Churn Prediction in Banking Sector

Pratik Vinayak Kamble[1], Ajesh Nair[2], Tina Saini[3] , Girish Vasant Patil[4]

[1-4]Department of Computer Engineering & Pillai College of Engineering, Navi Mumbai, India

---***---

Abstract - The membership of a banking system depends on the interest provided for the account , charges on the transactions and the services provided on the account. Now-a-days, many banks provide competitive services which give customers options to choose from. This might lead to customer dissatisfaction and result in the deactivation of the membership also known as churn. Churn is 'when a client cancels a membership to a company they have been using. In this project we will try to predict the churn from the previous databases of the banking systems and study it using data visualization. In current days, the customers are getting more attracted towards the quality of service (QoS) provided by the company. As a result, customer churn and engagement has become one of the main problems for most of the banks. In this project, we have used a voting approach of SVM and Random Forest to find the most successful and accurate way to predict customer churning or retention.

***Key Words*:** Churn, banking system, retention, deactivation, SVM, Random Forest.

1.INTRODUCTION

Churn prediction is a critical aspect of customer relationship management in the banking sector. Churn refers to the loss of customers or their discontinuation of business with a bank. Churn in banking can occur for a variety of reasons, such as poor customer service, high fees, low interest rates, or better offers from other banks. High churn rates can lead to decrease in revenue and market share, while retaining existing customers can lead to growth of revenue and customer loyalty. Therefore, predicting churn and taking proactive measures to retain customers can help banks reduce customer attrition, enhance customer loyalty, and increase revenue. In the Banking sector, this churn depends on features such as credit score, loans, investment, age, etc.

Some of the benefits of churn prediction in the banking sector are:

Cost savings: Acquiring new customers is more expensive than retaining existing ones. By predicting churn, banks can take proactive measures to retain customers and save on the costs associated with customer acquisition.

Increased customer satisfaction: Predicting churn and taking proactive measures to address it can help banks improve customer satisfaction. By addressing customer complaints and concerns in a timely manner, banks can enhance the overall customer experience.

Enhanced customer loyalty: Retaining existing customers through proactive measures can increase customer loyalty. Loyal customers are more likely to use more products and services from the bank and refer the bank to others.

Improved revenue: Retaining existing customers can lead to increased revenue. Loyal customers are more likely to use additional products and services from the bank, resulting in increased revenue for the bank.

Overall, churn prediction is an essential tool for banks to manage customer relationships effectively, reduce customer attrition, and increase revenue. In this project we will take these features into consideration for predicting the churn.

2.LITERATURE SURVEY

Here relevant literature is taken into consideration and reviewed. The brief information about the referred research papers is explained here with their methodology, uses and summary.

Literature Review

In paper [1], the analysis process was performed by applying machine learning techniques such as logistic regression, k-nearest neighbors, decision trees, random forests, SVM, Adaboost, multi-layered sensors, and naive Bayes method to the relevant datasets. We observed that random forest was the most successful method for both datasets considered.

In paper [2], k-nearest neighbor, SVM, random forest, and decision tree classifiers are used. We then use several feature selection methods to learn more about relevant features and check system performance. This study shows that using a random forest model after oversampling has advantages over other models when considering accuracy.

In paper [3], the emphasis is given more on improving the Quality of Service through churn prediction. Here, SVM algorithm is used which helps to add more value to the customer retention strategies

In paper [4], the churn analysis problem is addressed by considering a scenario in which a company with sensitive databases wants to perform churn analysis techniques on the joins of that database without revealing unnecessary information. The main purpose of this paper is to predict whether customers will churn or retain in the near future based on predictive analytics considering company billing data.

In paper [5], a machine learning approach for churn prediction in e-commerce that uses the Random Forest algorithm to predict whether a customer will be churned or be retained on various customer attributes such as purchase history, browsing behavior, and demographic data. The results of the study show that Random Forest algorithm outperforms other machine learning algorithms such as LogisticRegression, Decision Tree, and Naive Bayes in terms of accuracy, precision, F1-score and recall.

