Churn Prediction in Banking using Machine Learning Algorithms

***Abstract*** - In this paper, we apply prominent machine learning algorithms to predict customer churn in the banking industry. During the first phase of our paper, we applied the standardscaler before applying ML classification models, followed by a 10-fold cross-validation process. The second phase of the study involves evaluating the performance of each of these models using a variety of evaluation metrics. The best overall classifier was the Random Forest with accuracy of 86% and F-measure of 92%.

***Keywords – bank churn prediction, machine learning, logistic, SVM, Random Forest, KNN classification, ROC curve, DET Curve, Precision, Recall, Accuracy, F-measure***

1. **INTRODUCTION:**

As part of Customer Relationship Management (CRM), businesses or other organizations administer their interactions with customers using data analysis to analyse large amounts of information. Customers leaving their service providers is referred to as churn  [1] [2] .

Banking sector banks are concerned about customer churn prediction as it has a significant impact on their profit margins [3]. An effective customer retention program can target high-risk customers who are likely to discontinue their business and switch to a competitor. It is important to identify bank customers accurately and prior to marketing to them to minimize the costs of customer retention [2] [4].

Typically, customer churn is determined by estimating or analysing the percentage of customers who switch to an alternative service [5], [6]. It is one of the most common problems in any industry. Banking is one such industry that pays a great deal of attention to the behaviour of customers by tracking their activities. The cost of acquiring a new customer is very high compared to the cost of maintaining an existing customer [7] [2]. By handling these customers, companies can increase their profits. In this regard, it is important to maintain the existing customer base, which can only be achieved by understanding the grievances of the customer when they change banks. This paper presents a model to churn the bank customers using k-nearest neighbour (KNN) algorithm [2], logistic regression, random forest, and Support Vector Machines. Next section explains the literature survey based on different algorithms used in various papers [2]

1. **DATA SET DESCRIPTION:**

The dataset was retrieved from the Kaggle dataset repository. The dataset consists of 10000 instances with 14 attributes without any missing or null values. Each instance contains information about a single customer, including their unique ID, Credit Score, Balance, Estimated Salary, Demographic information, and whether they left the bank or not. In the case of a closed account with a bank, the binary flag 1 is set, whereas in the case of a retained account, the binary flag 0 is set. *Table 1.* DATASET DESCRIPTIONdisplays the summary of the data set describing features, feature types and value ranges for categorical features.

Table 1. DATASET DESCRIPTION

|  |  |  |  |
| --- | --- | --- | --- |
| **No** | **Attribute Name** | **Description** | **Type** |
| 1 | Row Number | Row Numbers from 1 to 10000 | Numeric |
| 2 | Customer Id | Unique Ids for bank customer identification | Numeric |
| 3 | Surname | Customer's last name | Categorical |
| 4 | Credit Score | Credit score of the customer | Numeric |
| 5 | Geography | The country from which the customer belongs | Categorical |
| 6 | Gender | Male or Female | Categorical |
| 7 | Age | Age of the customer | Numeric |
| 8 | Tenure | Number of years for which the customer has been with the bank | Numeric |
| 9 | Balance | Bank balance of the customer | Numeric |
| 10 | NumOfProducts | Number of bank products the customer is utilizing | Numeric |
| 11 | Has Cr Card | Binary Flag for whether the customer holds a credit card with the bank or not | Numeric |
| 12 | Is Active Member | Binary Flag for whether the customer is an active member with the bank or not | Numeric |
| 13 | Estimated Salary | Estimated salary of the customer in Dollars | Numeric |
| 14 | Exited | Binary flag 1 if the customer closed account with bank and 0 if the customer is retained | Numeric |

1. **DATA PRE-PROCESSING:**

ML algorithms only works with numerical data. In case of categorical data, it can be converted into numeric values before applying the classification models to the dataset. In our dataset, we have 3 categorical attributes such as “Surname”, “Geography” and “Gender”. As the surname column does not have any impact on the reason of customers exiting the bank, we will be dropping it. Hence, we will be considering the columns “Geography” and “Gender”.

Values France, Spain and Germany are mapped 0,1 and 2. Similarly, we have mapped female with 0 and Male with 1as shown below.

