Machine models for Predicting Credit Card Holder's Retention

Dr. Maria Rona L. Perez
FEU – Institute of Technology
Manila, Philippines
ronaloboperez@gmail.com

Dr. Ace C. Lagman
FEU – Institute of Technology
Manila, Philippines
aclagman@feutech.edu.ph

John Benedict C. Legaspi
FEU – Institute of Technology
Manila, Philippines
jclegaspi@feutech.edu.ph

Dr. Kirk Alvin S. Awat
FEU – Institute of Technology
Manila, Philippines
ksawat@feutech.edu.ph

Dr. Roman M. De Angel
FEU – Institute of Technology
Manila, Philippines
rmdeangel@feutech.edu.ph

Dr. Renato Maaliw III
Southern Luzon State University
Quezon, Philippines
maaliw@slsu.edu.ph

***Abstract*—Customer retention is not only a cost effective and profitable strategy, but in today's business world it is necessary specially in the credit card industry. Through the influence by multiple published researches about machine learning and its prediction models, this paper develops a predictive model that can proedict credit card holder retentin. The data from this research comes from the datasets of Google dataset search where it offers millions of publicly available data to use by researchers or scientist. These includes historical transactions, new merchant transactions, train and test files that contain the card IDs that was used for training and prediction. Thus, allowing the researcher to create loyalty score for each card ID for predicting retentions. The predictive models were validated through accuracy, precision and recall. While in predicting the customer's loyalty score based on the customer's transactions, the researcher used the overall performance metric that is root-mean squared error (RMSE). As for conclusions, loyalty scores can be used by credit card industry as a metric in predicting customer's retention.**

Keywords—machine learning, predictive model, customer retention, credit card holder

I. Introduction

Over the past two decades Machine Learning has become one of the mainstays of information technology and with that, a rather central, albeit usually hidden, part of our life [1]. With the ever-increasing amounts of data becoming available there is good reason to believe that smart data analysis will become even more pervasive as a necessary ingredient for technological progress.

Improving customer loyalty has become a popular area for managers, consultants, and academics [2]. The influences in support of loyalty are simple to apprehend. Loyal customers are reported to recommend others to become customers of that merchant, commit a higher share of their category spending to a merchant, and have more likely higher customer retention rates [3]. The credit card industry has maximized customer engagement processes to develop a deep understanding of their customer's life cycle in terms of their interests, passions, associations, and affiliations [4]. They need to integrate new features from customer insights to drive business probability.

Thus far, however, there have been no peer-reviewed, scientific investigations examining the relationship between customer loyalty and retention. To identify whether the scheme make customer experience more enjoyable and does bring repeated purchases to the merchant this research seeks to examine the cardholders' transactions to commonly used participating merchants from the datasets, including loyalty aspects, and their relationship to future customer preferences: purchasing and retentions.

The rest part of this paper is structured as follows. Section 2 gives the background of the study discussion an overview of customer loyalty and retention. Section 3 gives an existing research in the literature. Section 4 introduction of the major techniques employed. Experimental results are presented in Section 5. This paper is concluded in Section 6 with a list of recommendations.

II. Background of the Study

Credit card companies have its quest in competition on their very own industry. As billions of people around the world are credit card user, credit card providers reach out to their clients to motivate them to use their cards through vouchers, no annual fees and conversions to possibly increase retentions.

The major challenges of credit card industry are interpreting the huge dataset correctly which contains business process, the used of featured engineering on nested data and mitigating the effect of outliers [6]. The datasets are largely anonymized and the meanings of features are not elaborated.

In this paper, the researcher develops a predictive model for card holder's retention. This predictive model will provide high accuracy for identifying the most important aspects and preferences in a customer lifecycles. Superior customer insights will lead to expand profiling, segmentation, targeting, acquisition, maturation, recommendation and retention based on customer behavior and preferences.

The data for this study comes from datasets in the Google Datasets Search [7]. These includes historical transactions, new merchant transactions, train and test files that contain the card IDs were used for training and prediction. Allowing the researcher to create loyalty score for each card ID for predicting.

The major problem of transforming card holder's transactions which was also the major reason why this paper has been crafted. The need to find a way to aggregate or transform a card's purchasing history into a single representative vector in order to model properly is very challenging. Since it is difficult that both numeric and categorical features are present, the number of transactions varies from card to card, these is a inter temporal dependence between transactions and the non-unique and unevenly-spaced timestamps which standard time series cannot be applied.