In paper [6], a comparative study of various machine learning algorithms such as Logistic regression, Decision tree, Random Forest, and Gradient Boosting for customer churn prediction. The datasets from Telecom company are used and the performance of these models are evaluated. The paper concludes with Random Forest outperforming other algorithms in terms of accuracy and F1-score.

In paper [7], a systematic review of churn prediction models in subscription based businesses is shown. The study identifies various factors that contribute to customer churn such as price, content, and user experience. The review also identifies techniques such as Logistic Regression, Decision Tree, Neural Networks and Random Forest.

In paper [8], customer churn prediction is done on online food delivery platforms through machine learning approach. The study uses various customer attributes such as order frequency, order value, and delivery time to predict churn. The study uses approaches such as Logistic Regression, Decision Tree, Random Forest, Gradient Boosting to evaluate performance and concludes that Random Forest algorithm outperforms the others in terms of accuracy, precision, F1-score and recall.

3.PROPOSED WORK

The aim of this project is to predict the churn in the banking sector with the help of machine learning models. In this section the proposed system architecture is described. The proposed system architecture is divided into three parts: preprocessing, training and testing of data then creating the desired model.

The proposed system architecture of our model is shown in Figure 3.1

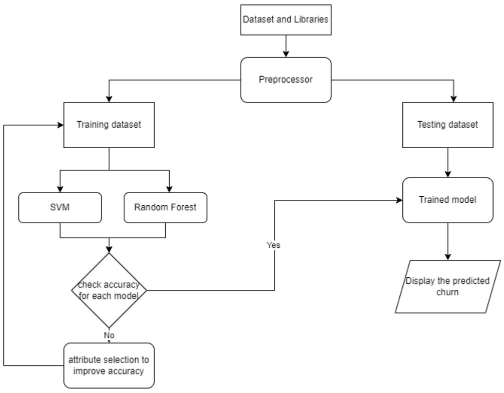


Fig: 3.1 Proposed System architecture

4.METHODOLOGY

In this section implementation details of a proposed work is described. It includes methodology, description of machine learning and hardware and software details.

The proposed work consists of six main steps:

Preparation of dataset

Here, we consider relevant data from the banking dataset. The relevant data are the ones that take part in the prediction of output of our model and the rest data is dropped. Then, we ensure that the data is clean, complete and in a format that can be used for modeling. Clean and preprocess the data to remove missing values, outliers, and redundant features.We relate the features that are dependent on each other and create a new common feature to reduce the number of features. (Note: Here, the feature refers to the columns of the dataset). Then, we scale the data columns having continuous values from 0 to 1 and the distinct column to binary values of 'Yes' and 'No' to '1' and '-1' for easier calculations.

Split Train and Test data

Now, this dataset is split into two different datasets: train data and test data. The train data is the one used to train a model while test data is used to evaluate a model. The split can be 50-50, 60-40, etc. But in our proposed system, the split is 80% training data to 20% testing data.

Creating the model

Support Vector Machine: Support Vector Machines(SVMs) are a generally used algorithm in supervised machine literacy, frequently applied to bracket and retrogression problems. They're particularly useful for bracket tasks, similar to prognosticating the feelings expressed in a piece of textbook. The SVM algorithm works by creating a hyperplane, or optimal decision boundary, in n- dimensional space that can directly classify data points into the applicable order.

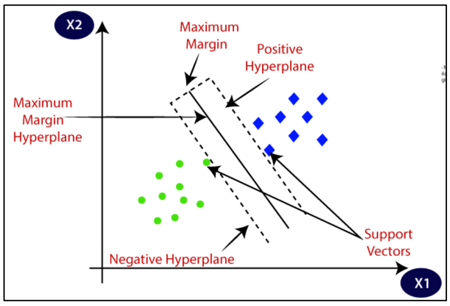


Fig 4.1 Support Vector Machine Algorithm

SVMs can use different types of kernel functions to transform the data into a higher-dimensional space, where it can be more easily separated. Two commonly used kernel functions in SVMs are the radial basis function (RBF) kernel and the polynomial kernel.

The RBF kernel is often used when the data is not linearly separable, and it can capture complex nonlinear relationships between the features. The polynomial kernel is often used when the data has a polynomial structure, and it can capture nonlinear relationships between the features that are polynomial in nature. In general, the choice between the RBF and polynomial kernel depends on the specific characteristics of the data and the problem at hand. Both kernels have their strengths and weaknesses, and it is important to choose the kernel that works best for the given data and problem.