**Geography:** France 🡪 0, Spain 🡪 1, Germany 🡪 2

**Gender:** Male 🡪 1, Female 🡪 0

In order to boost the value of accuracy we have standardized the dataset before applying the classification models using ***“Standard Scalar”*** from the Scikit learn. The model fitting and model learning function will not function equally when variables are measured at different scales. This may result in a bias when the variables are measured at different scales. In order to avoid this potential problem, feature-wise standardization (μ=0, σ=1) models are usually used before fitting the models [8].

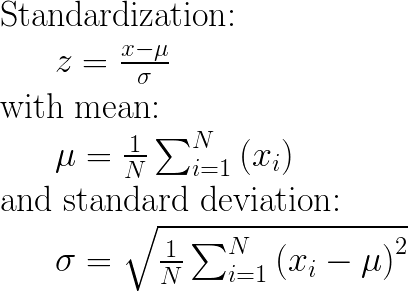


Figure 1 - Standardization Formula [9]

1. **MACHINE LEARNING CLASSIFICATION METHODS:**
2. **Support Vector Machines:**

Support Vector Machines (SVMs) are supervised learning models that analyse data and recognize patterns and are used in classification and regression analyses. Basically, the SVM method consists of constructing an optimal plane or hyperplane that categorizes the data. An optimal hyperplane is a field that separates data into classes, located at right angles to the nearest pattern. Datasets are described by patterns, which are dots [10]. Optimal hyperplanes are determined by finding the maximum margin. Margin means the distance between the hyperplane and the closest pattern in each class. The support vectors are patterns located closest to the optimal hyperplane and their position has an impact on it [11].

Line chart

Description automatically generated with low confidence Chart

Description automatically generated

Figure 2 - Possible Hyperplanes – SVM [12]

The main highlight of SVM is that even with complex datasets like non-linearly separable datasets, it converts the data into linear one using higher dimensions.

The kernels are a set of mathematical operations used by SVM algorithms. A kernel's job is to take data as input and change it into the required form. The most fundamental kind of kernel is a linear one, and it performs best when there are many characteristics. It is mostly preferred for text-classification problems as most of these kinds of classification problems can be linearly separated and they are faster than other functions. Linear Kernel (one dimensional) [13] is calculated as K(xi,xj) = xi.xj + c [14]

1. **Random Forests Algorithm:**

Random forest is an ensemble learning method and because of the number of decision trees participating in the process it considered as a highly accurate method. Following are the few advantages of the Random Forest method:

Firstly, it does not have the overfitting problem as it calculates the average of all the predictions to remove the percentage of bias.

Secondly, handling missing values. For a given dataset, it selects a sample range and constructs decision tree for each sample to get a prediction result from each decision tree. For this predicted result, it performs a voting. Finally, the prediction result with the highest vote will be selected as the final prediction [15].

1. ***Logistic Regression:***

The logistic regression model is used to understand the relationship between an outcome (dependent or response) variable and a set of independent (predictor or explanatory) variables. The predictor variables are called covariates. Logistic regression is a type of ‘classification’ algorithm because it tries to "classify" observations from a dataset into different categories. [16]

The function used in a logistic regression model is called the logistic function or sigmoid function which is a common S-shaped curve (sigmoid curve) and it is a mathematical function with the shape of the letter “S” that can transfer any real value to a range between 0 and 1. The sigmoid function and curve is as follows:

A picture containing text, line, font, diagram

Description automatically generated

Figure 3 - Logistic Function - Sigmoid Equation and Sigmoid Curve [17]

Figure 4 - Workflow of Logistic Regression

In machine learning, the sigmoid function is an activation function that introduces non-linearity to a model. The loss function is finding the difference between the true and predicted value for a single data point, whereas the cost function is finding the difference for the entire dataset. The model randomly assigns weight and bias values, and the cost function is calculated. If the cost function is high, gradient descent is used for minimizing the cost function in this algorithm by finding optimum weight and bias values. [18]

