CC_Default Dataset Desciption

Background:

The problem of default on credit card is on the rise. Although the US economy is improving and unemployment rate is record low, the big four US retail banks sustained approximately 20 per cent jump in losses from credit cards in 2017, which shows the deterioration in the ability of consumers to keep up with their debt service burden. Recently disclosed results showed Citigroup, JPMorgan Chase, Bank of America and Wells Fargo took a combined $12.5bn hit from soured card loans last year, about $2bn more than a year ago [1]. Missed payments on credit cards at small banks have risen sharply over the past year, a sign that their cardholders are taking on more debt than they can handle [2]. The following research is to identify the parameters to help banks minimize providing credit cards to such clients and hence, decrease the overall loss from credit card default.

Dataset Description:

The research is being carried out on public dataset [3]. The dataset has a binary variable: default payment, as the response variable and 23 other variables, as explanatory variables.

The details of the variables defined in the dataset are:

LIMIT_BAL:	Amount of the given credit (NT dollar): it includes both the individual consumer credit and his/her family (supplementary) credit.
SEX:	Gender (1 = male; 2 = female).
EDUCATION:	(1 = graduate school; 2 = university; 3 = high school; 4 = others).
MARRIAGE:	(1 = married; 2 = single; 3 = others).
AGE:	(year).
PAY_1 – PAY_6:	History of past payment. The past monthly payment records (from April to September 2005) as follows: PAY_1 = the repayment status in September 2005; PAY_2 = the repayment status in August 2005; . . .; PAY_6 = the repayment status in April, 2005. The measurement scale for the repayment status is: -1 = pay duly; 1 = payment delay for one month; 2 = payment delay for two months; . . .; 8 = payment delay for eight months; 9 = payment delay for nine months and above.
BILL_AMT1 - BILL_AMT6:	Amount of bill statement (NT dollar). BILL_AMT1= amount of bill statement in September 2005; BILL_AMT2 = amount of bill statement in August 2005; . . .; BILL_AMT6 = amount of bill statement in April 2005.
PAY_AMT1 - PAY_AMT6:	Amount of previous payment (NT dollar). PAY_AMT1 = amount paid in September 2005; PAY_AMT2 = amount paid in August 2005; . . .; PAY_AMT6 = amount paid in April, 2005.

DEF_PAY_NEXT_MONTH: Default payment (1=yes, 0=no).

The Categorical Variables defined in the dataset are SEX, EDUCATION, MARRIAGE, PAY_1, PAY_2, PAY_3, PAY_4, PAY_5, PAY_6.

Table 1. The Numerical Variables with Summary

Variable	Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
LIMIT_BAL	10000	50000	140000	167500	240000	1000000
AGE	21.00	28.00	34.00	35.49	41.00	79.00
BILL_AMT1	-165600	3559	22380	51220	67090	964500
BILL_AMT2	-69780	2985	21200	49180	64010	983900
BILL_AMT3	-157300	2666	20090	47010	60160	1664000
BILL_AMT4	-170000	2327	19050	43260	54510	891600
BILL_AMT5	-81330	1763	18100	40310	50190	927200
BILL_AMT6	-339600	1256	17070	38870	49200	961700
PAY_AMT1	0	1000	2100	5664	5006	873600
PAY_AMT2	0	833	2009	5921	5000	1684000
PAY_AMT3	0	390	1800	5226	4505	896000
PAY_AMT4	0	296	1500	4826	4013	621000
PAY_AMT5	0.0	252.5	1500.0	4799.0	4032.0	426500.0
PAY_AMT6	0.0	117.8	1500.0	5216.0	4000.0	528700.0

On successfully importing the dataset in RStudio [4] and pre-processing it was noted that there were no missing values nor any duplicate values in the dataset. However, the dataset was found to be imbalanced as the ratio of no to yes in the response variable, default payment was 77.88:22.12. Hence, the number non-defaulters were observed to be approximately 3.5 times more than the number of defaulters. Also, it was observed that the degree of correlation between BILL_AMT1, BILL_AMT2, BILL_AMT3, BILL_AMT4, BILL_AMT5 and BILL_AMT6 was very high whereas the degree of correlation between other variables was very low.

Literature Review:

Further, the aim of this research is to apply different algorithms like Decision tree, K Nearest Neighbors, Support Vector Machine, Logistic Regression, Artificial Neural Network and to find the best method to find the characteristics of credit card holders who are more likely to default. The theme of the problem is based on Classification, Predictive Analytics, Data mining and Knowledge Discovery. To understand the concepts of classification and dimensionality reduction books [5] & [6] will be referred. Predictive Analytics, Data mining and Knowledge Discovery topics will be referred from [7] & [8].

