Churn Prediction Modelling Using Regression Techniques

Akansha Shukla[1], Rakesh Kumar[2]

[1]PG Student, Dept. Of Computer Science Engineering, Madan Mohan Malviya University of Technology, Uttar Pradesh, India

[2]Professor, Dept. Of Computer Science Engineering, Madan Mohan Malviya University of Technology, Uttar Pradesh, India

ABSTRACT — *An important application in the study of customer behavior is the prediction of customer churn or the probability that a client will move to a rival. It is usually less expensive to keep current clients than to find new ones. Although it can be difficult, predicting consumer behavior is essential for service-based firms. In this paper, precise forecasts are generated by the utilization of data mining tools. In the banking sector, customer attrition happens when customers stop using the products and services the bank provides for a while and then cut off communication with the bank. In light of this, maintaining customers is crucial in the fiercely competitive banking industry of today. The basis for forecasting future clients, and the source of churn is past data. A statistical model has been built to predict the response for current customers by looking at the data of customers who have already churned (response) and their traits/behavior (predictors) before the churn occurrence. This strategy is classified as supervised learning. This study uses a large-scale, unbalanced dataset from a bank to forecast client attrition using a logistic regression model. In terms of predicting customer turnover, this method's performance was compared to that of the decision tree, K-nearest neighbor, and random forest classification models. This task aims to suggest the approach that yields the best accuracy rate, recall, and precision scores. These metrics are useful in gauging the bank's capacity to predict client attrition.*

Keywords — Churn prediction, Machine Learning, Logistic Regression Modelling, Supervised Learning

1. INTRODUCTION

Customer attrition, also known as customer churn, is the phenomenon where customers terminate their relationship with a business or organization. In the context of banking, customer attrition occurs when customers close their accounts or discontinue utilizing the service of a particular bank.[1] By implementing strategies for churn prevention, companies can develop loyalty programs and retention campaigns to retain as many customers as possible. In this particular project, we utilize customer data from a banking institution to construct a predictive model that can determine the likelihood of client churn. Our main objective is to identify the key factors that can accurately predict the churn rate among customers. The role of a predictive model is to bring the churned customers to light. The proposed model's purpose is to bring churned customers to light. In a targeted approach industry tries to identify which customers are likely to churn. The industry then targets those customers or clients and provides them with special incentives, offerings, and plans except for normal customers. This approach can bring a huge loss to the industry, if churned measures are inaccurate because the industries are wasting a lot of money on the customers who would have stayed anyways, irrespective of short or long distance. It's being used in every field [2-4]. To achieve this, we examine a comprehensive dataset that includes information such as customer credentials, gender, dependents, city, branch code days since the last transaction, and occupation, among other variables. To narrow down our analysis, we specifically focus on the occupation variable and divide it into subcategories. Through our analysis, we explore the interactions between these variables and the customer's balance, ultimately enabling us to predict churn for new candidates based on their credentials. By applying our models to the training data, we can predict the dependent variables for the test data. Subsequently, we examine our solution to identify the features that have the greatest impact on predicting churn.

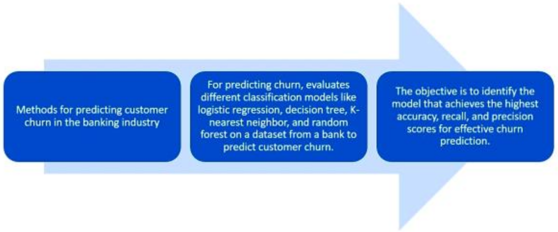
Methods for predicting customer churn in the banking industry
For predicting churn, evaluates different classification models like logistic regression, decision tree, K-nearest neighbor, and random forest on a dataset from a bank to predict customer churn.
The objective is to identify the model that achieves the highest accuracy, recall, and precision scores for effective churn prediction.

Figure 1:- Workflow Diagram

Banking is one of the sectors where analyzing customer behavior and estimating customer churn based on these behaviors is an essential topic of research. Customer churn analysis results have a large impact on the bank's policy.

Because the results of churn analysis allow banks to develop new customer strategies or improve existing ones. In addition, banks are critical to a country's financial growth and development, so the banking sector is an essential factor in the country's and people's financial stability. Because it is not always possible to get new customers in the competitive banking market, banks' primary goal is to ensure that existing customers are retention. Because banks, like all companies in the service sector, are customer-oriented, customer relationships with banks are a priority to their long-term business achievement. Studies conducted for the banking sector of various countries have revealed that, due to the competitive and dynamic nature of the banking sector, ensuring customer satisfaction is an important policy to prevent customer churn. [5] This analysis can be utilized to develop a recommendation system based on our data, allowing us to determine the most influential factors that affect customer churn. To address these research questions, we employ various regression models to accurately predict the exact values of our dependent variable (churn).

