IARJSET ISSN (O) 2393-8021, ISSN (P) 2394-1588

International Advanced Research Journal in Science, Engineering and Technology

Impact Factor 8.311 ⁜ Peer-reviewed & Refereed journal ⁜ Vol. 13, Issue 3, March 2026

DOI: 10.17148/IARJSET.2026.13340

CUSTOMER CHRUN PREDICTION USING MACHINE LERANING

Manoj A[1], Dr. K. Thenmozhi[2]

Student, Department of Information Technology, Dr NGP Arts and Science College, Coimbatore[1]

Professor, Department of Information Technology, Dr NGP Arts and Science College, Coimbatore[2]

Abstract: Customer churn prediction is an essential business tool that allows organizations to predict customers who will likely stop using their services. This prediction is essential in reducing revenue loss and improving customer retention. This research work presents a machine learning-based customer churn prediction system that combines the latest resampling methods and Natural Language Processing (NLP) techniques. To handle the issue of class imbalance, Random Oversampling, Random Undersampling, SMOTE, and ADASYN are applied. In addition, BERT (Bidirectional Encoder Representations from Transformers) is applied for the extraction of contextual features from customer feedback and text data. The preprocessed features are then used as input for classification models like Random Forest, XGBoost, and Logistic Regression. The experimental results based on accuracy, precision, recall, F1-score, and ROC-AUC show improved predictive accuracy. The combination of resampling techniques and BERT-based feature extraction is highly effective in improving the accuracy of customer churn prediction.

Keywords: Customer Churn, Machine Learning, Random Forest, XGBoost, Logistic Regression, BERT, SMOTE, ADASYN, Class Imbalance, NLP.

I. INTRODUCTION

Customer churn is defined as the situation where customers discontinue the use of a company's products or services. In competitive sectors like telecom, banking, and online platforms, customer retention is more economical than customer acquisition. Hence, customer churn prediction is critical for effective customer retention.

Customer churn prediction is affected by issues like class imbalance and the presence of multiple types of data, such as numerical, categorical, and text data. Conventional models tend to concentrate only on structured data sources and overlook the importance of customer feedback.

The proposed study aims to develop a hybrid model that incorporates analysis of structured data, extraction of textual features using BERT, and methods for handling class imbalance.

II. LITERATURE REVIEW

Customer churn prediction has been a popular topic in machine learning and data analytics because of its direct effect on business profits. Many traditional and modern methods have been suggested to improve the accuracy of churn predictions.

Early studies mainly used statistical and basic machine learning models, including Logistic Regression, Decision Trees, and Naïve Bayes, for churn classification. Logistic Regression was commonly chosen due to its simplicity and ease of understanding. However, it often had trouble capturing complex non-linear relationships between customer characteristics and churn behavior.

Decision Tree-based models offered better interpretability and managed non-linear patterns more effectively. Later, ensemble methods like Random Forest showed improved performance by combining several decision trees and reducing overfitting through majority voting. These models provided better predictive accuracy compared to single classifiers.

Another major challenge found in churn prediction research is class imbalance. In many real-world datasets, the number of churned customers is much lower than that of retained customers. Research has indicated that imbalanced datasets can harm model performance, especially recall for the minority class. To tackle this issue, resampling techniques such as Random Oversampling, Random Undersampling, SMOTE (Synthetic Minority Oversampling Technique), and ADASYN

IARJSET ISSN (O) 2393-8021, ISSN (P) 2394-1588

International Advanced Research Journal in Science, Engineering and Technology

Impact Factor 8.311 ⁜ Peer-reviewed & Refereed journal ⁜ Vol. 13, Issue 3, March 2026

DOI: 10.17148/IARJSET.2026.13340

(Adaptive Synthetic Sampling) have been proposed. These methods create synthetic minority samples or adjust class distribution to enhance classifier performance.

While structured numerical data has been used for churn prediction, recent research emphasizes the value of including unstructured textual data, like customer feedback and reviews. Traditional text feature extraction methods, such as Bag of Words (BoW) and TF-IDF, convert text into numbers but do not capture contextual meaning.

