A picture containing text

Description automatically generatedA picture containing text

Description automatically generatedPredict Customers Attraction

Student Name: Alaa Alkamel

Student Number:501143650

Supervisor’s Name: Dr. Tamer Abdou

Table of Contents

[Abstract 3](#_Toc121137479)

[Introduction 4](#_Toc121137480)

[Data Mining and Machine Learning 6](#_Toc121137481)

[Machine Learning: 7](#_Toc121137482)

[Classification 7](#_Toc121137483)

[Related Work 8](#_Toc121137484)

[Classification Methods 10](#_Toc121137485)

[Random Forest (RF) 10](#_Toc121137486)

[Logistic Regression (LR) 11](#_Toc121137487)

[Decision Tree (DT) 11](#_Toc121137488)

[Brief descriptive statistics of the datasets 11](#_Toc121137489)

[Table 1 presents a brief data description 12](#_Toc121137490)

[Methodology 17](#_Toc121137491)

[Experiments Results 19](#_Toc121137492)

[First experiment 20](#_Toc121137493)

[Second and the third experiments 24](#_Toc121137494)

[Conclusion 31](#_Toc121137495)

[Research problem 31](#_Toc121137496)

[Limitations and Future Work of the Research 32](#_Toc121137497)

[Reference 33](#_Toc121137498)

# Abstract

Customer attrition, also known as customer churn, is the loss of clients or customers.

All bank companies have credit card facilities, so keeping customers loyal and moving to another credit card is a challenge for each bank or company. They would appreciate it if one could predict for them who is going to get churned so they can proactively go to the customer to provide them better services and turn customers' decisions in the opposite direction. I analyze the data from a fictitious data source available on Kaggle [1], to see the reason for this attrition and predict it.

Target: 0: Existing Customer (Customer will stay) 1: Attrited Customer (Customer will churn).

The purpose of this analysis is to focus on the problem of customer churn in the credit card by solving these problems:

1. Build a model to help the company to predict the possibility that someone will churn or move to another company or not, which will be achieved by (Random Forest, Logistic Regression, and decision tree)
2. Use the AUC (Area Under the Curve), Accuracy, Sensitivity, Specificity, training time, predicting time, Brier score, MCC (Matthew’s correlation coefficient) and Stability to determine the best model.
3. Predict which factors make the customer churn and switch to another company and how these factors will affect customer churn.
4. Relationship between the churn rate and important features.

# Introduction

Customer attrition also known as customer churn, or customer turnover, or customer defection, is defined as the customer tend to leave the product and stop being making a relationship with a bank or company, it has become one of the major problems for many companies and one of biggest challenges that they are facing. Customer churn introduces not only some loss in income but also other negative effects on the operations of companies [2].

Many studies are done to predict customer churn in various areas such as the telecommunication industry, banking, insurance, etc. the following reasons show why the study on churn is very important:

* “Find new customers will be much harder and more expensive.
* Loss of customer will be loss of profits (natural consequence).
* Customer churn leads ta o negative impact on other customers will tarnish the brand of the organization” [3].

Customers shift from one company or bank to another for different reasons, such as the availability of the latest technology, poor customer service, interest rates, the proximity of the geographical location, the different services offered, etc.

Knowing your customers better will enable you to serve them better and keep them loyal forever. This is the main theme of Customer Relationship Management (CRM) [4].

An effective CRM decision support system increases the quality of customer relationships, thereby increasing retention in several ways:

* It supports predictive modelling to help banks identify who is likely to leave and why and what to do about it
* It enables a new level of personalization in service offers and marketing approaches, which fosters loyalty
* It brings greater richness to customer interaction, thereby increasing customer satisfaction because consistent information is shared across all customer touch points. [5]

Therefore, it is important from the point of the real-life market to sustain the large competition within the business world. And it's essential to manage the churning additionally. Churn prediction is important for the real-life market and to thrive the business competition additionally, it is a necessity to manage [6].

In this work, I use a dataset from a Data Science website, Kaggle [1]. This data set contains about 10,000 customers mentioning their age, salary, marital status, credit card limit, credit card category, number of dependents etc. This dataset has nearly 22 features. One of the features is Attrition\_Flag, which is the target variable. Using the Attrition\_flag feature is possible to identify if the customer will churn or not. In this data, the result of customer churned 16.1% is shown in figure 1.

