International Journal of Scientific Research and Management (IJSRM)
||Volume||12||Issue||09||Pages||1434-1446||2024||
Website: https://ijsrm.net ISSN (e): 2321-3418
DOI: 10.18535/ijsrm/v12i09.ec03

Machine Learning for Enhanced Churn Prediction in Banking: Leveraging Oversampling and Stacking Techniques

*Omar Faruq Shikdar[1], Fahad Ahammed[2], Arifa Sultana Mily[3], [4]Ashraful Islam

[1,2,3,4] Department of Computer Science and Engineering
East West University, Dhaka-1212, Bangladesh

Abstract:
Every sector of business is getting more competitive as time passes. More and more companies are offering services to people. Banking sector is no different. With the plethora option customer has in terms banking, holding on to customer may prove difficult for banks. The aim of this research is to build a system that will help banks to predict which customers are likely going to churn and allow them to take precaution to stop customers from leaving. In this study we have used classifiers such as: K-Neighbors, Random Forest, XGboost, Adaboost classifiers and Ensemble Model (Stacking Technique) that uses all of these models together. The experimentation was conducted on a dataset from Kaggle. The dataset used in this research was heavily imbalanced. So, different oversampling methods like Random Oversampling and SMOTE-ENN have been used. In data preprocessing, label encoding, as well as normalization method was done and for validation K-folding technique (K=5) have been used. The highest accuracy has been achieved by using Random Oversampling with Ensemble Model (Stacking) which is 97.31% (std: 0.0033, k=5). The 99% confidence interval for the model's accuracy is [0.956, 0.988].

Keywords: Customer churn Prediction, Machine Learning, Oversampling, Ensemble Model, K-fold Validation

1. Introduction:

Customer churn, the likelihood of customers discontinuing business with a company, is a critical concern for enterprises spanning various industries, including finance, telecommunications, and online services. The duration a customer stays with a company directly influences their lifetime value, highlighting the imperative nature of addressing churn issues. Retaining existing customers proves to be more cost-effective than acquiring new ones, emphasizing the pivotal role of customer relationship enhancement in business success [1]. This paper delves into customer churn prediction as the primary strategy for proactive churn prevention. Identifying customers prone to defection in advance allows companies to implement targeted initiatives, such as special programs or incentives, to mitigate churn risk.

The widely adopted binary classification model is employed for this purpose, categorizing customers as likely to churn or not [2]. Various methodologies, including AdaBoost classifier, Random Forest, KNN, XGB have been explored to address this challenge. While decision tree-based algorithms reveal classification rules and neural networks determine prediction probabilities, both lack explicit expression of uncovered patterns. Genetic algorithms, although accurate, face challenges in determining prediction likelihoods, hindering their applicability to churn prediction tasks requiring customer ranking. Scholars have introduced alternative methods, such as Bayesian classifiers, improved one-class SVM, sequential pattern association analysis, and survival analysis, each contributing to the growing landscape of churn prediction techniques [1]. Notably, Lemmens and Croux pioneered the application of ensemble learning algorithms, specifically bagging and stochastic gradient boosting, demonstrating significant improvements in prediction accuracy for a U.S. wireless telecom company's customer database [3]. Customers can readily switch from one organization, such as a bank, to another in pursuit of improved service quality or more competitive pricing. Organizations widely acknowledge that acquiring new customers is significantly more challenging and costly than retaining existing clients [4]. Organizations face the challenge of delivering reliable services to customers while maintaining a good working partnership with them. One of their primary focuses is on customer churn, which occurs when clients or subscribers discontinue their engagement with a company or service. Winning new business from clients involves going through the sales pipeline, while customer retention is typically more cost-effective. Therefore, the need for a system that can accurately predict customer churn in the early stages is crucial for any organization [5].

We use 5 different machine learning techniques to predict customer churn in the banking sector: AdaBoost classifier, XGB, k-nearest Neighbor (KNN) Random Forest (RF) and a stacking model. To find the most accurate model, conducted experiments using these classifiers under different conditions and performed feature selection methods to identify the most relevant features. The experimentation was conducted on the churn modeling dataset from Kaggle. Dataset was preprocessed using proper techniques such as label encoding and normalization. To prove model consistency K-fold validation was performed and 99% confidence interval was calculated after 1000 iteration. We believe that our research will make a significant contribution to the field of prediction of bank churn. This paper aims to build a framework that can predict client churn in the banking sector using some Machine learning techniques.

The remainder of this paper is structured as follows. In Section 2, we explain the related work. Dataset, preprocessing, machine learning models are presented in Research Methodology in Section 3. We explain the experiment results and discussion about them in Section 4. Some concluding remarks and ideas for future work are given in Section 5.

2. Literature Review:

Research on predicting churn for bank customers has garnered significant attention in recent years, particularly with the advancement of artificial intelligence techniques. This highlights the growing importance of customer retention in the banking industry, focusing on various methodologies and techniques for forecasting customer churn using data analytics and machine learning.

The Analytics Department et al. proposed a comprehensive study using statistical and machine learning models, including Random Forest, ANN, XG-BOOST,

Decision Tree, and Logistic Regression, to investigate churn prediction for savings account customers in Indian commercial banks[6]. With a 78% accuracy rate in predicting customer attrition, the Random Forest model performs better than the others. Customer vintage, age, average balance, occupation code, population type, average debit amount, and transaction frequency are important influencing factors. The study emphasizes how much less expensive it is to retain current customers as opposed to finding new ones and suggests tailored advertising campaigns to reduce attrition. The conclusion adds important context to the literature on machine learning models for predicting customer churn in the banking industry and emphasizes the need to improve customer satisfaction to reduce churn and preserve deposits.

Shao Jinbol et al. proposed a comprehensive study that examines the use of AdaBoost for customer churn prediction in CRM by contrasting its versions—Real, Gentle, and Modest AdaBoost—on a credit debt database from a Chinese bank [7]. It emphasizes how much less expensive it is to keep current clients than to find new ones. When it comes to prediction accuracy, AdaBoost beats SVM. The paper discusses attribute selection, sampling, and data processing, demonstrating the effectiveness of AdaBoost—particularly Real and Gentle AdaBoost—in managing unbalanced datasets. In identifying high-risk customers and improving CRM strategies, the conclusion highlights AdaBoost's usefulness.

