Maastricht University

School of Business and Economics

SHOULDICE CASE REPORT

Bernardo Oliveira	(I6115981)
Lilian Do Khac	(I6114941)
Alex De Vidal De St. Germain	(I6115568)

Supervisor:	Drs. Marianne Peeters
Issue due date:	**10. December 2015**

Contents

List of Figures iii

List of Tables iv

1 Appendix 1

1.1 Review the predictor variables and guess what their role in a credit decision might be. Are there any surprises in the data? 3

1.2 Devide the data into training and validation partitions, and develop classification models using the following data mining techniques: logistic regression, classification trees, and k-nearest neighbor 10

1.3 Choose one model from each technique and report the confusion matrix and the cost/gain matrix for the validation data. Which technique has the most net profit? 10

1.4 Let us try and improve our performance. Rather than accept the initial classification of all applicants credit status, use the predicted probability of sucess in logistic regression ..where success means 1.. as a basis for selecting the best credit risk first, followed by poorer risk applicants. . . . 10

1.4.1 Sort the validation on predicted probability of success. 10

1.4.2 For each case, calculate the net profit of extending credit. 10

1.4.3 Add another column for cumulative net profit. 10

1.4.4 How far into the validation data do you go to get the maximum net profit?(Often, this is specified as a percentile or rounded to deciles.) 10

1.4.5 If this logistic regression model is scored to future applicants, what 'probability of success' cutoff should be used in extending credit? 10

Bibliography 11

List of Figures

1.1 Distribution of $RESPONSE$, percentage share, n=1.000 observations . . 3
1.2 Distribution of $RESPONSE$ with other Categorical variables, percentage share of sub-categories of each category, n=1.000 5
1.3 Scatterplot Matrix of Numerical Variables, n=1.000 6
1.4 Correlation Matrix of German Credit Data Set, n=1.000 7

List of Tables

1.1 Variables for the German Credit Dataset (Shmueli et al., 2010) 2
1.2 Statistical Overview of German Credit Data Set, n=1.000 8

1 Appendix

In the following we will conduct the given data set in order to determine the optimal model and model specifications regarding credit risk classification prediction. The software programmes that are used for this purpose are: Microsoft Excel, IBM SPSS Modeler, and R. First of all the origin of the data set will be depicted and following this in section 1.1 the data set is being prepared for exploratory and analysis reasons. In section 1.2 we will develop classification models using the following data mining techniques: logistic regression, classification trees, and k-nearest neighbor. In section 1.3 the developed models' performances from section 1.2 will be evaluated using the confusion and cost/gain matrix. Finally the developed models' performances will be improved by altering the models' initial classification in section 1.4.

The German credit data set at hand was obtained from the homepage of Shmueli et al. (2010). Shmueli et al. (2010) obtained the original data set from Professor Dr. Hans Hofmann who owned the chair of statistics and econometrics at University of Hamburg until 2008. The original data set (german.data (*Machine Learning Database*, 2015)) was provided by Professor Hofmann in year 1994 and has served as important test data for the analysis and creation of credit-scoring algorithms. The original data set contained 7 numerical and 13 categorical attributes (in total 20 variables), these were transformed into numerical attributes in the data set of Shmueli et al. (2010) (GermanCredit (Shmueli et al., 2010)).

For the analysis, and determination of an appropriate model, in the following the data set *GermanCredit* from Shmueli et al. (2010) will be used (see table 1.1). This data consists of a set of n=1.000 multivariate observation with 7 categorical, 6 numerical, 19 binary attributes (in total 32 variables) in the area of finance. In the following analysis the variable $OBS\#$ describing the observation's number will be excluded because it does not any valuable information for the analysis. The dependent variable is $RESPONSE$.

1 Appendix

Table 1.1: Variables for the German Credit Dataset (Shmueli et al., 2010)

