Grenze International Journal of Engineering and Technology, January Issue

Bank Customer Churn Prediction using Machine Learning

Vishal Parikh[1], Priyanshu Savla[2], Jay Valaki[3] and Vatsal Vadher[4]
[1-4]Computer Science and Engineering Department, Nirma University, Ahmedabad, India
Email: vishalparikh@nirmauni.ac.in, {21bce265, 21bce312, 21bce306}@nirmauni.ac.in

***Abstract*—The banking sector constantly changes, and the growing customer churn problem forces banks to implement proactive customer retention measures. [1] To predict and lessen the impact of bank customer churn, this term paper conducts a thorough investigation into the application of various machine learning models, such as logistic regression, support vector machines (SVM), decision trees, neural networks, k- nearest neighbours (KNN), and Bayesian classification. Through thorough testing and comparative research,we examine the effectiveness of every model, revealing its unique advantages and disadvantages. Our research reveals the most effective methodology, offering a solid and sophisticated response to the financial industry's critical problem of client attrition. This study also highlights how early detection of possible churners enables banks to create retention strategies proactively. Institutions may detect clients who are about to leave by using sophisticated analytics, which enables the tactical execution of individualized retention programs. In addition to addressing the rising issue of customer turnover, this proactive strategy gives financial institutions insightful information that they can use to build enduring client relationships ina highly competitive market.**

***Index Terms*— Bank Customer Churn, Customer Attrition, Predictive Modeling, Stratified Sampling, and Churn Forecasting.**

I. Introduction

Businesses use data mining tools to do churn research in an increasingly competitive industry to survive [2]. Customer churn has been the subject of several research using various methods, such as support vector machines (SVM), genetic modelling, sequential patterns, neural networks, and classification trees. Through experiments and case studies, these studies have demonstrated the potential of data mining. Nevertheless, the extant churn analysis algorithms have intrinsic constraints stemming from the distinct features of the churn prediction issue. In this paper, we tackle these issues by putting forth a strategy that has been tried and tested when forecasting churn in the banking sector. The simplicity of data gathering, the availability of large amounts of customer behaviour data, and the capacity to forecast future consumer behaviour make banks suitable study subjects [3]. The approach is not limited to a single bank; it may be used for other service businesses and a wide range of technical applications. We use the stratified method to address unbalanced datasets, optimize parameters using methods such as grid search CV, and do extensive data pretreatment, exploratory data analysis, standard scaling, and feature engineering. Implementing several machine learning models at the study's conclusion positions our work as a comprehensive investigation of cutting-edge tactics for successful churn prediction and mitigation inthe ever-changing banking industry.

Grenze ID: 01.GIJET.11.1.372_1

II. LITERATURE SURVEY

In the extensive body of literature exploring *Bank Customer Churn Prediction*, a recurring dataset structure surfaces, encompassing 10,000 records and 12 insightful columns capturing diverse customer-centric attributes. The dataset features columns such as:

1) **customer id:** A unique identifier for each customer.
2) **credit score:** Reflects the creditworthiness of the customer.
3) **country:** Specifies the customer's geographical location.
4) **gender:** Captures the gender of the customer.
5) **age:** Represents the age of the customer.
6) **tenure:** Indicates the duration of the customer's association with the bank.
7) **balance:** Reflects the financial balance in the customer's account.
8) **products number:** Counts the number of products the customer is engaged with.
9) **credit card:** Binary indicator of the customer holding a credit card.
10) **active member:** Binary indicator of the customer's active membership status.
11) **estimated salary:** Offers an approximation of the customer's income.
12) **churn:** Binary label indicating whether the customer has churned.

Notably, the redundant *customer_id* column is judiciously eliminated for prediction purposes. Pre-processing stepsinclude replacing zero values in *tenure* and *balance* with the respective column medians, mitigating the impact of outliers. A critical normalization step involves applying standard scaling to homogenize values, ensuring an equitable influence of each variable during model training. Researchers consistently underscore the importance of this comprehensive approach to dataset manipulation for robust bank customer churn prediction models.

