Rattanakosin Journal of Science and Technology: RJST
2023, Volume 5 Issue 2: 1-6
Research Article
Online ISSN: 2673-0553

Using Data Mining Techniques to Develop Models for Credit Card Payment Amounts

Sarinna Maplook[1*], Neunghatai Chaiarporn[2], Siriporn Samutwachirawong[3]
[1,2,3] Program in Statistics and Information Management, Faculty of Science, Maejo University, Chiang Mai 50200, Thailand

[1]E-mail: sarinna@mju.ac.th

Abstract

The goals of this study were to create a credit card payment amount model utilizing data mining techniques by using the following four methods: regression analysis, artificial neural networks, support vector machines for regression, and maroon peak (Model Tree: M5P). The amount of credit card payments from March 2015 to August 2022 served as the study's source of data. The experimental findings indicate that, at 7.80% of MMRE, the maroon peak (Model Tree: M5P) was the model with the highest performance for prediction.

Keywords: data mining, Artificial Neural Network, Support Vector Machine for Regression, M5P

1. Introduction

Using a credit card is one of the conveniences of nowadays because the main goal of credit cards is not to carry large amounts of cash to purchase goods and services. Simply holding a credit card enables payment for goods and services, which is officially known as a cashless society that utilizes technology to make transactions more convenient. As a result, cash becomes less important, and credit cards, resembling mobile cash, are preferred by most consumers. They also serve as emergency money in times of need. In the situation of COVID-19, which affects many aspects, especially the economy, the value of credit card spending is observed to be higher. Cardholders use credit card funds for revolving expenses instead of paying with cash immediately. This information is derived from the collection of the Bank of Thailand.

Data mining involves searching for useful information from numerous sources to extract insights for analysis, identifying patterns or relationships within databases, and preparing the information for use in business management planning. This research proposes the application of data mining techniques, specifically Regression Analysis, Artificial Neural Network, Support Vector Machine for Regression, and Maroon Peak data mining approaches (Model Tree: M5P), to analyze domestic and international credit card purchases. The Bank of Thailand will collect data between March 2015 and August 2022, which serves as the basis for this research. By utilizing data mining techniques, the bank will be able to analyze data quickly and effectively, enabling them to forecast credit card spending on a monthly basis.

[*] Corresponding author, e-mail: sarinna@mju.ac.th

Submission date: 18 09 2022 / Revised date: 26 06 2023 / Published date: 28 08 2023

2.Methodology

2.1 Objectives

1. to evaluate the effectiveness of various methods and select the most effective one for calculating credit card spending.

2. to use data mining techniques to create a forecasting model for credit card usage.

2.2 Research methodology

Making forecasts about the traits or trends of future interests is known as forecasting. To be used as knowledge while making decisions. Before planning, forecasting must be completed as the first task for making decisions with accuracy and precision.

Searching is referred to as data mining. Information or knowledge stored in a sizable, complicated database that can be used to inform decisions. Data preparation is the first step in the data mining process. Data format conversion, data mining and data analysis.

2.2.1 Collect credit card spending data from March 2015 to August 2022 from the Bank of Thailand.

2.2.2 Prepare the information by sorting the data based on the monthly period and relevant factors. The data is organized in Excel.

2.2.3 Perform modeling using linear regression (LR), Artificial Neural Network (ANN), Support Vector Machine for Regression (SVR), and Maroon Peak (Model Tree: M5P).

The regression analysis technique, known as Linear Regression, is used to analyze regression. Regression analysis is divided into two types: Linear Regression and Non-linear Regression. Linear Regression analysis is further divided into Simple Linear Regression (SLR) and Multiple Linear Regression (MLR). Simple regression analysis (MLR) is a statistical method used to study the linear relationship between two variables, in this case, X and Y variables. It is employed to examine the nature of the relationship between two or more variables, which are classified as independent variables and dependent variables.

Mathematical models are discussed in relation to artificial neural networks. Tools that are capable of Pattern Recognition and Knowledge Extraction, as well as capabilities within the human brain, can be made by simulating the operation of the neural network in the human brain using Connectionist. The original idea behind this technique came from research on the "Neurons" and "Synapses" that make up the brain's bioelectric network. Each neuron has an input and a transmission terminal on its dendrite, which is a type of nerve cell. "Axon" is the name given to the nerve impulses, which are similar to a cell's output. Electrochemical processes power these cells. Neurons will pass through the dendrite and into the cell body when they are activated by external stimuli or by other cells. Electrochemical reactions are how these cells function. The ability of neurons to pass through dendrites and become triggered by external stimuli or by other neurons will determine whether further cells need to be stimulated. A neuron's axon will continue to excite other cells if the nerve fibers are robust enough.