With progress in Natural Language Processing (NLP), deep learning models like BERT (Bidirectional Encoder Representations from Transformers) perform better at extracting contextual meanings from text. BERT processes words from both directions, which helps it understand context and capture semantic relationships more effectively than older methods. Several recent studies show that combining contextual text embeddings with structured data improves predictions about customer behavior.

III. PROBLEM STATEMENT

Customer churn is a major issue for organizations in competitive industries like telecommunications, banking, e-commerce, and subscription services. When customers stop using a company's service, it results in significant revenue loss and added costs for gaining new customers. Thus, identifying customers likely to churn is vital for putting in place effective retention strategies.

However, predicting customer churn is complicated due to several challenges. First, churn datasets often have an imbalance, where the number of retained customers is much higher than those who have churned. This imbalance can hinder the performance of standard machine learning models, particularly when it comes to accurately identifying minority (churn) cases.

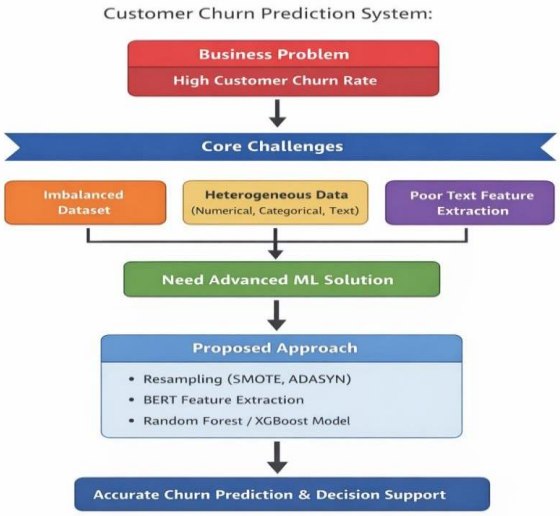
Second, customer data includes various types, such as numerical (usage patterns, tenure, charges), categorical (subscription type, payment method), and unstructured textual data (customer feedback, complaints, reviews). Traditional churn prediction models mainly focus on structured data and often overlook valuable insights found in textual feedback.

Third, standard feature extraction methods for text, like TF-IDF and Bag-of-Words, do not capture contextual meaning and semantic relationships well. This can lead to lower prediction accuracy.

Therefore, there is a need for a better churn prediction system that:

- Effectively manages class imbalance
- Integrates structured and unstructured data
- Uses advanced feature extraction techniques
- Provides reliable and accurate churn predictions

This project aims to develop a machine learning-based churn prediction system that combines resampling techniques (like Random Oversampling, Random Undersampling, SMOTE, and ADASYN), BERT-based feature extraction for textual data, and ensemble learning models such as Random Forest and XGBoost. This approach aims to improve prediction performance and assist in proactive business decision-making.



System Flow Diagram of Customer churn prediction using machine learning

IARJSET ISSN (O) 2393-8021, ISSN (P) 2394-1588

International Advanced Research Journal in Science, Engineering and Technology

Impact Factor 8.311 ⁜ Peer-reviewed & Refereed journal ⁜ Vol. 13, Issue 3, March 2026

DOI: 10.17148/IARJSET.2026.13340

IV. METHODOLOGY

The proposed Customer Churn Prediction system uses a clear process that includes data preprocessing, feature extraction, handling class imbalance, and machine learning classification. This workflow ensures reliable churn prediction. The methodology is divided into the following steps:

1. Data Collection

The dataset includes both structured and unstructured data gathered from customer interactions and business operations. The data consists of: - Demographic information (age, gender, location) - Account details (subscription type, tenure, payment method) - Transaction history and usage patterns - Customer feedback and textual reviews This combination allows for a complete understanding of customer behavior.

2. Data Preprocessing

Raw data often has missing values, inconsistencies, and noise. So, preprocessing is done to improve data quality. The preprocessing steps include: - Handling missing values - Removing duplicate records - Encoding categorical variables - Normalizing numerical features - Cleaning textual data, such as removing special characters and stop words if needed This step makes sure the dataset is ready for machine learning models.