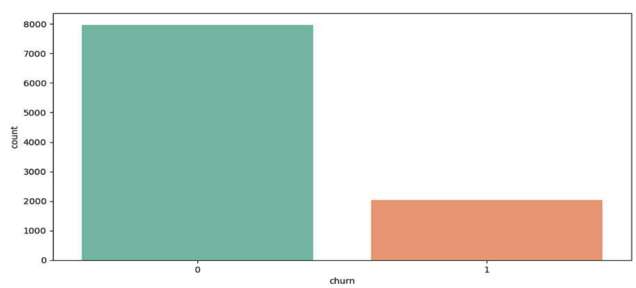

Figure 1: Customer Turnover Imbalance

Figure 1 shows a significant imbalance in the dataset's customer turnover, with about 8,000 users showing consistent loyalty and a noteworthy subset of about 2,000 customers who have churned. Understanding the imbalances in the dataset is crucial because of this notable disparity in turnover rates, which may cause skewed model predictions. The stratified approach corrects this imbalance and provides a more representative trainingset. This method maintains the target variable's (churn) proportionate distribution in training and testing datasets. By doing this, the model's capacity to generalize and produce precise predictions for both churn and non-churn events is improved since it is trained on a dataset that faithfully represents the incidence of churn in real life. Addressing the imbalance issue, this systematic application of the stratified approach results in a more reliableand fair prediction model for bank customer attrition.

Figure 2 effectively displays a notable dispersion of client attrition among various nations. France and Germany account for about 40% of the number of churned clients, with Spain making up a far smaller percentage(20%). This disparity calls for a more thorough investigation of the variables affecting customer attrition in Spain. Several contextual variables, including cultural quirks, prevailing economic situations, or differing service expectations, might explain Spain's reported reduced churn rate. Investigating these components in more detail using supplementary analysis and, if available, qualitative insights is crucial. Determining the reasons behind the lower customer attrition rate in Spain can yield important insights for improving retention tactics or customizing offerings better to suit the tastes and requirements of the regional clientele. This detailed analysis can direct focused efforts to lower customer attrition and improve the overall efficacy of churn prediction models.

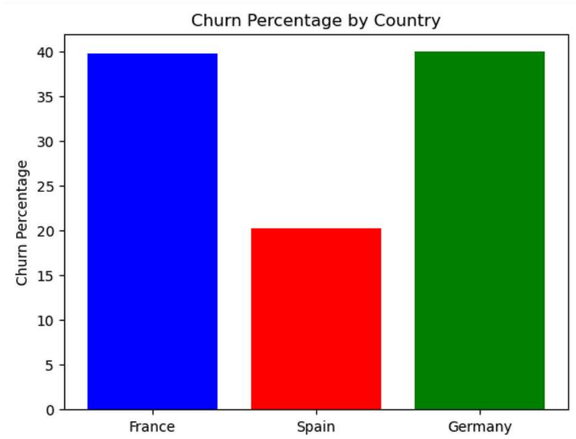

Figure 2: Dispersion of Client Attrition

III. PROBLEM STATEMENT

One of the biggest concerns in the banking and financial services industry is the ongoing difficulty in forecasting client turnover. This study uses a wide range of data points in the banking industry to solve this problem. These include essential columns such as active member, tenure, age, gender, credit score, balance, and other pertinent data. The main goal is to use this extensive dataset to create a solid machine learning-based solution [4]. The goal is to develop an automated system that can recognize future customers who are about to quit the financial institution by utilizing the insights contained within these diverse features. The institution could take proactive steps and facilitate focused activities for successful customer retention tactics with the help of such advanced technology.