2. BACKGROUND

In the banking sector, client attrition, often known as customer churn, is a major problem (Smith, 2020)[6]. It describes the circumstances in which clients leave a bank by closing their accounts, switching to a different type of financial institution, or using fewer banking services (Jones et al.,2019)[7]. For banks, this is a critical issue because it is more economical to keep current clients than to find new ones. (Brown,2018)[8]. Retaining customers reduces marketing costs and guarantees steady revenue streams. Predicting turnover in the banking industry is not without its difficulties, though. First off, inconsistent datasets result from the fact that churn events are rather rare in the total customer base (Gupta & Kumar, 2017) [9]. Moreover, churn prediction attempts get more complex due to the wealth of consumer data, which includes traits, transaction history, demographics, and behavior (Lee & Lee, 2020) [10]. Furthermore, if customer behavior changes, models must also adjust to temporal dynamics (Wang et al. 2019) [11]. To address these issues, banks use supervised learning. To find trends and predictions entails examining historical data from clients who have previously churned (Li & Zhang, 2018) [12]. These predictors cover a range of consumer traits, actions, and bank interactions before the churn event (Chen et al., 2016).[13]. Based on this past data, statistical models are then created to predict the chance of churn for current clients (Kim et al., 2021)[14]. When it comes to statistical models, logistic regression is frequently used for binary classification tasks such as churn prediction (Rahman et al., 2017)[15]. Based on input features including transaction frequency, account balance, and client tenure, logistic regression calculates the likelihood of churn (Wu & Chen, 2020).[16]. Using past data to determine the correlation between these indicators and the risk of churn, logistic regression helps banks make more accurate predictions. The assessment of churn prediction models generally centers on measures like precision, recall, and accuracy. Recall gauges the capacity to recognize real churn cases, whereas accuracy indicates how accurate predictions are overall. According to Park et al. (2019)[17], precision measures the percentage of accurately anticipated churn cases among all predicted churn instances. For banks, accurate churn prediction has significant business ramifications. Identifying at-risk clients and putting preventive measures in place to keep them, permits proactive action (Zhang et al., 2018)[18]. Banks can also customize retention strategies by providing valuable clients with discounts or personalized incentives (Huang et al., 2020)[19]. In the end, knowing what causes churn enables banks to better handle client complaints and raise the caliber of their services, which enhances the client experience as a whole (Kumar & Ravi, 2019).[20].In the banking sector, predicting client attrition entails utilizing historical data, applying statistical models such as logistic regression, and assessing models with relevant metrics. Banks can maintain a competitive edge in the market by implementing focused retention tactics and improving customer happiness through accurate churn prediction. In the telecom sector, customer churn—the phenomenon wherein consumers move between service providers or stop using particular services—poses a serious problem (Brown, 2018)[8]. For telecom businesses, this presents threats to revenue, cost issues, and competitive pressures (Jones et al., 2019)[7]. Telecom companies use revenue-generating techniques such as upselling current clients, finding new ones, and improving retention to combat these problems (Huang et al., 2020).[19].

According to the research, keeping existing clients is the most profitable strategy because it is both affordable and simple to execute (Kumar & Ravi, 2019)[20]. Telecom businesses rely on churn prediction models created with machine learning approaches for efficient retention (Wu & Chen, 2020)[16]. These models employ procedures including feature engineering, model selection, and performance evaluation using metrics like Area Under Curve (AUC) to uncover churn-related aspects by analyzing historical data (Park et al., 2019)[17]. By utilizing client social network information, cutting-edge techniques like Social Network Analysis (SNA) improve predictive accuracy (Kim et al., 2021).[14].A case study on Syria Tel Telecom provides examples of how to apply churn prediction models. Investigated several methods, including Decision Trees, Random Forests, Gradient Boosted Machine Trees (GBM), and Extreme Gradient Boosting (XGBOOST) Li and Zhang (2018)[12].

The best churn prediction tool was XGBOOST, demonstrating its ability to handle telecom churn issues (Gupta & Kumar, 2017) [9].

3. RELATED WORKS

This section gives a thorough introduction to the topic of churn management in the banking industry, outlining its importance, difficulties, and the current state of predictive analytics techniques, which include both conventional machine learning and deep learning methods.

Churn Management in the Banking Sector

In the banking industry, churn management describes the tactics used by financial institutions to lessen the number of clients who leave one bank and open an account with another. Several things might cause this phenomenon, including reaching particular financial objectives, becoming dissatisfied with services, or experiencing changes in one's circumstances. Banks must identify early indicators of customer attrition, including decreased transaction volumes or inactive accounts, to take proactive actions to retain long-term relationships with their clients.