1. ***K-Nearest Neighbours (KNN)***

KNN is a straightforward algorithm that classifies data points based on the supposition that nearby data points may be related by using distance metrics such as Euclidean and Manhattan distance. Finding the k data points in a data collection that are closest to a particular query data point is the goal of the KNN classification problem [19]. Since KNN relies on memory to keep all its training data, it stores all of the examples that are already accessible and categorizes new cases based on feature similarity. It is also known as a memory-based or instance-based learning approach. Euclidean distance is based on Pythagoras theorem. Manhattan distance metric is named so because the grid layouts look like the street of Manhattan. The distance formula for Euclidean and Manhattan are given below:

A picture containing text, clock, watch

Description automatically generatedA screenshot of a computer

Description automatically generated with low confidence

Figure 5 - Euclidean and Manhattan Distance Calculation [20]

For huge datasets, KNN is not advised because it degrades and affects the performance of the model as a whole. It is frequently used for data mining, recommendation systems, and other things.

1. **EVALUATION METRICS:**

To evaluate the performance of the classification models, we are using various metrics like precision, recall, accuracy, F-measure, ROC Curve, DET Curve and AUC. Precision, Recall, Accuracy and F-measure are calculated from the values of confusion matrix [21], for various parameters shown in Table 1. The data is considered true positive (TP) if it has a positive label and is categorized as such; otherwise, it is considered false negative (FN) (FN).

A data point counts as a true negative (TN) if it has a negative label and is categorized as such; a data point counts as a false positive (FP) if it is classed as such (FP) [10]. Precision is the ratio of the correctly predicted affirmative customers.

The percentage of affirmative customers that were correctly identified is known as ***recall***. Precision works with predictions as its base whereas recall has truth as its base for all the calculations. The percentage of total predictions that were accurate is known as ***accuracy***.

Since excellent performance in one of those indices does not always imply good performance in the other, precision or recall alone cannot adequately explain the effectiveness of a classifier. Because of this, F-measure is frequently employed as a single statistic to assess the performance of classifiers. F-measure is the harmonic mean of recall and precision. [21].

****

Figure 6 - Evaluation Metrics Equations [21]

Higher the value of precision and recall implies that the model is performing well. ***K – Fold Cross Validation*** is a very useful technique for assessing the effectiveness of the classification model. In our paper, we have implemented the 10 - fold cross validation technique. ***Balanced accuracy*** is a metric we can use to assess the performance of a [classification model](https://www.statology.org/regression-vs-classification/). It is calculated as:

Balanced accuracy = (Sensitivity + Specificity) / 2

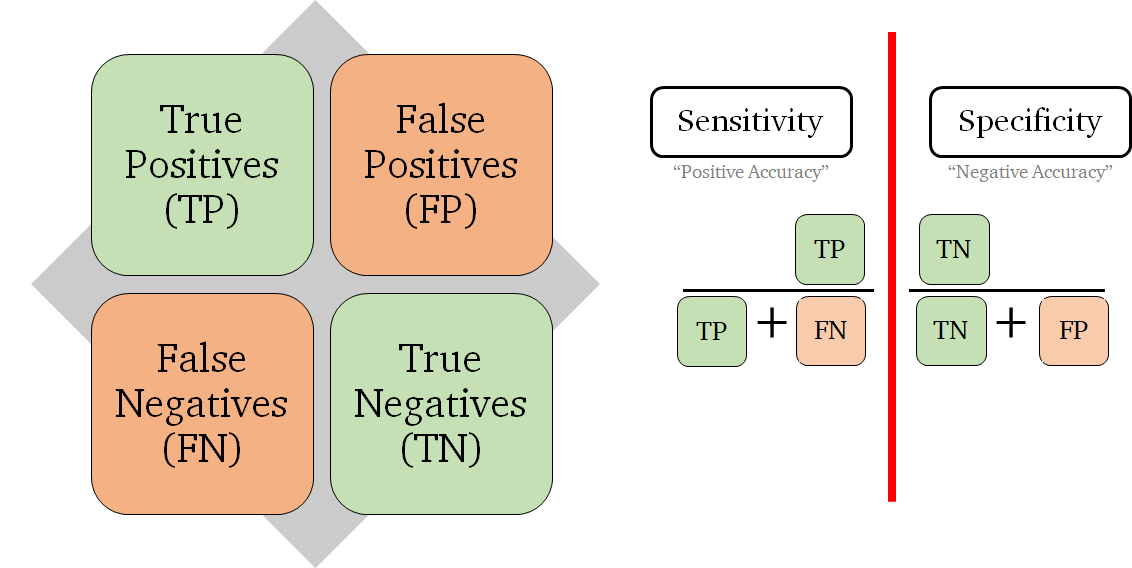


Figure 7 - Derivation of Sensitivity & Specificity from Confusion Matrix [22]

