Automation of Credit Card Customer Churn Analysis using Hybrid Machine Learning Models

R. P. Ram Kumar¹*, B. Sahithi¹, K. Neeharika¹, M. Shivaleela¹, Digvijay Singh², K. Rama Krishna Reddy³

¹Department of CSE (AI & ML), GRIET, Hyderabad, Telangana State, India
²Uttaranchal Institute of Technology, Uttaranchal University, Dehradun, 248007, India
³KG Reddy College of Engineering & Technology, Hyderabad, India

Abstract. Credit Card Customer Churn Analysis (C4A) is a phenomenon where customers stop using a specific business credit card service. Predicting customer churn is crucial for Credit Card (CC) companies because it enables them to spot at-risk customers and take precautions to retain them. The aim of the paper named C4A is to create a model that accurately predicts customers who are most likely to stop using CC. The paper involves gathering and analyzing customer information from Kaggle, including transaction history, demographics and credit card usage patterns for prediction. Machine learning algorithms namely, Logistic Regression, KNN, XGBoost Classifier, Decision Tree and Hybrid Models integrating Logistic Regression and KNN, Logistic Regression and Decision Tree are used to train to find patterns and correlations that point to customer churn. The accuracy of the proposed method is 0.846 with LR, 0.849 with KNN, 0.90 with a hybrid model integrating LR and KNN, 0.928 by integrating LR and DT, 0.91 with DT, and 0.93 with XGBoost.

1 Introduction

Churn is substantially, more important in market places in which competition is intense and winning new customers is more difficult than holding on to the ones you already have. Companies that provide financial services but have contracts that are not legally enforceable such as banks, credit card companies, insurance companies, and credit unions, are especially concerned about churn since it affects their ability to make money. It is not rare for these companies to have attrition rates as high as 25–30%, and even enterprises with some kind of annual contract may see attrition rates as high as 5–7%. Churn is the term given to the phenomenon that occurs when customers of a credit card company either completely stop using their cards or switch to a different provider. Churn can have a negative impact on the bottom line of a provider. Credit Card Customer Churn (CCCC) can be defined as the rate at which existing credit card holders either stop using their cards, move to a different provider, or leave their current provider. For this reason, it is absolutely necessary for credit card firms to have an understanding of the factors that contribute to the churning of customers and to formulate strategies for reducing churn rates. Research in the

* Corresponding author: ramkumar1695@grietcollege.com
banking business that focuses on analysing customer behaviour and predicting customer turnover based on that behaviour is an important and growing topic of research. The conclusions of the research on client retention and defection have a sizable bearing on the policies of the financial institution. Figure 1 shows the percentage of churn various industries suffered in 2022.

![Median Churn Rate (%)](https://example.com/churn-rate.png)

**Fig. 1.** Churn Rate in various Industries 2022 (Courtesy: Source [1]).

### 2 Literature survey

Amrita Doshi proposed a method utilising AutoML technologies to forecast the number of credit card customers who may churn out of an organisation [2]. H2O-GradientBoosting, H2O-RandomForest, H2O-DeepLearning, Auto-Sklearn, and Auto-keras were supposed to be the method of choice for predicting churn. Among all the employed algorithms, the algorithm Auto-Sklearn has the best accuracy, whereas the H20-deep learning algorithm has the lowest accuracy. CCCC Prediction was proposed by Xinyu Miao and Haoran Wang utilising Random Forest (RF) [3]. Predicting the CCCC with the help of Machine Learning (ML) methods such as RF, Logistic Regression (LR), and K-Nearest Neighbour was the goal (KNN). After making adjustments to the parameters, the Random Forest algorithm achieved an accuracy score on the testing set data of 95.68%.

Using machine learning methods to develop a customer churn prediction for credit cards was a suggestion made by Dana AL-Najjar, Nadia AL-Rousan, and Hazem AL-Najjar [4]. Using ML algorithms such as Bayesian Networks, Classification and Regression Trees (CRT), Neural Networks, C5 Trees, and Chi-Square Automatic Interaction Detection (CHAID) Trees, the goal was to make churn predictions. C5 Tree outperformed all the involved ML models involved namely Bayesian Networks, CRT and Neural Networks in training. Customer Churn Analysis (CCA) was proposed for use in the banking sector by Hasraddin Guliyev and Ferda Yerdelen Tatoğlu [5]. Using ML techniques such as LR, RF, Extreme Gradient Boosting, and Decision Tree (DT), the goal was to detect existing clients before losing them to the competitor. The XGBoost algorithm performed the best across all metrics, with an area under the curve (AUC) of 96.97 percent; the RF model came in
second. CCCC was proposed by Kamil Demirberk through the use of Support Vector Machine (SVM) in conjunction with Bayesian Optimization [6].

