Journal of Information Systems Engineering and Management
2025, 10(25s)
e-ISSN: 2468-4376
https://www.jisem-journal.com/

Research Article

Empowering Smart Cities with AI: Predictive Models for Customer Retention in Banking

Dr. K. Narasimhulu[1], J. Sivakumar[2], Y. Venkatrao[3], B. Sreevani[4], G. Charan Teja[5]

[1] Associate Professor, Department of Computer Science and Engineering, Rajeev Gandhi Memorial College of Engineering and Technology, Nandyala, Andhra Pradesh, India.

[2,3,4,5] Department of Computer Science and Engineering, Rajeev Gandhi Memorial College of Engineering and Technology, Nandyala, Andhra Pradesh, India.

[1]narsimhulu.kolla@gmail.com orcid:0000-0003-0756-379X , [2]jsivakumar421@gmail.com, [3]yalamanchivenkatraochoudary@gmail.com, [4]boleramayya@gmail.com, [5]charanvd123@gmail.com

ARTICLE INFO

Received: 05 Jan 2025

Revised: 24 Feb 2025

Accepted: 04 Mar 2025

ABSTRACT

Introduction: Smart cities thrive on innovative technologies, and artificial intelligence (AI) plays a pivotal role in enhancing customer-centric services. In the context of the banking sector, customer retention is vital for maintaining competitiveness, especially in the highly dynamic urban environments of smart cities.
Objectives: The main objective of this study is to investigate the application of supervised machine learning algorithms to predict customer churn, a critical factor in developing efficient retention strategies.
Methods: This work uses a dataset of 10,000 customer records, models such as Logistic Regression, K-Nearest Neighbors, Decision Tree, Random Forest, Gradient Boosting, LightGBM, XGBoost, and Naive Bayes were evaluated. Preprocessing and analysis were conducted with key metrics including accuracy, precision, recall, F1-score, cross-validation, and AUC-ROC.
Results: The results reveal that ensemble models, particularly Gradient Boosting, XGBoost, Random Forest, and LightGBM, deliver superior performance on unbalanced data, achieving accuracies of 85.65%, 85.65%, 85.25%, and 85.35%, respectively.
Conclusion: On balanced data, LightGBM outperformed others with an accuracy of 84.21%. These findings highlight the potential of AI-driven predictive models to empower banking institutions in smart cities, fostering better customer retention and contributing to sustainable urban development.
Keywords: Customer churn, Machine Learning, Ensemble models, Banking, Evaluation metrics, Boosting.

INTRODUCTION

Customer churn, which involves customers decision to leave a company, is one of the major problems that competitive industries like banking are facing today. In the case of banking, churn does not only affect profitability but also has a long-term impact on customer loyalty. Retaining an existing customer is relatively cheaper than trying to get a new customer. If the banks have a good customer base, then it not only helps in retaining present customers but also attracts new customers. While extensive research has been conducted on customer churn in different sectors, studies that focus on the banking sector are very less. However, banking data may involve complex customer behavior patterns which require advanced predictive modeling techniques. According to the study conducted by Amy Gallo, states that getting a new customer is from 5 to 25 times costlier than maintaining the existing ones. According to this, we can understand that maintaining the present customers is how much beneficial [1].

Ram Prabhakar states that reducing the churn by 5% double the profits. Banks generally offer similar products and services for similar prices, what differentiates is the quality of services provided by them. Great customer service means providing the right product or service to the right customer at the right time. About 40% of the customers switched banks because of poor service provided by banks. To do this, understanding the customer behaviors, satisfaction and preferences is important. Customer satisfaction is the main reason why customers either stay or leave [2].

There are various reasons for customer churn in banks. Some of them include poor customer service, lack of personalization, poor online banking services, limited access to branches/ATMs, and better offers from competitors. In today's world, many banks are going towards implementing predictive analytics which helps in identifying high-risk customers and implementing effective retention strategies. To do this, they are using advanced Machine Learning (ML) models by training with historical data.

This paper discusses the performance of several supervised ML algorithms, such as Logistic Regression, Decision Tree, Random Forest, Gradient Boosting, LightGBM, XGBoost, K-Nearest Neighbors, and Naive Bayes, in predicting customer churn with a dataset of 10,000 customer records. The performance of these models is judged based on accuracy, precision, recall, F1-score, and AUC-ROC, providing insights that are valuable to the customer retention strategies specific to the banking sector [3].

