Research on Predicting Credit Card Customers' service using Logistic Regression and Bp Neural Network

Xirui Wan

Flintridge Sacred Heart Academy, CA 91011, USA.

xiruiwan24@fsha.org

Keywords: Machine Learning, Data science, Logistic Regression, Bp Neural Network, credit card service

Abstract: As the number of credit card customer increases, the urgency for an efficient model to predict customer services became larger than ever. On the other hand, machine learning is also growing as a new branch of artificial intelligence. Its ability to process large amounts of data sets and do tasks with human monitoring well suits the market of predicting customer services. However, machine learning is sorted into many different kinds of models with different advantages and drawbacks. The Logistic regression model is used to analyze and explain the relationship between a nominal scale response variable and more than one explanatory variable. In addition, Logistic regression model is the best for credit card companies because of its clear and easily interpreted outputs. However previous research has shown that neural networks seem to have a higher precision and accuracy. In order to test for the truth, a comparison of Bp neural network and Logistic regression model is being made based on the data set provided by Kaggle. The data set included 23 kinds of information about 10128 customers and outliers had been previously eliminated.

1. Introduction

Introduced by Diners' Club, Inc in 1950, Credit cards had benefited customers of all kinds. As the credit card progressed, companies categorized each customers' credit card into Blue, silver and gold (not considering the outliers) and gave each customer a credit limit [3]. Credit limit symbolises not only the maximum amount of money charged, it also reflects the owner's credit score. There are mainly four factors that affect a user's credit score: credit history, capacity, capital, and collateral [4]. First of all, credit history reflects the user's ability to manage his or her own debt which includes payments, earnings, and is the user progressing towards debt free. Second, Capacity mainly focuses on the user's monthly earnings and tax returns. Third, Capital which represents not only the user's savings but also the history of the savings. At Last, Collateral is the sign of security for the lender which makes sure the lender will be paid back in some form [4]. On top of credit score, credit card services had not been needed by every customer. If the companies had better ability to predict the possibility a customer is going to leave credit card service, the result will not only benefit the providers but also credit card users. Through the process of eliminating the customer that abandons credit card services, more efficient services could be provided for customers in need of credit card service.

As technology progresses and the system settles, most credit limits could be decided by machines with minor or completely no human monitoring. Therefore the importance of a machine that could predict the user's need for customer services with self learning skills became larger and larger. Applying logistic regression and Bp neural network models to the area of credit cards is following the time's trend. It could be further applied not only for large companies like national banks but also for small lenders and loaners. This machine could provide its users highly accurate predictions in the area of credit card services which could be set as a standard for both lenders and loaners. The result would provide a stable and reasonable possibility of leaving customers for the bankers which could be further used into predicting a company's future. The high positive or negative correlation of leaving customers and certain aspects of the customers would also be a strong supporting evidence on which

  DOI: 10.25236/memssr.2021.010

customers are “strong customers” and which way of commercializing shall the company progress.

The rapid change of the Big data era had long been predicted. In Moore’s Law, Gordon Moore predicted that every year after 1965 will double in the amount of integrated circuits. In real world, the growth had been more than double, in fact the growth of every year seems to be exponential [2]. With the growing data, the challenge of sorting and using all the unstructured data in the digital universe seems to be an urgent problem. Correspondingly, data science also emerged from large amounts of data bases and the Big Data Era. The techniques of processing the data by human became more and more unrealistic due to the size of the data and many more [2]. Machine learning, a relatively new branch of the artificial intelligence area of study, focused on teaching machines the ability to process data and learn skills on their own [1]. This shift from human to machines also brought a huge impact on business and communication [2]. Even when performing a simple task, one might be interacting with several businesses. For example, when one is booking a ticket online, the customer will input their information and the verification for scanning (passport, Id card, QR code, etc.). The information will then be processed by the transportation company, the online booking website, the governmental database, and many more. The companies need to take the information they need and sometimes even process the information that had been preprocessed by other companies. Performing all the task of processing data information with expectation crisis is the end goal of machine learning.

