Analyzing customer churn in banking: A data mining framework

Aishwarya Saxena, Anushi Singh, Govindaraj M.

1. Introduction

Customer churn refers to the phenomenon where customers discontinue their relationship with a business or organization. In the context of the banking industry, customer churn occurs when individuals or businesses close their accounts, switch to another financial institution, or cease using certain banking services (Sun 2021). Analyzing customer churn in banking is crucial for financial institutions as it helps them understand why customers are leaving and enable them to take proactive measures to retain their valuable clientele. Customer churn, also known as customer attrition, is a critical challenge faced by banks and financial institutions (de Lima Lemos et al. 2022). It refers to the loss of customers who discontinue their relationship with a bank, such as closing their accounts or shifting their business to a competitor. Customer churn has significant implications for banks, affecting their profitability, market share, and overall customer satisfaction.

To address these issues, banks are increasingly using data mining techniques to analyze customer churn patterns and identify factors contributing to customer attrition. The practice of gathering helpful information and recurring patterns from massive amounts of data is known as data mining. By leveraging data mining approaches, banks can better understand customer behavior and develop targeted strategies to retain customers and reduce churn (Wu and Li 2021; Tao et al. 2020).

Authors of the study (Kaur and Kaur 2020) utilize many machine learning models to the bank dataset in an effort to forecast the likelihood of customer churn. These models include logistic regression (LR), decision trees (DT), K-nearest neighbor networks (KNN), random forests (RF), and others.

In the article Satria et al. (2020), a deep learning model was developed to forecast the loss of customers in the banking industry by employing ANN architecture to address a categorization challenge. In order to accomplish tasks like extraction of features, recognition of patterns, regression, and categorization, ANN makes use of its many layers and its numerous existing nodes/neurons.

The research Muneer et al. (2022) aims to create a model that provides valuable churn prediction for the financial sector. They use the three innovative models “random forest (RF), AdaBoost, and support vector machine (SVM) to develop a method for predicting customer attrition. Using the synthetic minority oversampling method (SMOTE)” to correct for under sampling and oversampling in an imbalanced dataset yields the most accurate results.

In order to identify the customers that provide the most significant risk of leaving the bank, the authors in the study Dalma et al. (2020) employed a supervised machine learning approach to develop a unique algorithm. With different datasets, various classifiers can provide varying degrees of accuracy. K-nearest neighbor (KNN) is a revolutionary new method for improving accuracy using weighted scales and the XGBooster algorithm. Using weighted scales and the KNN technique, the dataset is appropriately divided into training and testing models.
A study, Gholamiangonabadi et al (2019), suggested anticipating the loss of customers at an Iranian bank, where they presented a novel methodological strategy. They use data pre-processing to clean up their data first. Then, the k-medoids technique is used to group the data. Clustering efficacy is measured with the Davies-Bouldin index. Several types of neural networks (NNs) were used to analyze the data for patterns, including “radial basis function (RBFNN), generalized regression (GRNN), multilayer perceptron (MLPNN), and support vector machine (SVM).” MLPNN and SVM models were shown to have higher precisions with lower costs than other models.

In the study Amuda and Adeyemo (2019), a predictive model that utilizes the Multi-layer Perceptron of Artificial Neural Network structure was constructed with the purpose of predicting the degree of customer turnover in financial organizations.

The goal of the research Imron and Prasetyo (2020) was to improve the efficiency of the K-Nearest Neighbor technique for classifying data by determining how Z-Score normalization works and how best to determine the ideal K value parameters using Particle Swarm Optimization. Z-score and Particle Swarm Optimization were employed for data normalization in order to locate the best potential K value.

The study’s objective Sjarif et al (2019) was to provide a method for predicting customer turnover utilizing the “Pearson Correlation and the K Nearest Neighbor algorithm.” The outcome demonstrates that the K Nearest Neighbor algorithm outperforms other algorithms’ accuracy.

The objective of this study is to apply a data mining approach to analyze customer churn in banking. By examining historical customer data, including demographic information, transactional records, customer interactions, and other relevant factors, we aim to identify key predictors of churn and develop a predictive model to forecast customer attrition. This research will contribute to the existing body of knowledge by providing valuable insights into customer churn behavior in the banking industry. The findings can help banks develop proactive churn prevention strategies, improve customer retention efforts, and enhance overall customer experience.

The rest of this paper is as follows: part 2 explores the related works, part 3 explores the methodology used for predicting the customer’s chunk in the banking industry, and Part 4 shows the performance of the proposed and existing method. And part 5 concludes with the conclusion part.

2. Materials and Methods

This work aims to utilize efficient data mining techniques to generate accurate early forecasts regarding the loss of customers at a commercial bank. A diagrammatic representation of the recommended model may be seen in Figure 1.

2.1. Data sample

In order to model chucks, the dataset that was utilized for this investigation was taken from Kaggle. This dataset consists of ten thousand bank customer data records, each of which has fourteen variables, including sociodemographic factors, account level attributes, and behavioral aspects, as outlined in Table 1.
Table 1 Dataset description.

