Predicting Customer Churn In The Banking Sector: Using Ensemble And Machine Learning Techniques To Improve Business

A PROJECT REPORT

Submitted by

NAVYA BALASUNDARAM
(2116210701177)

in partial fulfillment for the award of

the degree of

BACHELOR OF ENGINEERING

in

COMPUTER SCIENCE AND ENGINEERING

RAJALAKSHMI ENGINEERING

COLLEGE ANNA UNIVERSITY,

CHENNAI

MAY 2024

RAJALAKSHMI ENGINEERING COLLEGE, CHENNAI

BONAFIDE CERTIFICATE

Certified that this Thesis titled **“Predicting Customer Churn In The Banking Sector: Using Ensemble And Machine Learning Techniques To Improve Business**” is the bonafide work of “**NAVYA BALASUNDARAM (2116210701177)”** who carried out the work under my supervision. Certified further that to the best of my knowledge the work reported herein does not form partof any other thesis or dissertation on the basis of which a degree or award was conferred on an earlier occasion on this or any other candidate.

SIGNATURE

Dr . S Senthil Pandi M.E.,Ph.D.,

PROJECT COORDINATOR

Professor

Department of Computer Science and Engineering

Rajalakshmi Engineering College

Chennai - 602 105

Submitted to Project Viva-Voce Examination held on_____________________________

Internal Examiner **External Examiner**

Predicting Customer Churn In The Banking Sector: Using Ensemble And Machine Learning Techniques To Improve Business

Navya Balasundaram – 210701177

Computer Science and Engineering

Rajalakshmi Engineering College

Abstract

Customer churn is the concept where customer suddenly leave a business to commit elsewhere. There will be a sudden reduction in the number of customers and will lead to fall in the business. This is most common in the banking and telecom industries. Hence, retaining the majority of existing customers, is a more ideal strategy than looking to find new customers in bulk. This paper aims to utilize several machine learning models including classification as well as clustering to predict this customer churn for an organisation suing customer data. Out of all the models used, K means clustering used with Random forest classifier seemed to give the highest accuracy.

1. Introduction

In the banking sector, the amount of customer data being utilized is very large. A lot of data is collected each time they visit the bank. While this data has the potential to get into the wrong hands, it can also be beneficially used to predict customer churn. Data such as their estimated salary, what type of card they own, how many points they have earned, what kind of complaints have they put forward, what issues they have faced will be crucial in predicting their attrition. A customer may have many reasons to leave a bank but if it can be predicted early on, the bank can enhance their customer service and prevent it from happening. It holds true for all banks and all customers. Reasons for customer attrition can be many. It could be any one of the following: competition banks could have better offers, bank location might be inconvenient, bank might have high fees, might have bad customer service or few customer benefits. If a prediction can be done early on using machine learning, all of this can be improved by bank.

Another important thing to consider is that, it is natural to wonder if predicting the coming of a new customer based on bank data could be done and facilities could be improved to attract new customers, But according to studies and real world data, it comparatively exponentially less expensive to predict customer churn using customer data and it is even more useful since customer data is being used directly, it will with providing personalised customer service and retention strategies.

In this paper, several methods are used to perform this prediction. Some of the best and most popular methods for this problem statements are, Random Forest Classification, XGBoost classification, Logistic Regression, Support Vector Machines and K means clustering, finally an ensemble model is also utilised combing many of the above models. It is safe to say that there are many ways to achieve high accuracy and hold on to existing customers, providing good service while also walking on a path to expect new customers.

Machine learning and may of its techniques have shown to be highly useful in any business and corporate sector. It's many techniques to analyse data on a large scale and also perform predictions to produce key points to improve the functioning of the organisation has been remarkable. This paper only highlights the models used here. A dataset with customer details from a bank is used. With big data becoming huge, machine learning puts forward many ways to attain key results from such huge data.