Rajamohamed and Manokaran [7] hypothesised ML methods and Rough Clustering to predict CCCC. The machine learning (ML) methods SVM, RF, DT, KNN, and Hybrid Models were used with the intention of achieving the goal of increasing the retention rate. Support vector machine mixed with rough k-means clustering method works well and has better accuracy than any other hybrid model. Predicting the data from the CCCC and the Automobile Insurance Fraud was the idea of Ganesh Sundarkuma, Vadlamani Ravi, and Siddeshwar [8]. The purpose of this study was to provide evidence of the usefulness of the One Class SVM (OCSVM) methodology that was developed. The research came to the conclusion that the proposed under-sampling methodology was effective in lowering the complexity of the construction system while simultaneously producing important findings.

ML solution to the problem of churn in the banking business was proposed by Amgad Muneer and his research team [9]. The use of ML algorithms, including RF, SVM, and AdaBoost, was intended to accomplish the goal of increasing the retention rate. According to the findings of the research, RF performed significantly better than the other algorithms used, achieving an F1 score of 0.91. The prediction of CCCC was suggested by Ning Wang and Dong-xiao Niu [10]. The goal was to increase retention rate utilising the machine learning algorithms known as Rough Set Theory (RST), DT, Ridge Regression (RR), Artificial Neural Network (ANN), and Least Square - SVM (LS-SVM).

Guangli Nie and the other members of his research team came up with the idea for Credit Card Churn [11]. Using ML algorithms, specifically DT and LR, the goal was to determine the factors that contribute to customer attrition. According to the findings of the study, factors such as the transaction amount and count during transaction information as well as card information had a substantial impact on the retention of clients. Churn of Customers in Banks was a concept that was proposed by Dudyala, Anil Kumar, and Ravi [12]. Using ML algorithms such as Multilayer Perceptron (MLP), Linear Regression (LR), Decision Trees - J48, Radial Basis Function (RBF) Network in Support Vector Machines (SVM), the goal was to identify major predictive variables for Customer Churn.

Data mining strategies were used in the Credit Card Churn proposal that Guoxun Wang and his research partner developed [13]. The goal was to improve customer retention by the utilisation of machine learning techniques such as Simple Cart, J48, Random Tree, LR, Bayes Networks, Naive Bayes, Decision Table, and PART. The concept of customer churn in banks was proposed by Renato Alexandre, Thiago, and Benjamin [14]. The machine learning (ML) methods known as DT, KNN, Elastic net, and LR as well as SVM were utilised with the intention of churn prediction. According to the findings of the study, both Transactions and Transactions_DIFF (also known as Transaction Difference) had a substantial impact on the retention of customers. Attrition of customers as a strategy for commercial banks was proposed by Benlan and team [15].

The concept of customer churn was proposed in banking by Manas and Kumar [16]. The goal was to estimate customer turnover using machine learning techniques, specifically KNN, SVM, RF, and DT. The conclusion of the study is that all of the employed algorithms, with the exception of SVM, increased their accuracy when doing oversampling. The details gathered from the 15 papers related to credit card churn were summarised and placed in the Table 1. This summary of all the papers in the table was intended to give a quick overview.

Authors [17] highlighted the significance of ML in prediction, pattern recognition and error reduction across diverse fields, emphasising the impact of AI in broad domain. The paper [18] discusses the importance of text summarization in online shopping and surveys various techniques, highlighting the use of seq2seq models with LSTM and attention mechanisms for improved accuracy. The approach [19] utilized Advanced Deep Learning
with global threshold to improve E-commerce product classification, achieving high accuracy and challenging existing technology. Author [20] presented text classification algorithms for various applications and explores the use of machine learning in detecting phishing attacks.

Table 1. Summary of existing approaches.