<table>
<thead>
<tr>
<th>Description</th>
<th>Attribute</th>
</tr>
</thead>
<tbody>
<tr>
<td>ID of customer</td>
<td>Customer ID</td>
</tr>
<tr>
<td>Number of customers</td>
<td>Row Number</td>
</tr>
<tr>
<td>Location of customer</td>
<td>Geography</td>
</tr>
<tr>
<td>Age of Customer</td>
<td>Age</td>
</tr>
<tr>
<td>Customer gender</td>
<td>Gender</td>
</tr>
<tr>
<td>Customer name</td>
<td>Surname</td>
</tr>
<tr>
<td>Score of credit card usage</td>
<td>Credit Score</td>
</tr>
<tr>
<td>No of products used by customer</td>
<td>No. of Products</td>
</tr>
<tr>
<td>The period of having the account in months</td>
<td>Tenure</td>
</tr>
<tr>
<td>Estimated salary of the customer.</td>
<td>Estimated Salary</td>
</tr>
<tr>
<td>Indicates customer leaved or not</td>
<td>Churn</td>
</tr>
</tbody>
</table>

2.2. Data pre-processing

Data pre-processing plays a crucial role in data mining. Since they impact task completion rates directly, it has to handle irrelevant, noisy, and unreliable information. And if that data conversion is required, too. In this analysis, these factors were used to determine churn rates.

- Irrelevancy: Relevant information is defined as facts or characteristics that do not affect the conversation topic. A classifier’s performance may occasionally be impacted by maintaining such features. Row number, Customer Id, Surname, and Geography are factors that have no bearing on the forecast when considering into account the churn dataset. Therefore, these traits were manually disregarded in our investigation.
- Transformation: The process of converting one set of data into another format is referred to as data transformation. The data quality is improved, and applications are protected against potential minefields when the data have been appropriately formatted and verified. Possible minefields include null values, undesired duplication, inaccurate indexing, and incompatible file formats. In the course of this study, a modification to the data will be carried out.

2.3. Model building

Step-1: Using clustering, you may divide an immense amount of data into manageable sets of related records. It’s able to pick up on broad themes discussed throughout the corpus. Document clustering has several applications beyond simply creating information maps from extensive document collections. The succeeding learning recoveries and accesses can be enhanced by this. For instance, document clustering has been used to improve the effectiveness of text layout and identify occurrence scenes in records that are only frequently used. Similarly, several early studies and emerging search engines use an archive clustering strategy to manage and consequently compose list items into vital classifications and thus offer cluster-based perusing instead of displaying query items in one not-significant list.

A parametric probability density function is referred to as a Gaussian mixture model, abbreviated as GMM. A weighted average of Gaussian densities is used to describe this model. A model based on the Gaussian mixture is a weighted average of the Gaussian densities of M different components, as shown in equation 1:

\[
(Y|\lambda) = \omega_j h(Y|\mu_j, \Sigma_j)
\]

In this case, \(Y\) is a D-dimensional continuous data vector, \(\omega_j\) and \(j = 1, \ldots, N\) is the values for the mixture weights \((Y|\mu_j, \Sigma_j), j = 1, \ldots, N\) are the parts of the Gaussian density functions. Each component density function is a D-variate Gaussian function of the form.

\[
h(Y|\mu_j, \Sigma_j) = \frac{1}{(2\pi)^{D/2}|\Sigma_j|^{1/2}} \exp \left\{ -\frac{1}{2} (y - \mu_j)^t \Sigma_j^{-1} (y - \mu_j) \right\}
\]

They were considering a covariance matrix of \(\Sigma_j\) and a mean vector of \(\mu_j\). The mixing weights are permitted to deviate from the requirement \(\sum_{j=1}^N \omega_j = 1\). The variables that comprise the parameters of a Gaussian mixture model are the densities of the individual components, their variances, and the weights assigned to each component in the final mixture. The notation \(\{\omega_j, \mu_j, \Sigma_j\} j = 1, \ldots, N\) can be used for these variables.

Step 2: Prediction: ASVM is used to create a prediction model for determining which customers are likely to churn based on the output of clustering findings. The use of binary classification is commonplace in model prediction. It is effective for many real-world problems and may resolve both linear and non-linear ones. In (3) above, \(f(y)\) is the function used to categorize data, \(w\) is the weight, \(U\) is the carriage, \(y\) is the input, and \(c\) is the bias.

\[
f(y) = wUy + c
\]
Decreasing the distance $2/\|x\|$ is comparable to optimizing $1/2\|x\|^2$, as illustrated in Figure 2, where the margin between the classification face is defined as $xUy + c = 1$ and $xUy + c = -1$ is $2/\|x\|$. Then, we may reframe the challenge of finding the best experience for categorization as the following optimization problem:

$$\min_{x,c} \frac{1}{2}\|x\|^2 \quad (4)$$

$$\text{s.t. } z_j((x, y + c)) \geq 1 \text{ for any } j = 1, ..., n \quad (5)$$

In this research, the non-linear classification problem is solved by employing the radius basis function (RBF) kernel method. RBF kernel applied to two vectors of features in a particular input space, $y$ and $y'$, representing two samples.