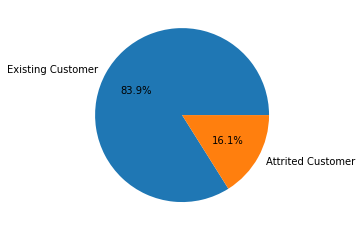


Figure 1 Pie chart shows the percentage of Existing Customer & Attrited Customer

To fix this problem I will use the over-sampling technique (SMOTE) and random over-sampling technique. Based on the dataset variable types, since the dependent variable is a binary type, the techniques that I will use are (1) Decision Tree (2) Logistic Regression (3) Random Forest.

# Data Mining and Machine Learning

“Data mining is a technique for finding information hidden in a data set. By using statistical techniques, mathematics, artificial intelligence, and machine learning to extract and identify potential and useful information stored in large data” [7].

Researchers show data mining techniques are found to be more effective in churn prediction. Especially Predictive modelling techniques are often found to be more accurate in churn prediction [8].

Data mining includes three major techniques:  artificial intelligence technologies, machine learning techniques, and statistical techniques [9].

## Machine Learning:

Machine Learning is a branch of Artificial Intelligence which helps in analytical model building. “The machine learning models learn from the data, identify general patterns in it and construct decisions with minimal human intervention” [10].

Diagram

Description automatically generated

Figure 2: Machine Learning Techniques [10]

### Classification

When the target variable is categorical then it is considered as Classification problems as it is in my dataset. The classification works to find a model that describes and recognizes a class or concept of data. This model is developed based on the analysis of training data and is used to predict the class label of an object whose label class is unknown. For the prediction of churn customers, data mining classification algorithms that are often used are Decision Tree, Logistic Regression, Support Vector Machine, and Naïve Bayes.[7]

# Related Work

The analysis of customer churn is a wide-ranging area. In this section first I will write about some studied related to customer churn in the bank industry.

Kaur’s study [2] this dataset contains 21 features and 28,382 customers; features are selected as baseline features using chi-square. For model-building Logistic regression, Decision tree, K nearest neighbour and Random Forest are used. The data is split into training 70% is used for model building and testing 30% for performance evaluation of the trained model. The performance comparison is made considering stratified sampling, without stratified sampling, and 8-fold cross-validation (CV). AUC-ROC, Recall, Precision and Accuracy are used for evaluating the performance of the classifiers on the dataset. This study shows comparison results between the performance of all features and after using the feature selection. Ensembling techniques are applied such as averaging and voting to improve the performance of the model with lesser accuracy. Random Forest performs better among all the models used with an accuracy of 85.22%.

The study [11] purpose customer churn in the bank industry, this study contains 10000 bank clients with 13 features, the target variable is a binary representing whether the customer will churn or not. Minimum Redundancy Maximum Relevance (mRMR) and Relief are used for the feature selection. This data is highly imbalanced, and the size of the available data sample is small. Because of that, the oversampling technique is used. If the undersampling is used the size of the data will decrease in a way that enough data will not be there to build the model. The 10 features obtained are taken for the remaining study. The dataset is split into 70% training data and 30% testing data. The classification methods were applied in the study k-Nearest Neighbor (KNN), k-Nearest Neighbor (KNN), Decision Tree (DT) and Random Forest (RF). Each model is evaluated by the accuracy which is obtained after 10-fold cross-validation. Also, random confusion matrix was applied for each model. Finally, a random forest with oversampling achieved the best result which is 95.74%.

Another study about customer churn in the bank sector [12]This study contains 274,542 customers 11.6% churned to solve the imbalance problem tree-based model is applied, and the dataset is split into two parts 80% training set and 20% testing set. The classification models used in this study are Logistic Regression, Decision Tree, Random Forest, and Xg-boost. To perform and evaluate the model Sensitivity, Specificity, Accuracy and Area Under the Curve (AUC) are used. The Xg-boost performed the best for each evaluation method, the best one was 96.67% with AUC. The Random Forest comes after Xg-boost.