Also, these companies are having difficulties in identifying the right metric to be used to model the collected data. There is a need of knowing whether their motivation schemes make creates a customer experience and drives a deciding factor to retain their cards.

The following questions will be answered in this research:

1. What are the preliminary data or main factors needed that may affects card holder's retention?
2. How will the Predictive Model be evaluated to obtain the results for card holder's retention in terms of:
    a. Accuracy
    b. Efficiency

How accurate and precise is the pre-defined prediction model in performing prediction?

III. Related Literature

This section presents related literature which comprise of machine language, data pre- processing techniques, predictive models and algorithms, topics on customer retentions and lifecycle that were associated with the development of predictive model for cardholder's retention.

A. Machine Learning

As mentioned in the study of [8], machine learning techniques has been widely applied in the banking and finance sector. In machine learning, classification is a supervised learning approach in which the classifier learns from the data input given to it and then uses this learning to classify new observation. In a paper written by [9] demand forecasting in restaurants using machine learning is proposed. Many researchers have been on demand forecasting technology using POS data. However, in order to make demand forecasting at a real store, it is necessary to establish a store-specific demand forecasting model in consideration of various factors such as the store location, the weather, events, etc. As stated by [10], machine Learning can be roughly categorized into three classes supervised learning, unsupervised learning and reinforcement learning. The labeled data are used to train in supervised learning, which behaves like a "teacher". On the contrary, unsupervised learning has to find the structure all by itself, which means no labels are given. In reinforcement learning, agent interacts with the environment, learns online and tries to maximize accumulated reward. All those three classes can go on extending into more specified algorithms.

B. Data Pre-processing Techniques

In the paper written by [11], stated that data mining algorithm may perform differently on datasets with different characteristics, e.g., it might perform better on a dataset with continuous attributes rather than with categorical attributes, or the other way around. The research of [12] stated that imbalanced domains are an important problem frequently arising in real world predictive analytics. A significant body of research has addressed imbalanced distributions in classification tasks, where the target variable is nominal. In the context of regression tasks, where the target variable is continuous, imbalanced distributions of the target variable also raise several challenges to learning algorithms. Imbalanced domains are characterized by: (1) a higher relevance being assigned to the performance on a subset of the target variable values; and (2) these most relevant values being underrepresented on the available data set.

C. Predictive Models

In the paper of [13], they develop advanced analytics tools that predict future customer behavior in the non-contractual setting. They establish a dynamic and data driven framework for predicting whether a customer is going to make purchase at the company within a certain time frame in the near future. In the article of [14], churn prediction plays a central role in churn management programs for customer- centric firms. Methodological advances have emphasized the use of customer panel data to model the dynamic evolution of a customer base to improve churn predictions. However, pressure from policy makers and the public geared to reducing the storage of customer data has led to firms' 'self-policing' by limiting data storage, rendering panel data methods infeasible.

D. Customer Retention

Luarn and Lin [15] use a theoretical model to explain how trust, customer satisfaction, perceived value and commitment influence attitudinal and behavioral loyalty for e-service context, and validate it empirically. Propose a structural equation model of customer loyalty in the retail banking market

to develop and test a model that aids further understanding of the determinants of customer loyalty in the wireless telecommunication industry. In the paper of [16] illustrates that a retail store' s customer behavioral loyalty can be predicted to a reasonable degree using the transactional database.

IV. Methodology and Techniques

This section presents details the research design and methodology used for collecting and analyzing on factors influencing cardholder's retention.

Experimental research design is to enable researcher to estimate the effect of an experimental treatment [17]. Experimental research can be done in the laboratory, in the class and in the field. In this study, the experimental research is done in the class with taking students as population. A researcher chooses the design to determine the validity of conclusions can be drawn from the study.

A. Data Science Workflow of the Study

The best solution can be provided to the problem using the general framework and adapt it to new problem.