2. Literature Review

Because it has a direct impact on profitability and client retention efforts, customer churn prediction is an important field of research for the banking sector. Numerous predictive modelling approaches have been investigated by researchers, including machine learning algorithms like artificial neural networks (ANN), logistic regression, decision trees, random forests, and support vector machines (SVM). For churn prediction, each method has specific benefits in terms of scalability, interpretability, and accuracy. The process of developing a model heavily relies on feature selection. Research has shown how important it is to choose important data including account activity, transaction history, customer demographics, and customer service interactions. By determining the most important elements causing customer attrition, feature importance analysis allows banks to concentrate their efforts on specific retention strategies.

As ensemble methods are good at mixing various learning algorithms to improve forecast accuracy, they have gained popularity in the field of churn prediction. Boosting, stacking, and bagging are some of the techniques used in order to catch the complex trends that underlie consumer behaviour. Furthermore, class imbalance—the presence of unequal numbers of churners and non-churners in the dataset—is a common issue in churn prediction. To solve this problem and increase the adaptability of prediction models while lowering bias, researchers have suggested techniques like oversampling, undersampling, and synthetic data generation. These methods are crucial for ensuring accurate forecasts and helping to make important decisions in the banking organisations.

In order to obtain better churn forecast accuracy, recent studies explore hybrid approaches that combine machine learning techniques with traditional statistical techniques. Hybrid models combine the best features of several approaches to enhance prediction performance and offer useful information for client retention strategies. Additionally, efficient and effective churn prediction models that can handle massive amounts of streaming data are made possible by developments in big data and real-time analytics. Based on dynamic behaviour patterns, real-time prediction enables banks to boost client interaction and use timely retention measures.

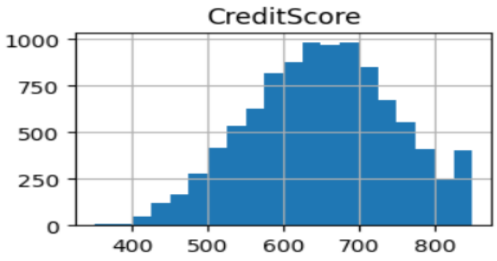

Fig 1. Analysis of customer data.