IV. PROPOSED MODEL

A. Logistic Regression

Logistic Regression is employed to model the probability of customer churn based on features. The logistic function transforms the linear combination of features into a probability between 0 and 1.

$$P(\text{Churn}) = \frac{1}{1 + e^{-(\beta_0 + \beta_1 \times \mathrm{F}_1 + \beta_2 \times \mathrm{F}_2 + \ldots + \beta_n \times \mathrm{F}_n)}}$$

F = Feature of the Dataset

B. Naive Bayes

Naive Bayes calculates the probability of churn given customer features, assuming independence between features.It leverages Bayes' theorem for probability computation.

$$P(\text{Churn Features}) = \frac{P(\text{Features}|\text{Churn}) \times P(\text{Churn})}{P(\text{Features})}$$

C. Support Vector Machines (SVM)

SVM constructs a decision function to differentiate between churn and non-churn customers. [5] In the case ofa linear kernel, the decision function is a linear combination of features.

$$\text{Decision\_Function} = \beta_0 + \beta_1 \times \mathrm{F}_1 + \ldots$$

D. Decision Trees

Decision Trees are utilized to make decisions based on customer features. The tree branches at nodes based onfeature thresholds, guiding decisions until reaching leaves.

(No specific formula, as decision trees are tree-based structures.)

E. Deep Learning Sequential Model

[6] The deep learning sequential model involves a neural network with multiple layers. The output is determinedby applying an activation function to the linear combination of features.
Prediction = A F($\theta_0 + \theta_1 \times F_1 + \theta_2 \times F_2 + ... + \theta_n \times F_n$)where A–F represents the activation function.

V. BANK CUSTOMER CHURN PREDICTION ALGORITHM

The purpose of the Bank Customer Churn Prediction Algorithm is to predict and manage customer attrition in the banking sector efficiently. The first step is importing the necessary Python libraries, which lay the groundworkfor further data manipulation and modelling assignments. Next, the algorithm imports a dataset with 12 columns and 10,000 rows, the foundation for predictive analysis.
The algorithm is very careful in the Data Preprocessing step to improve the data quality. Null entries are eliminated to manage missing values, outliers are found to maintain data integrity, and feature engineering methods are used to generate new, pertinent features.
A deeper exploration of the subtleties of the dataset is required for the Exploratory Data Analysis (EDA) stage. Correlation analysis yields valuable insights into the associations among variables, whilst the visual depiction of data distributions provided by histograms facilitates comprehension of the underlying patterns.
Data transformation guarantees consistency and modelling compatibility. Label encoding is used to express categorical variables in a numerical format, while standard scaling brings numerical characteristics to a consistentscale.
Finding the best predictive model depends on the Model Training and Evaluation step. Once the dataset is divided into training and testing sets, several models are implemented, such as Neural Network, k-Nearest Neighbors,and Logistic Regression [7]. Model performance is evaluated using F1-score, accuracy, precision, and recall measures. The algorithm chooses the Neural Network as the best model in a calculated move.

Algorithm 1 Bank Customer Churn Prediction

Import Libraries: Import necessary Python libraries (NumPy, Pandas, Scikit-learn, Joblib). **Load Dataset:**Load a dataset with 10,000 rows and 12 columns. **Data Preprocessing:**Handle missing values:$data\_cleaned = data\_original.dropna()$ Detect outliers:$outliers = find\_outliers(data\_cleaned)$ Feature engineering:$engineered\_features = create\_features(data\_cleaned)$

Exploratory Data Analysis (EDA):

1) Correlation analysis:
$\rho(X, Y) = \frac{cov(X,Y)}{\sigma_X \sigma_Y}$
2) Histograms:
$histogram = plot\_histogram(data)$

Data Transformation:

1) Standard Scaling:
$X_{\text{std}} = \frac{X-\mu}{\sigma}$
2) Label Encoding:
Encode categorical variables numerically.

Model Training and Evaluation:

1) Split data (75:25) into training and testing sets.
2) Apply models:
Logistic Regression, k-Nearest Neighbors, Naive Bayes, SVC, Decision Tree, Neural Network.
3) Evaluate models using accuracy, precision, recall, and F1-score metrics.
4) Select Neural Network as the best-performing model.

Model Persistence:

1) Save final model:
$joblib.dump(neural\_network\_model, 'final\_model.joblib')$
2) Save scaler:
$joblib.dump(scaler, 'scaler.joblib')$

The approach guarantees the preservation of the top-performing Neural Network model and its corresponding scaler in the last stages of Model Persistence. The banking industry may manage and lower client turnover using these preserved artifacts, which can be used for in-the-moment forecasts or additional research.