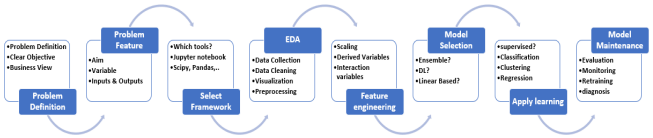

Figure 1: General Workflow of the Study

The figure above illustrates the different phases of the study. Credit card companies is one of the leading industries today. This study is predicting loyalty score for each card_id represented in test.csv and sample_submission.csv. In problem, develop an algorithm to identify and serve the most relevant opportunities to individuals by uncovering signal in customer loyalty. After problem definition and problem feature, the researcher will select a framework to solve the problem. This is the programming language to use and by what modules the problem will be solve? For the Exploratory Data Analysis, the researcher will follow the analytical and statistical operations: (1) Data Collection; (2) Visualization; (Data Cleaning); and (4) Data Preprocessing. Find the type of features dataset in the given dataset. It is the most important part in machine learning. Then, identify the predicting model to use, apply and evaluate.

B. Evaluation of the Predictive Models

To efficiently test the the algorithms, confusion matrix test was used.

The confusion matrix is a useful tool for analyzing how well your classifier can recognize tuples of different classes. It gives an overview of the numbers of True Positive (TP), True Negative (TN), False Positive (FP), and False Negative (FN) where the evaluation of correct TP and TN and incorrect FP and FN can be assessed.

This stage is divided into two aspects that include training and testing data. The extracted model derived from the training data was then tested using separate data sets called test data sets. Table 1 below indicates the classification rate e to determine the accuracy of the model.

TABLE I. Classification Table

	Performance Measure of the Algorithm		
		Yes	***No***
Predicted	Yes	True Positive	False Positive
	No	False Negative	True Negative

The true positive (TP) and true negatives (TN) are correct classifications. The accuracy can be computed by adding the correct predictions of true positive and true negative divided by the total number of items.

$$Accuracy = \frac{TN + TP}{TP + FP + TN + FN} \quad (10)$$

1. Accuracy – The accuracy of a measurement system is the degree of closeness of measurements of a quantity to that quantity's actual (true) value. It is the percentage of sum of all true positives and false negatives out of the sum of all the true positives, true negatives, false positives, and false negatives.

$$accuracy = \frac{(tp+tn)}{tp+tn+fp+fn} \quad (1)$$

2. Precision – The precision of a measurement system, also called as reproducibility or repeatability, is the degree to which the repeated measurements under unchanged conditions show the same results. It is formulated as the fraction of the number of true positives to the sum of true positives and false positives.

$$precision = \frac{tp}{tp+fp} \quad (2)$$

3. Recall – The recall measurement in this context is also referred to as the true positive rate or sensitivity. It is the ratio of true positives over the sum of true positives and false negatives, or the percentage of flows in an application class that are correctly identified. Equation 1. Recall

$$recall = \frac{tp}{tp+fn} \quad (3)$$

C. Statistical Analysis

This research used quantitative data analysis techniques through the following statistical methods:

The Likert Scale will be used to interpret the results from the computation of root-mean squared error. The results were arbitrarily scaled as follows:

TABLE II. LIKERT SCALE

Range	Verbal Interpretation
4.51 – 5.00	Outstanding
3.51 – 4.50	Very Satisfactory
2.51 – 3.50	Satisfactory
1.51 – 2.50	Unsatisfactory
1.00 – 1.50	Poor

The t-test for correlated between the average scores of a single sample of individuals who are assessed at two different times [18].

The formula for finding the t statistic is:

$$t_D = \frac{\overline{X}_D}{\left(\frac{s_{\overline{D}}}{\sqrt{n}}\right)} \quad (4)$$

Where:

$\overline{X}_D$ = mean difference

$s_{\overline{D}}$ = standard deviation

n = number of respondents

To predict any given customer's loyalty score based on the customer's past purchasing behavior, the researcher will used the overall performance metric that is root-mean squared error (RMSE).

$$\text{RMSE} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}, \quad (5)$$

Where:

ŷ – the predicted loyalty score for each card_id

y – actual loyalty score assigned to a card_id.

Using Python as programming language to build the prototype, it has a built-in functions that is used to find the minimum value and the maximum value of the given array of datasets

To find the minimum value : min(big_array);

To find the maximum value : max(big_array);

To perform logistic regression in Python, the formula below was used to predict the loyalty scores using the given datasets. Using the Sigmoid function (shown below), the standard linear formula is transformed to the logistic regression formula (also shown below). This logistic regression function will be useful for predicting the class of a binomial target feature.

$$p = \frac{1}{1 + e^{-y}} \quad (6)$$

V. EXPERIMENTAL RESULTS

This section is composed of information gathered that provides answers to the questions raised in the previous section.