3. Feature Extraction using BERT

Customer feedback is an important sign of churn behavior. To get useful insights from textual data, we use BERT (Bidirectional Encoder Representations from Transformers). Unlike traditional text feature extraction methods like TF-IDF, BERT captures contextual meaning by analyzing words in both directions. The steps include: Tokenizing the text, Generating contextual embeddings, Converting text into numerical feature vectors. We then combine these embeddings with structured numerical features to create a complete feature set.

4. Class Imbalance Handling

Churn datasets usually have an imbalance, with fewer customers who have churned than those who have been retained. To tackle this issue, we use the following resampling techniques: Random Oversampling Random Undersampling SMOTE (Synthetic Minority Oversampling Technique) ADASYN (Adaptive Synthetic Sampling) These techniques help balance the dataset and improve the model's ability to classify churn cases accurately.

5. Model Training and Classification

The processed and balanced dataset is used to train multiple machine learning models: Logistic Regression Random Forest XGBoost Random Forest uses ensemble learning and majority voting. XGBoost uses gradient boosting to improve predictive accuracy. Logistic Regression acts as a baseline model.

6. Model Evaluation

Model performance is evaluated using the following metrics: Accuracy Precision Recall F1-score ROC-AUC These metrics make sure that the model achieves high overall accuracy and also identifies churned customers effectively.

7. Decision Support and Recommendation

Based on the prediction results, the system identifies high-risk customers and gives useful insights for keeping customers. This helps businesses run targeted marketing campaigns, create personalized offers, and improve services.

8. Summary

This project presents a Customer Churn Prediction system that combines machine learning techniques, resampling methods, and Natural Language Processing (NLP) to improve predictive performance. The system tackles major challenges in churn prediction, such as class imbalance and different data types like numerical, categorical, and textual feedback. To improve prediction accuracy, techniques like Random Oversampling, Random Undersampling, SMOTE, and ADASYN are used to balance the dataset. Additionally, BERT-based feature extraction captures contextual meaning from customer feedback. This allows the model to use valuable insights from unstructured textual data. The extracted features are merged with structured data and processed by machine learning classifiers like Random Forest, XGBoost, and Logistic Regression. Experimental evaluation shows that combining BERT embeddings with ensemble learning models significantly boosts churn detection performance. The system not only achieves higher accuracy but also enhances recall and F1-score for identifying at-risk customers. In summary, the proposed solution offers a strong and flexible approach to customer churn prediction. It aids proactive business decision-making through effective decision support and recommendation tools.

IARJSET ISSN (O) 2393-8021, ISSN (P) 2394-1588

International Advanced Research Journal in Science, Engineering and Technology

Impact Factor 8.311 ⁜ Peer-reviewed & Refereed journal ⁜ Vol. 13, Issue 3, March 2026