**In this work, I will split the data into 80% training set and 20% testing as it most common ratio in the studies [12][13]. The data is imbalanced, and the observation is small, so I will use the random oversampling technique and SMOTE as I read in the previous study [11] for small data observation is better to use oversampling.**

**Based on what read about studies related to churn most of those who used Random Forest gave highly accurate results [2][11][14], according to that I will use Random Forest to build the model. In addition, I will use Decision Tree and Naïve Byes as [7] mentions some of the algorithms that are often used to predict customer churn are Decision Tree and Logistic Regression.**

**To evaluate the model, I will use Area Under the Curve, Sensitivity, Specificity and Accuracy which most of the studies used some or all of these evaluation measures [2][11][12][14].**

# Classification Methods

The classification problem uses supervised learning, The dataset is spilt into two phases training and testing. Training set for model building and the result will be applied to the test set. I will briefly describe the classification algorithms such as Random Forest, Logistic Regression, and decision tree which I will use to build the model.

## Random Forest (RF)

RF is an ensemble classifier for tree learners. The method uses several decision trees so that each tree relies on the values of an individually selected random vector with the same distribution for all trees. The right choice for the tendency of decision trees to overfit their training collection. In short, Random forests are a way to combine many deep decision trees which are learned on various sections of the same dataset with the target of decreasing the variance. The advantage of using RF is that it comes with quite high dimensional data, with no need to perform dimensionality reduction and feature selection. The training rate is also higher and easy to use in parallel models [15].

## Logistic Regression (LR)

Logistic Regression is a machine learning algorithm which is used for classification problems, and it can deal with different combinations of variables and could be perfect in predicting the churn with high accuracy and it is a type of probability statistical classification model [16].

## Decision Tree (DT)

The decision tree is a supervised-learning algorithm where a target variable is also defined, and it is mostly used in classification problems. For both Types of variables, i.e., categorical and continuous the algorithm works fine. The tree is split into internal nodes and leaf nodes based on splitting criteria. The splitting criteria are named Gini-based, or Information Gain based. The internal nodes represent features or attributes where testing is done, and the leaf node represents the class [2]. There are different algorithms used in the decision trees: ID3, C4.5, CART, C5.0, CHAID, QUEST and CRUISE.

# Brief descriptive statistics of the datasets

In this section I will show a brief desecration of the dataset, this work was done in python.

First, I dropped the unnecessary columns from the dataset. Then I checked the data type as shown in table1. The next step, seeing the unique values of columns, helps me to understand the data and visualize data as much as easily. I found there is no missing value or duplicate.

## Table 1 presents a brief data description

|  |  |  |
| --- | --- | --- |
| Attribute | Data Type | Description |
| CLIENTNUM | Numerical | Client number. Unique identifier for the customer holding the account |
| Attrition\_Flag | Categorical(binary) | Internal event (customer activity) variable - if the account is closed then 1 else 0 |
| Customer\_Age | Numerical | Demographic variable - Customer's Age in Years |
| Gender | Categorical(binary) | Demographic variable - M=Male, F=Female |
| Dependent\_count | Numerical | Demographic variable - Number of dependents |
| Education\_Level | Categorical | Demographic variable - Educational Qualification of the account holder (example: high school, college graduate, etc.) |
| Marital\_Status | Categorical | Demographic variable - Married, Single, Divorced, Unknown |
| Income\_Category | Numerical | Demographic variable - Annual Income Category of the account holder (< 40K, 40K-60K, 60K-80K, 80K-120K, > 120k) |
| Card\_Category | Categorical | Product Variable - Type of Card (Blue, Silver, Gold, Platinum |
| Months\_on\_book | Numerical | Period of relationship with the bank |
| Total\_Relationship\_Count | Numerical | Total no. of products held by the customer |
| Months\_Inactive\_12\_mon | Numerical | No. of months inactive in the last 12 months |
| Contacts\_Count\_12\_mon | Numerical | No. of Contacts in the last 12 months |
| Credit\_Limit | Numerical | Credit Limit on the Credit Card |
| Total\_Revolving\_Bal | Numerical | Total Revolving Balance on the Credit Card |
| Avg\_Open\_To\_Buy | Numerical | Open to Buy Credit Line (Average of last 12 months) |
| Total\_Amt\_Chng\_Q4\_Q1 | Numerical | Change in Transaction Amount (Q4 over Q1) |
| Total\_Trans\_Amt | Numerical | Total Transaction Amount (Last 12 months) |
| Total\_Trans\_Ct | Numerical | Total Transaction Count (Last 12 months) |
| Total\_Ct\_Chng\_Q4\_Q1 | Numerical | Change in Transaction Count (Q4 over Q1) |
| Avg\_Utilization\_Ratio | Numerical | Average Card Utilization Ratio |