3. Proposed methodology

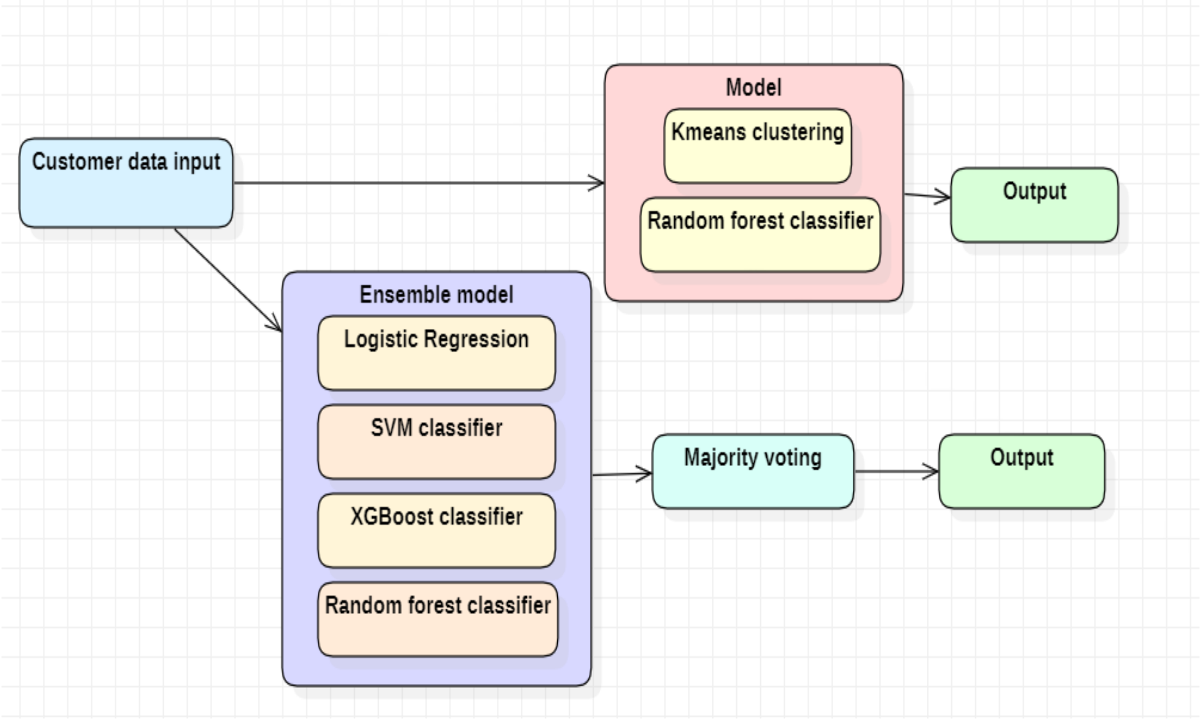

Fig 2. Working of the models

Prediction

Two different techniques are used, one is an ensemble model, utilizing four other models: Logistic Regression, SVM classifier, XGBoost classifier and Random forest classifier. All of them work separately on the dataset and give each of them give their own outputs for the prediction. Then a majority voting classifier, combines all of their outputs, takes the majority vote and produces the final output. While this seems to be very efficient, the accuracy is not that great due to the fact that none of these models perform clustering, only classification. While on the other hand, the other model produces results with significantly higher accuracy. This is because, this model utilizes clustering and also classification.

Dataset Description

This dataset consists of many useful features. It contains the usual data that a bank would maintain after a customer visit.

This dataset contains a total of 10,001 instances and 15 features. It contains the following features: CreditScore, Gender, Age, Geography, Tenure, Balance, NumofProducts,HasCreditCard,IsActiveMember, EstimatedSalary, Exited, Complain, Satisfaction, CardType and PointsEarned. Some of these features and a little useless. The models use most of these features to produce best results.

Dataset Preprocessing

A dataset must be cleaned, consistent, and ready for analysis before it can be used to train machine learning models. This process is known as data preparation. It requires a number of activities, including managing outliers, scaling numerical characteristics, encoding categorical variables, and handling missing values.

Keeping track of missing values in the dataset is a crucial component of data preparation. For example, there can be missing entries in columns like "Balance" or "CreditScore." Based on the available data, these missing values can be replaced using statistical methods like mean or median restoration. This makes it easier to guarantee that the dataset will continue to be fair and full for further research.

Numerical feature scaling is crucial to bringing all features to a similar scale and avoiding dominance of some features over others throughout the model-training process. Features such as "Age" and "Balance," for example, may have varying scales. These features can be normalized into a consistent range that is appropriate for modeling using methods such as standardization or Min-Max scaling.

Additionally, in order to maintain the dataset's reliability and integrity, outliers must be addressed. Model performance can be greatly affected by outliers in numerical features like "Age" or "Balance." By identifying and managing outliers properly, statistical techniques like Z-score and IQR (Interquartile Range) can help to guarantee that the dataset is fair and supports precise model predictions.

Applying Machine Learning

Firstly, the useless columns such as RowNumber, CustomerId, Surname, etc are dropped, since they are not useful to the model. This is called feature selection. Only the needed features are even sent into the model. The first model is the ensemble model, which consists the four models. They together form an ensemble. Each of the models are trained separately. The dataset is broken into train and test sets. The train set is consumed by all four models. They all are fit over the dataset after label encoding takes place. A cross validation has the capability to produce better results.

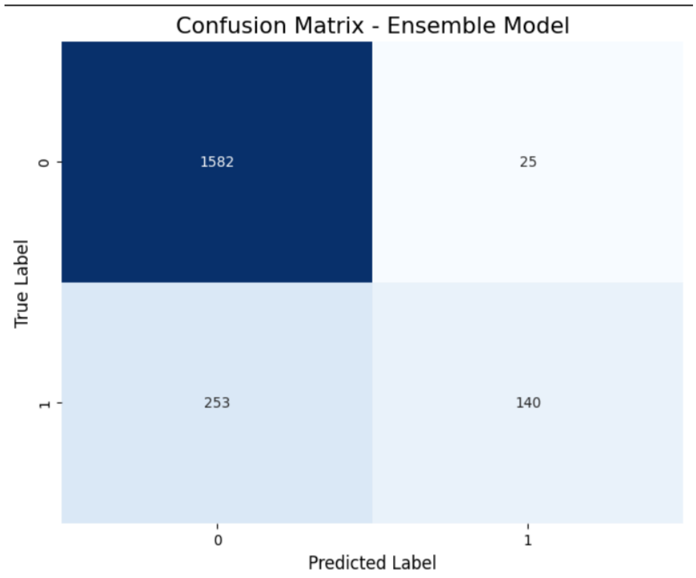

Fig 3. Confusion matrix of ensemble model.