Random Forest : The random forest algorithm is a type of ensemble learning technique that uses a set of decision trees to achieve high prediction accuracy and model stability. This technique can handle both regression and classification tasks. Each tree classifies (or votes for) a data instance based on its attributes, and the forest chooses the classification that receives the most votes. For regression tasks, the decisions of different trees are averaged.

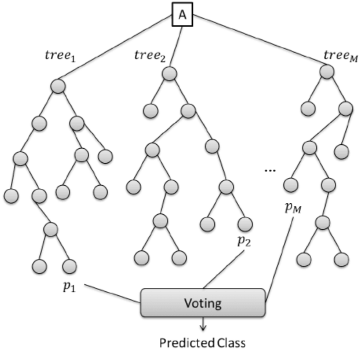


Fig 4.2 Random Forest Algorithm

Voting Approach: Voting is a common ensemble method in machine learning that combines the predictions of multiple models to improve the overall prediction accuracy. In the voting approach, each model makes its own prediction, and then a majority vote is taken to determine the final prediction. In this paper, we are using the hard approach. In this approach, each model's prediction is counted as a single vote, and the class with the highest number of votes is chosen as the final prediction. This approach works well when the models have similar accuracies.The voting approach can be used with any type of model, including decision trees, neural networks, and support vector machines. It is often used in classification problems, but it can also be used in regression problems by taking the mean or median of the predictions instead of the majority vote. One of the advantages of the voting approach is that it can help to reduce the risk of overfitting, especially when the base models are diverse and complementary.

Creation of the model: Now, we create an ensemble learning model of SVM and Random Forest by using the Voting approach. Using a voting approach on Support Vector Machines (SVMs) and Random Forest (RF) models is a powerful way to improve the overall performance of the ensemble model. Using a voting approach on SVM and RF models helps to leverage the strengths of both models and improve the overall prediction accuracy. SVMs are good at handling non-linear data and high-dimensional data, while RFs are good at handling noisy data and can capture interactions between features. By combining the predictions of both models, the ensemble model can potentially overcome the weaknesses of each model and achieve higher performance.

Training and Testing of model

Now that the model is created, we start training the model by the use of the training dataset. We keep tuning the hyperparameters till the required accuracy is acquired. Now, we validate the model on the testing dataset to evaluate its performance and measure its prediction accuracy. Once the model is validated and has achieved satisfactory results, deploy it in a production environment to predict customer churn in real-time.

Web Application

Now that the model is ready, we have created a web application form that takes the user information and displays the predicted output. This web application is created in HTML, CSS and Flash framework.

Block Diagram

The below is the block diagram of our web application and proposed system. The user interacts with ‘index.html’ to fill the details on the provided form. Then, in the backend these values are taken as inputs and are processed for the model. then the model uses these values to predict an output and that output is displayed in 'result.html'. The 'churn.py' is the file where the proposed model is created and then that model is stored in 'voting_clf1.pkl.

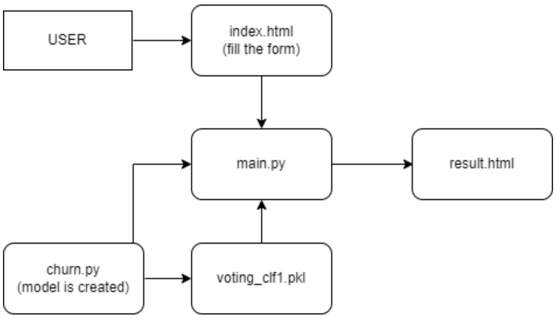
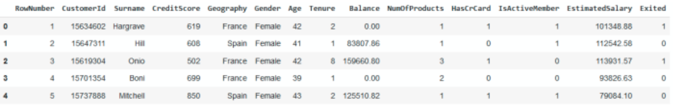


Fig 4.3 Block Diagram

5.DATASET DESCRIPTION

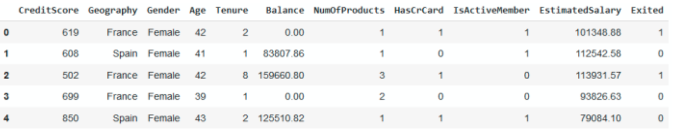
The below figure is a snapshot of the dataset that is taken as an input to train and test our model. The dataset has 10000 rows and 14 columns. The rows distinguish the no. of customers while the columns provide us with the details of those customers.