<table>
<thead>
<tr>
<th>Ref. No.</th>
<th>Description</th>
<th>Dataset</th>
<th>Accuracy</th>
</tr>
</thead>
<tbody>
<tr>
<td>[2]</td>
<td>Predicting CCCC using AutoML tools</td>
<td>Dataset was collected from Kaggle</td>
<td>H2O-GradientBoosting-82, H2O-Random Forest-41, H2O-DeepLearning-61.09, Auto-sklearn-95.9, Auto-Keras-81.11</td>
</tr>
<tr>
<td>[3]</td>
<td>Predicting the CCCC using the ML algorithms namely, RF, LR and KNN</td>
<td>Dataset approximately consists 10000 individual pieces that are constrained to form 21 different features</td>
<td>Random Forest-95.6, LR-90.36, KNN-90.32</td>
</tr>
<tr>
<td>[4]</td>
<td>Bayesian networks, Classification and RT, Neural Networks, C5 trees, and Chi-Square automatic interaction detection (CHAID) trees</td>
<td>The dataset is made up of approximately 10000 individual parts that are organised into 21 distinct features</td>
<td>C5 Tree-96.3</td>
</tr>
<tr>
<td>[5]</td>
<td>CCA using LR, RF, Extreme Gradient Boosting, DT</td>
<td>Approximately 274542 pieces can be assembled into 14 characteristics using the dataset.</td>
<td>RF-96, SVM-94</td>
</tr>
<tr>
<td>[6]</td>
<td>CCCC using SVM and hyper-parameter tuning using Bayesian Optimization</td>
<td>The dataset has around 10127 pieces that are constrained to form 20 features</td>
<td>SVM Linear Kernel-91, Polynomial Kernel-88, Radial Kernel-86</td>
</tr>
<tr>
<td>[7]</td>
<td>Predict Credit Card Churn (CCC) using SVM, RF, DT and KNN</td>
<td>The dataset has 23 features made up of approximately 30000 pieces from UCI repository</td>
<td>SVM-96.85, Bayesian Optimization-93</td>
</tr>
<tr>
<td>[9]</td>
<td>Predict Churn in Banking Industry</td>
<td>The dataset has about 10127 pieces that are constrained to make 23 features.</td>
<td>RF-88.7 with 91.90 F1 score</td>
</tr>
<tr>
<td>[10]</td>
<td>Predict CCC using RST, RR, ANN and LS-SVM</td>
<td>The dataset was confined to 13 features with 1500 pieces.</td>
<td>DT-80.50, Ridge Regression-75.10, ANN-84.30, LS-SVM - 89.90</td>
</tr>
<tr>
<td>[11]</td>
<td>Predict Churn using ML algorithms DT and LR</td>
<td>The dataset is made up of approximately 8 million pieces confined to make 135 features</td>
<td>DT-80.50</td>
</tr>
<tr>
<td>[12]</td>
<td>To outline important predictor</td>
<td>The dataset has around</td>
<td>DT-85, Rough Set</td>
</tr>
</tbody>
</table>
3 Problem statement and objectives

3.1 Problem statement

Automated Credit Card Customer Churn Analysis (C4A) is loss of customers to competition. A problem faced by credit card companies because it is expensive to acquire a new customer and companies want to retain existing customers.

3.2 Objectives

The following section depicts the objectives of the proposed work.

- In the context of C4A analysis, the goal is to construct a model that can effectively predict which consumers are at risk of churning as well as determine the causes that are most probable to lead to churn.
- The purpose of this effort is to provide assistance to credit card firms so that they may better build focused retention tactics and personalized incentive programmes to keep clients from defecting to competing businesses.
- This paper's goal is to determine the different important driving factors that have an effect on the percentage of customers who leave a company.
- The motivation behind this paper is to evaluate churn in its most minute elements. As if you evaluate business conducted by different companies, you won't be able to see the churn in its most minute elements. As a result, conducting customer analysis based on data enables us to gain a deeper understanding.