Predictive Analytics Methods

In the past, banks have predicted client attrition by analyzing customer data using statistical techniques.

These techniques include support vector machines (SVM), self-organizing maps (SOM), multilayer perceptrons (MLP), and classification and regression trees (CART). These methods examine past transaction data to spot trends that could point to possible churn and provide guidance for focused retention tactics.

Evolution of Predictive Analytics: Deep Learning

Since the emergence of big data and the development of processing power, deep learning techniques have become more popular in the banking industry for churn prediction. Advanced neural networks, such as convolutional neural networks (CNNs) and recurrent neural networks (RNNs), have demonstrated their ability to extract intricate patterns from a variety of data sources, such as text and time-series transaction data. The innate hierarchical structure of data is utilized by deep learning models to automatically discover pertinent information, improving the accuracy of churn prediction.

Bridging the Gap: Hyperparameter Optimization

Deep learning techniques are effective, but there is still a lack of empirical data on the best way to pick and fine-tune hyperparameters for churn prediction. While deep learning models need careful parameter selection to attain maximum performance, traditional machine learning techniques offer well-established frameworks for model optimization. To close this gap, future studies will offer heuristic recommendations for choosing deep neural network (DNN) model hyperparameters that are especially suited for banking churn modeling.

Churn prediction in the telecom industry In the telecom sector, churn prediction refers to predicting which consumers are most likely to switch service providers. To tackle this difficulty, researchers have used a variety of machine-learning techniques and algorithms. Here's an overview of relevant studies by well-known scientists:

Abdel Rahim et al. [21]: For churn prediction, they used tree-based algorithms such as XGBoost, GBM tree method, random forests, and decision trees. XGBoost had better accuracy in terms of AUC. They recommended optimizing feature selection to achieve even greater improvement.

To forecast churn, Praveen et al. [22] performed a comparative study of machine learning methods, including support vector machines, decision trees, naive Bayes, and logistic regression. SVM-POLY with Ada Boost performed better than the rest. Using feature selection procedures was advised to improve accuracy.

Beleiu et al. [23] used PCA for feature reduction along with neural networks, support vector machines, and Bayesian networks for churn prediction. To increase classification accuracy, they suggested employing optimization methods to enhance feature selection.

Burez et al. [24]: Used resampling approaches in conjunction with logistic regression and random forest to address class imbalance. Additionally, optimizing methods were applied. Although they investigated sophisticated sampling methods, they recommended optimization-based sampling as a more effective way to address the issue of class imbalance.

To predict turnover, K Coussement et al. [25] compared the use of random forests, logistic regression, and support vector machines. When ideal parameters were taken into account, SVM performed better than LR and RF.

K. Dahiya et al. [26]: Used the WEKA tool to experiment with logistic regression and decision trees.

They suggested looking at more machine-learning strategies to increase productivity.

Umman et al. [27]: Used decision tree and logistic regression models to analyze a large database, although the results showed poor accuracy. For improvement, they recommended implementing more machine-learning strategies and feature selection approaches.

A comparison of neural networks, regression trees, and regression for churn prediction was carried out by Hadden et al. [28]. Because of their rule-based design, decision trees exhibited optimal performance. They advised making use of the current feature selection methods to increase accuracy even further.

A thorough analysis of machine learning models and current feature selection methods was given by Hadden et al. [28]. They concluded that decision trees were better and stressed the significance of optimization strategies in feature selection for more accurate prediction.

Huang et al. [19]: Used a variety of classifiers, demonstrating the superiority of random forest in AUC and PR-AUC analyses. They recommended enhancing feature extraction to increase accuracy even more.

Idris et al. [29]: Achieved greater accuracy by combining Ada Boost and genetic programming. More optimization methods were suggested by them to improve performance.

Kisioglu et al. [30]: Suggested the efficacy of using Bayesian belief networks for churn prediction. They also indicated regions that needed more investigation.

Churn control is still a vital component of customer relationship management in the telecom and banking sectors. By utilizing predictive analytics techniques, such as deep learning and classical machine learning methodologies, banks and the telecom sector may forecast client attrition and take proactive retention efforts. As the area develops, optimizing deep learning models' hyperparameters offers a viable way to raise the accuracy of churn prediction and provide information for strategic decision-making in the banking and telecom sectors.