VI. Model selection

When it came to forecasting loan customer turnover, the Artificial Neural Network (ANN) outperformed other models such as logistic regression, Gaussian Naive Bayes, decision trees, Support Vector Machines (SVM), and others, based on the assessment criteria of accuracy, recall, precision, and F1-score [8]. It is the best optiondue to its exceptional performance in all parameters, outperforming other models. The ANN was chosen above other models because of its capacity to manage complicated patterns and categorize loan approval, which fully matches the project's goals and stakeholder needs."

VII. Results

Figure 3 shows the Accuracy, precision, Recall and F1-Score for the different Machine Learning models. The performance metrics of many machine learning models used to forecast bank customer turnover are compiled in the displayed Table I. A particular model, such as Neural Network, KNN, GaussianNB, BernoulliNB, SVC, Decision Tree, and Logistic Regression, correlates to each row [9]. The performance measurements, expressed as % values, comprise precision, recall, F1 score, and total accuracy. Higher numbers indicate greater performance. These metrics give insights into how well the algorithms forecast client attrition. The findings can help choosethe best model for predicting bank customer attrition by weighing the trade-offs between recall, precision, and overall accuracy concerning the particular needs and goals of the organization.

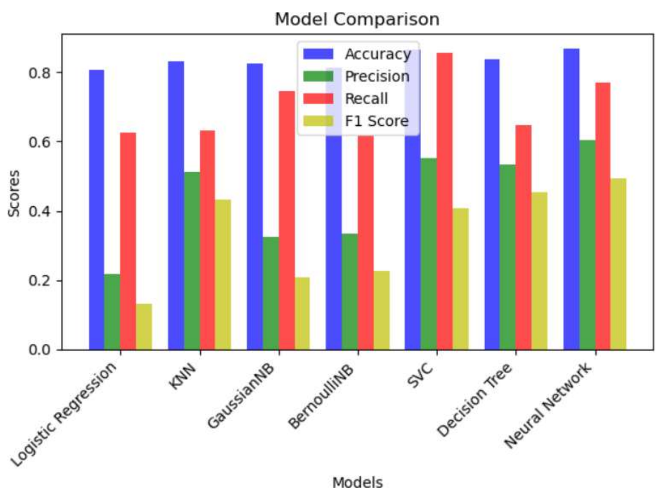

Figure 3: Model Comparison

TABLE I: PERFORMANCE METRICS OF DIFFERENT MODELS

Model	Precision (%)	Recall (%)	F1 (%)	Accuracy (%)
Logistic Regression	21.75	62.62	13.16	**80.72**
KNN	51.34	63.22	43.22	**83.32**
GaussianNB	32.57	74.65	20.83	**82.44**
BernoulliNB	33.29	61.70	22.79	**81.40**
SVC	55.13	85.54	40.67	**86.52**
Decision Tree	53.06	64.25	45.19	**83.72**
Neural Network	60.67	77.85	49.71	**86.88**

VIII. Model Evaluation

A. Accuracy

Accuracy measures the ratio of correctly predicted churn and non-churn observations to the total observations.

$$Accuracy = (TP + TN)/(TP + TN + FP + FN)$$

Churn Predictions: It assesses how accurately the model predicts both churned and non-churned customers.

B. Precision

Precision measures the ratio of correctly predicted churned customers to the total predicted churned customers.

$$Precision = TP/(TP + FP)$$

Churn Predictions: It assesses the model's accuracy in identifying customers as churned out of all the customersit predicted as churned.

C. Recall (Sensitivity)

Recall measures the ratio of correctly predicted churned customers to all actual churned customers.

$$Recall = TP/(TP + FN)$$

Churn Prediction: It evaluates how well the model captures all the actual churned customers among all customerswho churned.

D. F1 Score: It is the harmonic mean of precision and recall, balancing the two metrics

$$F1Score = 2 * (Precision * Recall)/(Precision + Recall)$$

Churn Prediction: It balances the precision and recall of churned customers, offering an overall assessment ofthe model's ability to identify churn while minimizing false predictions correctly.