4 Proposed method

4.1 Phases in the proposed work

Figure 2 shows the architecture diagram. The schematic of the system's architecture starts with a block on the top left corner that indicates the dataset. The dataset is a representation of the credit card churn dataset that was obtained from Kaggle. The data is

<table>
<thead>
<tr>
<th>Variables for Customer Churn using ML algorithms namely Multilayer Perceptron, LR, DT, RF, RBF and SVM</th>
<th>14814 pieces confined to make 22 features, taken from Business Intelligence Cup organized by Chile</th>
<th>Algorithms-90, MLP-89.09, LR-89.80, RF-96.69, RBF-93.04</th>
</tr>
</thead>
<tbody>
<tr>
<td>To predict credit card customer churn using data mining</td>
<td>The dataset contains 6 million credit card holders with 6 tables defining various features</td>
<td>MLP-93.25, RBF-93.11 SVM-93.22, LR-93.25 DT-95.47, RF-95.59</td>
</tr>
<tr>
<td>Predict Customer Churn in Banks using DT, KNN, LR, Elastic net, and SVM</td>
<td>The dataset contains around 500000 pieces confined to make 35 features</td>
<td>DT-78.2, KNN-77.9, Elastic Net-76.2, LR-76.2 SVM-80.3, RF-82.8</td>
</tr>
<tr>
<td>Predict customer Attrition in Commercial Banks using SVM</td>
<td>The dataset is made up of approximately 50000 pieces confined to make 11 features</td>
<td>Linear SVM-99.09 RBF SVM-98.95</td>
</tr>
<tr>
<td>Predict Customer Churn in Banks KNN, SVM, RF and DT</td>
<td>The dataset was confined to 11 features with 10000 pieces</td>
<td>KNN-83.97, SVM-79.63 DT-78.32, RF-83.6</td>
</tr>
</tbody>
</table>
given as an input to the first phase in the architecture diagram, which corresponds to the problem statement being defined. Here, an attempt to execute C4A analysis, which is a challenge that is faced by firms that deal with credit cards. The missing values can be managed in a few different ways: removing the tuple that corresponds to the missing value from the dataset; the missing value can also be filled with the mean or median of the particular attribute; alternatively, or fill it with a global constant. Exploratory data analysis, also known as EDA, is a process that consists of looking at the data, obtaining understanding about the data, and deriving insights or primary qualities from the data.

In order to carry out the implementation of the paper, several different ML models were utilized. The fourth part of this process is attempting to train the models. Within this paper, we have successfully implemented six different models. KNN, DT, and LR, as well as XGBoost, are some of the traditional methodologies that are utilized in ML models. The KNN approach is based on the premise that, in the feature space, objects that are similar to one another would often be situated near to one another. This theory forms the basis of the KNN method. When presented with a dataset that contains both input features and a target variable, the DT algorithm will segment the dataset into smaller subgroups by basing their composition on the input features. As soon as a stopping criterion, such as a maximum depth or a minimum amount of data points, is met, it divides the subsets into smaller subsets on a recurring basis. XGBoost is a specific type of gradient boosting approach that builds a number of DT and combines the findings of those trees to provide final predictions. To find the ideal set of hyper parameters for an XGBoost model, the hyper parameter optimization technique known as Cross-Validation (CV) can be used.
When modelling and analyzing data that contains a binary outcome attribute, LR is the technique of choice. It is a strategy for predicting the likelihood that an event will occur based on a collection of predictor factors. In LR, the dependent variable is categorical, meaning that there are only two possible outcomes—typically represented as 0 and 1—and it can only take on one of those values. Logical regression attempts to establish the degree of association that exists between the independent factors and the dependent variable in order to make a prediction regarding the probability that the dependent variable will take one of the two possible values.

An effort to implement two hybrid models, which are essentially a combination of several different traditional methodologies that are already in use. One of the hybrid models that has been implemented is a combination of LR and KNN, while the other hybrid model is a combination of LR and DT. The objective of this particular type of hybridization is to improve the overall performance of the model by mitigating the negative effects of LR. The range of values for the expected output in LR is between 0 and 1. Everything that lies within the range of 0 to 0.5 is considered to be 0, and everything that lies within the range of 0.5 to 1 is considered to be 1. However, with the assistance of hybridization, we are working on developing a new method of prediction. The hybrid model's new rules for categorization state that anything that lies within the range of 0 to 0.25 is to be classed as 0, and anything that falls within the range of 0.75 to 1 is to be categorized as 1. In order to categorize values in the range 0.25 to 0.75, we are further employing the use of one more associated classification model. The related model for the first hybrid model is a KNN model, whereas the corresponding model for the second hybrid model is a DT model.