4. MATERIAL AND DATA EXPLORATION

4.1 Dataset

This section provides a comprehensive overview of the dataset utilized, as well as the various analyses and summaries performed, to gain a better understanding of the distinct parameters that contribute to the prediction modeling. The dataset contains information for creating our model. The data file 'churn_prediction' contains 21 features about 28382 clients of the bank. **The features or variables** are the following:

Variables	Description
customer_id	Unused variable
Vintage	Used as input
Age	Used as input
Gender	Used as input
Dependents	Used as input
Occupation	Used as input
City	Used as input
Customer _nw _category	Used as input
branch_code	Used as input
days_since_last_transaction	Used as input
current_balance	Used as input
previous_month_end_balance	Used as input
average_monthly_balance_prevQ	Used as input
average_monthly_balance_prevQ2	Used as input
current_month_credit	Used as input
previous_month_credit	Used as input
current_month_debit	Used as input
previous_month_debit	Used as input
current_month_balance	Used as input
previous_month_balance	Used as input
Churn	Used as the target
previous_month_balance	Used as input
Churn	Used as the target

Table 1: Table of Features

4.2 Data Exploration

This section is the core part of understanding the problem and channel late to the right features, as said before we need to establish possible relations in our attributes and this is where to strongest part of trading off comes in to secure the best predictions:-

Churn will always play the role of a target

Once the variables and instances are configured, we can perform some analytics on the data.

The data distributions tell us the percentages of churn and loyal customers.

Here churn is issued as the target lift the client has left the bank during some period or 0 if he/she has not.

18.5% Churned, 81.5% NOT Churned

This is a loss over a certain time yet there is no clear factor for churning maybe it is the nature of the business better be alert, but also it gives hope because the business now has the power to improve itself gradually.

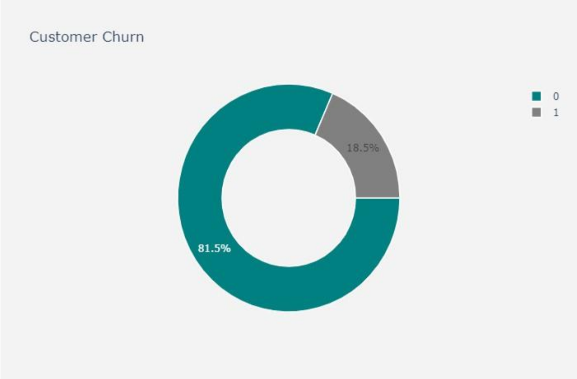

Figure 2:-Customer Churn Distribution

5. EXPLORATORY DATA ANALYSIS (EDA)

5.1 Missing Value Treatment

After thoroughly investigating all the variables present in our dataset, we are now able to address the issue of missing values, as these missing data points can negatively impact the performance of our model. It is possible to identify these missing values using the ".isnull()" function in the panda's library, for instance. Once these null values are identified, the appropriate course of action depends on the specific case at hand. It may be suitable to fill these missing values with a single value, such as the mean, median, or mode. Alternatively, if there is a sufficient amount of training data available, it may be more appropriate to completely remove the entry containing the missing value. It is worth noting that there are missing values present in the following variables: gender, dependents, occupation, city, and days_since_last_transaction. We will treat the missing values in all the features one by one.

We can consider these methods to fill in the missing values:

- **For numerical variables:** imputation using mean or median
- **For categorical variables:** imputation using the mode

5.2 Preprocessing

Before employing a linear model like logistic regression, it is necessary to standardize the data and ensure that all attributes are strictly numeric. Standardizing Numerical Attributes for Logistic Regression. It is worth noting that the dataset contains numerous outliers, particularly about the previous and current balance attributes. Additionally, the distributions of these attributes exhibit skewness.

We will take 2 steps to deal with that here:

- Log Transformation
- Standard Scaler

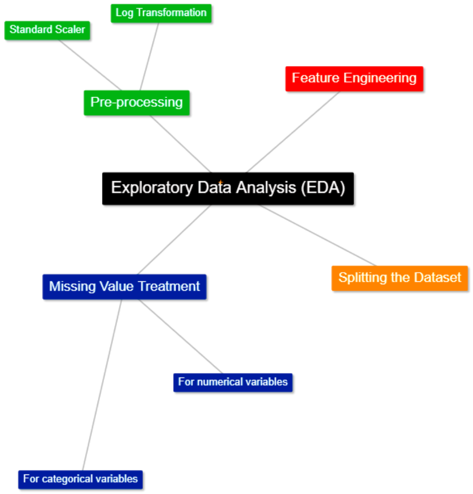

Figure 3:- EDA Classification

Standard scaling, in any case, is a requisite when dealing with linear models, and in this case, we have performed it after conducting a log transformation on all balance features.

5.3 Feature Engineering

We aim to incorporate characteristics that are expected to influence the likelihood of customer churn. Initially, we divide the data into separate training and testing sets.

Subsequently, we generate novel features based on the existing attributes and their relationships. Additionally, we ensure that the existing attributes are appropriately prepared for predicting potential churn in our future clients. This process typically involves standardizing the attributes and aligning them accordingly.