There are techniques that can be utilized to get better results from this model but, there is actually an alternative method that produces remarkable better results.

It is the Kmeans clustering algorithm working together with random classifier. This model produces magically good results. This is because, Kmeans clustering is performed first and the instances are clustered and each cluster is given a label. Now this cluster label, is given as an extra feature into the Random Forest Classifier. This exponentially improves the performance and very high accuracy. This shows that the best way to tackle problem statements relating to a huge number of people is clustering.

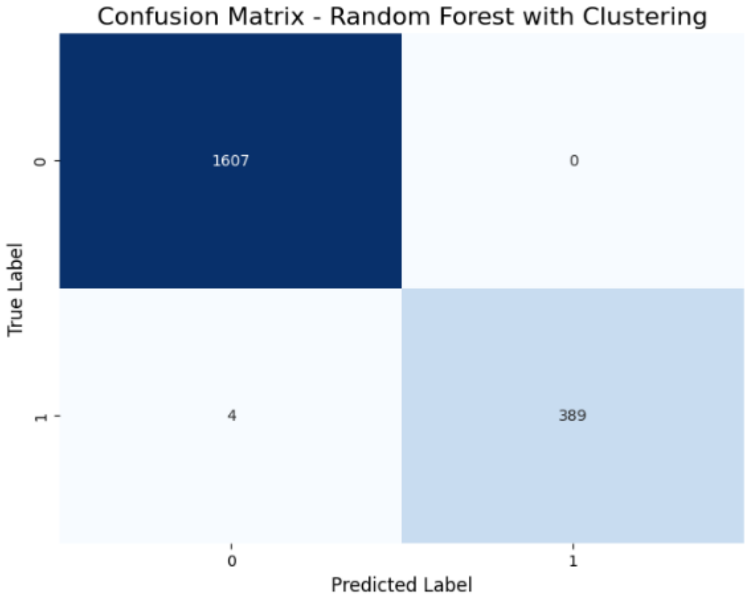

Fig 4. Confusion matrix of Kmeans and Random Forest model.

4. Results

Ensemble model

Classification Report

Precision	Recall	F1-score	Support
0.86	0.98	0.92	1607
0.85	0.36	0.50	393

Accuracy

Logistic Regression: 0.815

Random Forest: 0.864

SVM: 0.8555

XGBoost: 0.8605

Ensemble model: 0.861

The accuracy of the ensemble model can also be looked at as the average of accuracy of all the models separately.

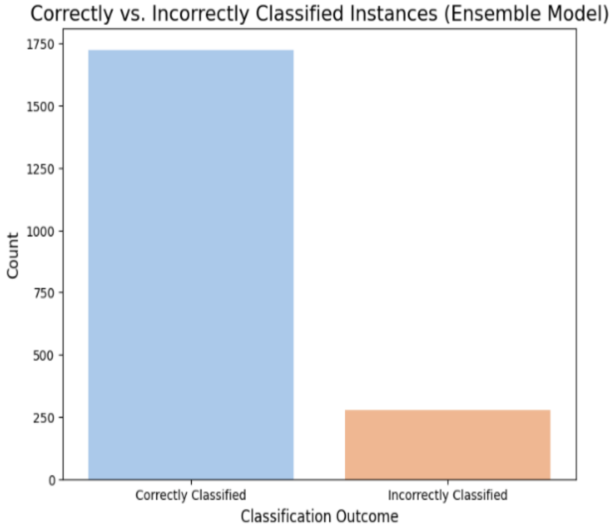

Fig 5. Correctness of classification for ensemble model.