METHODOLOGY

This section encompasses the methodology of this study. It has several important steps such as getting the data, preprocessing data, feature selection, and preparing for model training and evaluation. The methodology also describes how classes are unbalanced and how it helps in improving the model performance using Synthetic Minority Oversampling Technique (SMOTE). Moreover, k-fold cross-validation is present in the final to test the overfitting of models. All these steps have been done in Python, Jupyter Notebook, and using pandas for manipulating data; matplotlib, plotly.express, and seaborn for visualizing data; scikit-learn to apply machine learning techniques to the data. The workflow used in this study is shown in Figure 1.

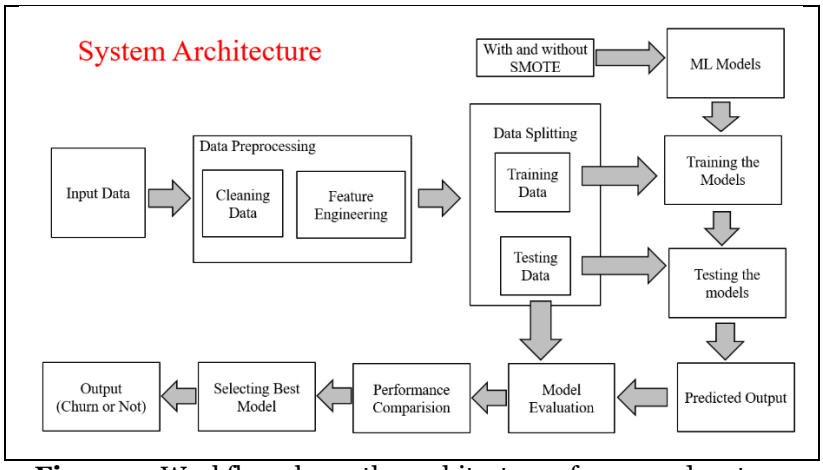

Figure 1: Workflow shows the architecture of proposed system.

Each step involved in this study is explained one by one.

1. Data Collection and Dataset Description

We tried to collect the dataset from the banks but due to privacy and confidentiality issues, we are not able to do that. So, we have taken the dataset from Kaggle [12]. The dataset includes customer demographics (e.g., age, gender), account information (e.g., balance, tenure, number of products), and behavioral data (e.g., credit score, exit status). It contains 10,000 rows and 14 columns. The 14 columns in the dataset are RowNumber, CustomerId, Surname, CreditScore, Geography, Gender, Age, Tenure, Balance, NumOfProducts, HasCrCard, IsActiveMember, EstimatedSalary, and Exited.

2. Data Preprocessing

Preprocessing of data is a necessary step to ensure that the dataset is clean and ready for analysis. These steps include:

- **Feature Encoding:** Gender, card type, and geography were all categorical variables that were encoded by Label Encoding.
- **Feature Scaling:** Continuous features in this experiment were age, credit score, estimated salary, and balance. These features are scaled to equate data value ranges by applying the StandardScaler method.

3. Feature Selection and Data splitting

Feature selection is an important step in machine learning, especially in predicting customer churn. It identifies the most relevant features that would contribute to improving the model performance. Feature selection is done using Recursive Feature Elimination (RFE) which works through the recursive elimination of a feature and builds a model with the remaining features. RFE sorts the features according to their importance to the model through RFE ranking which helps to determine the features that best predict churn. The splitting of data sets into different segments is one of the very basic methodological machine learning approaches: training and testing data. At this point the model is subjected to real-life testing on data that it has not seen, thereby replicating real life and avoiding overfitting. Usage ratio 80-20 between training and testing data.

4. Selection of Model

Below are some of the supervised machine learning models that helped us to predict whether a customer would either churn or stay in the bank.