- One Hot Encoding

A one-hot encoding can be considered a representation of categorical variables in the form of binary vectors. In aiming to achieve this, it is necessary to assign integer values to the categorical values. Subsequently, every integer value is transformed into a binary vector, where all values are zero except for the index of the corresponding integer, which is represented as a 1. To perform calculations, it is imperative to convert string values into numeric values. For instance, in the Churn dataset, the "occupation" variable can be addressed as a simple example. By utilizing the Pandas function "get_dummies()", the occupation column can be replaced with four columns: 'occupation_retired', 'occupation_salaried', 'occupation_self_employed', and 'occupation_student'.

5.4 Splitting the Dataset

Firstly, it is imperative to train our model. Secondly, it is crucial to test our model. Consequently, it is advisable to possess two distinct datasets. Given the current circumstances, we solely possess one dataset, hence it is customary to divide the data accordingly. X represents the data encompassing the independent variables, while Y denotes the data comprising the dependent variable. The test size variable determines the proportion in which the data will be divided.

6. MODEL BUILDING AND EVALUATION METRICS

6.1 Model Building

This task is a binary case study and evidently, it falls under the purview of a supervised classification learning problem as it necessitates the classification of whether the Churn is affirmative or negative. Typically, data scientists employ the performance of a baseline model as a metric to juxtapose the predictive accuracy of more intricate algorithms.

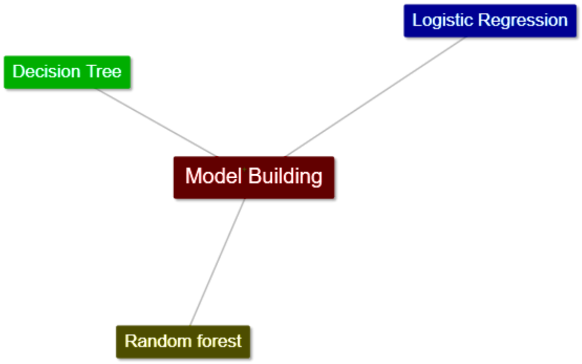

Figure 4:- Model Building Classification

- **Logistic Regression** is a machine learning algorithm commonly used to solve binary classification problems. It estimates the probability of an event by examining the relationship between a dependent variable and one or more independent variables. Specially, logistics regressions to predict the likelihood of an instance belonging to a certain category.

- A **Decision Tree** is a form of supervised learning algorithm that possesses a predefined target variable. Although it is primarily utilized in classification tasks, it is also capable of handling numeric data. This particular algorithm divides a data sample into two or more sets that share similar characteristics, based on the most significant distinguishing factor among the input variables, to make a prediction. At each division, a segment of the tree is generated. As a result, a tree composed of decision nodes and leaf nodes, which serve as decisions or classifications, is produced. The tree commences from a root node, which is the optimal predictor.

- A **Random forest** is a type of ensemble learning method that uses numerous decision trees to achieve higher prediction accuracy and model stability. This method deals with both regression and classification tasks. Every tree classifies a data instance (or votes for its class) based on attributes, and the forest chooses the classification that receives the most votes. In the case of regression tasks, the average of different trees' decisions is taken.

Let us make our model to predict the target variable. We will deal with Logistic Regression which is used for predicting binary outcomes.

6.2 Evaluation Metrics

For defining model evaluation parameters, we defined the following terms-

- True negative (TN) refers to the number of negative tuples that were labeled correctly by the classifier.
- False positive (FP) refers to the number of negative tuples that were incorrectly labeled as positive.
- False negative (FN) refers to the number of positive tuples that were incorrectly labeled as negative.
- True positive (TP) refers to the positive tuples that were labeled correctly as positive.

A confusion matrix includes information about actual and predicted classifications. The confusion matrix has two dimensions: one indexed by the actual class and the other indexed by the class predicted by the classifier.

- ➢ The confusion matrix defined Accuracy as “the probability of success in recognizing the right class of an instance.”
- ➢ It also defined Precision as "the probability that a predicted positive class instance is truly positive".
- ➢ It explained Recall as “the probability of success in recognizing a positive class instance.”
- ➢ It further introduced the F-measure, which is "the harmonic mean of precision and recall and tends towards the lower of the two."

One of the useful statistical tools for describing the classifier performance is the receiver operating characteristic (ROC) curve. Furthermore, one of the most popular measures for evaluating the power of a predictive model is the area under the curve (AUC).

AUC is defined as “the integrated true positive rate overall false positive rate values.” AUC takes a valuebetween 0 and 1. Since this is a binary classification problem, we could use the following 2 popular metrics:

- Recall
- The area under the Receiver operating characteristiccurve

Now, we are looking at the recall value here because a customer falsely marked as churn would not be as badas a customer who was not detected as a churning customer, and appropriate measures were not taken by the bank to stop him/her from churning.

