Predicting default of credit card clients

Roger Bukuru (BKRROG001)

University of Cape Town
Multivariate Statistics (STA50697Z)
Project Proposal

Introduction

One of the many functions of a bank is the ability to provide its customers with loans or credit facilities. These channels provide banks with the opportunity to generate additional revenues by means of gaining interest off the back of these financial instruments.

However these instruments have an intrinsic risk once awarded to a customer, customers may default in their repayments of such loans or credit facilities. In this research, we will focus on credit card facilities, specifically credit card clients in Taiwan. We will be exploring the probability of clients defaulting on their credit card facility given a set of features such as their demographic information (like gender, education, marital status, and age) to financial behaviors (including payment history, bill statement amounts, and previous payment amounts). The exploration will aim to provide statistical techniques to better manage credit risk default. Credit risk here means, the probability of a delay in the repayment of the credit granted.

Furthermore, we aim to explore given the above-mentioned features, which features/factors are more likely to affect a customer defaulting e.g. are younger customers more likely to default than older clients, and given these factors what is the probability of a customer defaulting? This will be done by exploring established techniques such as employing customer risk segmentation by conducting clustering techniques such as k-means and Linear Discriminant Analysis (LDA) techniques[2]. To assess the effectiveness of these techniques we will explore and compare them to Support Vector Machines (SVMs). As our data has many input variables, feature extraction methods such as PCA and Autoencoders will be explored before performing the above clustering and classification techniques. The repository of this project is available at https://github.com/rogerbukuru/Predicting-Default-of-Credit-Card-Clients

Keywords: credit card; default; bank; risk profiles; support vector machines, principal component analysis, linear discriminant analysis; autoencoders;

---

1 Hypotheses

Assessing the default of credit card clients is a non-linear problem as factors that can affect a client's defaulting are not linearly independent e.g. if one one has a high income this might generally reduce their probability of defaulting however coupled with high existing debt, the risk might not decrease linearly. Therefore we expect the non-linear methods presented to outperform the linear methods in better classifying or predicting customers defaulting as the non-linear methods will better understand the underlying non-linear effects of explanatory variables that contribute to default.

2 Literature Review

Credit card lending is a widely researched subject, our focus in this research focuses more particularly on credit risk prediction. An important risk management technique for banks and credit institutions in the hopes of reducing default risk with such financial instruments[2].

[1] illustrates that a number of statistical methods have been applied to develop reliable credit risk prediction models. In their research on credit card defaults, [1] implement feature selection and extraction techniques such as Principal Component Analysis(PCA) and Linear Discriminate Analysis (LDA). They proceeded to apply classification methods including K-nearest neighbor (KNN), Naive Bayesian (NB), Decision Tree (DT), and Support Vector Machines (SVM) to analyze credit card client defaults, identifying the most effective technique for assessing credit card default risk. Their findings indicated that SVM when combined with feature selection and extraction methods, achieved the highest predictive accuracy.

[2] went with the approach of placing customers within segments that represent various default risk profiles. This is conducted by using clustering techniques such as Density-Based Spatial Clustering of Applications with Noise (DBSCAN) and Self-Organzaing Maps (SOMs). For credit default prediction they utilized a tree-based method, Gradient Boosting. Their segmentation results were unsatisfactory due to issues such as too few feature variables, with the credit card they utilized. On the other hand, their prediction model produced high enough certainty that the bank for which the study was conducted for could utilize it in its functions.

[4] point out that from the perspective of risk control, estimating the probability of default will be more mean-

ingful than classifying customers into binary results– risky and non-risky. This is however highlighted to be a difficult task given credit issuers do not have a real probability of default. This is however attempted with 6 classification techniques including a k-nearest neighbor, logistic regression, discriminant analysis, naïve-Bayes, neural networks, and classification trees. Artificial neural networks proved to be the most reliable both in terms of classification and predictive accuracy regarding credit risk.

3 Aims and Objectives

Our aim as eluded to earlier will be to present various models that can help predict the default of credit card clients by means of clustering and classification techniques which will include the following:

1 Clustering (K-Means)

2 Linear Discriminant Analysis (LDA)

3 Support Vector Machines (SVM)

The analysis of the above techniques will be evaluated after the implementation of feature selection techniques or dimension reduction techniques, which will include both linear and non-linear techniques, these techniques include the following:

- Principal Component Analysis
- Autoencoders

4 Data Description

The data we will be performing our analysis consists of 30,000 observations that represent default payments of distinct credit card holders from a bank (a cash and credit card issuer) in Taiwan from April 2005 to September 2005[4]. The dataset utilizes a binary variable, default payment (Yes =1, No=0), as the response variable.