Alisa Bilal Zorić et al. Express uses data mining, specifically neural networks, to predict customer churn in the banking industry [2]. It identifies customers at risk of leaving and determines the value of retaining them. The study uses the Alyuda NeuroIntelligence software package and real-world data from a small Croatian bank. The key findings suggest that clients using more bank services are more loyal, emphasizing the importance of retaining customers with fewer than three products. The methodology involves iterative data mining, including problem definition, data gathering, model building, and knowledge deployment. The study emphasizes the importance of tailoring services to different customer segments, such as students, and the limitations of neural networks.

Manas Rahman et al. proposed The study uses KNN, SVM, Decision Tree, Random Forest classifiers with mRMR and Relief feature selection, oversampling for data imbalance, and other techniques to address customer churn in the banking sector [8]. The Random Forest model with post-oversampling attains the highest accuracy of 95.74% when it is applied to a Kaggle dataset containing 10,000 clients. In support of larger datasets, the conclusion advocates for customer retention with more products while acknowledging data limitations. Certain classifiers are impacted by feature selection, but tree classifiers are not. Model

accuracy in churn prediction is not specifically discussed in the study; instead, it is mentioned in passing in the literature review.
Shaoying Cui et al. proposed a comprehensive study that highlights the value of big data in customer relationship management and focuses on using an improved Fuzzy C-
Means (FCM) algorithm to predict customer attrition in the banking industry [9]. The suggested algorithm is effectively implemented on actual data from a commercial bank and includes an improved validity function that takes into account separation and compactness measures. The findings show that customer clusters including stable, lost, gold, and churn groups can be identified with a high degree of accuracy. The paper, however, would benefit from a more in-depth analysis of the predictive accuracy of the algorithm. The literature review is also mentioned in passing, and investigating different data mining techniques would improve the paper's value to the field.
Vijayakumar Bharathi S et al. proposed the study uses machine learning techniques to predict youth customer defection in retail banking[10]. To predict churn, the authors gathered data from 602 young adult bank customers in India and used a variety of machine learning algorithms, including ensembles. The effectiveness of the ExtraTreeClassifier model was demonstrated by its 92% accuracy rate and 91.88% AUC. According to the study, accessibility to ATMs, zerointerest personal loans, the lack of mobile banking, and customer service are all important churn drivers. The study highlights the value of customer retention for the banking sector, especially for the younger, tech-savvy demographic. The results fill a research gap in the retail banking industry regarding churn prediction, and they provide banks with useful information to improve customer retention.
Amgad Muneer et al. addressed the critical problem of forecasting customer attrition in the banking sector, highlighting the importance of client retention for bank expansion and profitability [11]. They suggested utilizing three intelligent models—Random Forest (RF), AdaBoost, and Support Vector Machine (SVM)—in a machine learningbased approach. The study demonstrated that the Random Forest model outperformed the other models with an astounding 91.90 F1 score and an overall accuracy of 88.7%, especially when used in conjunction with the Synthetic Minority Oversampling Technique (SMOTE) to handle imbalanced data. The authors pre-processed the data, carried out a comprehensive exploratory data analysis, and used SMOTE to balance the dataset. The suggested Random Forest predictor had an advantage over other models in predicting customer attrition in the banking industry, as demonstrated by a literature review and comparison. The study's results provide valuable insights for banks to strategize and implement customer retention measures effectively.

Seyed Hossein Iranmanesh et al. focus on predicting customer churn for a retail bank in Iran using advanced data analysis techniques [12]. The model uses machine learning tools and deep learning techniques, specifically a neural network model, to classify customers based on their churn rate. The study highlights the importance of understanding customer dynamics and behavior for effective retention, particularly in the banking sector. Job types, age, and occupation demographics play a significant role in customer churn, with food services and technical-economic sectors showing the highest rates. The model's validation involves presenting results to the bank's business analysis unit

Table 1. Summary of existing literature in Predicting Churn for Bank Customers

Reference	Title of the paper	Models	Accuracy	Limitations
[5]	Analysis and prediction of bank user chum based on ensemble learning algorithm	- Catboost, Lightgbm, and Random Forest	highest accuracy of 90%	Uses quarterly data, but considering temporal dependencie s could improve predictions due to potential changes in customer behavior over time.
[6]	Churn Prediction for Savings Bank Customers: A Machine Learning Approach	Random Forest, ANN, XG-BOOST, Decision Tree, and Logistic Regression	highest accuracy of 78%	Details about the parameters used in the models, especially for Random Forest, are not provided.

[7]	The Application of AdaBoost in Customer Churn Prediction	AdaBoost, SVM	highest accuracy of 80%	The paper briefly touches upon the issue of imbalanced datasets and the use of balanced sampling
[8]	Machine Learning Based Customer Churn Prediction in Banking	KNN, SVM, Decision Tree, Random Forest	highest accuracy of 95.74%	Sensitivity analysis of interval variables and the potential computation al requirement s associated with the proposed method.
[10]	An Ensemble Model for Predicting Retail Banking Churn in the Youth Segment of Customers	ExtraTreeClassifier	highest accuracy of 92%	The study's results may be biased due to the small sample size of 1524 customers from a single bank.
[11]	Predicting customers churning in banking industry: A machine learning approach	Random Forest (RF), AdaBoost, and Support Vector Machine (SVM)	highest accuracy of 88.7%	Lack of specifics on the preprocessing steps used in the data makes it challenging to evaluate their influence on the results.
[13]	Customer churn prediction using improved balanced random forests	decision trees, artificial neural networks, and class-weighted core support vector machines		The study focuses on the banking industry. It is unclear how well IBRF would perform in other industries or with

Table-1 shows the summary of the extensive literature review that has been done. It shows the highest accuracy among existing research is 95.74%.