Var.#	Variable Name	Description	Variable Type	Code Description
1.	OBS#	Observation No.	Categorical	
2.	CHK_ACCT	Checking account status	Categorical	0: $<$ 0 DM 1: 0$<$..$<$ 200 DM 2: $\geq$ 200 DM 3: no checking account
3.	DURATION	Duration of credit in months	Numerical	
4.	HISTORY	Credit history	Categorical	0: no credits taken 1: all credits at this bank paid back duly 2: existing credits paid back duly till now 3: delay in paying off in the past 4: critical account
5.	NEW_CAR	Purpose of credit	Binary	car (new) 0: No, 1: Yes
6.	USED_CAR	Purpose of credit	Binary	car (used) 0: No, 1: Yes
7.	FURNITURE	Purpose of credit	Binary	furniture/equipment 0: No, 1: Yes
8.	RADIO/TV	Purpose of credit	Binary	radio/television 0: No, 1: Yes
9.	EDUCATION	Purpose of credit	Binary	education 0: No, 1: Yes
10.	RETRAINING	Purpose of credit	Binary	retraining 0: Yes, 1: No
11.	AMOUNT	Credit amount	Numerical	
12.	SAV_ACCT	Average balance in savings account	Categorical	0: $<$ 100 DM 1: 11$\leq$..$<$ 500 DM 2: 500$\leq$..$<$ 1.000 DM 3: $\geq$ 1.000 DM 4: unknown/ no savings account
13.	EMPLOYMENT	Present employment since	Categorical	0: unemployed 1: $<$ 1 year 2: 1 $\leq$..$<$ 4 years 3: $\geq$ 7 years
14.	INSTALL_RATE	Installment rate as % of disposable income	Numerical	
15.	MALE_DIV	Applicant is male and divorced	Binary	0: No, 1:Yes
16.	MALE_SINGLE	Applicant is male and single	Binary	0: No, 1: Yes
17.	MALE_MAR_WID	Applicant is male and married or a widower	Binary	0: No, 1: Yes
18.	COAPPLICANT	Application has a coapplicant	Binary	0: No, 1: Yes
19.	GUARANTOR	Applicant has a guarantor	Binary	0: No, 1: Yes
20.	PRESENT_RESIDENT	Present resident since - year	Categorical	0: $\leq$ 1 year
21.	REAL_ESTATE	Applicant owns real estate	Binary	0: No, 1: Yes
22.	PROP_UNKN_NONE	Applicant owns no property (or unknown)	Binary	0: No, 1: Yes
23.	AGE	Age in years	Numerical	
24.	OTHER_INSTALL	Applicant has other installment plan credit	Binary	0: No, 1: Yes
25.	RENT	Applicant rents	Binary	0: No, 1: Yes
26.	OWN_RES	Applicant owns residence	Binary	0: No, 1: Yes
27.	NUM_CREDITS	Number of existing credits at this bank	Numerical	

28.	JOB	Nature of job	Categorical	0: unemployed/unskilled - non-resident 1: unskilled - resident 2: skilled employed/official 3: management/self-employed/Highly qualified employee/officer
29.	NUM_DEPENDENTS	Number of people for whom liable to provide maintenance	Numerical	
30.	TELEPHONE	Applicant has phone in his or her name	Binary	0: No, 1: Yes
31.	FOREIGN	Foreign worker	Binary	0: No, 1: Yes
32.	RESPONSE	Credit rating is good	Binary	0: No, 1: Yes

1.1 Review the predictor variables and guess what their role in a credit decision might be. Are there any surprises in the data?

As follows the data set is being explored in four steps. First of all we will look into the dependent variable $RESPONSE$ (see figure 1.1) and then all categorical variables (see figure 1.2) excluding the $OBS\#$ variable will be explored. Following this we will look at all numerical variables (see figure 1.3) and finally a correlation matrix as well as statistical overview of all variables will be conducted (see figure 1.4 and 1.2 respectively). The German Credit data set contains n=1.000 observations with 300 responses with bad credit rating and 700 responses with good credit rating.

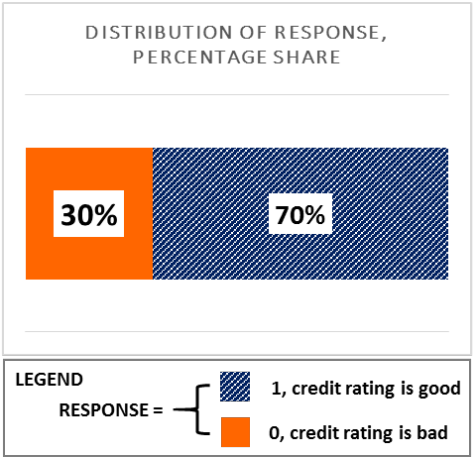

Figure 1.1: Distribution of $RESPONSE$, percentage share, n=1.000 observations

In figure 1.2 we can see the distribution of $RESPONSE$ by all other categorical variables

$CHECK\_ACC$, $HISTORY$, SAV_ACCT, $PRESENT\_RESIDENT$, $EMPLOYMENT$, and JOB. According to (Shmueli et al., 2010, pp.76-77) categorical variables can inflate the dimension of the dataset which we can see by comparing the original dataset which had 20 variables in total and the preprocessed data set which has 32 variables (due to categorical and binary variables). The preprocessing however is necessary in order to be able to conduct data mining methods on the data set. (Shmueli et al., 2010, pp.76-77) suggest to reduce the number of sub-categories through combination of similar sub-categories.