The following are the preliminary data used to generate results that recommends card holder's loyalty behavior:

TABLE III. DATASETS TO DETERMINE CARD HOLDER'S RETENTION

File	Description	Purpose
train.csv	This is the training set	It contains the card_ids that was used for training and prediction.
test.csv	This is the test set	It contains the card_ids that was used for training and prediction.
historical_transactions.csv	Up to three months' worth of historical transactions for each card_id (card identifier)	It contains information about each card's transactions.
merchant.csv	Additional information about all merchants or mechant_id (unique merchant identifier) in the dataset.	It contains aggregate information for each merchant_idrepresented in the data set.
new_merchant_transactions.csv	Two months' worth of data for each card_id containing ALL purchases that card_id made at merchant_ids that were not visited in the historical data	It contains the transactions at *new* merchants (merchant_ids that this particular card_id has not yet visited) over a period of two months.

Table 3 shows the data to be collected to standardized and established manner that enables the researcher to answer or problem stated and evaluate outcomes of the datasets.

In the field of statistics and for classification purposes, four key terms including true positives (tp), true negatives (tn), false positives (fp), and false negatives (fn) are usually computed for comparing the results and assessing the performance of the classifier utilized. The terms positive and negative refer to the classifier's prediction, also known as the expectation, and the terms true and false refer to whether that prediction corresponds to the external judgment, also known as the observation. These terms and their associations are illustrated in Table 3 for classification of test cases.

Accordingly, three major measurement metrics: accuracy, precision, and recall are usually used to assess how well a binary classification is performed.

TABLE IV. CLASSIFICATION OF TEST CASES

Test Cases	Effective Test Case			Non- Effective Test Case		
	Accuracy	Precision	Recall	Accuracy	Precision	Recall
Negative Loyalty Score						
Test Case 1	69.95	69.18	100	63.09	63.06	100
Test Case 2	68.64	65.43	100	65.95	61.67	100
Positive Loyalty Score						
Test Case 1	69.93	68.23	100	67.24	66.40	100
Test Case 2	72.26	70.01	100	69.57	64.47	100
Average	70.21%	68.21%	100%	66.46%	63.9%	100%

Table 4 reports the performance of the predictive model in predicting loyalty scores. For effective test cases, the overall accuracy (70.21%) and recall (100.00%) ratios are considerably good. The precision is measured as 68.21% and it is not so significant when compared to the other two metrics. The non-effective test case; similarly, the accuracy (66.9%) and the recall (100.00%) ratios are considerably good. The precision is measured as 63.9%. It is important to note that the failing test cases were always assigned into different clusters.

For the two predictive models, the recall ratio is 100% demonstrating that it can effectively identify almost all of the actual effective test cases. In other words, the value 100% indicates that all test cases, which can expose a fault, were classified into the effective category by the model. High accuracy indicates that for most cases, both the actual effective and actual non-effective test cases are classified properly and, thus, minimizing the cost of not running non-effective test cases.

Model Comparison Techniques

The confusion matrix table illustrates a tabular display that evaluates the forecasting precision of a predictive model.

The main objective of a predictive model is to maximize the correctly classified instances. For binary classification scenarios, the misclassification rate gives the overall model performance with respect to the exact number of categorizations in the training data.

To determine the accuracy level of the classification table of the algorithms the accuracy formula was used.

TABLE V. SUMMARY OF CLASSIFICATION ALGORITHMS ACCURACY RESULT

Classification Technique	Accuracy
Logistic Regression	87.8
Naïve Bayes	86.5
Decision Tree	88.2
KNN	87.9

The Predictive Model for Cardholder's Retention is explained below:

```
'learning_rate': 0.01,
'subsample': 0.9855232997390695,
'max_depth':
'top_rate': 0.9064148448434349,
'num_leaves': 63,
'min_child_weight': 41.9612869171337,
'other_rate': 0.0721768246018207,
'reg_alpha': 9.677537745007898,
'colsample_bytree': 0.5665320670155495,
'min_split_gain': 9.820197773625843,
'reg_lambda': 8.2532317400459,
'min_data_in_leaf': 21,
'verbose': -1,
'seed':int(2**fold_),
'bagging_seed':int(2**fold_),
'drop_seed':int(2**fold_)
```

Figure 2: Decision Tree (Generated Values)

These values were generated using a decision tree classifier which is a type of classification algorithm. Using Feature engineering and Bayesian ridge, the model was refined to define its probabilistic capability. This means that arriving at such values; the model can predict positive loyalty and negative loyalty.