	RowNumber	CustomerId	Surname	CreditScore	Geography	Gender	Age	Tenure	Balance	NumOfProducts	HasCrCard	IsActiveMember	EstimatedSalary	Exited
0	1	15634602	Hargrave	619	France	Female	42	2	0.00	1	1	1	101348.88	1
1	2	15647311	Hill	608	Spain	Female	41	1	83807.86	1	0	1	112542.58	0
2	3	15619304	Onio	502	France	Female	42	8	159660.80	3	1	0	113931.57	1
3	4	15701354	Boni	699	France	Female	39	1	0.00	2	0	0	93826.63	0
4	5	15737888	Mitchell	850	Spain	Female	43	2	125510.82	1	1	1	79084.10	0

Fig 5.1 Original dataset

The irrelevant columns for the models are dropped and the new dataset is considered as input which is shown below.



	CreditScore	Geography	Gender	Age	Tenure	Balance	NumOfProducts	HasCrCard	IsActiveMember	EstimatedSalary	Exited
0	619	France	Female	42	2	0.00	1	1	1	101348.88	1
1	608	Spain	Female	41	1	83807.86	1	0	1	112542.58	0
2	502	France	Female	42	8	159660.80	3	1	0	113931.57	1
3	699	France	Female	39	1	0.00	2	0	0	93826.63	0
4	850	Spain	Female	43	2	125510.82	1	1	1	79084.10	0

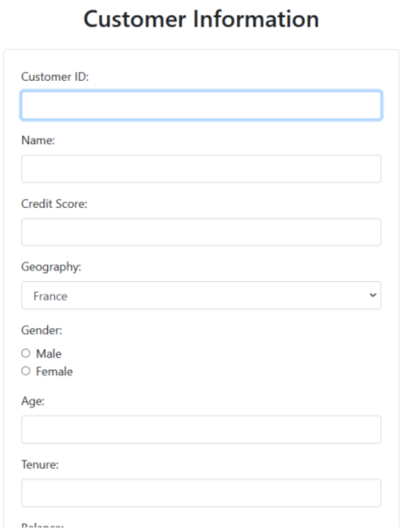
Fig 5.2 Refined dataset

6.RESULT AND DISCUSSION

The experimental results and the sample screenshots of the proposed system are provided in this given section.

Web Application Interface

Form



Customer Information

Customer ID:

Name:

Credit Score:

Geography:
France

Gender:
○ Male
○ Female

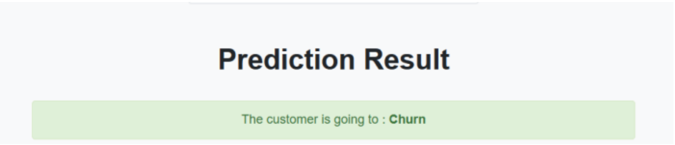
Age:

Tenure:

Balance:

Fig. 5.3 Home Page

Result



Prediction Result

The customer is going to : **Churn**

Table 5.4 Result

Performance Evaluation: Result Analysis

In order to evaluate the proposed system, the test dataset is used. This test dataset is used by all the models to create a ROC (Receiver Operating Characteristic) AUC graph that compares the performance of the models.