In this paper, two models will be compared in their accuracy, precision, recall, f1 score, and roc in order to obtain the best model for predicting services for credit card customers. Inside the Logistic regression model, L1 and L2 logistic regression models are being focused on. On the other hand, Bp neural network was chosen out of the neural networks as a typical example to be compared towards. Previous research had often shows the neural networks seem to perform better at classification and identification, this paper will test if Bp neural network is better at prediction. In one of the cases, a team of Chinese people conducted an analysis of Logistic Regression and Bp neural network in the area of classifying heart sound signals and identifying normal heart sounds. In their abstract, they clearly wrote that the results of the neural network at classifying heart sounds is better than those of the logistic regression model [10]. In this paper, the process of why logistic regression model, especially L1, is the most efficient model for predicting services for credit card customers.

2. Data Research

The following data set consists of multiple factors of the customer’s personal information. A total of 23 different categories of personal information of more than 10,000 customers give us a detailed overview of nowadays credit card customers.

Most important of all, the Attrition Flag represents the depletion marks for each customer on their customer service. In the research, the main task is to predict the attrition marks with the following values. The figure below shows the processed data used in our model which highlights the mean, standard deviation, and the min to max. In addition to the end product, the data set itself also showcases interesting results. First of all, the Gender mean is 0.47 which means women and men were almost balanced in this data. This may be used to further support the prediction that the end result will be applicable for both subgroups. Second, the range of the customer age was relatively small which supports the data set to be free of outliers. Two other supporting factors of this statement is the minimum age of applying for a credit card is 18 and the global average life expectancy is 72.6 which is close to 73. Lastly, the average credit limit is relatively higher than the American average credit limit which shows the data set didn’t include the lowest or highest income people which brought up the average since there are less high income people than there are low income people in the United States. Besides all the interesting side factors the most important usage of the data set is to train the models with efficient data so the model could reach its best function. Referring to most of the Literature reviews, 80% of data was randomly chosen for training the model while the rest (20%) will be left for testing the functions of the models.

Table 1: processed data of the training set [5]

	mean	std	min	25%	50%	75%	max
Attrition_Flag	0.84	0.37	0	1	1	1	1
Customer_Age	46.33	8.02	26	41	46	52	73
Gender	0.47	0.5	0	0	0	1	1
Dependent_count	2.35	1.3	0	1	2	3	5
Education_Level	3.1	1.83	0	2	3	5	6
Marital_Status	1.46	0.74	0	1	1	2	3
Income_Category	2.86	1.5	0	2	3	4	5
Card_Category	0.18	0.69	0	0	0	0	3
Months_on_book	35.93	7.99	13	31	36	40	56
Total_Relationship_Count	3.81	1.55	1	3	4	5	6
Months_Inactive_12_mon	2.34	1.01	0	2	2	3	6
Contacts_Count_12_mon	2.46	1.11	0	2	2	3	6
Credit_Limit	8631.95	9088.78	1438.3	2555	4549	11067.5	34516
Total_Revolving_Bal	1162.81	814.99	0	359	1276	1784	2517
Avg_Open_To_Buy	7469.14	9090.69	3	1324.5	3474	9859	34516
Total_Amt_Chng_Q4_Q1	0.76	0.22	0	0.63	0.74	0.86	3.4
Total_Trans_Amt	4404.09	3397.13	510	2155.5	3899	4741	18484
Total_Trans_Ct	64.86	23.47	10	45	67	81	139
Total_Ct_Chng_Q4_Q1	0.71	0.24	0	0.58	0.7	0.82	3.71

3. Models

3.1 Logistic Regression

Linear regression models are a popular statistical method in quantitative analysis studies.