It successfully classifies that the card holder's loyalt on the

merchant with 96% precision. This was performed using reducing memory consumption function. Then perform binarization of the category, and include it as a feature to train the model. As the researcher extracts the features, the processes follow and remove the outliers. The visualize model is displayed in Figure 3 and figure 4.

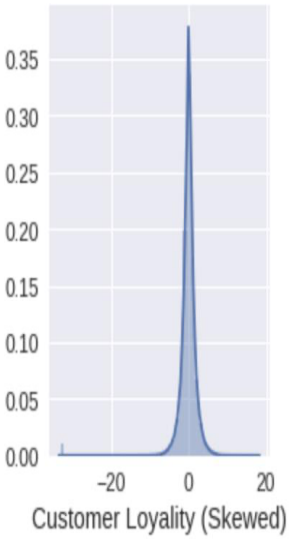

Figure 3: Visualization of Customer Loyalty using Python

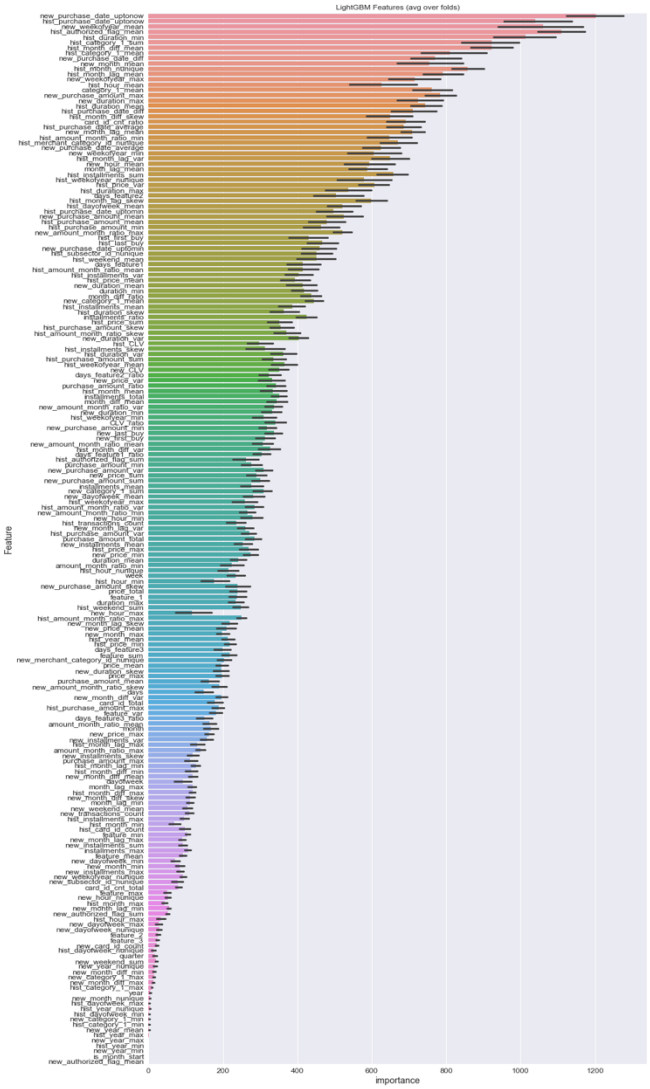

Figure 4: Rank of Feature Importance

VI. Conclusion and Future Works

The data pre-processing technique is an important step in extracting machine learning model as it improves the accuracy of prediction. In terms of predicting credit card holders retention, the decision tree algorithm has the highest accuracy with 88.2 performance measures.