A tree model used to forecast numerical data is referred to as the M5P tree model technique. In the process of choosing the nodes of each tree layer, which was formed from the decision tree, but will be distinct. The output values are numerical, but the input values can be both continuous and non-continuous. The data will be evaluated from the root node down to the leaf node, which will receive the features of the route selection from each layer of node. Additionally, at the leaf nodes of the tree, there is a linear model that predicts the groups of Numeric value data.

A separating hyperplane serves as the formal definition of a Support Vector Machine (SVM), a discriminative classifier. In other words, the method generates an optimum hyperplane that classifies fresh samples given labeled training data (supervised learning). This hyperplane is a line that divides a plane into two halves in two-dimensional space, with each class located on each side.

Regression analysis is a popular technique used to create linearly correlated predictive models. The ANN model and SVM model are widely utilized in research due to their high accuracy, especially for non-linearly correlated data. Additionally, the SVM method is known for its fast processing capabilities and suitability for small data sets.

2.2.4 Measure the efficiency using the MAE and RMSE values for all four techniques.

2.3 Data analysis

In order to evaluate data and create forecasting models for credit card spending volume both domestically and abroad, the researcher employs the WEKA program. By using data mining techniques to create data sets for predicting models using data from March 2015 to August 2021, with the end goal of testing the performance of the data set between September 2021 and August 2022.

Table 1 Information on credit card usage from March 2015 to August 2022

Months	The amount of credit cards used for payment								
	Training Data set							Testing Data set	
	2015	2016	2017	2018	2019	2020	2021	2021	2022
January		101	106	107	115	120	131		139
February		90	96	101	114	114	119		126
March	90	101	109	116	121	123	132		136
April	87	97	104	104	113	114	125		135
May	87	97	101	105	115	117	125		130
June	84	93	101	109	114	116	125		129
July	90	99	103	109	112	117	129		138
August	88	99	104	109	111	118	129		134
September	83	97	104	103	107	116		123	
October	97	101	111	110	114	119		112	
November	99	105	113	118	115	130		112	
December	120	125	145	144	144	151		130	

The researcher considers the RMSE and MAE values as indicators for prediction models that are suitable for each month of forecasting for the model performance testing, separating each month in the test data set. From the data, the researchers used all data to test, create models by using 4 data mining techniques to test the effectiveness of credit card spending predictions. To choose a useful model, the researchers considered the RMSE values.

Mean Absolute Error (MAE):

The MAE measures the average magnitude of errors in a set of forecasts, regardless of their direction. It assesses accuracy for continuous variables. The equation for MAE is provided in the library references. In simpler terms, the MAE is the average of the absolute differences between the forecasts and the corresponding observations in the verification sample. The MAE is a linear score, meaning that all individual differences are equally weighted in the average.

$$\text{MAE} = \frac{\left|y_i - y_p\right|}{n} \tag{1}$$

y_i = actual value
y_p = predicted value
n = number of observations

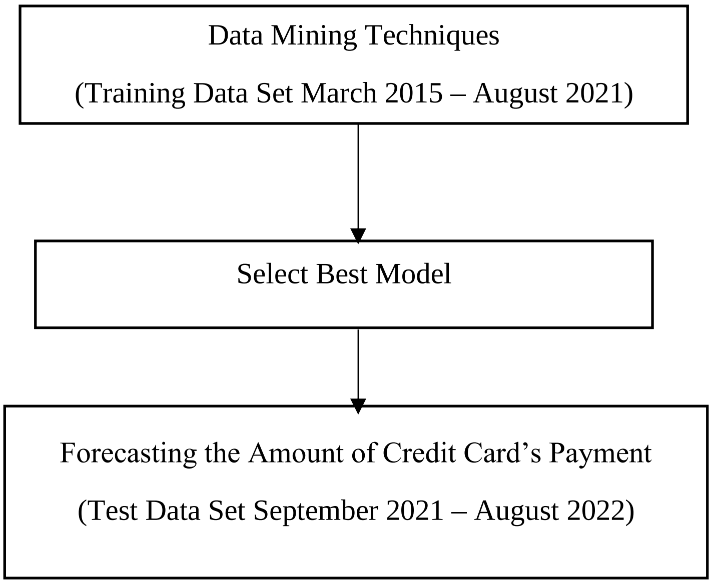
Root Mean Squared Error (RMSE):

The RMSE is a quadratic scoring rule that measures the average magnitude of errors. The equation for RMSE is given in both references. To explain the formula, each difference between the forecast and corresponding observed values is squared, then averaged over the sample. Finally, the square root of the average is taken. Since the errors are squared before averaging, the RMSE assigns relatively higher weight to large errors. This makes the RMSE most useful when large errors are particularly undesirable.

$$\text{RMSE} = \sqrt{\frac{\sum\left(y_i - y_p\right)^2}{n}} \tag{2}$$

y_i = actual value
y_p = predicted value
n = number of observations

The MAE and RMSE can be used together to assess the variation in errors in a set of forecasts. The RMSE will always be larger than or equal to the MAE; the greater the difference between them, the higher the variance in the individual errors in the sample.