However, in the real world, linear regression models do not provide a good explanation for categorical output variables, so a common approach to setting categorical variables is to use log-linear regression models [8]. A log-linear model becomes a logistic regression model when the dichotomous variables in the log-linear model are treated as output variables and defined as a function of a set of independent variables. Logistic regression models have many advantages: they can handle dichotomous dependent variables; they do not require the assumptions required by other multivariate techniques; they automate variable selection; and the model results are clear and easy to interpret for business units [8].

More specifically, the Logistic regression model is distinguished as two kinds. Binary classification and multiple classifications. In general the Binary classification involves less steps since multiple classification is basically repeating binary classification over every group added. Out of different ways of multiple classification, “one vs rest” is commonly used as it is efficient and easy for future classification. On the other hand, “Many vs Many” is also an efficient way of multiple

classification but it involves selection from the different sets of data sets. In the end, of all the methods of “Many vs Many”, “one vs one” is used most commonly as it could be used in both binary classification and multiple classification. When one is thriving for speed, Ovr is the correct choice. But if one wants a model with high precision MvM will do it with at a slow rate [6].

Regularization of Logistic Function is also an important factor of ensuring a promising result. Regularizing a Logistic Function usually means lowering the precision factor. This is because a model with too much precision will result in overfitting which only ensures the model will function on the data it practiced, not with any new data [6]. For example if the model looked at one brand of shoe for the training too much and focused on specific details, the model will fail to detect other brands of shoes. Moreover, in order to adjust the complexity of the model to also avoid overfitting, one need to also adjust the loss function or cost function. The purpose of controlling the model complexity is to allow our model to have a better generalization ability without getting a completely inconsistent error performance with the training set in the actual test set. Thus, this problem becomes the frequently mentioned problem of preventing model overfitting in machine learning [9].

$$\log\left(\frac{p(X)}{1-p(X)}\right)=\beta_0+\beta_1 X$$

$$-(\mathrm{y}\log(p)+(1-y)\log(1-p))$$

3.2 BP neural network

BP neural network is a multi-layer feed-forward neural network which is mainly differentiated for the fact that the signal is forward propagated while the error is backward propagated. The process of BP neural network is mainly divided into two stages, the first stage is the forward propagation of the signal, from the input layer through the hidden layer, and finally to the output layer; the second stage is the backward propagation of the error, from the output layer to the hidden layer, and finally to the input layer [7].

The transfer function used in BP networks is a nonlinear transformation function called Sigmoid function. Sigmoid is praised for both the function itself and its derivatives are continuous which made further processing the graph relatively easy and efficient. In addition, Sigmoid function is also categorized into two different types: unipolar S-functions and bipolar S-functions. When looking at the graph of Sigmoid function, one could notice that when when the derivative is positive and increasing between net 0 to -5, f(x) is increasing at an increasingly speed. To train the neural network, we should try to keep the value of net in a range where convergence is relatively fast. This two functions work collaboratively to make it clear when and how the function went wrong when the result is not ideal [7].

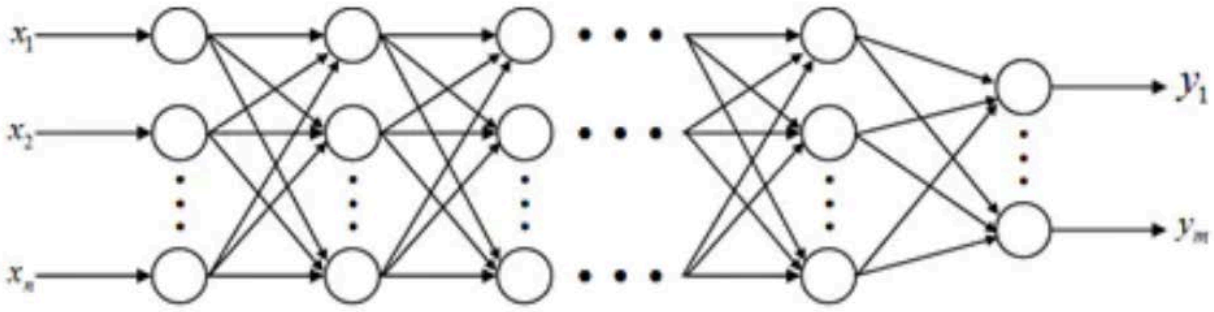

Figure 1: A showcase graph of how the Bp network works

4. Results

According to the Figure X shown below, both the accuracy and precision rate of the L1 logistic regression model has the most promising result. The results are collected by processing the true positive, true negative, false positive, and false negative. True and False meant if the result is