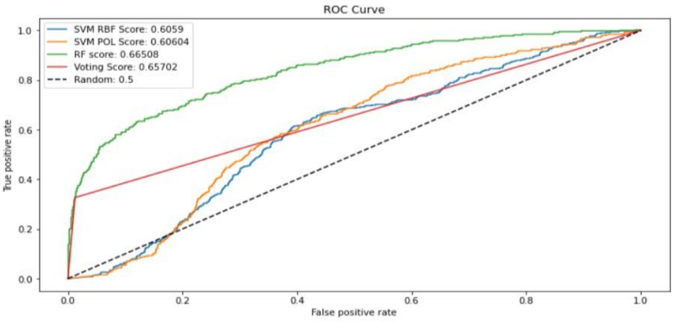
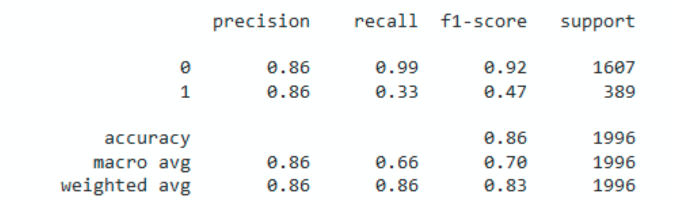


Fig 5.5 ROC AUC graph

The ROC AUC graph is a useful tool for evaluating the performance of binary classifiers and comparing different models.The x-axis represents the false positive rate (1-specificity), which is the proportion of negative samples that are incorrectly classified as positive. The y-axis represents the true positive rate (sensitivity), which is the proportion of positive samples that are correctly classified as positive.An AUC score of 0.5 indicates that the classifier is randomly guessing, while an AUC score of 1 indicates perfect classification. A higher AUC score indicates better classification performance. The ROC curve can help visualize the trade-off between sensitivity and specificity. The key insights from the ROC AUC graph, is that voting score and random forest models are having similar scores and are efficient models for binary classifiers. The below table shows us about the precision, recall, f1-score, support, accuracy, macro avg, weighted avg of our proposed model.



```
              precision    recall  f1-score   support

           0       0.86      0.99      0.92      1607
           1       0.86      0.33      0.47       389

    accuracy                           0.86      1996
   macro avg       0.86      0.66      0.70      1996
weighted avg       0.86      0.86      0.83      1996
```

Fig 5.6 Classification Report

7.CONCLUSION AND FUTURE SCOPE

In this report, we have compared SVM and Random Forest models, given them the best parameters and then combined them to make our proposed model through a voting approach. We can conclude that our proposed model gives us a more accurate and precise prediction than the existing and other models. Different evaluation parameters like precision, recall, accuracy, f1-score, support, weighted average, and macro average are described. The study reveals that our system can be improved by combining more accurate models using the same voting approach.

In future, this proposed system can also add other machine learning models like gradient boosting, decision tree, neural networks, whichever is more accurate to improve the accuracy of our model. Experimenting with these models may reveal additional insights or better performance. We can consider expanding the scope of our project to include other strategies for retaining customers, such as personalized marketing campaigns or loyalty programs.

REFERENCES

[1] Hamdullah Karamollaoğlu, Düzce University, Düzce, İbrahim Yücedağ, İbrahim Alper Doğru. Customer Churn Prediction Using Machine Learning Methods: A Comparative Analysis, 6th International Conference of Computer Science and Engineering, October 2021.

[2] Manas Rahman, V Kumar. Machine Learning Based Customer Churn Prediction In Banking. 4th International Conference on Electronics, Telecommunication and Aerospace Technology. December 2020.

[3] RajaGopal Kesiraju VLN, P. Deeplakshmi. Dynamic Churn Prediction using Machine Learning Algorithms - Predict your customer through customer behaviour. International Council on Computer Technology and Informatics. 21 April 2021.

[4] Navid Forhad, Md. Shahriar Hussain, Rashedur M Rahman. Churn analysis: Predicting churners. 9th International Conference on Digital Information Management. 18 December 2014 .

[5] Gaurav Kumar, Nikhil Kumar Singh, Ashish Mishra. Churn Prediction in E-Commerce: A Machine Learning Approach. 6th International Conference on Science and Information Technology. 2020.

[6] Arpan Pal , Kunal Roy. A comparative study of machine learning algorithms for customer churn prediction. 7th International Conference on Computer Science and Information Technology. 2020.

[7] Arash Alimohammadzadeh, Mohammadreza Khalizad Amir Hosein Azimi. Churn prediction in subscription-based business: A systematic review. 4th International Conference on Computer Science and International Technology. 2020.

[8] Vignesh P.G., Thirumalai Selvan M. Customer Churn Prediction for Online Food Delivery. 6th International Conference on Computer Science and Engineering. 2021.