The data consists of the following 23 predictor variables:

The initial set of variables comprises details regarding the client's personal information:

- Limit Balance: A customer credit limit
- Sex: (1 - Male; 2 - Female)
- Education: 1 = graduate school; 2 = university; 3 = high school; 4 = others
- Marriage: 1 - Married; 2 - Single; 3 - Other
- Age: Customer age

The subsequent attributes provide details on the delay in past payments corresponding to specific months:

- Pay0 - Pay11: History of past payment. We tracked the past monthly payment records (from April to September 2005) as follows: Pay0 = the repayment status in September 2005; Pay2 = the repayment status in August 2005; . . .; Pay6 = 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.

The following variables focus on information concerning the bill statement amount (that is, a monthly statement issued by credit card companies to cardholders for a particular month):

- BillAmt1 - BillAmt6: Amount of bill statement (NT dollar). BillAmt1 = amount of bill statement in September 2005; BillAmt2 = amount of bill statement in August 2005; . . .; BillAmt6 = amount of bill statement in April 2005.

The last set of variables takes into account the amount of previous payments made in a specific month.

- PayAmt1 - PayAmt6: Amount of previous payment (NT dollar). PayAmt1 = amount paid in September, 2005; PayAmt2 = amount paid in August, 2005; . . .; PayAmt6 = amount paid in April 2005.

The dataset originates from a study aimed at predicting credit card defaults, by [3], where the data came from an important bank in Taiwan. The study was aimed at comparing various classification techniques to predict default payments by credit card clients. Given it is sensitive information the data is completely anonymized and aggregated, which means that individual identities are not revealed and the data is presented in a summary form e.g. clients are identified in an anonymized chronological order from 1 to 30 000. Each observation is ordered according to the features described above.

4.1 Response Variable Analysis

Among the 30,000 observations, 6636 observations (22.12%) are cardholders with default payment which means that 77.88% of observations are clients who did not default, this is shown in Figure 1. This discrepancy might stem from significant bias, thus not accurately reflecting the bank's clientele. Nonetheless, it's important to recognize that the data collection occurred during a period of debt crisis, reinforcing the view that the data constitutes an unbiased sample of the customer base.[2]. Therefore, it is crucial to consider the high rate of defaults when drawing conclusions from the research presented.

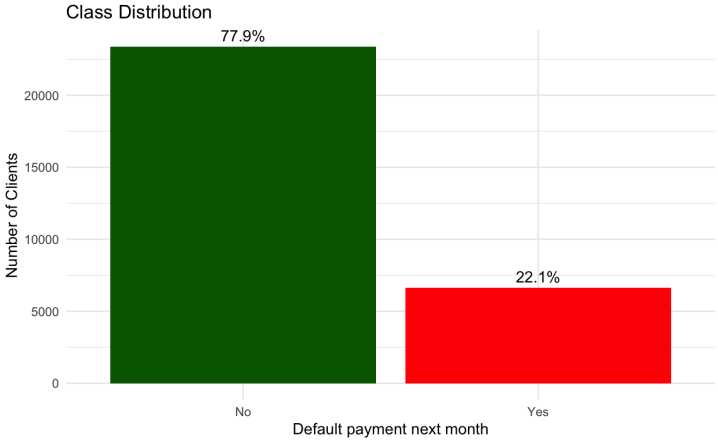

Fig. 1: Response Variable Distribution

5 Bibliography

[1] B. Emil Richard Singh and E. Sivasankar. Risk analysis in electronic payments and settlement system using dimensionality reduction techniques. In *2018 8th International Conference on Cloud Computing, Data Science Engineering (Confluence)*, pages 14–19, 2018.

[2] M. Merikoski, A. Viitala, and N. Shafik. Predicting and preventing credit card default. 2018.

[3] I.-C. Yeh. Default of credit card clients. UCI Machine Learning Repository, 2016. DOI: https://doi.org/10.24432/C55S3H.

[4] I.-C. Yeh and C. hui Lien. The comparisons of data mining techniques for the predictive accuracy of probability of default of credit card clients. *Expert Systems with Applications*, 36(2, Part 1):2473–2480, 2009.