Accuracy, classification report, precision, and recall are the terms that describe each of these components. The percentage of right forecasts to total predictions is what we mean when we talk about accuracy. A model's precision can be measured by the fraction of relevant examples that are present within the total number of occurrences that are predicted by the model to be positive. The percentage of relevant instances that a model was able to accurately identify is recall.

4.2 Modules description

In order to have a successful implementation of the paper, we attempted to incorporate all three modules as shown in Figure 4. The first two of the three modules were used to clean and acquire important information for the third module's training of the model. Module three was the most recent module. In its own right, the first module is concerned with the organizing and cleansing of data. It is necessary to reformat the data into a format that is more appropriate for it. This is something that can be accomplished in the very first module. During the second module, we will work on trying to understand each characteristic on its own, as well as determining the relationships between the various data qualities. The primary goal of univariate analysis is description; it collects data, summarizes that data, and looks for patterns in the data. Finding correlations between variables is part of what's involved in multivariate analysis. At the conclusion of the first two modules, we receive cleansed data that can be applied during the training phase of the third module. In this third module, we will implement several different ML modules.

5 Results and discussions

The Kaggle dataset includes 10127 rows and 23 columns altogether. The many characteristics of the data are represented in the 23 columns. The 23 columns correspond to the different features of data. The accuracy for LR is 0.846, for KNN it is 0.849, for DT it is 0.916, for XGBoost it is 0.931. For the proposed hybrid models integrating LR and
KNN, the accuracy attained is 0.9, and for the other integration of LR and DT, it is 0.92. The model built was successfully implemented and that is able to reliably forecast which consumers are at risk of cancelling their subscription. After putting a plan into action, the built model achieved a level of accuracy that was a maximum of 0.96. Through the use of EDA, were able to determine which characteristics of churn were the most significant.

Fig. 3. Flow of proposed method.

Fig. 4. Classification Report of LR Model.

Fig. 5. Classification Report of KNN Model.
Finding correlations between variables and also finding correlations between target variables. EDA was able to evaluate how strongly each of them is related with both themselves and the objective. This demonstrates that the objective “to provide assistance to credit card firms so that they may better build focused retention tactics and personalized incentive programs to keep clients from defecting to competitive business” was successfully implemented. This also contributes to the achievement of the objective of “determining the factors that effect on the percentage of customers who leave the company.” Without statistical analysis to extract valuable insights from the data, simply gathering feedback is insufficient. By building, the percentage of customer churn if the prevailing system and customer care of credit card service provider can be determined through analysis. Because, there is high chance of successfully implementing the associated objective.

### 6 Conclusion and future enhancements

Depending on the percentage of churn the companies can use any one the following tactics. The tactics are improving service and product offerings, improving overall performance of organization and help customer reduce their expenses. One could easily find percentage of churn after using classification by the model built. It is just the ratio of the total customer who are to churn to the ratio of the total customers multiplied by 100. Hence, from this a conclusion of the objective of early detect the customer at churn and help companies in retaining customers, enhancing customer loyalty, improving service and product offerings, improving reputation of brand, improving overall performance of organization, reduce expenses and so on. By implementing model which forecasts churn it help companies increase their profitability which is one of the primary objectives being implemented. This
can also help companies pro-actively approach customers at churn before losing them to competition and regretting later.

When companies have a model to predict customer at churn. Companies can take risk of trying various combinations of providing services for example, a combination providing customer services are 90 percent and expenses of customers are reduced by 10 percent. Similarly, another combination would be like providing services as 80 percent good as earlier but expenses of customers are reduced by combination.

References

8. S. Kumar, V. Ravi, V. Siddeshwar, One-class support vector machine based under sampling: Application to churn prediction and insurance fraud detection, in the Proceedings of 2015 IEEE International Conference on Computational Intelligence and Computing Research (ICCC) (2015)
10. Ning Wang, Dong-xiao Niu, Credit card customer churn prediction based on the RST and LS-SVM, in the Proceedings of the 6th International Conference on Service Systems and Service Management, Xiamen, China (2009)
11. Nie, Guangli, Rowe, Wei, Zhang, Lingling, Tian, Yingjie, Shi, Yong, Exp. Syst. Appl. 38 (2011)