Regarding the distribution of $RESPONSE$ by $CHECK\_ACC$ there are four sub-categories with each having different distributions so that combining categories would not be a good choice. Sub-category *0:< 0 DM* is the worst and sub-category *3:no checking account* is the best according to the given data set. Putting the attribute *3:no checking account* as best outcome is strange because if there is no checking account then for instance a bank does not have any business and therefore it is actually a rather bad attribute. Since we do not have domain knowledge the order of $CHECK\_ACC$ will be kept as it is.

Looking at the distribution of $RESPONSE$ by $HISTORY$ there are five sub-categories with categories *0:no credits taken* and *1:all credits at this bank paid back* as well as *2:existing credits paid back duly* and *3:delay in paying off in the past* having similar distributions. With respect to their similarities in distribution these two pairs of categories could be combined, however with regard to the rather different attributional message of these variables as well as the missing domain knowledge, these variables are not combined. Regardless of this the order has to be reversed to stay constistent with $CHECK\_ACC$ (and for the interpretation) because the worst attribute is assigned the best category at the moment. With regard to SAV_ACCT there is the same issue as with $HISTORY$. Here two categories *0:< 100 DM* and *1:100≤..<500 DM* can be combined as well as the three other categories. However since each sub-category describes different attributions and due to missing domain knowledge the order will be kept as it is.

Regarding $PRESENT\_RESIDENT$ all four sub-categories have a similar distribution which leads to the assumption that they could be combined. However combining these categories would create a non-sense category. Furthermore there is an error in the data set regarding this variable with respect to the inconsistency of the coding which starts at 0 and ends with 3. The raw data set however starts with the value 1 and ends with the value 4. We could assume that during the transformation process the values were

shiftet by 1 unit. Since it is not retraceable how the error occured we decide to omitt this variable for the later analysis.

The variable *EMPLOYMENT* is sub-divided into five sub-categories with each having a slightly different distribution. We decide not to combine categories in this category because a long employment status implies higher income security and thus it makes sense to leave the categories as they are.

Finally the variable *JOB* is sub-divided into four sub-categories with each having a slightly different distribution. Sub-category *0:unemployed/unskilled - non-resident* is surprising because an unskilled nature of the job is sub-divided wether its an resident or not with non-residence being worse. The categories as they currently are in the right ascending order and due to their individual attributional message should remain in different categories.

Figure 1.2: Distribution of *RESPONSE* with other Categorical variables, percentage share of sub-categories of each category, n=1.000

Figure 1.3 displays all numerical pairwise scatterplots: $DURATION$, $AMOUNT$, $INSTALL\_RATE$, AGE, $NUM\_CREDITS$, and $NUM\_DEPENDENT$. According to the scatterplot there is a positive correlation between $DURATION$ and $AMOUNT$. Other correlations, where the trend is not clear, are between $DURATION$ and AGE as well as $AMOUNT$ and AGE. From the only few numbers and thus resulting patterns of the other variables $INSTALL\_RATE$, $NUM\_CREDITS$, and $NUM\_DEPENDENT$ on can conclude that these variables are not numerical in nature.