3. Research Methodology:

The methodology of our research is presented in Figure 1. The very first step was collecting a dataset from Kaggle. After that the data was pre-processed using proper pre-processing technique like label encoding, normalization etc. Then among the available features; some features that holds more significance were select to train the models. The models that were used in this research are Random Forest, AdaBoost, KNeighbors, XGBoost and an Ensemble model that have Random Forest, AdaBoost, XGBoost as base models and Random Forest for final model. These model's performances are than evaluated based on different metrics to find the best models. The metrics that we used were accuracy, precision, recall and f1-score.

Data Preparation**:** Split dataset D into D_{train} and D_{test} . Apply oversampling techniques to D_{train} to get $D_{smote\text{-}enn}$ and D_{random}

Training: For each model M∈{RF,AB,KNN,XGB,Stack} where Stack= Ensemble(RF(RF,AB,XGB)) and each dataset $D_{train} \in \{D_{train}, D_{smote\text{-}enn}, D_{random}\}$.train using M and evaluate

Evaluation and Selection: Perform k-fold validation for each model on D_{test} a compute CVScore. Select the best model as Best Model =arg max {accuracy, recall, precision, f1-score}.

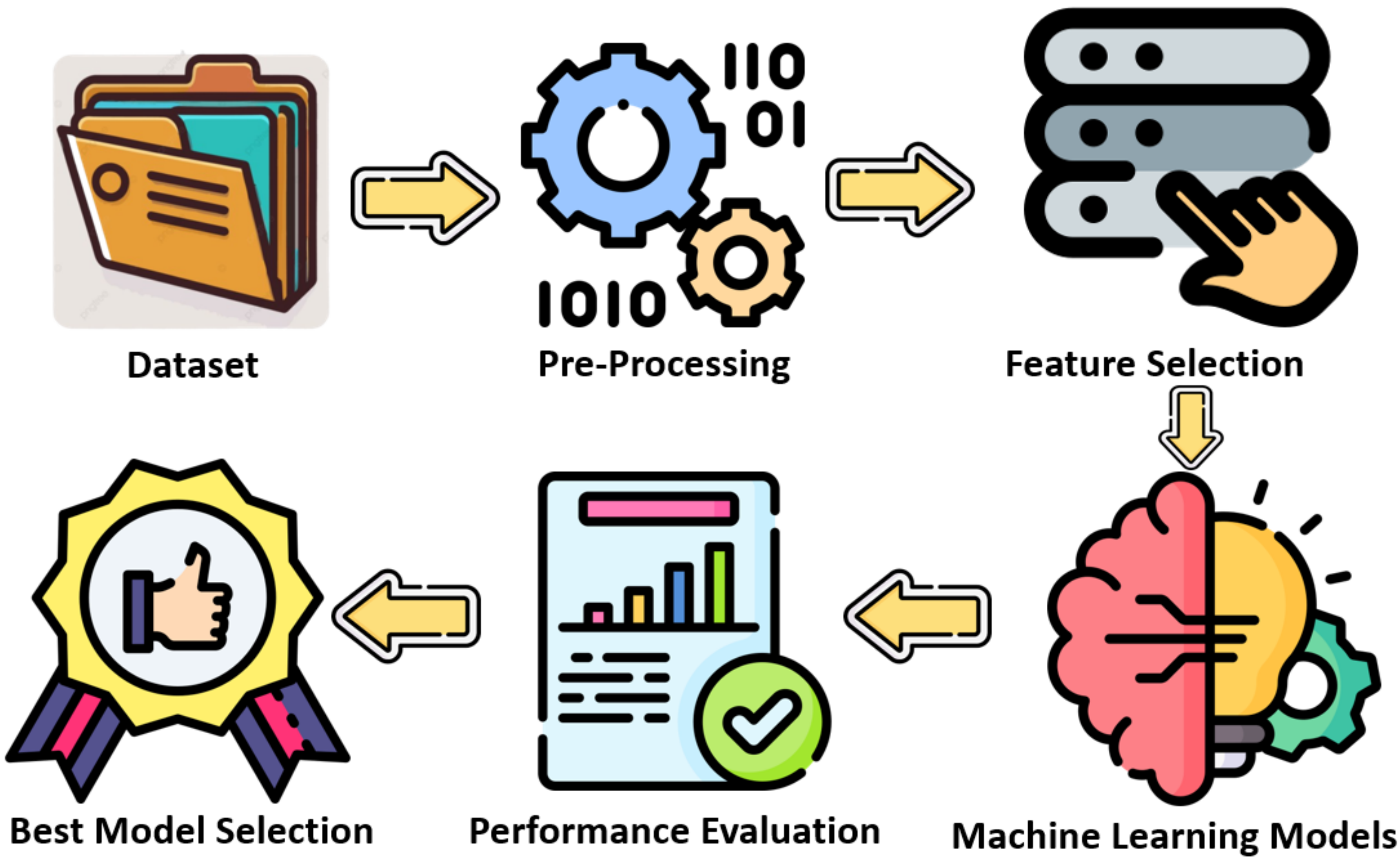

Fig 1. Methodology

Dataset and Pre-processing:

We have collected a "Predicting Churn for Bank Customers" dataset from Kaggle. The dataset has 14 columns and 10000 data points. All of these 14 features will not be used to train the models. To understand the significance of each feature we can look correlation matrix. For summarizing the dataset and identifying the patterns and relationships between feature value and target value correlation matrix was generated.