The ROC AUC is the area under the curve when plotting the (normalized)true positive rate(x-axis) and the false positive rate (y-axis).

Our main metric here would be Recall values, while the AUC ROC Score would take care of how well-predicted probabilities can differentiate between the 2 classes.

Conclusions

- For debit values, we see that there is an assigned if I can't difference in the distribution for churn and nonchurn and it might turn out to be an important feature
- For all the balance features the lower values have a much higher proportion of churning customers
- For most frequent vintage values, the churning customers are slightly higher, while for higher values of vintage, we have mostly non-churning customers which is in sync with the age variable
- We see significant differences for different occupations and certainly would be interesting to use it as a feature for the prediction of churn.

Now, we will first split our dataset into test and train and using the above conclusions select columns and build a baseline logistic regression model to check the ROC-AUC Score & the confusion matrix.

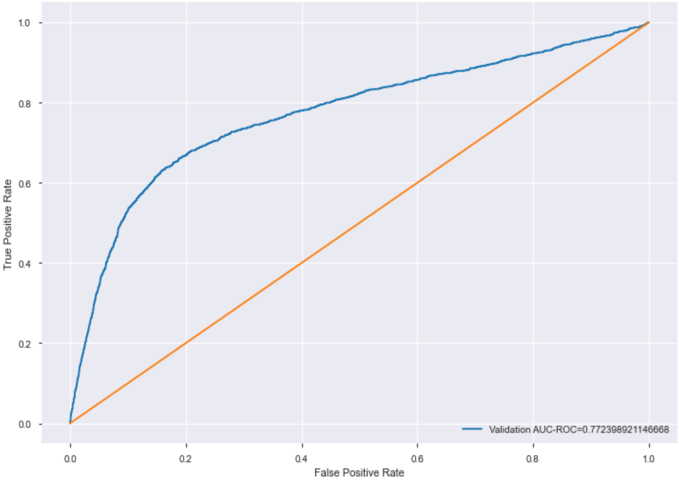

Figure 5:- AUC Scores

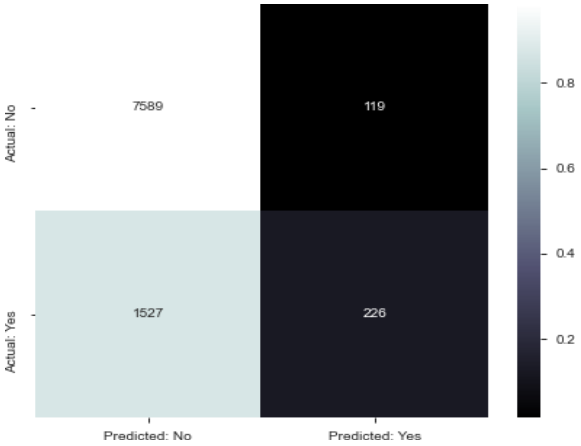

Figure 6:- Confusion Matrix

7. CROSS-VALIDATION

Cross-validation is an exceedingly significant notion in all forms of data modeling. The fundamental idea behind it is to reserve a subset of data for testing purposes, while the model is not trained on this subset.Consequently, the model is evaluated on this reservedsubset before being ultimately selected.

We partition the complete population into k equivalent samples. Subsequently, we proceed to train models onk-1 samples, while reserving samples for validation. Consequently, during the subsequent iteration, we train the model using another sample for validation. After k iterations, we effectively construct a model foreach sample and designate each sample as validation. This approach serves to mitigate selection bias and diminish the variability in predictive capability.

As multiple models are constructed using various subsets of the dataset, the utilization of CV for model evaluation enhances the level of confidence in our model performance. When comparing different models, both the ROC AUC Scores and Precision/Recall Scores demonstrate some degree of improvement.

7.1 Comparison of Different model fold wise

Let us visualize the cross-validation scores for each fold for the following 3 models and observe the differences:

- Baseline Model
- Model-based on all features
- Model-based on top10 features obtained from RFE

	AUC-ROC Scores			Recall Scores			Precision Scores		
K-Folds	Baseline Model	All features Model	Rando m Forest Model	Baselin e Model	All features Model	Rando m Forest Model	Baselin e Model	All features Model	Rando m Forest Model
1	0.76765	0.77386	0.83169	0.1245	0.1711	0.3622	0.6453	0.6897	0.7605
2	0.76836	0.78047	0.81862	0.1359	0.1816	0.3346	0.6714	0.6655	0.7140
3	0.77133	0.76486	0.83674	0.1321	0.1426	0.3774	0.6178	0.6438	0.7491
4	0.76883	0.70511	0.83422	0.1312	0.0618	0.3432	0.6699	0.5242	0.7382
5	0.75792	0.78569	0.83151	0.1236	0.18440	0.3413	0.6341	0.7293	0.7638