Since the data is found to be imbalanced, it is observed that Support Vector Machine does not perform well on imbalanced data sets [9]. To use SVM for this data, and to increase the accuracy of the results the data must be balanced. Hence, to transpose the data as balanced dataset, oversampling technique like SMOTE will be applied.

Feature selection generally is believed to enhance the accuracy of the resulting classifier and often constructs a model that generalizes better to unseen points as referred in the research paper [10].

For decision tree classifier, it is not required to convert the imbalanced dataset to balanced dataset. Class

Confidence Proportion Decision Tree (CCPDT), which is robust and insensitive to size of classes generates rules which are statistically significant as discussed in research paper [11].

K-NN classifier requires an equal number of good and bad sample cases for better performance [12]. Therefore, the dataset will be transposed to balanced dataset and then K-NN classifier will be applied.

According to Asrin KARIMI [13], logistic regression model has high accuracy in estimating good customers compared with ANN model and also ANN models in a comparison with LRM has high accuracy in estimating bad customers.

According to the research by Yeh and Lien [14], artificial neural network is the only one that can accurately estimate the real probability of default.

Therefore, all the algorithms defined above will be applied and outputs will be compared to see which performs better giving us more accurate results.

References:

[1] Alistair Gray in New York January 21, 2018, US banks suffer 20% jump in credit card losses https://www.ft.com/content/bafdd504-fd2c-11e7-a492-2c9be7f3120a

[2] AnnaMaria Andriotis, Credit-Card Losses Surge at Small Banks , March 4' 2018 https://www.wsj.com/articles/credit-card-losses-surge-at-small-banks-1520198436

[3] I-Cheng Yeh, Default of Credit Card Clients Data Set, Department of Information Management, Chung Hua University, Taiwan. Department of Civil Engineering, Tamkang University, Taiwan. https://archive.ics.uci.edu/ml/datasets/default+of+credit+card+clients

[4] RStudio, https://en.wikipedia.org/wiki/RStudio

[5] Ethem Alpaydın, Introduction to Machine Learning, Second Edition, 2010 Massachusetts Institute of Technology.

[6] Bishop, Pattern Recognition And Machine Learning, Springer 2006.

[7] Oded Maimon, Lior Rokach, Data Mining And Knowledge Discovery Handbook, Second Edition, Springer 2010.

[8] Vijay Kotu Bala Deshpande, Predictive Analytics and Data Mining Concepts and Practice with RapidMiner, 2015 Elsevier Inc.

[9] Rehan AkbaniStephen KwekNathalie Japkowicz, ECML 2004: Machine Learning: ECML 2004 pp 39-50 | Cite as Applying Support Vector Machines to Imbalanced Datasets. https://link.springer.com/chapter/10.1007/978-3-540-30115-8_7

[10] Ajay, Ajay Venkatesh, Shomona Gracia Jacob, Prediction of Credit-Card Defaulters: A Comparative Study on Performance of Classifiers, International Journal of Computer Applications (0975 – 8887) Volume 145 – No.7, (July 2016). https://pdfs.semanticscholar.org/4c24/5233a5f19abb4336979d4c5a65e502f443ee.pdf

[11] Wei Liu, Sanjay Chawla, David A. Cieslak, Nitesh V. Chawla, A Robust Decision Tree Algorithm for

Imbalanced Data Sets, (2010), https://epubs.siam.org/doi/abs/10.1137/1.9781611972801.67

[12] D.J. Hand, J.E. Henley, Statistical Classification Methods in Consumer Credit Scoring: a Review, J.R. Statist. Soc. A (1997) 160, Part 3, pp. 523 - 541 https://pdfs.semanticscholar.org/fa58/5ac49b37a801ccd1b2e49118518414c810e2.pdf

[13] Asrin KARIMI, Credit Risk Modeling for Commercial Banks, International Journal of Academic Research in Accounting, Finance and Management Sciences, Vol. 4, No.3, July 2014, pp. 187–192, E-ISSN: 2225-8329, P-ISSN: 2308-0337 http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.679.1812&rep=rep1&type=pdf

[14] I. Yeh, C. Lien, The comparisons of data mining techniques for the predictive accuracy of probability of default of credit card clients, Expert Syst. Appl. 36 (2009) 2473–2480. doi:10.1016/j.eswa.2007.12.020 https://dl.acm.org/citation.cfm?id=1465163