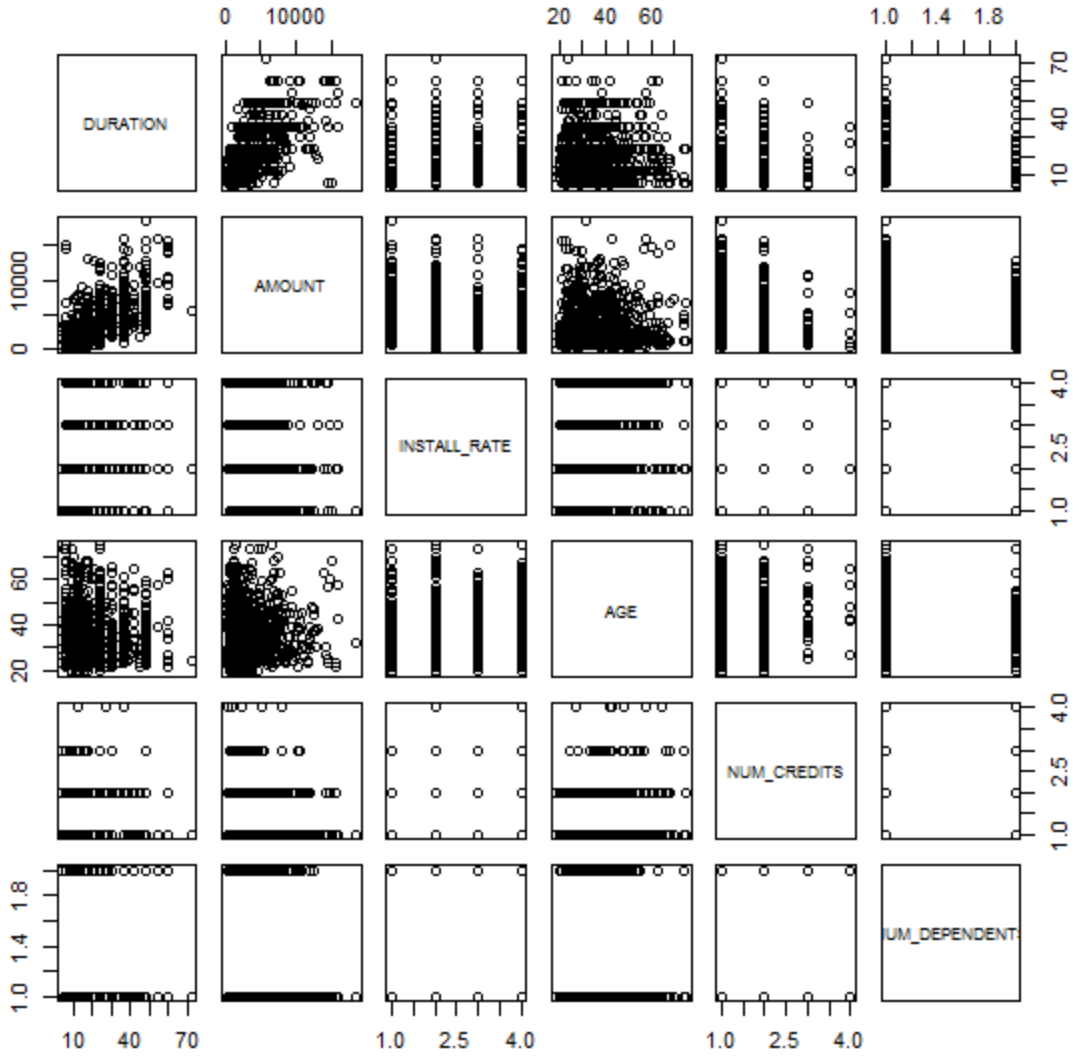

Figure 1.3: Scatterplot Matrix of Numerical Variables, n=1.000

In figure 1.4 the correlation matrix of the German Credit data set is displayed. According to the correlation matrix there is one pair of variables which has a high correlation coefficient of -0.74[1]. This implies that there is multicollinearity regarding this pair.