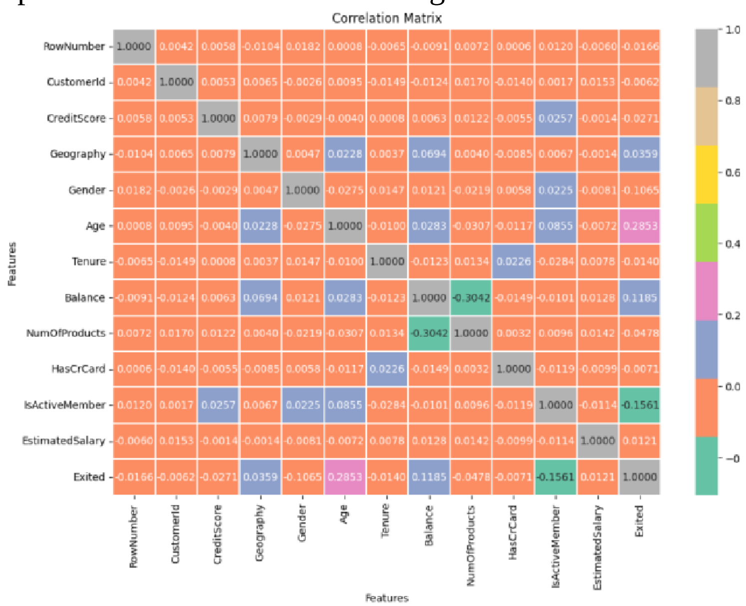

Fig 2. Correlation Matrix

From the correlation matrix shown in Figure 2, we can say no individual feature affects too much the target variable but together it makes a result. In order to create an accurate prediction model, we need to utilize multiple features.

Only the surname, rownumber and customerId these three columns have been dropped. To train the model CreditScore, Geography, Gender, Age, Tenure, Balance, NumOfProducts, HasCrCard, IsActiveMember, and Estimated Salary these features were used. In table-2, the description of the selected features has been given

The target column is excited. If exited is 0 that means customers exit on the other hand 1 means customers will continue with this bank.

Table 2. Selected Features for Model Training

Features	Description
Credit Score	The credit score of a customer can provide insights into their creditworthiness and financial stability.
Geography	The geographical location of a customer may play a role in their likelihood to churn due to cultural or economic factors
Gender	Gender can sometimes be a relevant factor, as certain industries may have different churn patterns for male and female customers.
Age	Age is often a significant predictor, as younger or older customers may have different behaviors and preferences
Tenure	The length of time a customer has been with the company can be indicative of loyalty and satisfaction.
Balance	The account balance of a customer can be a relevant factor, as high balances might indicate engagement, while low balances might suggest dissatisfaction.
Number of Products	The number of products a customer has with the company may impact their likelihood to churn.
Has Credit Card (HasCrCard)	Whether a customer has a credit card with the company might influence their decision to stay or leave.
Active Member	Whether a customer is an active member or not can be crucial, as engaged customers are less likely to churn.
Estimated Salary	The estimated salary of a customer can provide additional context about their financial situation.

Standard pre-processing steps has been applied before training the model. Since the dataset was a clean dataset pre-processing became a bit easier. Label encoder on the features that are categorical. Normalization was implanted for numeric features. Since the dataset was extremely imbalanced 3.91:1 oversampling was important. For data oversampling we used both Random over sampling and SMOTE-ENN.A statistical technique is what the synthetic minority oversampling technique (SMOTE) is. The objective of this strategy is to progressively add more examples to our collection in a balanced way. To feed our model, we take our current minority situations and create new instances of them. Because the technique takes a sample of the feature space for each target class and its closest neighbors, new instances are not just duplicates of existing minority cases; rather, they incorporate features from the target case and those from its neighbors. On the other hand, Random oversampling is much simpler. It just randomly generates more sample in the minority class. It is the easiest oversampling method. The oversampling technique is presented in Figure 3.

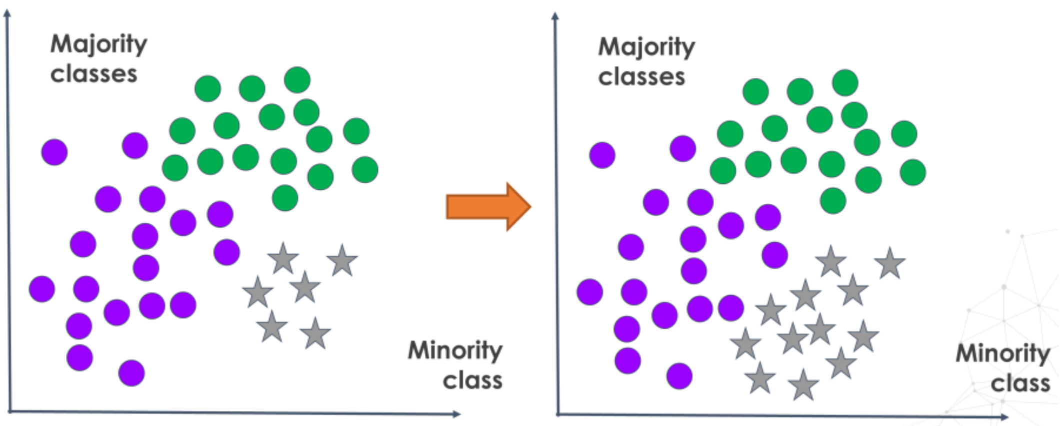

Fig 3. Over Sampling Technique

Machine Learning Models:

A. Random Forest

Random forest is an algorithm technique that combines multiple decision trees to get the optimal result [20]. It is an ensemble model. For training the decision trees random forest uses bootstrapping method. To predict when predicting class of a datapoint; each decision tree makes a prediction and final prediction is made after voting. The datapoint is assigned to the class that got the most vote. To enhance classification performance, especially for ambiguous data, a broad granular random forest technique is suggested [15]

Let D be the original dataset with N samples. Random Forest creates T bootstrap samples $D_1, D_2, \ldots, D_T$ from D. Each bootstrap sample D_t is obtained by randomly sampling with replacement.

D_t=BootstrapSample(D)

Each decision tree Tree_t is trained on its respective bootstrap sample D_T using the subset of features F_t at each node split. $\text{Tree}_t = \text{TrainDecisionTree}(D_t, F_t)$

For a new sample x, each tree Tree_t makes a prediction. The final prediction y^is obtained by aggregating the predictions from all trees

$$\hat{y} = \underset{c}{\text{argmax}}\left(\frac{1}{T}\sum_{t=1}^{T} I(\text{Tree}_t(x) = c)\right)$$

where I (·) is an indicator function that equals 1 if $\text{Tree}_t(x)$ equals class c, and 0 otherwise. The prediction is the class with the most votes.

