A·P·U
ASIA PACIFIC UNIVERSITY
OF TECHNOLOGY & INNOVATION

INDIVIDUAL ASSIGNMENT

TECHNOLOGY PARK MALAYSIA

CT045-3-M-ABAV

ADVANCED BUSINESS ANALYTICS AND VISUALIZATION

APDMF2112DSBA(DE)(PR)

JUNE 2022

TITLE: IMPROVING BANKING CUSTOMER RETENTION BY CUSTOMER CHURN PREDICTION

LEE KEAN LIM
TP065778

LECTURER: DR. PREETHI SUBRAMANIAN

TABLE OF CONTENTS

LIST OF ABBREVIATIONS

ANN	Artificial Neural Networks
GB	Gradient Boosting
LR	Logistic Regression

SECTION 1

INTRODUCTION

The rise of fintech and adoption of newer technologies has introduced greater competition to the traditional banks. According to de Lima Lemos et al. (2022), two thirds of existing banking customers are intending to or has already shifted to the use of fintech as compared to using traditional banks. In addition, bank managers and executives foresee the rise of fintech as a threat to the traditional banking. Therefore, a new challenge exists for the banks to retain the existing customers.

Improving the customer relationship management thus becomes a critical task for the banks to reduce customer churn rate. Which according to Tang et al. (2020), retaining existing customers cost much less than trying to attract new customers in using the banking services. Various studies adopted the machine learning approach in predicting the customer churn rate to provide targeted campaigns or services in order to retain the customers (Broby, 2022; Wang et al., 2020).

SECTION 2

PROBLEM STATEMENT

Predicting customer churn rate in the banking sector carries a few challenges. Generally, banks are working with millions of customers, which utilizing methods of human interventions in predicting the churn rate is not feasible. In addition, customers may have a change in requests at any time which is to be adapted by the banking executives in a very quick manner. However, each banking executive may be handling a large customer pool at any given time which may introduce error in the process of amending the requests by customers. Therefore, there is a need to automate the process of customer churn prediction which allows the detection of early signs of potential customer churn. The use of machine learning would facilitate the customer churn prediction, providing a high predictive accuracy and quick results. Which, targeted strategies can be adopted to such customers to improve the retention rate.

SECTION 3

AIM & OBJECTIVES

3.1 AIM

The aim of this study is to predict bank customer churn by utilizing different machine learning algorithms to improve the customer retention rate of a bank.

3.2 OBJECTIVES

The objectives of this study are as followed:

1. To develop predictive models using different machine learning algorithms for predicting bank customer churn.
2. To evaluate performance of different predictive models in predicting bank customer churn.
3. To identify causes leading to the likelihood of customer churn.

SECTION 4

SCOPE OF STUDY

1. The study is conducted based on a fictitious banking dataset. The dataset provided comprises of customer data and a binary target variable indicating whether customer is churned or retained. There are 10,000 observations and 14 features available from the dataset. The dataset is retrieved from Kaggle and can be accessed from the following link: https://www.kaggle.com/datasets/mathchi/churn-for-bank-customers
2. The data mining process will be performed using SAS enterprise Miner application. The application provides a drag and drop graphical user interface in the form of a process flow diagram that can be easily navigated when performing the data mining process. The visual view of the data mining process facilitates high interpretability and faster model development time. In addition, various modeling techniques are supported by the application which allow variation and comparison of performance from using different models.
3. The causes of customer churn will be determined based on the information derived only from the descriptive and predictive analysis of the dataset.
4. The selection of predictive algorithms will be limited to what is provided by the SAS Enterprise Miner application.

SECTION 5

METHODOLOGY

This study adopts the SEMMA methodology as a guide to the process of data mining. The methodology comprises of five stages namely **S**ample, **E**xplore, **M**odify, **M**odel, and **A**ssess. The following shows the description of tasks involved in each stage of the SEMMA methodology applicable in this study.

Stage 1: Sample

This stage involves splitting the dataset into two groups, namely the training data and the testing data. The dataset would be partitioned, 80% as the training data and 20% as the testing data. This is performed as an evaluation step for the fitted model. The target variable from the testing dataset will be dropped to allow the fitted model to predict the outcome and evaluate the performance of the prediction.

Stage 2: Explore

Exploration of the data can be performed through exploratory data analysis with the aid of visualization and statistics. Through data exploratory analysis, trends and anomalies in the dataset can be observed which provides a better understanding of the dataset. Visualization methods such as bar chart, scatter plot, box and whisker plot, etc., can be utilized to understand the relationship between variables and identifying anomalies. The use of visualization provides a graphical means of interpreting the dataset. In addition, complementing the visualization, performing descriptive statistics such as measuring central tendency and identifying frequency distribution would provide further information regarding the dataset.

Stage 3: Modify

The dataset contains a mix of categorical and numerical variables. Encoding of the categorical variables are expected due to many algorithms are not able to work directly with character labels (Tang et al., 2020). Example, label encoding can be performed for the “Gender” feature to convert the character labels into numeric format. Feature scaling is to be performed for numerical variables due to the different ranges of values of each feature. This is to ensure all features are weighed equally by the model thus minimizing the bias.

Several features can be removed from the dataset as it does not provide any impact to the predictability of the model. Such features identified from the dataset are “RowNumber”, “CustomerId”, and “Surname”.

Missing values are not identified from the dataset. However, outliers are present and should be identified and processed.

Stage 4: Model

Multiple prediction models will be developed using different machine learning techniques, namely Logistic Regression (LR), Artificial Neural Networks (ANN), and Gradient Boosting (GB). The LR model is chosen due to its quick computation and provides high explainability. This allows the identifying of features with high impact. In contrast, ANN and GB model does not provide high explainability. However, expectation that ANN and GB model would yield higher prediction accuracy due to their better fitting ability. Hyperparameter tuning will be performed on the models to attain better prediction results.

Stage 5: Assess

The models will be evaluated by the testing dataset to determine the fitness of model. An iterative process of evaluation and hyperparameter tuning will be performed to achieve a good fitted model for each algorithm. The accuracy, precision, recall, F1-score, and confusion matrix will be used as the evaluation metrics for the prediction models. Multiple evaluation metrics are used to provide a deeper evaluation of the models, due to the imbalanced data classes typically observed in churn dataset (Tang et al., 2020). The model attaining the highest accuracy will be deemed as the best model for customer churn prediction.

REFERENCES

Broby, D. (2022). The use of predictive analytics in finance. *The Journal of Finance and Data Science, 8*, 145-161. doi:https://doi.org/10.1016/j.jfds.2022.05.003

de Lima Lemos, R. A., Silva, T. C., & Tabak, B. M. (2022). Propension to customer churn in a financial institution: a machine learning approach. *Neural Computing and Applications*. doi:10.1007/s00521-022-07067-x

Tang, Q., Xia, G., Zhang, X., & Li, Y. (2020). *A Feature Interaction Network for Customer Churn Prediction*. Paper presented at the Proceedings of the 2020 12th International Conference on Machine Learning and Computing, Shenzhen, China. https://doi-org.ezproxy.apu.edu.my/10.1145/3383972.3384046

Wang, X., Nguyen, K., & Nguyen, B. P. (2020). *Churn Prediction using Ensemble Learning*. Paper presented at the Proceedings of the 4th International Conference on Machine Learning and Soft Computing, Haiphong City, Vietnam. https://doi-org.ezproxy.apu.edu.my/10.1145/3380688.3380710