- **Logistic Regression (LG):** Basic model for binary classification. It's a pretty simple, easy, and interpretable model that does an excellent job on linearly separable data. Prediction is based on the logistic function (sigmoid function)
- **Decision Tree (DT):** Decision Tree gives branches in data with feature thresholds. It is easy to interpret and works with smaller datasets but risks overfitting if unpruned. To perform well, it has to minimize Gini impurity/Entropy by choosing the best split.
- **Random Forest (RF):** Random Forest is an ensemble learning approach that builds several decision trees and uses voting concepts to predict output. It is robust to overfitting and works well with datasets having many features.
- **K-Nearest Neighbors (KNN):** KNN is the technique that classifies new data points according to the majority class represented by its K-nearest neighbors.
- **Gradient Boosting Classifier (GBC):** The process of creating models in a sequence is called Gradient Boosting where each next model's efforts go into correcting the error made by that last model.
- **XGBoost Classifier (XGBC):** A formidably excellent gradient-boosting framework, that can very efficiently handle larger-to-big datasets or handle imbalanced datasets. Those trees are constructed in the form of sequences to contravene the error produced by the preceding trees corrected. Its loss function resembles that of GBC but has an additional relative term for regularization:
- **LightGBM (LGBM):** Gradient boosting that is optimized for efficiency and scale. Histogram-based learning is comparatively much faster than other methods on this topic without the slightest result loss. It similarly has a penalty complexity regularization term, like in XGBoost, and applies leaf-wise growth instead of level-wise growth for trees.
- **Gaussian Naive Bayes (GNB):** Naive Bayes are probabilistic models that assume the feature independent. It is simple and fast, but good to apply under independence. It calculates the class probability using Bayes Theorem.

5. Model Training and Evaluation

SMOTE (Synthetic Minority Oversampling Technique): It is a technique for data augmentation used in machine learning datasets to tackle class imbalances. In this method, some synthetic samples are generated for the minority class in such a way that the classes become balanced, and thus the performance of the model is enhanced on this minority class.

(i) Without SMOTE: This model is just trained on the raw data and in that scenario, the class imbalance present is mainly that most of the customers are not churned in comparison to churned customers. Models have been evaluated based on metrics that include accuracy, precision, recall, F1-score, cross-validation, and AUC-ROC. Although accuracy typically is high, however, the recall for the minority class (churn) is generally low reflecting the not-good performance of the model in detecting the churned customers, mostly because the model is biased towards the majority class.

(ii) With SMOTE: Before training the model, SMOTE is applied to balance the class distribution by generating synthetic samples of the minority class data. Hence, a model will be adequately trained to classify customers that churned and customers that did not churn so fairly. Balanced data improves the metrics like recall and F1-score for the minority class. AUC-ROC also improves, which means better discrimination of churned against non-churned customers. Although SMOTE reduces a little precision because synthetic samples add noise.

RESULTS

Data is divided into two parts, one part is for training (80% of the data), and the other part is testing (remaining 20% of the data). We have applied various supervised machine learning algorithms and in that also mainly focused on ensemble models. The models are trained and tested with unbalanced and balanced data. The data is balanced using SMOTE.

(i) Evaluation results of models on unbalanced data (without SMOTE)

With the unbalanced data, gradient-boosting, random forest, XGBoost, and LightGBM performed well with accuracy scores of 85.65%, 85.25%, 85.65%, and 85.35% respectively. We have faced difficulty with the imbalanced data. With this imbalanced data, Logistic Regression and GNB models performed worse when compared to other models that were applied.

We have evaluated the models using various evaluation metrics like accuracy score, precision, recall and F1-score. The Logistic Regression model has the least precision value of 62.5%. The precision value is low due to an increase

in false positives. Logistic regression and GNB models have the lowest recall values with 17.2% and 24.57% respectively. The recall value is low because there is an increase in false negatives.

To evaluate the robustness and generalizability of the models, we performed 10-fold cross-validation on the dataset. All the results with imbalanced data are summarized in Table 1. XGBoost and LightGBM have an average accuracy of 86%. Decision Tree, Random Forest, and Gradient Boosting have an average accuracy score of 85%. LightBGM and XGBoost have low standard deviation values..

Models	Accuracy (%)	Precision (%)	Recall (%)	F1 Score (%)	AUC value(%)	CV mean(%)	Standard Deviation(%)
Logistic Regression	81.05	62.50	17.20	26.97	75.49	80.82	0.66
Decision Tree	84.50	69.48	42.51	52.74	82.72	85.19	0.82
Random Forest	85.25	80.11	36.61	50.25	85.10	85.69	0.89
KNN	84.15	68.15	41.52	51.60	79.22	83.60	0.96
Gradient Boosting	85.65	76.09	43.00	54.95	85.56	85.92	1.02
Naive Bayes	83.55	81.97	24.57	37.81	80.95	82.85	0.77
XGBoost	85.65	73.62	45.95	56.58	86.01	86.25	0.83
LightGBM	85.35	74.15	43.00	54.43	86.25	86.15	0.91

Table 1: This table describes the evaluation results of various machine learning models trained on unbalanced data (without the use of SMOTE)