The predictive models were validated through accuracy, precision and recall. While in predicting the customer's loyalty score based on the customer's transactions, the researcher used the overall performance metric that is root-mean squared error (RMSE). In terms of future works, the researcher aims to include additional features and attributes to improve the accuracy of the prediction . Several feature selection technique can be used to improve the quality of the data. In addition, other classification and clustering algorithms can be used to determine the optimal accuracy result of the classification.

VII. References

[1] Martinez, A. et al,. (2018). A machine learning framework for customer purchase prediction in non-contractual setting. European Journal of Operation Research. https://doi.org/10.1016/j.ejor.2018.04.034.

[2] Wang, Y., Lu, X. & Tan Y., (2018). Impact of product attributes on customer satisfaction: An analysis of online reviews for washing machines. Electronic Commerce Research and Applications. https://doi.org/10.1016/j.elerap.2018.03.003.

[3] Gonzalez-Carrasco, I. et al., (2019). Automatic detection of relationships between banking operations using machine learning. Information Sciences, Vol. 485, pp 319-346. https://doi.org/10.1016/j.ins.2019.02.030.

[4] Henrique B., Sobreiro V., & Kimura H. (2019). Literature review: Machine learning techniques applied to financial market prediction. Expert System with Applications, Vol. 124, pp. 226-251.

[5] Song, Y., Cao, Q., & Zhang C. (2018). Towards a new approach to predict business performance using machine learning. Cognitive Systems Research, Vol. 52, pp. 1004-1012.

[6] Bilalli, B., Abello, A., Banet, T., & Wrembel, R. (2018). Intelligent assistance for data pre-processing. Computer Standards & Interface, Vol. 57, pp. 101-109. https://doi.org/10.1016/j.csi.2017.05.004.

[7] P. Branco, L. Torgo & R.P. Riberio. (2018). Pre-processing approaches for imbalanced distributions in regression. Neurocomputing, Vol. 343, pp. 79-99. https://doi.org/10.1016/j.neucom.2018.11.100.

[8] Symeonidis, S., Effrosynidis, D., & Arampatzis A. (2018). A comparative evaluation of pre-processing techniques and their interactions for twitter sentiment analysis. Expert System with Applications, Vol. 110, pp. 298-310. https://doi.org/10.1016/j.eswa.2018.06.022.

[9] Shirazi, F., (2018). A big data analytics model for customer churn prediction in the retiree segment. International Journal Management. https://doi.org/10.1016/j.ijinfomgt.2018.10.005.

[10] Caigny, A., Coussement K., & Bock, K. (2018). A new hybrid classification algorithm for customer churn prediction based on logistic regression and decision trees. European Journal of Operation Research, Vol. 296, pp. 760-772.

[11] Ansari, A and Riasi, A. (2016). Modelling and evaluating customer loyalty using neural networks: Evidence from startup insurance companies. Future Business Journal, Vol. 2, pp. 15-30. https://doi.org/10.1016/j.fbj.2016.04.001.

[12] Ambrus, R., Izvercian, M., Ivascu, L., Artene, A. (2017). The link between competitiveness and sustainability of enterprises, 4th BE International Conference on Business & Economics.

[13] Hamidi, H. and Safareeyeh, M. (2019). A model to analyze the effect of mobile banking adoption on customer interaction and satisfaction: A case study of m-banking in Iran. Telematics and Informatics, Vol. 38, pp. 166-181. https://doi.org/10.1016/j.tele.2018.09.008.

[14] Shirazi, F. and Mohammadi, M. (2018). A big data analytics model for customer churn prediction in the retiree segment. International Journal of Information Management. https://doi.org/10.1016/j.ijinfomgt.2018.10.005.

[15] Alalwan, A. (2018). Investigating the impact of social media advertising features on customer purchase intention. International Journal of Information Management, Vol. 42, pp. 65-77.

[16] Ivascu, L., and Cioca, L. I. (2014). Opportunity Risk: Integrated Approach to Risk Management for Creating Enterprise Opportunities, The 2nd International Conference on Psychology, Management and Social Science, Psychology, Management and Social Science. Advances in Education Research, 49, 77-80.

[17] Daras, G., Agard, B., & Penz. (2018). A spatial data pre-processing tool to improve the quality of the analysis and to reduce preparation duration. Computer & Industrial Engineering, Vol. 119, pp. 219-232.

https://doi.org/10.1016/j.cie.2018.03.025.