A ***detection error trade off (DET)*** graph is a graphical plot of error rates for binary classification systems, plotting the false rejection rate vs. false acceptance rate. DET curves give the user direct feedback of the detection error tradeoff to aid in operating point analysis.

The user can deduct directly from the DET-curve plot at which rate false-negative error rate will improve when willing to accept an increase in false-positive error rate (or vice-versa).

***ROC curve*** is a performance measurement for the classification problems at various threshold settings. ***ROC is a probability curve*** ***and AUC represents the degree or measure of separability.*** The ROC curve is plotted with TPR against the FPR where TPR is on the y-axis and FPR is on the x-axis [23]. It tells how much the model is capable of distinguishing between classes.

***AUC (Area Under the Curve)*** helps compare different models since it summarizes the data from the whole ROC Curve. Higher the AUC, the better the model is at predicting 0 classes as 0 and 1 classes as 1. By analogy, the Higher the AUC, the better is the performance of the model.

Chart, line chart

Description automatically generated

1. **EXPERIMENTAL RESULTS:**

In this paper, we have applied the classification methods to the churn prediction problem of banking customers based on a Kaggle dataset: Logistic Regression, SVM, KNN, and Random Forest. At first, all methods were tested without standardizing data, and the accuracy scores were relatively low. As a result, we have standardized the data before applying the models, with the ***Random Forest*** classifier method achieving ***highest accuracy of 86% and 92% F-measure***.

From the ROC Curve, we can see that the Random Forest model curve is highest curve plotted above the random guessing classifier line with an **AUC score of 0.86** and it helps us to conclude our decision that it is considered as a best model for our dataset.

A booster method reduces the bias of the combined estimator by sequentially applying the base models. These are very powerful in terms of performance and accuracy. To further enhance the accuracy of the model, we can implement boosting algorithms such as Gradient and Adaptive Boosting Algorithms.

TABLE 2 provides a summary of all metrics for each of the four methods of classification mentioned above. According to our analysis of the accuracy and the other metric scores, *Random Forest is the best performing model for predicting churn rates in our banking dataset.*

1. **DISCUSSIONS AND CONCLUSION:**

In this paper, we have used the banking dataset to perform the churn prediction with the ML classification methods. Data pre-processing techniques like converting categorical data into numerical data and data standardization has been applied to the dataset. After that we have split the data into train and test sets and then, the customers of the bank are classified into loyal, or churn based on their activities. Also, from the EDA analysis it is evident that the customers between the age 40 and 60 were most likely to exit the bank. Once we have the processed dataset, we applied 4 ML algorithms and found that Random Forest has the highest accuracy score of 86%. In addition to this, we have implemented both adaptive and gradient boosting methods to further optimize the churn prediction analysis. Both methods achieved an accuracy of 86% and 85.2%, respectively, with an F-measure of 92% and 91%.

TABLE 3 contains the different evaluation metric scores of the gradient and adaptive boosting algorithms. While comparing these metrics with the TABLE 2, we can observe that there are no significant differences between the Random Forest model and the boosting algorithms. It is recommended that further research be conducted with a large-scale banking dataset to determine whether there is significant variation in the performance of these ML algorithms. It is to be noted that this research study is pertaining to the dataset used in this study; results may vary with other datasets. Taking this prediction into account, bank officials gain valuable insight into the bank's customers and the institution's operation. To assess the performance of the prediction model, customers' exit condition must be accurately identified [2].