B. Adaboost Classifier

For classification problems, the Adaboost classifier has demonstrated great accuracy in different applications [18]. It is model that utilizes boosting technique; it trains a lot of weak learners to get a strong learner [21]. The final prediction is a weighted vote of the predictions from all weak learners

It starts by assigning equal weights to all training samples. If there are N samples, each sample i

$wi = \frac{1}{N} w_i = N_1$

In each iteration t, A weak learner is trained on the weighted training data and he goal is to find a weak learner that minimizes the weighted error ϵ_t.

$$\epsilon_t = \frac{\sum_{i=1}^{N} w_i \cdot I(y_i \neq h_t(x_i))}{\sum_{i=1}^{N} w_i}$$

Compute the weight α_t of weak learners based on its error:

$$\alpha_t = \frac{1}{2}\ln\left(\frac{1-\epsilon_t}{\epsilon_t}\right)$$

Increase the weights of misclassified samples and decrease the weights of correctly classified samples:

$$w_i \leftarrow w_i \cdot \exp(\alpha_t \cdot I(y_i \neq h_t(x_i)))$$

After training a fixed number of weak learners or reaching a stopping criterion, combine them into a final strong classifier.

$$H(x) = \text{sign}\left(\sum_{t=1}^{T} \alpha_t \cdot h_t(x)\right)$$

C. KNeighborsclassifier

KNeighborsClassifier is a type of instance-based learning where the model does not explicitly learn a training function [22]. It actually remembers the training instances and makes prediction based on how similar the new datapoint is to previously trained instances. It gets assigned to the class of the datapoint that it is closet to.To classify a new sample, the algorithm calculates the distance between the sample and all points in the training dataset.

$$\text{Distance}(x, x_i) = \sqrt{\sum_{j=1}^{d} (x_j - x_{i,j})^2}$$

where x is the new sample, x_i is a training sample, and d is the number of features. The algorithm identifies the k closest training samples to the new sample based on the calculated distances.
For classification, the class label of the new sample is determined by a majority vote among its k nearest neighbors.

$$\hat{y} = \underset{c}{\text{argmax}}\left(\sum_{i=1}^{k} I(y_i = c)\right)$$

D. XGBoost Classifier
XGBoost (Extreme Gradient Boosting) is a highly efficient and scalable implementation of gradient boosting designed to improve the performance and accuracy of machine learning models [23]. It works similar to Adaboost, it combines weak learners to create a new strong learner. t optimizes the model by minimizing a loss function using gradient descent. The algorithm iteratively adds trees that reduce the residual errors (differences between actual and predicted values) of the previous model.

Loss = Objective Function + Regularization Term

XGBoost constructs trees sequentially, with each tree trying to correct the errors of the combined ensemble of previous trees.

$$\hat{y} = \sum_{t=1}^{T} \text{Tree}_t(x)$$

The algorithm uses gradient descent to minimize the objective function. During each iteration, the algorithm computes gradients of the loss function and updates the model accordingly.

$$g_i = \frac{\partial \text{Loss}}{\partial \hat{y}_i}$$

$$h_i = \frac{\partial^2 \text{Loss}}{\partial \hat{y}_i^2}$$

Where g_i, h_i are the gradient and Hessian (second derivative) of the loss function.

XGBoost includes L1 (lasso) and L2 (ridge) regularization to prevent overfitting

$$\text{Regularization Term} = \lambda \sum_{t=1}^{T} \text{Complexity of Tree}_t + \gamma \sum_{t=1}^{T} \text{Number of Leaves in Tree}_t$$

E. Ensemble Model (Stacking)
A variety of systems can function better when Ensemble models are employed. The Ensemble Model Output Statistics (EMOS) is a popular parametric technique for calibrating ensemble forecasts in the field of weather forecasting. Ensemble learning combines multiple machine learning models into a single model. For prediction multiple techniques can be used like bagging, boosting and stacking. We have used the stacking method. In our stacking model we used Random Forest, XGBoost and Adaboost as base models and Random Forest as the meta (final) model since it's the algorithm that achieved highest accuracy among all the algorithms. The final models take the base model's prediction as input and as output predict the class.

Let fi(X) denote the prediction of the i-th base model, where $i \in \{RF, XGB, AB\}$ For input features X, the predictions from each base model can be written as:

$$P_{RF} = f_{RF}(X)$$
$$P_{XGB} = f_{XGB}(X)$$
$$P_{AB} = f_{AB}(X)$$

$$X_{\text{stacked}} = \sum_{i=1}^{3} f_i(X)$$

Where $f_i(X)$ represents the predictions from the base models

Let $gRF_{meta}(X_{stacked})$ denote the meta-model (Random Forest) function. The final prediction y^ is given by:

$$\hat{y} = g_{\text{RF}_{\text{meta}}}\left(\begin{bmatrix} f_{RF}(X) \\ f_{XGB}(X) \\ f_{AB}(X) \end{bmatrix}\right)$$

Validation:

Validation is an important factor to determine how consistent a model is. To evaluate model performance multiple validation techniques can be utilized. K-fold validation was implanted in this research. In k-fold validation dataset is split into k-parts, then only one of those parts are used as test data and the rest are used for training the model. This process continue until all the parts has been used as test data. K-fold validation can determine whether a model is truly a high performance. For k fold validation we used k=5. Figure-4 shows k-fold validation technique.