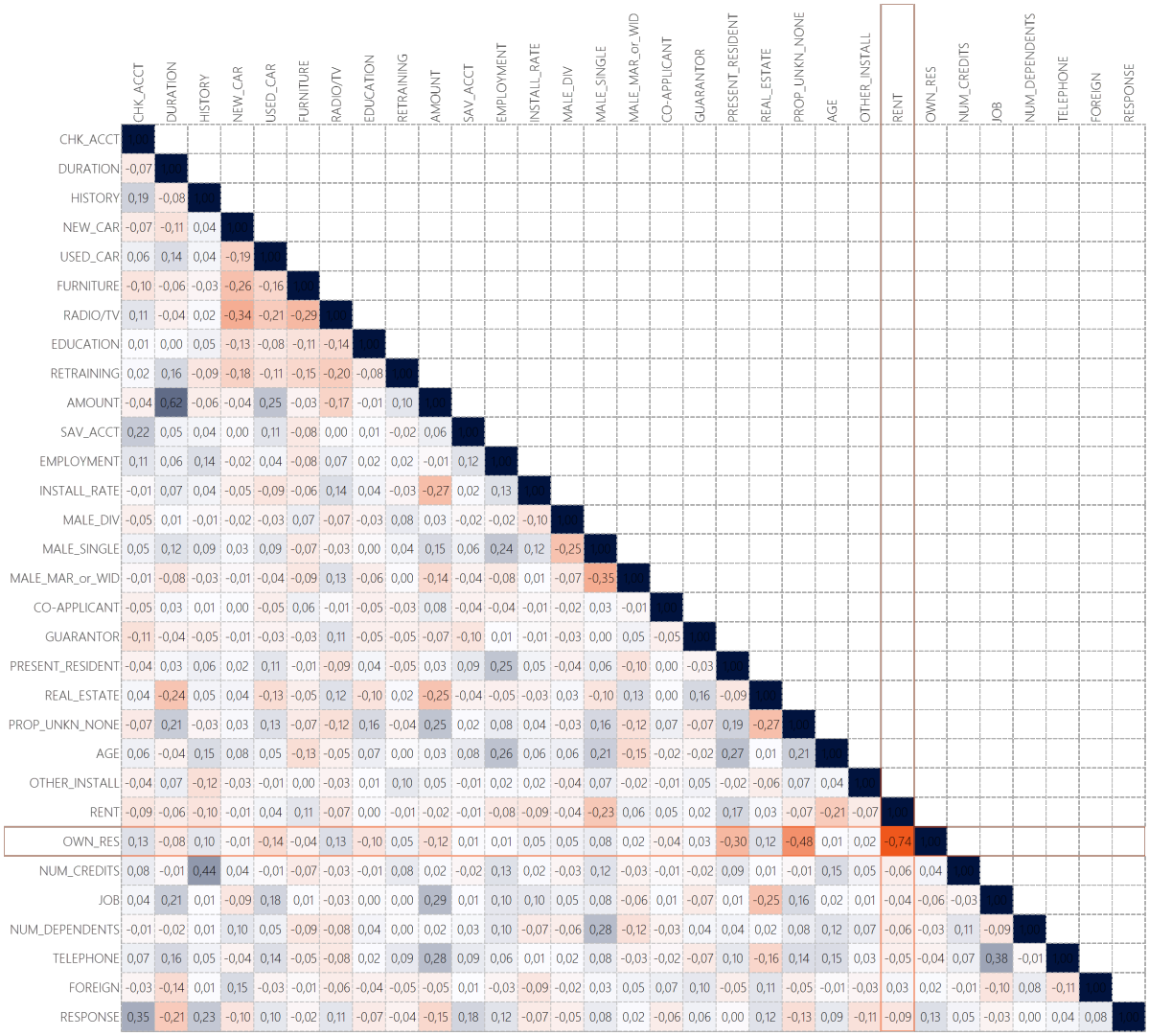

Figure 1.4: Correlation Matrix of German Credit Data Set, n=1.000

Finally table 1.2 provides an overview about basic statistical values for all variables. We can see that there is a great variety in terms of the variables' range which has to be taken into consideration.

[1]A high correlation coefficient is set in this analysis when the coefficient is $\geq |0.7|$