KMeans and Random Forest classifier model

Classification Report

Precision	Recall	F1-score	Support
1.00	1.00	1.00	1607
1.00	0.99	0.99	393

Accuracy

The accuracy of this model is 0.9975.

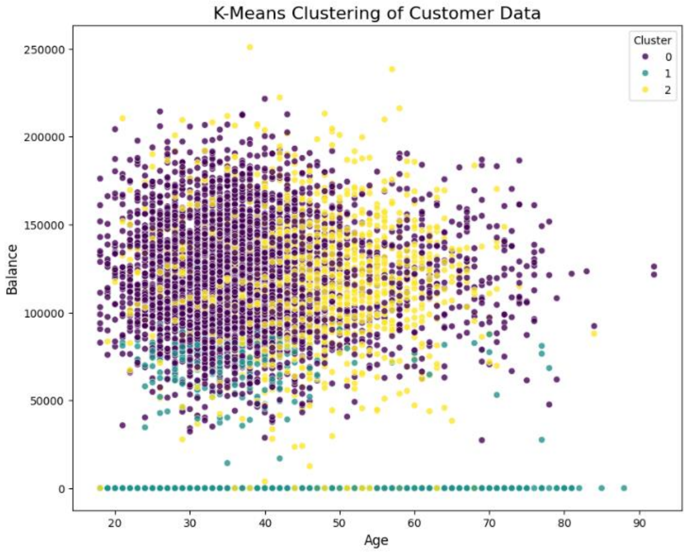

Fig 6. Kmeans Clustering od dataset

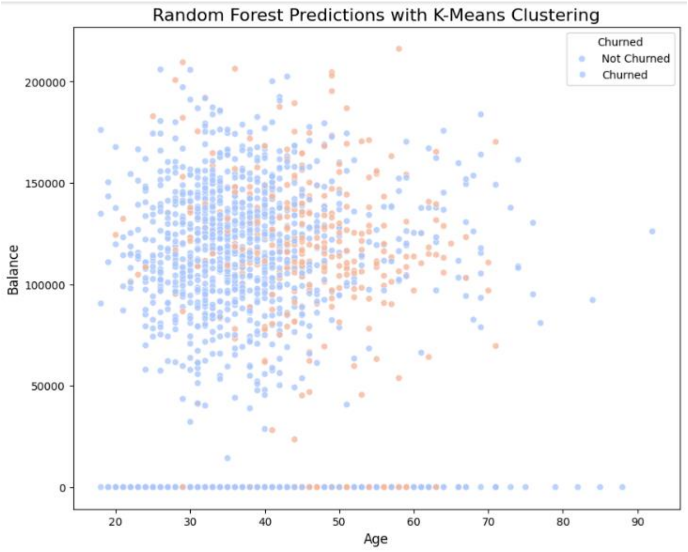

Fig7. Random Forest Predictions

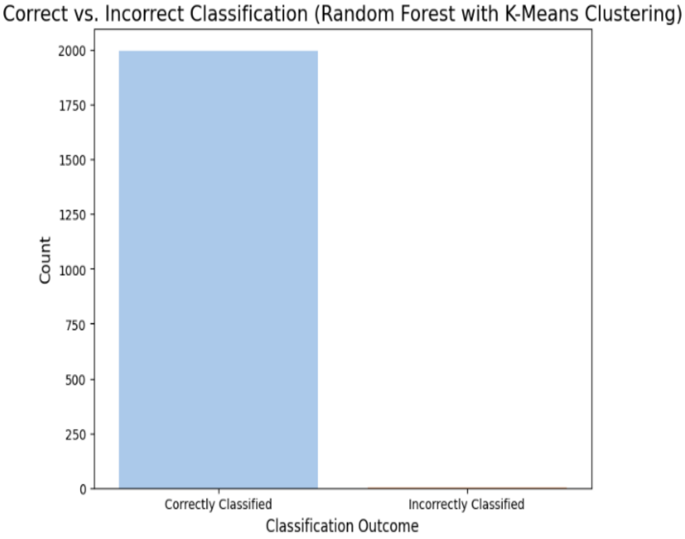

Fig 8. Correctness of Kmeans and Random forest model.