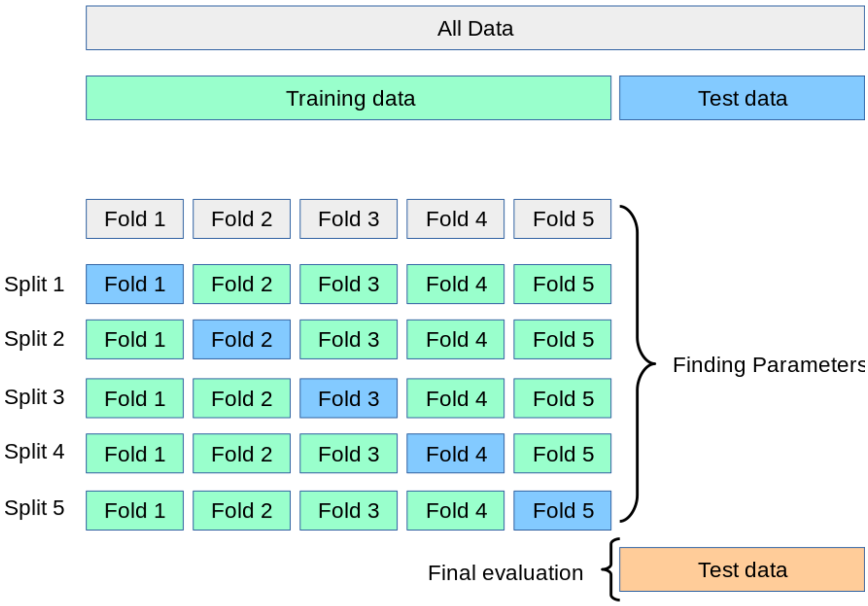

Fig 4. K-fold Validation

4. Result and Discussion:

In the following section, Firstly, we discuss the results obtained from the experiments conducted in this study. Then, the comparative analysis was provided to provide the readers with a clear comparison between the proposed classifiers in this study and the state of the art. Accuracy, Precision, Recall and F1-Score these 4 popular evaluation metrics were used for comparison.

Evaluation Metrics:

a) Accuracy:

accuracy $= \frac{TP+TN}{TP+FP+TN+FN}$

b) Precision:

Precision $= \frac{TP}{TP+FP}$

c) Recall:

Recall $= \frac{TP}{TP+FN}$

d) F1 Score:

F1 score $= \frac{2*precision*recall}{precision+recall}$

Table 3. The performance of proposed models on original data before oversampling

Classifiers (%)	Recall	F1-Score	Precision	Accuracy (%)
Random Forest	0.45 (std:0.0020)	0.56 (std: 0.0272)	0.77 (std: 0.0229)	86.04% (std: 0.0053)
Adaboost Classifier	0.47 (std: 0.0300)	0.57 (std: 0.0229)	0.72 (std: 0.0259)	85.46% (std: 0.0060)
K-Neighbors Classifier	0.08 (std: 0.0154)	0.12 (std: 0.0214)	0.22 (std: 0.0343)	75.61% (std: 0.0029)
XGBoost classifier	0.49 (std: 0.0259)	0.58 (std: 0.0179	0.71 (std: 0.0189)	85.46% (std: 0.0046)
Ensemble Model (Stacking Technique)	0.56 (std: 0.0162)	0.47 (std: 0.0105)	0.72 (std: 0.0250)	85.19% (std: 0.0071)

Table 4. The performance of proposed after SMOTE-ENN oversampling

Classifiers (%)	Recall	F1-Score	Precision	Accuracy (%)
Random Forest	0.91 (std: 0.0118)	0.89 (std: 0.0091)	0.87 (std: 0.0082)	87.43% (std: 0.0101)
Adaboost Classifier	0.87 (std: 0.0159)	0.86 (std: 0.0119)	0.85 (std: 0.0113)	83.95% (std: 0.0129)
K-Neighbors Classifier	0.94 (std: 0.0052)	0.9038 (std: 0.0073)	0.87 (std: 0.0151)	88.82% (std: 0.0094)
XGBoost classifier	0.90 (std: 0.0077)	0.89 (std: 0.0050)	0.88 (std: 0.0127)	88.01% (std: 0.0065)
Ensemble Model (Stacking Technique)	0.89 (std: 0.0113)	0.89 (std: 0.0086)	0.89 (std: 0.0094)	87.68% (std: 0.0043)

Table 5. The performance of proposed models after Random oversampling

Classifiers (%)	Recall	F1-Score	Precision	Accuracy (%)
Random Forest	0.98 (std: 0.0018)	0.96 (std: 0.0028)	0.93 (std: 0.0063)	95.37% (std: 0.0030)
Adaboost Classifier	0.74 (std: 0.0068)	0.76 (std: 0.0050)	0.78 (std: 0.0043)	76.90% (std: 0.0044)
K-Neighbors Classifier	0.81 (std: 0.0088)	0.73 (std: 0.0066)	0.66 (std: 0.0071)	69.36% (std: 0.0078)
XGBoost classifier	0.92 (std: 0.0068)	0.89 (std: 0.0058)	0.87 (std: 0.0086)	89.15% (std: 0.0062)
Ensemble Model (Stacking Technique)	0.97 (std: 0.0048)	0.97 (std: 0.0025)	0.98 (std: 0.0033)	97.31% (std: 0.0037)

Table 3 represents the performance of the models on original data. As we can see in the table-3 Random Forest has achieved the best accuracy with 86.04%. But we can see that just like the other models it has terrible recall, f1-score and even the precision is not great. The other model’s performance was also terrible. After applying SMOTE-ENN we see a far better result shown in table-4. The most improvement can be noticed in K-Neighbors Classifier which achieved the highest accuracy of 88.82%. All the model’s accuracy was over 83% which indicates a big improvement. After applying Random Oversampling, we can see a similar improvement of result illustrated in table-5. Only Adaboost and K-Neighbors Classifier accuracy have become lower but there is massive improvement in precision, recall and f1-score just like the other models. The other three models’ performance went up on all four metrics

In summary, for Random Forest and XGboost and Ensemble Model their performance become better for both Random Oversampling and SMOTE-ENN. But they were significantly better after applying Random Oversampling. For Adaboost its accuracy drops a lot after Random Oversampling but for SMOTE-ENN it’s accuracy stays almost similar. In both cases there are massive improvements in the other three metrics. In case of K-Neighbors its accuracy drops but other metrics have improved after Random oversampling. But after applying SMOTE-ENN we see improvement in all metrics. We have done K-fold validation on all of them. Ensemble Model has achieved the highest accuracy of 97.31% after random oversampling.

