Abstract No: SO-02

Machine learning model to predict bank customer's next expenditure with relevant merchant category

A. M. K. H. Umayanga[1*] and D. M. L. M. Dissanayake[1]

[1]Department of Statistics & Computer Science, University of Kelaniya, Sri Lanka
hirushani199@gmail.com*

The banking industry's increasing reliance on debit card transactions has generated a wealth of valuable data for understanding consumer behaviour. This study aims to develop a machine learning model to predict a customer's next expenditure and the corresponding merchant category using 50 customers' debit card transaction data for 11 years. Unlike existing research focused on bankrupt users and fraud detection, this study addresses the next expenditure prediction with merchant categories. For the bank, predicting a customer's next expenditure and merchant category enables targeted marketing efforts. The bank can send alert messages with discount offers specifically to each customer's spending habits, reducing marketing costs by only targeting relevant customers for relevant merchant types. Additionally, customers benefit from early reminders, allowing them to manage their finances effectively. For instance, a customer can receive a reminder about an upcoming insurance payment and allocate funds, accordingly, avoiding unnecessary expenses. This proactive approach can help reduce the number of bankrupt customers and long-term customer relationships. Challenges in this study include obtaining a dataset that is not readily available on the internet. The dataset was provided by the Digital Banking Department at the Head Office of the People's Bank while ensuring data privacy. Data preprocessing involved removing null values and unnecessary columns and creating customer IDs instead of account numbers. Then, identified 36 customers who consistently used debit cards and categorised merchant names into 11 groups. The dataset was split into training and testing sets using a specific date. Three machine learning algorithms, gradient boosting regressor, random forest regressor, and random forest classifier, were employed. Gradient boosting regressor is used to predict expenditures and merchant categories after encoding the categories using one-hot encoding. Random forest regressor is for expenditure prediction, and random forest classifier is used for merchant category prediction. Ordinal encoding was used to convert categories into numerical values. Model performance was optimised through hyperparameter (learning rate, number of trees, maximum depth of each decision tree, minimum number of samples required to split an internal node, minimum number of samples required to be at a leaf node, and fixed random seed for reproducibility) tuning using grid search, evaluating various combinations of hyperparameters through cross-validation. Models run through each customer's unique dataset since expanding patterns are different from each other. The results showed that the random forest regressor and random forest classifier-based method achieved higher accuracy compared to the gradient boosting regressor. This was evident from R^2 scores (0.9866 and 1.0605) and mean squared error values (MSEs are 313165.9622 and 5.6257). However, the method yielded R^2 scores exceeding 1 and a high MSE value due to an unbalanced dataset, where customers' debit card usage frequency varied. Obtaining a balanced dataset with an equal number of transactions for each customer is challenging, especially when requesting data from a bank. In the future, this study could be extended to predict the exact time and date of transactions using techniques like long short-term memory (LSTM) with a larger dataset like 1000 customers.

Keywords: Expenditure prediction, Gradient Boosting Regressor, Personalized Financial planning, Random Forest Classifier, Random Forest Regressor