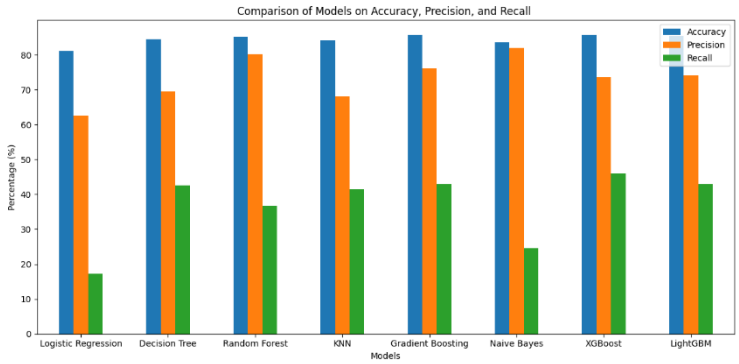

Figure 2: Bar Graph showing Accuracy, Precision, and Recall Values for all models (with imbalanced data)

The data shown in Table 1 is graphically represented in Figure 2 for accuracy, precision, and recall values for all models. By observing Figure 2, it is clear that accuracy values for all models are good, but recall values for all models are very less. All recall values are below 50%. This is because of the imbalanced data.

(ii) **Evaluation of models on balanced data (with SMOTE)**

After balancing the data using SMOTE, the precision and recall values are increased. It means the number of false negatives and false positives is decreased by increasing the efficiency of the models. For example, consider the linear regression model, the value of recall increased from 17.2% to 72.56% and for the KNN model, it increased from 41.52% to 91.74% and for GNB model, it increased from 24.57% to 76.4%. For the decision tree model, the precision value increased from 69.48% to 78.31%. The precision value for all other models also increased. The accuracy scores for XGBoost and LightGBM are 83.65% and 84.21% respectively. Overall SMOTE increased the performances of the models which initially did not perform well, it increased the recall values.

After performing k-fold cross-validation, we observed that XGBoost, LightGBM, and KNN have an average accuracy of 83.68%, 84.09%, and 85.19% respectively. All the results for balanced data are summarized in Table 2.

Models	Accuracy (%)	Precision (%)	Recall (%)	F1 Score (%)	AUC value (%)	CV mean (%)	Standard Deviation (%)
Logistic Regression	71.31	69.40	72.56	70.95	78.24	70.44	1.38
Decision Tree	77.62	78.31	74.19	76.19	86.28	76.53	1.56
Random Forest	79.72	78.48	79.91	79.19	88.83	79.41	1.58

KNN	83.74	78.30	91.74	84.49	92.02	85.19	2.03
Gradient Boosting	80.19	78.15	81.86	79.96	89.64	81.16	1.81
Naive Bayes	75.14	73.25	76.40	74.79	83.06	74.19	1.28
XGBoost	83.65	82.29	84.27	83.26	91.91	83.68	2.72
LightGBM	84.21	83.76	83.49	83.62	92.46	84.09	4.22

Table 2: This table describes the evaluation results of various machine learning models trained on balanced data (with the use of SMOTE).

The data shown in Table 2 is graphically represented in Figure 3 for accuracy, precision, and recall values for all the models. By observing Figure 3, it is clear that accuracy values for all models are slightly decreased due to noise added by the SMOTE, however, recall values for all models increased significantly. All recall values are increased below 50% to above 70%. This is because of the balanced data.

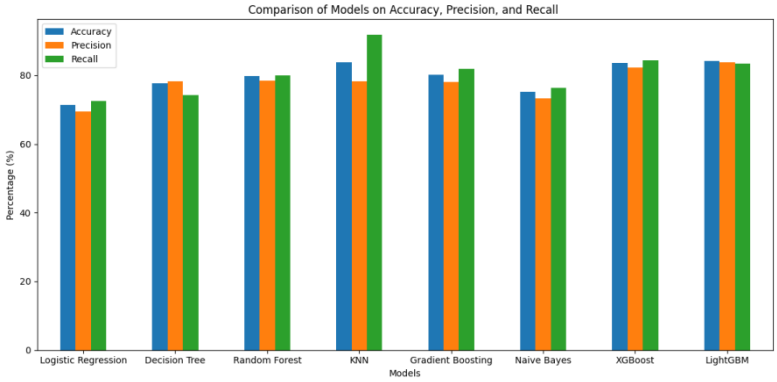

Figure 3 Bar Graph showing Accuracy, Precision, and Recall Values for all models (for Balanced data)