TABLE 2 - EVALUATION MEASURES AND RESULTS OF LOGISTIC, SVM, KNN AND RANDOM FOREST MODELS

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Evaluation Metrics** | **Logistic** | **SVM** | **KNN** | **Random Forest** |
| **Precision** | 0.82 | 0.81 | 0.79 | **0.88** |
| **Recall** | 0.97 | 0.98 | 0.80 | **0.96** |
| **F1 Score** | 0.89 | 0.89 | 0.80 | **0.92** |
| **10-Fold Cross Validation Scores** | | | | |
| **Fold 1** | 0.8142 | 0.81 | 0.7885 | 0.8671 |
| **Fold 2** | 0.8028 | 0.8071 | 0.7828 | 0.8528 |
| **Fold 3** | 0.8271 | 0.82 | 0.7885 | 0.8757 |
| **Fold 4** | 0.8142 | 0.81 | 0.7714 | 0.85 |
| **Fold 5** | 0.7828 | 0.7828 | 0.79 | 0.8342 |
| **Fold 6** | 0.8157 | 0.8171 | 0.8085 | 0.8485 |
| **Fold 7** | 0.8414 | 0.8385 | 0.8042 | 0.87 |
| **Fold 8** | 0.8114 | 0.8028 | 0.7857 | 0.8385 |
| **Fold 9** | 0.7985 | 0.7985 | 0.79 | 0.8414 |
| **Fold 10** | 0.8157 | 0.8014 | 0.7971 | 0.8671 |
| **Best Accuracy** | 0.812 | 0.809 | 0.791 | **0.855** |
| **Balanced Accuracy Score** | 0.5850 | 0.5536 | 0.6738 | **0.7247** |

|  |  |  |
| --- | --- | --- |
| **Evaluation Metrics** | **Gradient Boosting** | **Adaptive Boosting** |
| **Precision** | 0.87 | 0.85 |
| **Recall** | 0.97 | 0.98 |
| **F1 Score** | 0.92 | 0.91 |
| **10-Fold Cross Validation Scores** | | |
| **Fold 1** | 0.884 | 0.86 |
| **Fold 2** | 0.852 | 0.85 |
| **Fold 3** | 0.878 | 0.86 |
| **Fold 4** | 0.851 | 0.845 |
| **Fold 5** | 0.84 | 0.825 |
| **Fold 6** | 0.864 | 0.854 |
| **Fold 7** | 0.864 | 0.862 |
| **Fold 8** | 0.841 | 0.844 |
| **Fold 9** | 0.85 | 0.844 |
| **Fold 10** | 0.868 | 0.867 |
| **Best Accuracy** | 0.86 | 0.852 |
| **Balanced Accuracy Score** | 0.7139 | 0.6644 |

