Project 2
FYS-STK4155

Mina Tangen

October 14, 2019

1 Methods

1.1 Logistic regression

Logistic regression is a method used in classification problems. It gives a binary output value.

2 Data

2.1 Credit card data

This dataset is retrived from the Machine Learning Repository of the University of California, School of Informaton and Computer Science [ref.]. It is a collection of payment data from a bank i Taiwan in 2005 (Yeh, I-C. & Lien, C-H. (2009)).

The dataset has 23 explanatory variables ($X_1 - X_{23}$) in addition to client IDs and the response variable (Y). The explanatory variables holds information on the different consumers, such as information on their gender, age and education. The Appendix gives a detailed description of alle the explanatory variables. The response variable Y decribes whether or not the client defaults in the next month, i.e. if the client fails to pay by the deadline or not. Y can have the value 1 (meaning that the client did fail to pay) or 0 (meaning that the client payed on time).

Our aim using this data set is therefore to use the explanatory variables X_i to predict the response variable Y. Or, stated in more colloquial terms, how can a client's gender, age, education, previous payment history etc. explain whether or not he/she pays their credit card bill on time?

3 Appendix

3.1 Explanatory variables

Variable X_1 holds the amount of credit (given in NT dollars), and includes both the individual and supplementary credit (i.e., the consumers own credit and the credit of the family of that consumer).
Variable X_2 describes the gender, and has a value of 1 if the consumer is male and 2 if the consumer is female.
Variable X_3 describes the consumer's level of education, and takes on values from 1 to 4, where 1 is graduate school, 2 university, 3 high school and 4 others.

Variable	Description	Unit / value range	Name
X_1	Given credit	NT dollars	LIMIT_BAL
X_2	Gender	1 - 2	SEX
X_3	Education	1 - 4	EDUCATION
X_4	Marital status	1 - 3	MARRIAGE
X_5	Age	Year	AGE
$X_6 - X_{11}$	History of past pay	-1, 1 - 9	PAY_1[1] - PAY_6
$X_{12} - X_{17}$	Amount of bill statement	NT dollars	BILL_AMT1 - BILL_AMT6
$X_{18} - X_{23}$	Amount of previous payment	NT dollars	PAY_AMT1 - PAY_AMT6

Table 1: Overview of explanatory variables of the Credit card data set

Variable X_4 describes wether the consumer is married ($X_4 = 1$), single ($X_4 = 2$) or some other marital status ($X_4 = 3$).
Variable X_5 holds the age of the consumer in years.
Variables $X_6 - X_{11}$ holds past monthly payment records, from April (X_{11}) to September (X_6) 2005. These variables take on values from 1-9, which describes the delay of the past payment in months. E.g., if one of the variables has value 1, then the payment was done one month after the deadline and 9 months (or more) if the variables has the value 9. If the payment was done on time, these variables has value -1.
Variables $X_{12} - X_{17}$ holds the amount on the bill statement (given in NT dollars) for the same months as for the previous variables, whereas variables $X_{18} - X_{23}$ holds the amount of the previous payments (given in NT dollars), also for those same months. As for the past monthly payment records, the variables are orderer backwards (starting in September (X_{12}/X_{18}) going to April (X_{17}/X_{23})).

Reference Yeh, I-C. & Lien, C-H. (2009)

4 References

https://archive.ics.uci.edu/ml/datasets/default+of+credit+card+clients#

Yeh, I-C. & Lien, C-H. (2009) *The comparisons of data mining techniques for the predictive accuracy of probabilty of default of credit card clients.* Expert Systems with Applications, 36(2), 2473-2480.