Table 1.2: Statistical Overview of German Credit Data Set, n=1.000

Variable	**Av. Value**	**SE**	**Med.**	**Mod.**	σ	**Sample** σ^2	**Curto-sis**	**Skew-ness**	**Value Range**	**Min.**	**Max.**	**Sum**	**Conf. lvl**
OBS#	500,50	9,13	500,50	#N/A	288,82	83.416,67	(1,20)	(0,00)	999	1	500.500	1.000	17,92
CHK_ACCT	1,58	0,04	1,00	3	1,26	1,58	(1,66)	0,01	3	-	3	1.577	0,08
DURATION	20,90	0,38	18,00	24	12,06	145,42	0,92	1,09	68	4	72	20.903	0,75
HISTORY	2,55	0,03	2,00	2	1,08	1,17	(0,58)	(0,01)	4	-	4	2.545	0,07
NEW_CAR	0,23	0,01	-	-	0,42	0,18	(0,42)	1,26	1	-	1	234	0,03
USED_CAR	0,10	0,01	-	-	0,30	0,09	4,85	2,62	1	-	1	103	0,02
FURNITURE	0,18	0,01	-	-	0,39	0,15	0,76	1,66	1	-	1	181	0,02
RADIO/TV	0,28	0,01	-	-	0,45	0,20	(1,04)	0,98	1	-	1	280	0,03
EDUCATION	0,05	0,01	-	-	0,22	0,05	15,13	4,14	1	-	1	50	0,01
RETRAINING	0,10	0,01	-	-	0,30	0,09	5,45	2,73	1	-	1	97	0,02
AMOUNT	3.271,26	89,26	2.319,50	1.393	2.822,74	7.967.843,47	4,29	1,95	18.174	250	18.424	3.271.258	175,16
SAV_ACCT	1,11	0,05	-	-	1,58	2,50	(0,68)	1,02	4	-	4	1.105	0,10
EMPLOYMENT	2,38	0,04	2,00	2	1,21	1,46	(0,93)	(0,12)	4	-	4	2.384	0,07
INSTALL_RATE	2,97	0,04	3,00	4	1,12	1,25	(1,21)	(0,53)	3	1	4	2.973	0,07
MALE_DIV	0,05	0,01	-	-	0,22	0,05	15,13	4,14	1	-	1	50	0,01
MALE_SINGLE	0,55	0,02	1,00	1	0,50	0,25	(1,97)	(0,19)	1	-	1	548	0,03
MALE_MAR_or_WID	0,09	0,01	-	-	0,29	0,08	6,01	2,83	1	-	1	92	0,02
CO-APPLICANT	0,04	0,01	-	-	0,20	0,04	19,54	4,64		-	1	41	0,01
GUARANTOR	0,05	0,01	-	-	0,22	0,05	14,36	4,04	1	-	1	52	0,01
PRESENT_RESIDENT	2,85	0,03	3,00	4	1,10	1,22	(1,38)	(0,27)	3	1	4	2.845	0,07
REAL_ESTATE	0,28	0,01	-	-	0,45	0,20	(1,06)	0,97	1	-	1	282	0,03
PROP_UNKN_NONE	0,15	0,01	-	-	0,36	0,13	1,69	1,92	1	-	1	154	0,02
AGE	35,55	0,36	33,00	27	11,38	129,40	0,60	1,02	56	19	75	35.546	0,71
OTHER_INSTALL	0,19	0,01	-	-	0,39	0,15	0,61	1,62	1	-	1	186	0,02
RENT	0,18	0,01	-	-	0,38	0,15	0,81	1,68	1	-	1	179	0,02
OWN_RES	0,71	0,01	1,00	1	0,45	0,20	(1,11)	(0,94)	1	-	1	713	0,03
NUM_CREDITS	1,41	0,02	1,00	1	0,58	0,33	1,60	1,27	3	1	4	1.407	0,04
JOB	1,90	0,02	2,00	2	0,65	0,43	0,50	(0,37)	3	-	3	1.904	0,04
NUM_DEPENDENTS	1,16	0,01	1,00	1	0,36	0,13	1,65	1,91	1	1	2	1.155	0,02
TELEPHONE	0,40	0,02	-	-	0,49	0,24	(1,85)	0,39	1	-	1	404	0,03
FOREIGN	0,04	0,01	-	-	0,19	0,04	22,18	4,91	1	-	1	37	0,01
RESPONSE	0,70	0,01	1,00	1	0,46	0,21	(1,24)	(0,87)	1	-	1	700	0,03

1.2 Devide the data into training and validation partitions, and develop classification models using the following data mining techniques: logistic regression, classification trees, and k-nearest neighbor

Logistic Regression

Classification Trees

K-Nearest Neighbor

1.3 Choose one model from each technique and report the confusion matrix and the cost/gain matrix for the validation data. Which technique has the most net profit?

1.4 Let us try and improve our performance. Rather than accept the initial classification of all applicants credit status, use the predicted probability of sucess in logistic regression ..where success means 1.. as a basis for selecting the best credit risk first, followed by poorer risk applicants.

1.4.1 Sort the validation on predicted probability of success.

1.4.2 For each case, calculate the net profit of extending credit.

1.4.3 Add another column for cumulative net profit.

1.4.4 How far into the validation data do you go to get the maximum net profit?(Often, this is specified as a percentile or rounded to deciles.)

1.4.5 If this logistic regression model is scored to future applicants, what 'probability of success' cutoff should be used in extending credit?

Bibliography

Machine learning database [Web Page]. (2015). Retrieved from `https://archive.ics.uci.edu/ml/machine-learning-databases/statlog/german/` (22.11.2015)

Shmueli, G., Patel, N. R., & Bruce, P. C. (2010). *Data mining for business intelligence* [Book]. John Wiley & Sons.