IX. Conclusion and Future Work

Our Bank Customer Churn Predictor marks a significant step forward in anticipating customer behaviour within the banking industry. We've created a practical tool that empowers users to make informed decisions by combiningvarious smart techniques and user-friendly features.

Including machine learning algorithms like Logistic Regression and Decision Trees, along with a deep learning component, elevates our predictive capabilities [10]. This ensures we capture the apparent patterns and delve deeper into subtle nuances in customer behaviour.

Our commitment to quality data is reflected in the meticulous preprocessing of a comprehensive dataset from Kaggle. The Flask web app interface adds simplicity, allowing users to input data and receive valuable predictionsfor retention strategies easily.

As we share our work, we extend an open invitation for collaboration. Contributions from the broader communityare vital for refining and advancing the effectiveness of our churn prediction system.

In summary, the Bank Customer Churn Predictor is not just a tool; it's a testament to the possibilities when smart technology meets real-world challenges. We anticipate its impact will extend beyond our project, inspiring further advancements and insights in predictive analytics.

We intend to optimize model performance by investigating more sophisticated deep learning architectures and fine-tuning machine learning algorithms via ensemble approaches. To improve churn prediction insights, we want to go further into feature engineering and prioritize feature selection for model interpretability while looking for new behavioural cues and outside data sources. Streaming data analytics will allow for quick interventions for clients in danger, and real-time prediction skills will be a major area of research in this regard. Meanwhile, we will collect user input and apply it to improve the Flask web app's user interface, making it more clear and educational for users. Collaboration with the data science community is essential for model validation and benchmarking against varied datasets to ensure ethical concerns and fairness in predictions across demographic groups. The model's effective scalability, cloud deployment, and documentation are our ultimate goals to provideinsightful methods and new insights to the broader predictive analytics field.

References

[1] I. Huseyinov and O. Okocha, "A machine learning approach to the prediction of bank customer churn problem," in 2022 3rd InternationalInformatics and Software Engineering Conference (IISEC), 2022, pp. 1–5.

[2] P. Gopal and N. B. MohdNawi, "A survey on customer churn prediction using machine learning and data mining techniques in e-commerce," in 2021 IEEE Asia-Pacific Conference on Computer Science and Data Engineering (CSDE), 2021, pp. 1–8.

[3] M. Galal, S. Rady, and M. Aref, "Enhancing customer churn prediction in digital banking using ensemble modeling," in 2022 4thNovel Intelligent and Leading Emerging Sciences Conference (NILES), 2022, pp. 21–25.

[4] A. Raj and D. Vetrithangam, "Machine learning and deep learning technique used in customer churn prediction: - a review," in 2023 International Conference on Computational Intelligence and Sustainable Engineering Solutions

(CISES), 2023, pp. 139–144.
[5] T. Vafeiadis, K. I. Diamantaras, G. Sarigiannidis, and K. C. Chatzisavvas, "A comparison of machine learning techniques for customerchurn prediction," Simulation Modelling Practice and Theory, vol. 55, pp. 1–9, 2015.
[6] S. J. Haddadi, M. O. Mohammadi, M. Bahrami, E. Khoeini, M. Beygi, and M. H. Khoshkar, "Customer churn prediction in the iranian banking sector," in 2022 International Conference on Applied Artificial Intelligence (ICAPAI), 2022, pp. 1–6.
[7] S. Osowski and L. Sierenski, "Prediction of customer status in corporate banking using neural networks," in 2020 International JointConference on Neural Networks (IJCNN), 2020, pp. 1–6.
[8] Y. Wang, S. Zheng, G. Liu, and J. Li, "Research on bank customer churn model based on attention network," in 2023 IEEE 2nd International Conference on Electrical Engineering, Big Data and Algorithms (EEBDA), 2023, pp. 346–350.
[9] Y. Xie, X. Li, E. Ngai, and W. Ying, "Customer churn prediction using improved balanced random forests," Expert Systems withApplications, vol. 36, no. 3, pp. 5445–5449, 2009.
[10] X. Li and Z. Chen, "Customer churn prediction in bank based on different machine learning models," in 2022 2nd International Signal Processing, Communications and Engineering Management Conference (ISPCEM), 2022, pp. 274–279.