The results based on Ensemble Model are substantially higher than those based on other models, as tables 3,4 and 5 demonstrate. Ensemble model accuracy is high in all three methods. To predict customer churn in the banking sector, we consequently chose the Ensemble model (stacking).

After a thousand iterations, the 99% confidence interval for the model's accuracy is [0.956, 0.988]. We know that,

$$SE = \frac{s}{\sqrt{n}}$$
$$ME = Z \times SE$$
$$\mathrm{CI} = \bar{x} \pm ME$$

For 99% confidence interval z=2.576
Our Model's Standard Error, SE=0.62, Mean, x=97.15% Margin of error, ME= 2.576×0.62= 1.6
Confidence Interval, CI= 97.15 ± 1.6 = [0.956,0.988]

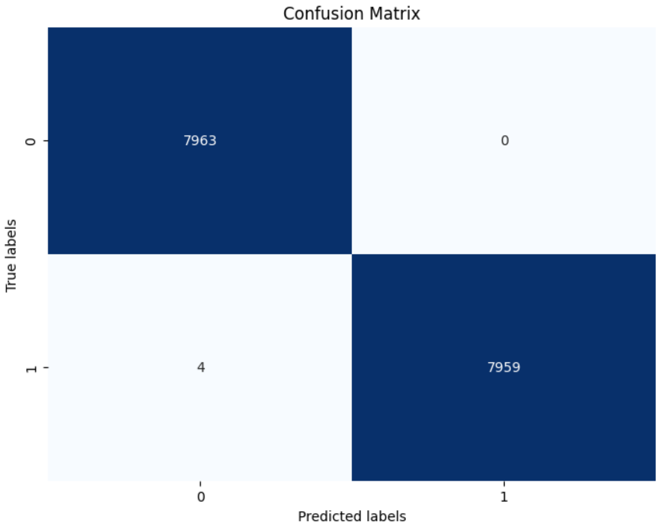

Fig 5. Confusion Matrix for Ensemble Model

The confusion matrix in Figure 5 illustrate how impressive the ensemble model is. It made only 4 wrong prediction. It's accuracy for identifying both class accuracy is an amazing feat. This confusion matrix further proves the effectiveness of ensemble modeling technique and random oversampling technique. Figure 6 on the other hand shows the ROC curve of confusion matrix. As expected AUC (Area Under the Curve) is 1, the highest possible value for ensemble modeling. It is a further testamant to the effectiveness of our proposed methodology. It handled the issue of imbalanced dataset emaculately.

The comparison is limited to the available metrics, but it essentially demonstrate that the proposed method potential for predicting customer churning in the banking industry. The best model is Ensemble Model (Stacking) with an accuracy of 97.31% (std: 0.0037), a Precision Score of 0.98 (std: 0.0033), a F1-score of 0.97 (std:0025) and a recall score of 0.97 (0.0048), using random oversampling , validated by k-fold validation (k=5)

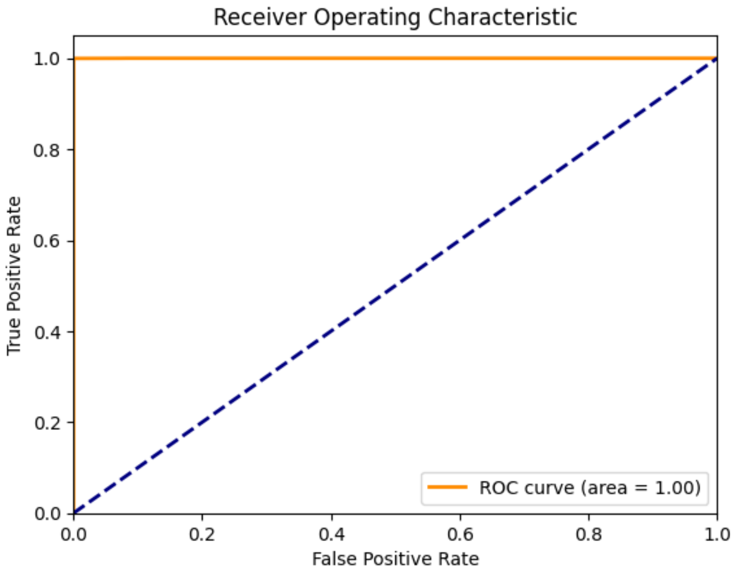

Fig 6. ROC Curve for Ensemble Model

5. Conclusion:

An important development in the financial industry, the study of churn prediction for bank customers provides useful information and instruments to help banks take proactive measures to manage their client base. The researchers employed five different models and concluded that Ensemble Model (Stacking Technique) that uses Random Forest, AdaBoost Classifier, and XGBoost classifiers as base models and Random Forest as the final model returns the best result. The issue of imbalanced dataset is resolved by random oversampling. With an accuracy of 97.31% (std:0.0037), the results show that the suggested methodology performs better than alternative approaches. This is further validated by the k-fold validation that was performed (k=5). The 99% confidence interval for the model's accuracy is [0.956, 0.988]; which a further testament of the model consistency. This superior performance suggests that Ensemble Modeling (stacking) is a highly effective technique for identifying customers at risk of churn, enabling proactive interventions and improved customer retention strategies. Furthermore, the application of the Random Oversampling technique demonstrated significant improvement in the performance of all models, highlighting its effectiveness in addressing imbalanced datasets. The study offers valuable insights into churn prediction in the banking sector, enabling early identification and proactive customer retention strategies, thereby enhancing banks' long-term success and customer relationships.