I visualize the distribution of the numerical variables shown in figure 3.

## Distribution of the Numerical

general overview of all the histogram so we can further describe the nature of people who got their accounts attrited.

A picture containing window, shoji, train, building

Description automatically generated

Figure 3 distribution of the numerical

## Correlation Matrix with Heat map

To discover the related attributes to customer attrition Correlation technique can be used. The correlation between the independent and dependent numeric variables are shown as a correlation matrix in figure 4 for numeric data using the Pearson correlation coefficient.

Chart

Description automatically generated

Figure 4 correlation matrix for numeric data

Strong Positive Relationship

* Avg\_open \_to\_Buy and Credit\_Limit =1
* Total\_Trans\_Amount and Total\_Trans\_Ct =0.81
* Customer\_Age and Months\_on\_Book =0.79
* Total\_Revolving\_Balance and Avg\_Utilization\_Ratio = 0.62

Negative Relationship

* Credit\_Limit and Avg\_Utilization\_Ratio = -0.48
* Avg\_Open\_To\_Buy and Avg\_Utilization\_Ratio =-0.5

And this the link GitHub website where codes and results are uploaded for EDA

<https://github.com/AlaaAli968/Bank-Churn/blob/main/Untitled9%20(1).ipynb>

# Methodology

In this section, I used various techniques on the dataset. These techniques are shown and discussed below.

Diagram

Description automatically generated

Figure 5 Methodology graph

First, I start with **Data Preparation** I checked the missing values and duplicated values, there is no ant missing values in my dataset. Then with **Exploratory Data Analysis** (EDA) to understand the data by seeing the types of the attributes. The basic quantitative analysis was done with the data, displaying the distribution, and visualizing some columns.

Also, the unique values for some columns, especially for the categorical columns to converted them in the next step as shown in figure 6.

Table

Description automatically generated with medium confidence

Figure 6 values for categorical attributes

Third step **Data Processing** in the section I converted the categorical columns to numeric ones because machine learning algorithms work best with numerical data. the target column converted to 0(Existing Customer) and 1 (Attrited Customer), and the rest of the categorical columns are converted by get dummies (convert each unique element in the object to a column heading).

The fourth step **Split the Data** into 70% training and 30% testing as it's the most popular ratio.

The fifth step **Balancing** the data to address the class imbalance oversampling the minority class is applied. I applied two techniques for random oversampling and SMOTE.

The sixth step **Build the Models** in the training set, I train different models (with imbalanced and balanced data) to see the best model. I compare three models Random Forest, Decision Tree, and Logistic Regression before SMOTE, after SMOTE and after Random Oversampling. Accuracy, Specificity, Sensitivity, MCC, etc. used to **Evaluate the Models** performance. As well as I checked the stability of the models by using the brier score. Finally, **Feature Importance** from the best models to determine which attributes are most important and effected the decision of credit card churn.

# Experiments Results

In this section I will summarize the results, as I explained earlier in my work, I compared the model’s evaluation scores with 3 experiments

1. Imbalanced data

2. After applying SMOTE with the training set

3. After applying Random Oversampling with the training set.

## First experiment

I started to train the models with the imbalanced data to see the difference after sampling.

1. **Confusion Metric**

A confusion matrix is a technique for summarizing the performance of a classification algorithm. It is a table with combinations of predicted and actual values.