predicted to be true or false, positive and negative meant if the actual value is true or false. During the process of distributing the data, 20% of the data was randomly collected and set for testing while the models used the following 80% for practicing. Accuracy, precision, recall, f1_score, and roc are all calculated upon this data set. Looking from a larger picture L1 logistic regression model is also the most promising model considering all the feature since its f1 score is the highest. F1 score calculation involves both precision and recall which meant even though the recall for L1 logistic regression model is not the best, the overall score is not overweighted.

Table 2 A comparison chart of the performance of the different models

model	accuracy	precision	recall	f1_score	roc
L1	0.904903	0.930879	0.95935	0.9449	0.77792
L2	0.892728	0.908137	0.972125	0.939043	0.707554
Bp	0.849951	0.849951	1	0.91889	0.5

5. Conclusions

In an overall statement, this paper focused on the process of predicting service for credit card customers. Technology had been in substitution of Human predictions in this area ever since the “Big Data” Era begun.Not only are human mainly biased and the human brain also could not process large amount of data as efficient as machines. Though machine learning is a relatively new area of study, it had proven itself to be useful at classification, prediction and all sorts of data managing. In additional, different models had shown to be better at different areas which enable them to learn various different kinds of tasks using databases that could “enflame” a human brain. In this paper, the two models compared are the logistic regression and the Bp neural network. The result clearly fortifies my statement that the logistic regression is the best model for predicting credit card services. However, previous research has shown that Bp and other neural network models tend to work better than logistic regressions which meant the parameter for the Bp neural network might not be adjusted to its best function. Future improvements for this model will include adjusting the parameters and comparing the accuracy and precision of the three models again. If the result still shows L1 logistic regression model has the best result, the statement that logistic regression is the best suited model will be left with no questioning. If the result shows otherwise, it will prove that the conclusion here is malfunctioned due to its poorly chosen parameters.

Acknowledgement

I am especially grateful to the Kaggle website for presenting the open database for the paper’s usage.

Meanwhile, I would also like to thank all the teachers of CMU research team for their help in the whole process of completing my thesis and finally publishing it.

References

[1] Mitchell, T. M. . Machine Learning. McGraw-Hill, 2003.

[2] van der Aalst W. Data Science in Action. In: Process Mining. Springer, Berlin, Heidelberg, 2016.

[3] Jason.steele. “The History of Credit Cards.” Experian, 21 Jan. 2020.

[4] Murray, Jean. “4 C's of Credit Needed to Get a Business Loan.” The Balance Small Business, Raveh Gill-More. “Bank Churners.” Kaggle, 18 July 2021.

[5] A Colloquial Introduction to Logistic Regression Algorithm (2) Sklearn Practical. ” Jianshu(*Fly Title)*, 13 November 2019.

[6] Yefeng Qiu. “Deep Understanding of BP Neural Network.” Jianshu (*Fly Title)*, 1 June 2018.

[7] Xiaoming Lin & Ye Chen. “Artificial intelligence stock selection framework and classical algorithm introduction.” [Huatai Metalworking] Artificial intelligence 1:, 1 June 2017.

[8] Koliverpool. “The Essential Purpose of The Supervised Learning Process-(Loss Function).” Jianshu (*Fly Title)*, 11 Jan. 2017.

[9] L. Li et al.,Classification of heart sound signals with BP neural network and logistic regression, 2017 Chinese Automation Congress (CAC), 2017.