5. Performance Metrics

Accuracy: This metric assesses how well the model predicts things overall. It is determined by dividing the total number of cases in the dataset by the ratio of correctly predicted instances (including true positives and true negatives).

Precision: Precision measures the proportion of true positive predictions among all positive predictions made by the model. It is calculated as

$$TP / (TP + FP)$$

where TP is true positives and FP is false positives.

Recall (Sensitivity): Recall quantifies the ability of the model to correctly identify all positive instances. It is calculated as

$$TP / (TP + FN)$$

where FN is false negatives.

F1 Score: The F1 score is the harmonic mean of precision and recall, providing a balanced measure of the model's performance.

It is calculated as

$$2 * (Precision * Recall) / (Precision + Recall).$$

Silhouette Score: This statistic compares the distances within and across clusters to assess how well a cluster was clustered. A higher silhouette score (which goes from -1 to 1) denotes clusters that are more clearly delineated.

Adjusted Rand Index (ARI): Taking into account all sample pairs, ARI calculates the degree of similarity between actual and predicted cluster labels. It has a range of -1 to 1, with 1 denoting full agreement between clusters.

Error checking

Confusion Matrix: The number of true positives (TP), false positives (FP), true negatives (TN), and false negatives (FN) is displayed in this tabular form. Numerous performance measures are computed using it.

Cross-Validation: To evaluate the model's stability and generalizability, methods such as k-fold cross-validation can be used. The process is dividing the dataset into k subsets, using k-1 subsets for training the model, and assessing it on the remaining subset. After k repetitions of this procedure, the average performance metrics are calculated.

Cluster Validation: To make sure that the data points are meaningfully divided into different groups, the cluster centroids and assignments can be checked after implementing K-means clustering.

6. Conclusion

The banking sector till date is one of the sectors with the highest number of customers. Not only this, the competition is very high in this sector. New schemes, new insurances, new loan policies, new plans etc are keeping customers on their toes, moving from one bank to another. Customers keep looking for better and better as they go. In a world that is so fast moving and full of competition, customer retention has been the main aim of every organisation in this sector.

This paper how customers can be retained and customer churn can be reduced through machine learning's prediction algorithms. This paper showed one model using KMeans clustering with Random Forest Classifier, which showed an accuracy of 99%. This is a very good result. With such good results, such a model can be utilised excellently in predicting customer churn in any banking organisation. A bank could use these results to create better customer retention strategies.

Since this model is using KMeans clustering, the separation of customers into different clusters can also help the organisation provide better, more personalised customer services. The bank would have more resources to understand their customers better and know what kind of service would be useful to a particular customer. Not only would this help in customer retention, but might also pull in more new customers. This clustering will always help the organisation be more attentive and approachable. This model is effective on the current dataset and will continue to be efficient and flexible to accommodate even more bigger data.

References

[1]Ullah, I., Raza, B., Malik, A. K., Imran, M., Islam, S. ul, & Kim, S. W. (2019).”

A Churn Prediction Model using Random Forest: Analysis of Machine Learning Techniques for Churn Prediction and Factor Identification in Telecom Sector.” IEEE Access, 1–1.

[2] C. Geppert, *“Customer churn management: Retaining high-margin customers with customer relationship management techniques”*

KPMG & Associates Yarhands Dissou Arthur/Kwaku Ahenkrah/David Asamoah, 2002.

[3] W. Verbeke, D. Martens, C. Mues, and B. Baesens,

“Building comprehensible customer churn prediction models with advanced rule induction techniques,” Expert Syst. Appl., vol. 38, no. 3, pp. 2354–2364, Mar. 2011.

[4]Noman Ahmad, Mazhar Javed Awan, Haitham Nobanee, Azlan Mohd Zain, Ansar Naseem, Amena Mahmoud,
“ *Personality Analysis for Churn Prediction Using Hybrid Ensemble Models and Class Balancing Techniques”,*IEEE Access, Year: 2024