Table

Description automatically generated

Figure 7 confusion matric

For my case the target 0: Existing Customer 1:  Attrited Customer

True Positive: We predicted positive and it’s true. We predicted that customer will churn and actually churn.

True Negative: We predicted negative and it’s true. We predicted that customer will not churn and actually not churn.

False Positive (Type 1 Error)- We predicted positive and it’s false. We predicted that customer will churn and actually not churn.

False Negative (Type 2 Error)- We predicted negative and it’s false. We predicted that customer will not churn and actually churn

Chart, treemap chart

Description automatically generated

Figure 8 RF Confusion Metric

Chart, treemap chart

Description automatically generated

Figure 9 LR Confusion Metric

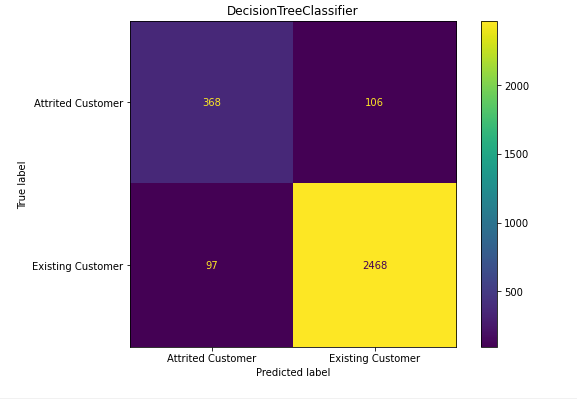


Figure 10 DT Confusion Metric

My target is to reduce the False Negative (Type 2 Error).

1. **Evaluation and build the models**

As I mentioned earlier, I used Random, Forest Decision and Tree Logistic Regression. Here I will explain some of the evaluation scores for the models and show the results

* MCC [-1,1] Matthew’s correlation coefficient is a more reliable statistical rate which produces a high score only if the prediction obtained good results in all the four confusion matrix categories (true positives, false negatives, true negatives, and false positives) 1 indicates perfect agreement (FP = FN = 0), 0 is expected for a prediction no better than random, -1 (TP = TN = 0) Indicates total disagreement between prediction and observation.
* Brier score [0,1] smaller is better, 0 means perfect accuracy.
* Accuracy evaluation metric that measures the number of correct predictions made by a model.

Graphical user interface

Description automatically generated

Figure 11 Evaluate the Models for Imbalanced data

From Figure 11 we can see that Random Forest is the best model with high values for Accuracy, MCC, AUC etc. Also, the lowest value with Brier.

Also, I showed plot for AUC to see the True positive rate and False positive rate for each model.

In additional, I checked the over/under fit for each model by checking the range of Brier scores for train and test set.

## Second and the third experiments

I applied SMOTE and Random Oversampling

I used the default strategy for SMOTE which is mean increase the percentage of minority class to be equal to the percentage of majority class. For random oversampling I used minority strategy. I tried to compare the evaluation results between these techniques figures 12 & 13.

Graphical user interface

Description automatically generated with low confidence

Figure 12 Evaluate the Models after SMOTE

Graphical user interface

Description automatically generated with medium confidence

Figure 13 Evaluate the Models after Random Oversampling

As it shown in figures 12 & 13 Random Oversampling gave me the best results with **MCC** ≈ 85.6% which is mean the prediction obtained good results in all the four confusion matrix categories (true positives, false negatives, true negatives, and false positives)

Brier ≈ 3.6% which is very low and close to zero, means the model moris e accurate(1-ACC).

Accuracy ≈ 96.3 % very high value that means the number of correct predictions made by a model, (focuses on TP &TN)

Sensitivity≈ 84% means (True Positive Rate) When it’s actually yes, how often does it predict yes. TP / (TP+FN)

Specificity ≈ 98% tells us the (True Negative Rate) TN / (TN+FP).

F1≈ 87% F1 Score high which means both precision and recall are high, F1=2\*(Recall \* Precision) / (Recall + Precision)

AUC≈ 99 high value close to 1.0 which means the model made all predictions perfectly.

For prediction and training time Random Forest was the highest one, but still best in another evaluation score.