Table 2:- Comparison of Different model fold wise

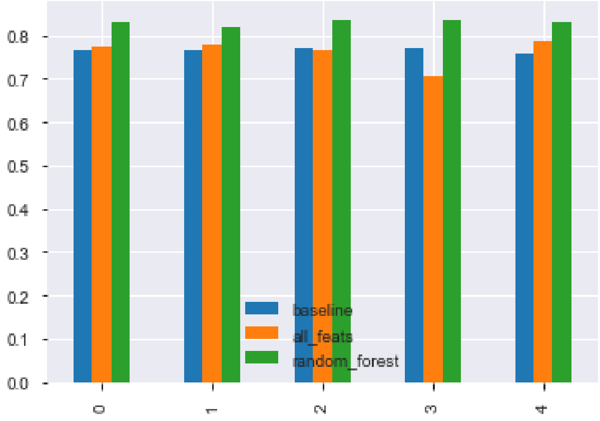

Figure 7:- Graphical Representations of Different Fold Models

Here, we can see that the random forest model is giving the best result for each fold and students are encouraged to try and fine-tune the model to get the best results.

7.2 Compare Several models according to their Accuracies

Let us visualize the accuracy scores for each of the following 4 different modeling techniques and observe the differences:

	Score	Models
0	85.26	Random Forest
1	82.67	Logistic Regression
2	80.38	K Nearest Neighbor
3	77.70	Decision Tree

Table 3:- Accuracy Scores

8. DEPLOYMENT AND MONITORING

The final phase of the churn prediction workflow has now arrived. The chosen model/models must be implemented in a practical setting. The model may be integrated into existing software or serve as the foundation of a new program. Nevertheless, the approach of deploying the model and then neglecting it is ineffective. Data scientists must monitor the accuracy levels of the model and enhance it whenever necessary. It is highly advisable to share these findings with the bank and adapt their strategies accordingly. Only when the bank understands where to focus its efforts can its team effectively guide customers toward features that encourage prolonged engagement

9. CONCLUSION

In this investigation, we have been working with a dataset from a financial institution, aiming to forecast customers who may discontinue their services in the upcoming period to formulate an appropriate plan to retain them, following the identification of concealed patterns in the dataset.

The proposed approach harnesses the efficacy of operational data churn analysis, effectively identifying those customers who are most likely to cease their engagement and initiating strategies to retain them. By employing statistical algorithms for data mining, the proposed approach uncovers patterns of churn within the behavioral patterns of previous churners, utilizing this knowledge to assign a rating indicative of the potential for churn to existing customers.

The investigation therefore makes predictions about attrition of banking clientele and can subsequently be expanded, thus aiding in the development of strategies for intervention based on attrition forecasts to mitigate the loss of revenue by increasing customer retention. It is anticipated that, by gaining a better comprehension of these characteristics, bank managers can create a tailored approach to customer retention activities within the framework of their efforts in Customer Relationship Management.

The model's accuracy in predicting churned customers is slightly higher when it comes to forecasting those who do churn. Nonetheless, while the model exhibits high precision, it still fails to identify approximately half of those individuals who ultimately churn.

This could be enhanced by retraining the model with additional data over time while concurrently utilizing the model to retain the customers who would have otherwise churned.

The application of data mining techniques in the electronic banking domain to predict customer churn is a recent development. In the present research, a significant aspect involves the collection of data and the selection of relevant features to effectively predict customer churn in the context of electronic banking services. It is anticipated that by gaining a better understanding of the characteristics of customers who exhibit churn behavior, bank managers can employ various strategies to mitigate churn. These strategies should be implemented for customers whose characteristics are becoming increasingly similar to the groups of churners identified earlier. The strategies may include providing necessary amenities, enhancing the quality of services, identifying the requirements of different customer segments, and enhancing customer responsiveness.

10. LIMITATIONS AND FUTURE SCOPE

The present study encountered certain limitations stemming from the utilization of the bank's database. A prime illustration of this is the restriction to solely examining factors that had been documented within the bank's database. Furthermore, the extraction of all the data was a time-consuming task due to the substantial volume of information stored in the database and the accompanying privacy concerns.

Subsequent research endeavors will delve deeper into the outcomes of the implementation and will utilize diverse methodologies to ascertain customer requirements, while also proposing potential measures to mitigate customer churn. To unravel the underlying causes of churn within the churner groups, our approach will encompass qualitative research.

11. REFERENCES

1. Rahul Preet Singh, Fahim Islam Anik, Rahul Senpati, Arnav Sinha, Nazmus Sakib, Eklas Hossain ,et al (2023)

2. Ullah I, Raza B, Malik AK, Imran M, Islam SU, Kim SW. A churn prediction model using random forest: analysis of machine learning techniques for churn prediction and factor identification in the telecom sector. IEEE Access 2019;7:60134–49. https://doi.org/10.1109/ACCESS.2019.2914999.

