do that, we have a sample of 1000 records (or instances) with 32 variables (or attributes). We intend not to use all these attributes to reduce the overlap of information in order the data mining algorithms operate more efficiently.

The aseline of this dataset or the result of applying a naive rule that classifies all records as members of the majority class is 70%. Therefore, our objective is to find a classification model that surpasses this benchmark. However, the ultimate goal is to find a model that maximizes the profit of approving credit based on the information available. Pre-processing steps We start with a few surprising information that we noticed about the data, followed by some techniques used to clean the dataset and reduce the number of attributes. 1 .

Attribute roles in a credit decision: After comparing the available data against our revious general knowledge about how credit institutions decide who gets a credit, we have found the interesting facts that are listed below. We made use of tools such as summary statistics, correlation matrix (see annexes), pivot tables, and graphics created in applications Weka, Excel, and Matlab to assist this task. O We expected applicants with good checking account and savings account status to have a higher likelihood of having a positive response.

Surprisingly, our data analysis has proven that our assumption was wrong. This is clearly shown in the following pivot table, which shows the average response according o attributes CHK ACCT and SAV ACCT: This table shows that 88. 3% of the applicants who do not have a checking account (CHK_ACCT= 3) got a positive response, which is an average that is higher than any o Attributes HISTORY and NUM_CREDITS should account positively for the applicant’s response. In other words, we expected that people with a good credit paying history had more positive responses.

On the other hand, bad credit paying history would increase the negative responses. In both cases, the number of credits taken should potentiate this result. Unexpectedly, nearly 83% of the applicants with a critical ccount (HISTORY= 4) had a positive response, as shown in detail in the following pivot table: In accordance to the Correlation Matrix (Annex 2), attributes AMOUNT and DURATION are clearly correlated (0. 625). This is expected up to a certain extent, because it is valid to assume that bigger amounts should take longer to pay.

We hope that the Principal Component Analysis will be able to Join these highly correlated attributes. At the same time, longer duration increases 2 risk and the higher the risk, the less likely should be positive response. Therefore, this improvement should have a direct impact in the response. O We expected that owning a real estate attribute REAL_ESTATE would increase the chances of a positive response, as well as having a COAPPLICANT or a GUARANTOR, because understand that these attributes provide the bank with some kind of guarantee.

However, none of those attributes have shown a relevant correlation to the response. 2. Exploration: In this part we are going through the data in order to find outliers, missing and inconsistent data. Outliners The Summary Statistics (Annex 1) showed some attributes with maximum value distant more than one and a half standard deviation from the median. Therefore, we roceeded with the following box plot analysis: i. Amount: based on the following box plot created for amount attribute, in which the edges are the 25th and 75th percentiles, we see 73 outliers with the whiskers ending value 7882.

In order to resolve this problem, we have deleted those 73 outliers. it. Age: based on the following box plot created for age attribute, in which the edges are the 25th and 75th percentiles, we see outliers with the whiskers ending at value 64. The solution in this case was also to delete the outliers. 3 Missing data: We have found no attributes with blank values. Inconsistent data The provided data contains instances with inconsistent data with regards to attributes (time of) EMPLOYMENT and JOB (nature).

If the value of one of those attributes is zero, the other attribute should have the same value. However this constraint is not valid, as shown in the following pivot table which contains the record count for any value combination of those attributes. Values with red background show the number of inconsistencies that combination. Since the number of inconsistent records is low in comparison to the total population (0. 52%), we decided to delete those 52 unreliable records. . Data Reduction: In this section we start to check the attributes.

We are going to combine and transform them to reduce the overlap of information. There are many ways to do it depending on the type of the variables. The first step consists in a naive approach based on the domain knowledge. Later on we are going to exploit the Correlation Matrix and the Principal Component Analysis for those attributes that are numerical. Attribute OBS# is clearly an identifier which is not relevant for the evaluation of the credit risk. Using this attribute would surely create an over fitting model. Therefore, The three attributes related to gender and marital status i. . MALE_DIV, MALE_SINGLE, and MALE_MAR_or_WID do not provide us with enough explicit information about gender, but we can assume that instances with value zero in all of them relate to female applicants. As a result, we created a new attribute called MALE which contains true in case the applicant is a man. The marital status information is partial in many aspects. Firstly, it sums up married and widowed, so it is impossible to say if the applicant has a partner. Besides that, it is not possible to identify the female marital status from those three aforementioned attributes.

Since this partial information might lead us to errors, we decided not to take the marital status into account for our analysis. We noticed a strong relationship among the attributes related to the purpose of the credit NEW CAR, USED CAR, FURNITURE, RADIO, TV, EDUCATION, RETRAINING. We assume the credit response is not related to 4 the type of good or type of education the applicant wants, but to the general purpose of the credit. Therefore, we grouped similar purposes into two new attributes, “ goods” and “ self-improvement. ” After that, we noticed that there would be a high negative correlation between those two new attributes (0.). There is no full

correlation only because some instances have no data about the purpose, i. e. all six relevant attributes are false for 55 records. Therefore, we decided to keep self- improvement attribute only because we assume that the 0. 55% of the cases in which the purpose is unknown are related acquiring some non-listed good. Although the number of foreigners in the sample is low less than 0. 04% -it is clear that being a foreigner raises the probability of getting a positive response almost every (89. 19%) foreigner applicant got the credit. Therefore, we decided to keep the attribute as is.