References:

1. S Neslin, S Gupta, W Kamakura, J Lu, C Mason,"Defection Detection: Improving Predictive Accuracy of Customer Chum Models", Working Paper, Teradata Center at Duke University, 2004.3
2. Zorić, A. B. (2016). Predicting Customer Churn in Banking Industry using Neural Networks. Interdisciplinary Description of Complex Systems, 14(2), 116–124. https://doi.org/10.7906/indecs.14.2.1
3. A Lemmens, C Croux, "Bagging and Boosting Classification Trees to predict churn", Journal of Marketing Research, Vol. 43, No. 2, pp 276286.2006
4. M.-K. Kim, M.-C. Park, and D.-H. Jeong, “The effects of customer satisfaction and switching barrier on customer loyalty in Korean mobile telecommunication services,” Telecommunications policy, vol. 28, no. 2, pp. 145–159, 2004
5. Y. Deng, D. Li, L. Yang, J. Tang and J. Zhao, "Analysis and prediction of bank user churn based on ensemble learning algorithm," 2021 IEEE International Conference on Power Electronics, Computer Applications (ICPECA), Shenyang, China, 2021, pp. 288-291, doi: 10.1109/ICPECA51329.2021.9362520.
6. Verma, Prashant (2020) "Churn Prediction for Savings Bank Customers: A Machine Learning Approach," Journal of Statistics Applications & Probability: Vol. 9: Iss. 3, Article 10.
7. S. Jinbo, Li Xiu and L. Wenhuang, "The Application ofAdaBoost in Customer Churn Prediction," 2007 International Conference on Service Systems and Service Management, Chengdu, China, 2007, pp. 1-6, doi:
8. 10.1109/ICSSSM.2007.4280172.M. Young, The Technical Writer’s Handbook. Mill Valley, CA: University Science, 1989.
9. M. Rahman and V. Kumar, "Machine Learning Based Customer Churn Prediction In Banking," 2020 4th International Conference on Electronics, Communication and Aerospace Technology (ICECA), Coimbatore, India, 2020, pp. 1196-1201, doi: 10.1109/ICECA49313.2020.9297529
10. S. Cui and N. Ding, "Customer churn prediction using improved FCM algorithm," 2017 3rd International Conference on Information Management (ICIM), Chengdu, China, 2017, pp. 112-117, doi: 10.1109/INFOMAN.2017.7950357.
11. Bharathi S V, Pramod D, Raman R. An Ensemble Model for Predicting Retail Banking Churn in the Youth Segment of Customers. Data. 2022; 7(5):61. https://doi.org/10.3390/data7050061M. Young, The Technical Writer’s Handbook. Mill Valley, CA: University Science, 1989.
12. Muneer, A., Ali, R. F., Alghamdi, A., Taib, S. M., Almaghthawi, A., & Ghaleb, E. a. A. (2022). Predicting customers churning in banking industry: A machine learning approach. Indonesian Journal of Electrical Engineering and Computer Science, 26(1), 539. https://doi.org/10.11591/ijeecs.v26.i1.pp539-549M. Young, The Technical Writer’s Handbook. Mill Valley, CA: University Science, 1989.

13. Iranmanesh, Seyed Hossein, et al. "Customer Churn Prediction Using Artificial Neural Network: An Analytical CRM Application." 3rd European International Conference on Industrial Engineering and Operations Management. 2019,M. Young, The Technical Writer's Handbook. Mill Valley, CA: University Science, 1989.
14. Xie, Y., Li, X., Ngai, E. W., & Ying, W. (2009). Customer churn prediction using improved balanced random forests. Expert Systems
15. With Applications, 36(3), 5445–5449.
16. https://doi.org/10.1016/j.eswa.2008.06.121
17. Dierckx, G. (2004). Logistic Regression model. Encyclopedia of Actuarial Science. https://doi.org/10.1002/9780470012505.tal017M. Young, The Technical Writer's Handbook. Mill Valley, CA: University Science, 1989.
18. Slimani, C., Wu, C., Rubini, S., Chang, Y., & Boukhobza, J. (2023). Accelerating random forest on Memory-Constrained Devices through data storage optimization. IEEE Transactions on Computers, 72(6), 1595–1609. https://doi.org/10.1109/tc.2022.3215898M. Young, The Technical Writer's Handbook. Mill Valley, CA: University Science, 1989.
19. Kutschenreiter-Praszkiewicz, I. (2023). Development of a neural network structure for identifying begin-end points in the assembly process. Journal of Machine Engineering. https://doi.org/10.36897/jme/163318M. Young, The Technical Writer's Handbook. Mill Valley, CA: University Science, 1989.
20. Medvedev, A., Delvenne, J., & Lambiotte, R. (2018). Modelling structure and predicting dynamics of discussion threads in online boards. Journal of Complex Networks, 7(1), 67–82. https://doi.org/10.1093/comnet/cny010M. Young, The Technical Writer's Handbook. Mill Valley, CA: University Science, 1989.
21. Zhang, Y., & Wang, L. (2023). An AdaBoost Method with K′KMeans Bayes Classifier for Imbalanced Data. Mathematics, 11(8), 1878. https://doi.org/10.3390/math11081878.
22. Singh, P. (2023). Intepretable Deep Gaussian Naive Bayes Algorithm (IDGNBA) based task offloading framework for Edge-Cloud computing. *International Journal of Data Informatics and Intelligent Computing*, *2*(2), 1–10. https://doi.org/10.59461/ijdiic.v2i2.57
23. Breiman, L. (2001). Random forests. Machine Learning, 45(1), 5-32. https://doi.org/10.1023/A:1010933404324
24. Freund, Y., & Schapire, R. E. (1997). A decision-theoretic generalization of on-line learning and an application to boosting. Journal of Computer and System Sciences, 55(1), 119-139. https://doi.org/10.1006/jcss.1997.1504
25. Cover, T., & Hart, P. (1967). Nearest-neighbor pattern classification. IEEE Transactions on Information Theory, 13(1), 21-27. https://doi.org/10.1109/TIT.1967.1053964
26. Chen, T., & Guestrin, C. (2016). XGBoost: A scalable tree boosting system. In Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (pp. 785-794). https://doi.org/10.1145/2939672.2939785
27. Wolpert, D. H. (1992). Stacked generalization. Neural Networks, 5(2), 241-259. https://doi.org/10.1016/S0893-6080(05)80023-1
28. Kohavi, R. (1995). A study of cross-validation and bootstrap for accuracy estimation and model selection. In Proceedings of the 14th International Joint Conference on Artificial Intelligence (pp. 1137-1143). https://dl.acm.org/doi/10.5555/1643031.1643047