Data Mining Techniques
(Training Data Set March 2015 – August 2021)
Select Best Model
Forecasting the Amount of Credit Card's Payment
(Test Data Set September 2021 – August 2022)

3. Results

3.1 Evaluation of the prediction model's effectiveness in relation to the data set

The researcher chooses the data set from which to extract the credit card spending data. Used to develop prediction models to compare the mistakes using MAE and RMSE methodologies to measure the effectiveness attained from the 4 data mining strategies. Table 2 contains the outcomes.

Table 2 Comparison of the effectiveness of predictive models utilizing past data

Lagged (Months)	Data Mining Techniques							
	Regression Analysis		Artificial Neural Network		Model Tree: M5P		SVMs	
	MAE	RMSE	MAE	RMSE	MAE	RMSE	MAE	RMSE
3	6.2156	8.2165	7.9851	10.3641	5.9821	9.2151	6.2546	9.8416
6	6.0406	8.2614	7.6691	10.1255	5.4626	9.1460	6.5161	9.3641
9	6.3051	8.1249	7.8116	9.5845	5.7723	9.1472	6.5423	9.5136
12	6.2534	6.2145	6.2513	9.9895	5.4231	8.2156	5.1265	8.2942
15	4.5332	6.3228	8.2661	8.6945	4.3256	7.2561	5.2156	8.5994
18	4.5289	6.1046	7.3115	7.6451	4.1246	4.2152	5.1692	6.5971
21	4.9561	6.4231	5.9210	6.9413	3.1568	3.9985	4.2566	7.3665
24	4.2215	5.9862	4.9518	6.6416	3.6421	3.7486	3.9711	7.0646

Table 2 shows that the 24-month historical data set is the most useful for all 4 data mining approaches when creating historical data.

3.2 A comparison of each month's prediction model effectiveness

Table 3. Comparison of the effectiveness of predictive models

Data Mining Techniques	RMSE
Regression Analysis	12.05%
Artificial Neural Network	19.62%
Model Tree: M5P	7.80%
SVMs	8.56%

Table 3 shows that the Model Tree Technique (M5P), which is the most effective method for forecasting credit card spending, has an RMSE of 7.80%.

4. Conclusion and Discussion

In this study, the effectiveness of several data mining techniques for creating models of credit card payment amounts was compared in order to select the best one. The amount of credit card purchases made by Thailand's bank domestically and abroad between March 2015 to August 2022 made up the data for this study. Training data and testing data were the two categories created from the data. The 24-month historical data collection was used to create historical data for all 4 data mining approaches, and it was discovered that the Model Tree Technique (M5P), which forecasts credit card expenditure, is the most successful method with an RMSE of 7.80%.

5. References

[1] Chitra, K., and B. Subashini. (2013). Data Mining Techniques and its Applications in Banking Sector. International Journal of Emerging Technology and Advanced Engineering. 3(8), 219-226.

[2] Chu, Wesley, and Tsau Young Lin. (2015). Foundations and advances in data mining. Springer Science & Business Media. 180,125-136.

[3] Desai, V. S., Crook, J. N., Overstreet, G. A. (1996). A comparison of neural networks and linear scoring models in the credit union environment. European Journal of Operational Research. 95, 24-37.

[4] Ghodselahi A. (2019). A Hybrid Support Vector Machine Ensemble Model for Credit Scoring. International Journal of Computer Applications. 17, 75-87.

[5] N. keadyam. (2017). Forecasting the amount of credit card's Payments Using Time series Data Mining Techniques. Siam University.

[6] Mustaffa Z Zainal NA. (2016). Developing a gold price predictive analysis using Grey Wolf Optimizer. IEEE Student Conference on Research and Development 2016.

[7] Rousseeuw, P. J. (1987). Silhouettes: A graphical aid to the interpretation and validation of cluster analysis. Journal of Computational and Applied Mathematics. 20, 53–65.

[8] R.A. Johnson and D.W Wichern. (2018). Applied Multivariate Statistical Analysis.

[9] Sven F. Crone and Rohit Dhawan. (2007). Forecasting Seasonal Time Series with Neural Networks: A Sensitivity Analysis of Architecture Parameters. IEEE Trans. Neural Networks (IJCNN), 2099-2104.

[10] Weiss, S. M., Kulikowski, C. A. (2018). Computer Systems That Learn: Classification and Prediction Methods from Statistics, Neural Nets, Machine Learning, and Expert Systems; 1991st San Francisco, Calif, USA Morgan Kaufmann

[11] Zurada, Jozef, and Martin Zurada. (2011). How Secure Are "Good Loans: Validating Loan-Granting Decisions and Predicting Default Rates on Consumer Loans. Review of Business Information Systems (RBIS). 6(3), 65-84.