The correlation matrix including response reveals that there is a low, yet significant correlation between response and check account (0. 35) so we assume that check account is an important variable. As expected, attributes RENT and OWN_RES have a clear inverse relation -correlation -0. 736. This is because any applicant who owns his or her house does not pay rent. We consider that the trade-off of keeping both attributes is more expensive than choosing Just one. Besides that, we consider more representative to know if the applicant owns residence because it could be assumed that it is a warrant to get a redit.

Therefore, we have chosen to remove attribute RENT. According to what we were expecting, the variables telephone and number of dependents, present a very low correlation with response. We decided to remove these from our analysis. A principal components analysis based on age, amount, and duration reveals that we can replace these three variables with Just two new attributes named PCI and PC2, losing less than 14% of the information. We see that the first principal component is most affected by the weights of

duration and amount, while the second measures the balance between two quantities, uration versus age.

To conclude, the third exploits the balance between duration versus amount. According to this table we decided to 5 replace duration, amount, and age using Just the first two principal components. The resulting data set after removing attributes, combining attributes using PCA, and removing instances due to inconsistent or outlined data, consists of 859 instances with the following 20 attributes, which means we have reduced the number of variables by more than 30%: PCAI PCA2 HISTORY SELF-IMPROVEMENT SAV ACCT EMPLOYMENT JOB INSTALL RATE MALE CO-APPLICANT GUARANTOR PRESENT RESIDENT REAL ESTATE

PROP UNKN NONE OTHER INSTALL OWN RES NUM CREDITS FOREIGN RESPONSE 6 Classification Models and Results For our next step, we divided the data randomly into training (60%) and validation (40%), and tried different models by omitting some attributes while creating them in XLMiner. After a number of trials, we have not found relevant improvement by doing that with any combination or attributes. Therefore, we decided to keep all of the attributes from our previous step for the models we will report below. ??? Logistic regression We have created a classification model based on logistic regression using XLMiner ith cutoff 0. . The results are shown in the table below: Although the error rate is relatively low 21. 80% if compared to the following classification models, the cost/gain matrix reveals that using this model would incur in a loss of 400 DM. ??? Classification trees terminal nodes parameter as too high. Therefore, we created our model using minimum five records per terminal node, which is around 1% of our training

set. The resulting model had an accuracy of nearly 75%, which is about the average of other models, and a final cost 19000 DM, which is the worst net loss of all scenarios. ??? Naive Bayes

Since XLMiner is not able to deal with numeric values in Naive Bayes method, we have created an exceptional model by disregarding the continuous numeric attributes PCAI and PCA2. The results are shown in the table below: 7 On the other hand, Weka is able to deal with numeric attributes while using Naive Bayes method by automatically discretizing those attributes using Fayyad-Irani's MDL method (BOUCKAERT, 2013). We then produced a model in Weka using the exact same training and validation sets, considering all attributes available. This model results were Just slightly better, as one can see from the table below.

However, this slight difference provided us with gain greater than cost. ??? Discriminant Analysis For the discriminant analysis, we used XLMiner with its default parameters. The results of such experiment are shown in the tables below: This model had the worst accuracy error rate 29% but surprisingly it gave us the most net profit on the validation data with the default cut-off of 0. 5 5100 DM. 8 Cut-off Analysis In order to find the best cut-off for the logistic model as instructed in the assignment, we have taken our logistic regression model, sorted the validation data on " Prob. or 1 success)", and calculated the cumulative net profit based on the actual cost/gain of extending credit. By doing that, we were able to find the optimal cut-off for this model, which is the probability of success when the cumulative net profit is maximized, wherein the probability of success is 82. 28%. A relevant excerpt of the resulting table is shown below to illustrate this result. Since

the validation dataset contains 344 instances and the optimal value was reached at the top 171st instance, it is correct to say that this cut-off would result in extending credit for less than 50% of the applicants. 9

Conclusions The first important aspect we noticed in our analysis is that as soon as we used PCA the performances of classifications slightly improved from 73% to 76% in average. achieved combining Just three numerical attributes and choosing only two of the resultant variables. This event has shown the importance of data reduction during the pre-processing step in data mining process. Thus, In order to build a great model, the structure of the data set and the choices made in terms of dimension reduction are critical. We noticed that, in general, models created using Naive Bayes and logistic regression ad the best accuracy.

However, the final result is not only depending on the data and the model. The asymmetric misclassification costs put in evidence a further crucial aspect: “ Implicit in our discussion of the lift curve, which measures how effective we are identifying the members of one particular class, is the assumption that the error on misclassifying a case belonging to one class is more serious than for the other class” (SHMUEL’, G. ; PATEL, N. R. ; BRUCE, P. C. , 2010). For instance, the discriminant analysis provided us the worst model between those e tried in terms of accuracy.

Nevertheless, it gives the best net result in terms of profit if using the default probability of success, which is 0. 5. We have shown that the optimal solution for the logistic regression model in terms of profit is achieved with setting the cut-off value to around 0. 8. This result implies that, in order to

minimize the risk and increase the profit the bank, we should decrease the percentage of positive response passing from 70% of the baseline to around 50% of the logistic regression. This means setting higher standards in the policy of credits.

To conclude, these result should not be consider the global optimum for this case but Just one of possible approaches. Extend the cut off analysis, exploiting more trees techniques or simply changing the model we could reach different results. 10 References BOUCKAERT, R. et al. WEKA Manual or verston 3-7-10. Hamilton, New zealand: University of Waikato, 2013. Available at http://prdownloads. sourceforge. net/weka/WekaManual-3-7-10. pdf SHMUEL', G. ; PATEL, N. R. ; BRUCE, P. C. Data Mining for Business Intelligence. Concepts, techniques, and applications in Microsoft Excel with XLminer. Hoboken, USA: Wiley, 2010.