CONCLUSION

This study demonstrates the potential of AI-driven predictive models for customer churn in the banking sector, contributing to the vision of smart cities where customer-centric services are key. The machine learning algorithms employed—Logistic Regression, Random Forest, Decision Tree, K-Nearest Neighbors, Naive Bayes, Gradient Boosting, XGBoost, and LightGBM—proved effective in identifying at-risk customers. Treating class imbalance with SMOTE significantly enhanced recall and F1-scores, particularly for churned customers. Among the models, LightGBM, XGBoost, and Gradient Boosting emerged as the top performers, with LightGBM showing superior accuracy on balanced data. The insights gained can support proactive customer retention strategies, reducing churn and fostering stronger relationships in the banking sector within smart cities.

REFERENCES

[1] “The value of keeping right customers” by Amy Gallo (2014), editor at Harvard Business Review. Available at https://hbr.org/2014/10/the-value-of-keeping-the-right-customers

[2] “Customer Retention in Banking” by Ram Prabhakar (2024), Head of Solutions and Content. Available at https://www.xerago.com/xtelligence/customer-retention-in-banking

[3] “An Improved Random Forest Algorithm (ERFA) Utilizing an Unbalanced and Balanced Dataset to Predict Customer Churn in the Banking Sector” by Sultan Yahya Al-Sultan and Ibrahim Ahmed Al-Baltah (2024). DOI 10.1109/ACCESS.2024.3395542

[4] “Machine Learning Based Customer Churn Prediction in Banking” by Manas Rahman and V Kumar (2020). DOI: 10.1109/ICECA49313.2020.9297529. ISBN: 978-1-7281-6387-1.

[5] H Guliyev and F Yerdelen Tatoglu, “Customer churn analysis in banking sector: Evidence from explainable machine learning models”,2021. Journal of Applied Microeconometrics (JAME). 1(2), 85-99, DOI: 10.53753/jame.1.2.03

[6] Seyed Hossein Iranmanesh, Mahdi Hamid, Mahdi Bastan, Hamed Shakouri G., Mohammad Mahdi Nasiri, “Customer Churn Prediction Using Artificial Neural Network: An Analytical CRM Application”,2019. Proceedings of the International Conference on Industrial Engineering and Operations Management Pilsen, Czech Republic, July 23-26, 2019.

[7] Pahul Preet Singh, Fahim Islam Anik, Rahul Senapati, Arnav Sinha, Nazmus Sakib, Eklas Hossain research on “Investigating customer churn in banking: a machine learning approach and visualization app for data science and management” (2024). https://doi.org/10.1016/j.dsm.2023.09.002

[8] “Development of a Customer Churn Model for Banking Industry Based on Hard and Soft Data Fusion” by Masoud Alizadeh, Danial Sadrian Zadeh, Behzad Moshiri, and Allahyar Montazeri (2023). DOI: 10.1109/ACCESS.2023.3257352
[9] Saurabh Mhaske, “Use of Machine Learning for Customer Churn Analysis in Banking” (2022). DOI: https://www.doi.org/10.56726/IRJMETS29877
[10] “Churning of Bank Customers Using Supervised Learning” by Hemlata Dalmia, Ch V S S Nikil, and Sandeep Kumar (2020). DOI: 10.1007/978-981-15-3172-9_64. Available at https://www.researchgate.net/publication/340855263
[11] Micheal Arowolo, Bilkisu Jimada Ojuolape, Saheed Yakub, Abdulsalam Sulaiman Olaniyi, “Customer Churn Prediction in Banking Industry Using K-Means and Support Vector Machine Algorithms”, 2020. International Journal of Multidisciplinary Sciences and Advanced Technology, Vol 1 No 1 (2020) 48–54. Available at https://www.researchgate.net/publication/346910130
[12] Bank customer churn modelling dataset from Kaggle by Amir Motefaker (https://www.kaggle.com/datasets/amirmotefaker/churn-modeling-bank)
[13] Jason Brownlee, “SMOTE for Imbalanced Classification with Python”, 2021. Available at https://machinelearningmastery.com/smote-oversampling-for-imbalanced-classification/
[14] Scikit-learn official documentation [online], https://scikit-learn.org/stable/
[15] Geeks for Geeks [online], “Evaluation Metrics in Machine Learning”. Available at https://www.geeksforgeeks.org/metrics-for-machine-learning-model/
[16] Analytics Vidhya [online], https://www.analyticsvidhya.com