Also, I showed the confusion mastics after applying Random Oversampling

A picture containing text, monitor, electronics, screen

Description automatically generated

Figure 14 Confusion mastics after applying ROS

From the figure 14 we can FP and TN have been reduced.

In Additional I plot the ROC curve (TPR vs. FPR) and showing the performance of classification models at all classification thresholds

Chart

Description automatically generated

Figure 15 ROC Curve (Random Oversampling)

As we can see in figure 15 RF gave the highest value for TPR verses low value for FPR.After that I checked the stability again by checking the range of Brier score for train set after applying ROS and test set, the models are stabile as figures 16 & 17 & 18 shown.Chart, line chart

Description automatically generated

Figure 16 Stability for LR model

Chart, line chart

Description automatically generated

Figure 17 Stability for RF model

Chart, line chart

Description automatically generated

Figure 18 Stability for DT model

The ranges for brier scores are small in train and test sets, which means the models are stabile.

The final step for these experiments I applied the feature importance, I used Random Forest with Random Oversampling as shown in figure 19, to extract the top 10 features for the dataset. These features are ['Total\_Relationship\_Count', 'Months\_Inactive\_12\_mon', 'Credit\_Limit', 'Total\_Revolving\_Bal', 'Avg\_Open\_To\_Buy', 'Total\_Amt\_Chng\_Q4\_Q1', 'Total\_Trans\_Amt', 'Total\_Trans\_Ct', 'Total\_Ct\_Chng\_Q4\_Q1', 'Avg\_Utilization\_Ratio']

A picture containing table

Description automatically generated

Figure 19 Feature Importance with the values of importance

After I chose the top 10 important features from Random Forest, I showed the positive & negative correlations for the top 10 features with churn rate as shown in figure 23.

# Conclusion

This section mentioned the answer for research problem Limitations and Future Work of the Research

## Research problem

1. Build a model to help the company to predict the possibility that someone will churn or move to another company or not, which will be achieved by (Random Forest, Logistic Regression, and decision tree)

Above problem was achieved by building Random Forest, Logistic Regression, and decision tree in the dataset 3 times first time with imbalanced data, sconed time after applying SMOTE, and third time after applying Random Oversampling.

1. Use the AUC (Area Under the Curve), Accuracy, Sensitivity, Specificity, training time, predicting time, Brier score, MCC (Matthew’s correlation coefficient) and Stability to determine the best model.

Second problem was achieved for the 3 experiments as shown earlier in figures 11 & 12 & 13.

Also, as I mentioned the best model was Random Forest with Random Oversampling regarding all evaluation scores except training time and predicting time was the highest time compared to other models which is mean RF was slower than the other model.

I showed each model in the 3 experiments and how its performed in figure 20 & 21 & 22.

Chart, bar chart

Description automatically generated

Figure 20 Random Forest in the three experiments

Chart, bar chart

Description automatically generated

Figure 21 Decision Tree in the three experiments

Chart, bar chart

Description automatically generated

Figure 22 Logistic Regression in the three experiments

1. Predict which factors make the customer churn and switch to another company and how these factors will affect customer churn.

This problem was done by using the features important from the Random Forest classifier and choosing the top 10 important features which are ['Total\_Relationship\_Count', 'Months\_Inactive\_12\_mon', 'Credit\_Limit', 'Total\_Revolving\_Bal', 'Avg\_Open\_To\_Buy', 'Total\_Amt\_Chng\_Q4\_Q1', 'Total\_Trans\_Amt', 'Total\_Trans\_Ct', 'Total\_Ct\_Chng\_Q4\_Q1', 'Avg\_Utilization\_Ratio']

The above attributes are the factors which mostly impact the customer’s decision to churn.

1. Relationship between the churn rate and important features.

The problem above was achieved by showing the positive & negative correlations for the top 10 features with churn rate.

Chart, waterfall chart

Description automatically generated

Figure 23 Visualize the correlation of the important features with churn

From figure 23 we can see the Attrition flag (churn) increase when the customer stays inactive for more months. On other hands churn decrease when total\_relationship\_count increase which is mean when the customer has many products such as credit cards, and personal lines of credit. Home equity lines of credit. Also churn has another negative relationship with churn total revolving balance, total transactions count, total transaction count difference of Q4 and Q1, average card utilization ratio, total transaction amount last 12 months, etc. These attributes have a negative relationship with churn which is mean if one of them increases the chance of churn will be less.