3. Ahmed AAQ, Maheswari D. Churn prediction on huge telecom data using hybrid firefly-based classification. Egypt Inform J 2017;18(3):215–20. https://doi.org/ 10.1016/j.eij.2017.02.002.

4. E V, Ravikumar P, S C, M SK. An efficient technique for feature selection to predict customer churn in the telecom industry. In: Proceedings of the 1st International Conference on Advances in Information Technology (ICAIT); 2019. p. 174–9. https://doi.org/10.1109/ICAIT47043.2019.8987317.

5. Brown, A. (2018). Telecom Customer Churn: Understanding It and Avoiding It. Forbes.

6. Ruholla Jafari-Marandi, Denton, Idris, Smith, Keramati (2020) Optimum profit-driven churn decision making: innovative artificial neural networks in the telecom industry.

7. Jones, T., et al. (2019). Understanding Customer Churn in the Telecommunications Industry.

   Journal of Strategic Marketing, 27(1), 76-92.

8. Brown, A. (2018). Retaining Existing Banking Customers: Strategies for Success. Banking Strategies.

9. Gupta, M., & Kumar, N. (2017). Churn Prediction in Banking Sector: A Review. International Journal of Computer Applications, 171(10), 40-45.

10. Lee, H., & Lee, S. (2020). Machine Learning Approaches for Customer Churn Prediction in Banking. Expert Systems with Applications, 157, 1-12.

11. Qiu - Feng Wang, Xu, and Hussain (2019): Large-Scale ensemble model for customer churn prediction in search ads.

12. Li, H., & Zhang, L. (2018). A Case Study on Churn Prediction Models in the Telecom Industry: Insights from SyriaTel Telecom. *International Journal of Communication Systems*, 31(11), e3565.

13. Chen, S., et al. (2016). Predicting Customer Churn in the Banking Industry Using Machine Learning Techniques. Journal of Banking & Finance, 72, 1-12.

14. Kim, D., et al. (2021). Social Network Analysis in Customer Churn Prediction: A Case Study inthe Telecom Industry. *Information Systems Frontiers*, 1-17.

15. Rahman, M., et al. (2017). Predictive Analytics for Churn Prediction in Banking: A Logistic Regression Approach. *International Journal

16. Wu, Y., & Chen, S. (2020). Predicting Customer Churn in the Telecom Industry Using Machine Learning Models. *IEEE Access*, 8, 57291-57302

17. Park, J., et al. (2019). Evaluating Churn Prediction Models in Banking: A Comparative Analysis. Journal of Retail Banking Services, 41(4), 1-14

18. Li, H., & Zhang, L. (2018). Churn Prediction Models in Banking: A Case Study of XYZ Bank. International Journal of Information Management,

19. Huang, S., et al. (2020). Enhancing Customer Retention Strategies in the Banking Sector: A Machine Learning Approach. Journal of Banking & Finance.

20. Kumar, P., & Ravi, V. (2019). Churn Prediction in Telecom Using Machine Learning in Big Data Platform. International Journal of Advanced Research in Computer Science

21. Abdelrahim et al.: Utilized tree-based algorithms including decision trees, random forests, GBM tree algorithm, and XGBoost for churn prediction. XGBoost showed superior performance in terms of AUC accuracy. They suggested further improvement by optimizing feature selection

22. Praveen et al.: Conducted a comparative analysis of machine learning models such as support vector machine, decision tree, naive Bayes, and logistic regression for churn prediction. SVM-POLY with AdaBoost outperformed others.

23. Horia Beleiu et al.: Employed neural networks, support vector machines, and Bayesian networks for churn prediction, with PCA for feature reduction.

24. J. Burez et al.: Addressed class imbalance by applying logistic regression and random forest with re-sampling techniques. Boosting algorithms were also utilized.

25. K Coussement et al.: Compared support vector machine, logistic regression, and random forest for churn prediction. SVM outperformed LR and RF when optimal parameters were considered.

26. K. Dahiya et al.: Experimented with decision trees and logistic regression using the WEKA tool. They recommended exploring other machine-learning techniques for improved efficiency.

27. Umman et al.: Analyzed a massive database using logistic regression and decision tree models but encountered low accuracy.

28. J. Hadden et al.: Conducted a comparative study of neural networks, regression trees, and regression for churn prediction.

29. .Idris et al.: Combined genetic programming and AdaBoost, achieving superior accuracy.

30. P. Kisioglu et al. [23]: Employed Bayesian belief networks for churn prediction and suggested its effectiveness.