TABLE 3 - EVALUATION METRICS FOR GRADIENT AND ADAPTIVE BOOST ALGORITHMS

# **References:**

|  |  |
| --- | --- |
| [1] | V. Mahajan, R. Mishra and R. Mahajan, “Review of data mining techniques for churn prediction in telcom,” *JIOS 37(2),* pp. 183-197, 2015. |
| [2] | Dalmia, H, Nikil, C. V and Kumar, S., “Churning of Bank Customers Using Supervised Learning,” *In Innovations in Electronics and Communication Engineering,* no. Springer, Singapore., pp. pp. 681-691, 202. |
| [3] | Yan, L, Wolniewicz, R. H and Dodier, R, “Predicting customer behavior in telecommunications,” *IEEE Intelligent Systems,* vol. 19(2), pp. 50-58, 2004. |
| [4] | B, Kadeřábková and P, Maleček, “Churning and labour market flows in the new EU member states,” *Procedia Economics and Finance,* vol. 30, pp. 372-378, 2015. |
| [5] | S.A. Qureshi, A.S. Rehman, A.M. Qamar and A. Kamal, “Telecommunication Subscribers’ Churn Prediction Model Using Machine Learning,” no. IEEE, pp. 131-136, 2013. |
| [6] | B. Mishachandar and K.A. Kumar, “Predicting customer churn using targeted proactive retention,” *Int. J. Eng. Technol.,* vol. 7(2.27), pp. 69-76, 2018. |
| [7] | K. Mishra and R. Rani, “Churn prediction in telecommunication using machine learning,” *International Conference on Energy, Communication, Data Analytics and Soft Computing (ICECDS),* pp. 2252-2257 IEEE, 2017, August. |
| [8] | S. Loukas, “How Scikit-Learn’s StandardScaler works. Medium,” *https://towardsdatascience.com/how-and-why-to-standardize-your-data-996926c2c832,* 2020, June 10. |
| [9] | Anon, “Imgur.com,” *https://i.stack.imgur.com/obywE.png,* 2022. |
| [10] | Gupta, B, Kumar, A, Panda, K and Sahu, S, “Support Vector Machine for Selecting Features for Elimination Breast Cancer,” *International Journal of Research in Engineering and Science (IJRES) ISSN, 7,,* no. http://www.ijres.org/papers/Volume%207/Is, p. 13–19, 2019. |
| [11] | Bustamam, A, Bachtiar, A and Sarwinda, D, “ Selecting features subsets based on support vector machine-recursive features elimination and one dimensional-naïve bayes classifier using support vector machines for classification of prostate and breast cancer.,” *Procedia Computer Science,* pp. 157,450-458, 2019. |
| [12] | R. Gandhi, “Support Vector Machine — Introduction to Machine Learning Algorithms,” *Towards Data Science; Towards Data Science,* no. https://towardsdatascience.com/support-vector-machine-introduction-to-machine-learning-algorithms-934a444fca47. |
| [13] | “SVM Kernel Functions - 'Coz your SVM knowledge is incomplete without it,” *TechVidvan ,* Vols. https://techvidvan.com/tutorials/svm-kernel-functions/, 2020, March 13. |
| [14] | S. Awasthi, “Seven Most Popular SVM Kernels,” *Dataaspirant,* no. https://dataaspirant.com/svm-kernels/. |
| [15] | “Sklearn Random Forest Classifiers in Python Tutorial. (n.d.),” *Www.datacamp.com,* no. https://www.datacamp.com/tutorial/random-forests-classifier-python. |
| [16] | Hosmer, W and Lemeshow., “Applied logistic regression (2nd ed.),” *New York ; Chichester : Wiley,* 2004. |
| [17] | “Andymath.com,” *Logistic Function,* no. https://andymath.com/logistic-function/, 2022, August 15. |
| [18] | D. (. Kumawat, “Introduction to Logistic Regression - Sigmoid Function, Code Explanation | Analytics Steps,” *Www.analyticssteps.com,* no. https://www.analyticssteps.com/blogs/introduction-logistic-regression-sigmoid-function-code-explanation. |
| [19] | M. Schott, *K-Nearest Neighbors (KNN) Algorithm for Machine Learning. Medium. ,* no. https://medium.com/capital-one-tech/k-nearest-neighbors-knn-algorithm-for-machine-learning-e883219c8f26, 2020, February 27. |
| [20] | D. Subramanian, “A Simple Introduction to K-Nearest Neighbors Algorithm,” *Medium, Towards Data Science,* no. https://towardsdatascience.com/a-simple-introduction-to-k-nearest-neighbors-algorithm-b3519ed98e, 2019, June 8. |
| [21] | T. Vafeiadis, K. I. Diamantaras, Sarigiannidis, G and Chatzisavvas, K.Ch, “A comparison of machine learning techniques for customer churn prediction.,” *Simulation Modelling Practice and Theory,* pp. 55,1-9, 2015. |
| [22] | Y. Serpa, *What are Classification Metrics? Medium.,* no. https://towardsdatascience.com/the-illustrated-guide-to-classification-metrics-the-basics-cf3c2e9b89b2, 2021, October 11. |
| [23] | S. Narkhede, “Understanding AUC - ROC Curve.,” *Medium; Towards Data Science,* no. https://towardsdatascience.com/understanding-auc-roc-curve-68b2303cc9c5, 2018, June 26. |
| [24] | Nisbet, Miner and Yale, “Handbook of Statistical Analysis and Data Mining Applications (2nd ed.),” no. Academic Press, 2018. |
| [25] | J. Huneycutt, “Implementing a Random Forest Classification Model in Python. Medium.,” no. https://medium.com/@hjhuney/implementing-a-random-forest-classification-model-in-python-583891c99652, 2018, May 21. |