## Limitations and Future Work of the Research

The dataset contains unknown values, but I didn’t remove those rows because the dataset is small, in addition these unknown values will be changed in a future version of this dataset. For now, I considered all the unknown values as a separate category.

The dataset contains values for one year only, it would be better if I have the data values for couple of years. Also, I need to know more details, for example, the annual fees, and type of rewards for the customer for example cash back or airline miles, etc.

For future work ensemble model can be used to achieve better predictive performance on a predictive modeling problem than a single predictive model.

# 

# Reference

[1] S. Goyal, "Credit Card customers," 19 November 2020. [Online]. Available: https://www.kaggle.com/sakshigoyal7/credit-cardcustomers. [Accessed 18 January 2022].

[2] Kaur, I., & Kaur, J. (2020, November). Customer churn analysis and prediction in banking industry using machine learning. In *2020 Sixth International Conference on Parallel, Distributed and Grid Computing (PDGC)* (pp. 434-437). IEEE.

[3] Dolatabadi, S. H., & Keynia, F. (2017, July). Designing of customer and employee churn prediction model based on data mining method and neural predictor. In *2017 2nd International Conference on Computer and Communication Systems (ICCCS)* (pp. 74-77). IEEE.

[4] Rababah, K., Mohd, H., & Ibrahim, H. (2011). Customer relationship management (CRM) processes from theory to practice: The pre-implementation plan of CRM system. *International Journal of e-Education, e-Business, e-Management and e-Learning*, *1*(1), 22-27.

[5] Joyner, E. (2002). Customer Relationship Management in Banking. *SAS White paper, North Carolina offices*.

[6] Hudaib, A., Dannoun, R., Harfoushi, O., Obiedat, R., & Faris, H. (2015). Hybrid data mining models for predicting customer churn. *International Journal of Communications, Network and System Sciences*, *8*(05), 91.

[7] Karvana, K. G. M., Yazid, S., Syalim, A., & Mursanto, P. (2019, October). Customer churn analysis and prediction using data mining models in the banking industry. In *2019 International Workshop on Big Data and Information Security (IWBIS)* (pp. 33-38). IEEE.

[8] Umayaparvathi, V., & Iyakutti, K. (2016). A survey on customer churn prediction in telecom industry: Datasets, methods and metrics. *International Research Journal of Engineering and Technology (IRJET)*, *3*(04).

[9] Jassim, M. A., & Abdulwahid, S. N. (2021, March). Data Mining preparation: Process, Techniques and Major Issues in Data Analysis. In *IOP Conference Series: Materials Science and Engineering* (Vol. 1090, No. 1, p. 012053). IOP Publishing.

[10] Wadikar, D. (2020). Customer Churn Prediction.

[11] Rahman, M., & Kumar, V. (2020, November). Machine learning based customer churn prediction in banking. In *2020 4th International Conference on Electronics, Communication and Aerospace Technology (ICECA)* (pp. 1196-1201). IEEE.

[12] Guliyev, H., & Tatoğlu, F. Y. (2021). Customer churn analysis in banking sector: Evidence from explainable machine learning models. *Journal Of Applied Microeconometrics*, *1*(2), 85-99.

[13] Charandabi, S. E. (2020). Prediction of Customer Churn in Banking Industry. *Age*, *18*(92), 38-92.

[14] Mishra, A., & Reddy, U. S. (2017, November). A comparative study of customer churn prediction in telecom industry using ensemble based classifiers. In *2017 International Conference on Inventive Computing and Informatics (ICICI)* (pp. 721-725). IEEE.

[15] Breiman, L., J. Friedman, R. Olshen, and C. Stone. "Classification and regression trees–crc press." *Boca Raton, Florida* (1984).

[16] Nie, G., Rowe, W., Zhang, L., Tian, Y., & Shi, Y. (2011). Credit card churn forecasting by logistic regression and decision tree. *Expert Systems with Applications*, *38*(12), 15273-15285.