$$L(y, y') = \exp \left( -\frac{\|y - y'\|^2}{2\sigma^2} \right) \quad (6)$$

$$L(y, y') = \exp \left( -\|y - y'\|^2 \right) \quad (7)$$

Where $\|y - y'\|^2$ the squared Euclidean distance among the two features is vectors and $\sigma$ is an arbitrary constant. Then, GridSearchCV is utilized in adaptive support vector machines (ASVM) to locate appropriate hyper-parameters, such as (to use C or gamma values), that enhance accuracy and prediction outcomes. It thoroughly explores the parameter grid utilizing all possible parameter permutations. The initial task we need to do is to compile a dictionary of all the parameters and the ranges of values for them that we intend to try out. In order to proceed, a new instance of the “GridSearchCV-class” must be constructed. The final step is to invoke the fit class method of the “GridSearchCV class” and supply it with the training and test sets. As soon as the technique finishes operating, we look for the parameters that yield the maximum accuracy.

3. Results

We include a comparison of the suggested method’s performance to that of current approaches in this section. The existing method such as convolutional neural network (CNN), decision tree and multinomial regression (DT-MR), and artificial neural network (ANN). The parameters used for the comparison are accuracy (%), precision (%), and recall (%), F1-score (%).

An accurate forecast is one in which the sum of the positive and negative samples is proportional to the entire sample size. It primarily evaluates the accuracy of the model’s global predictions. Figure 3 and Table 2 depict the accuracy result. This demonstrates that our suggested technique, GMM-ASVM, has a greater accuracy in forecasting customer churns in the banking industry than the existing methods, CNN, DT-MR, and ANN. When comparing it with the currently employed methods.

The proportion of the number of positive samples that were accurately predicted in relation to the total number of positive samples is known as precision. This metric is primarily used to indicate the reliability of the positive samples. Figure 4 and Table 3 depict the precision result. This demonstrates that our suggested technique, GMM-ASVM, is superior in forecasting customer churns in the banking industry than the existing methods, CNN, DT-MR, and ANN. When comparing it with the currently employed methods.
The recall is the proportion of precisely anticipated positive sample size to the actual number of positive samples, and it represents the coverage of the prediction model. Recall may be expressed as a percentage. Figure 5 and Table 4 depict the recall result. This shows that when compared to the currently in use techniques, CNN, DT-MR, and ANN, our proposed methodology, GMM-ASVM, has a higher value in recall for forecasting the customer churns in the banking business.
The amount of time required by our suggested algorithm to complete its assigned work is referred to as the computation time. It is represented as seconds. Figure 6 and Table 5 depict the outcome of computational time. This shows that when compared to the currently in use approaches, CNN, DT-MR, and ANN, our recommended methodology, GMM-ASVM, takes less time to anticipate customer chруnсs in the banking business. This demonstrates that the model we offered will be effective.

<table>
<thead>
<tr>
<th>No. of sample</th>
<th>CNN (De Caigny et al 2020)</th>
<th>DT-MR (Rouhani and Mohammadi 2022)</th>
<th>ANN (Yahaya et al 2021)</th>
<th>GMM-ASVM [Proposed]</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>80.15</td>
<td>85.25</td>
<td>84.75</td>
<td>88.25</td>
</tr>
<tr>
<td>2</td>
<td>83.12</td>
<td>82.25</td>
<td>85.15</td>
<td>89.14</td>
</tr>
<tr>
<td>3</td>
<td>85.25</td>
<td>83.85</td>
<td>86.25</td>
<td>91.22</td>
</tr>
<tr>
<td>4</td>
<td>84.75</td>
<td>85.15</td>
<td>89.75</td>
<td>93.17</td>
</tr>
<tr>
<td>5</td>
<td>86.88</td>
<td>86.1</td>
<td>90.15</td>
<td>94.24</td>
</tr>
</tbody>
</table>

Table 4 Values of recall.

4. Conclusion

The rapid growth and arrangement of many administrations within the financial sector increase the probability of the industry losing profitable customers. Rapid advancement in data innovation in a variety of organizations, such as the financial sectors, which create enormous datasets, enables reasonable investigation to be performed to anticipate the behaviour of consumers and build up the connections of customers, with the goals of satisfying customers, attracting customers, and
retaining customers. Data mining techniques are used to effectively discover previously interred information and to learn from customers’ data. In order to predict customer losses in the banking sector, this study employs a Gaussian mixture model clustering-based adaptive support vector machine (GMM-ASVM). This research predicts customer behaviour using a clustering method by examining consumer competence and loyalty to the banking business using GMM. The clustering results were classified with 98% accuracy using ASVM. This research provides executives at banks with the knowledge they need to assess customer behaviour, which might lead to the implementation of strategies based on the engaging quality of customer contacts and boost the appropriate actions of administrator capacity.

**Ethical considerations**

Not applicable.

**Declaration of interest**

The authors declare no conflicts of interest.

**Funding**

This research did not receive any financial support.

**Reference**