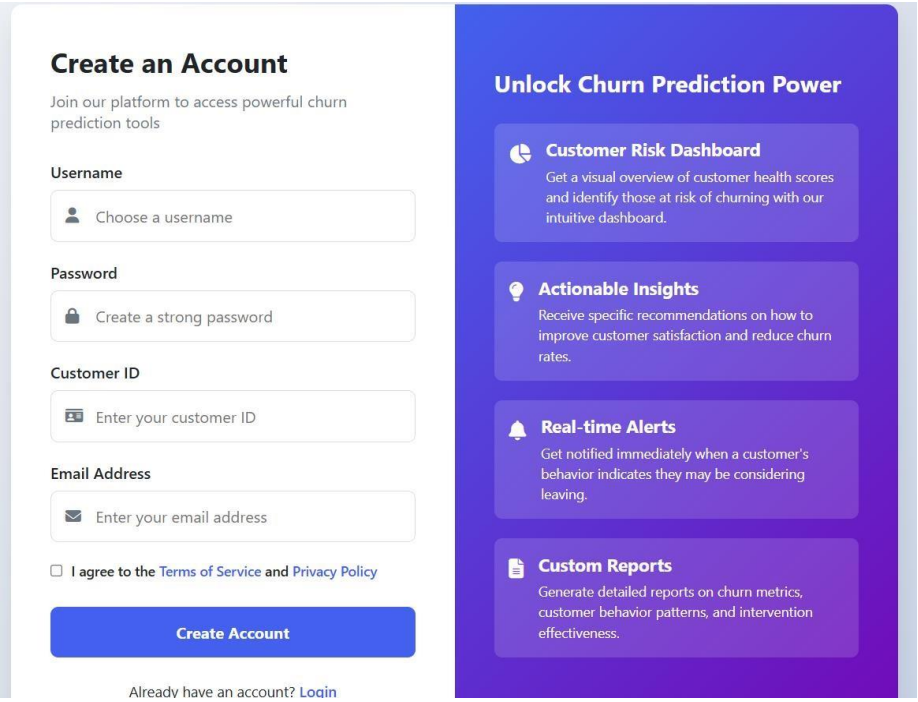
DOI: 10.17148/IARJSET.2026.13340

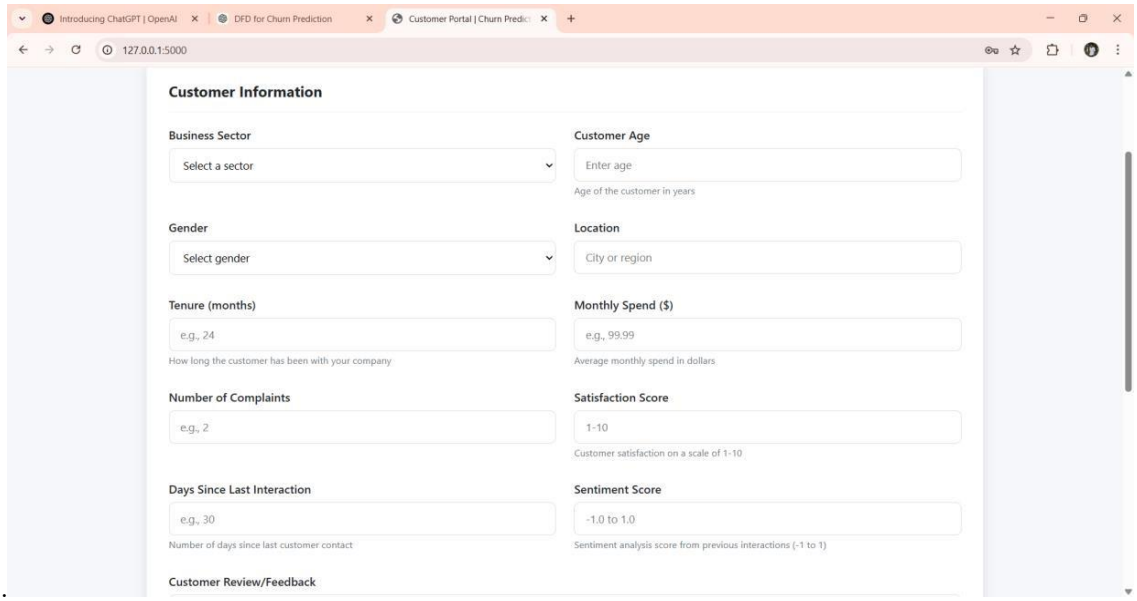
V. RESULTS AND DISCUSSION

The results of the proposed Customer Churn Prediction system show that combining BERT-based text feature extraction with resampling techniques greatly improves model performance. After using class balancing methods like SMOTE and ADASYN, the models achieved better recall and F1-score for the smaller churn class.

This reduced the misclassification of at-risk customers. Among the classifiers we tested, ensemble models like Random Forest and XGBoost had higher accuracy and better overall performance than Logistic Regression.

Adding contextual text embeddings from customer feedback further improved the reliability of predictions. In summary, the combination of resampling, BERT feature extraction, and ensemble learning offers a strong and effective solution for reliable churn prediction and supporting business



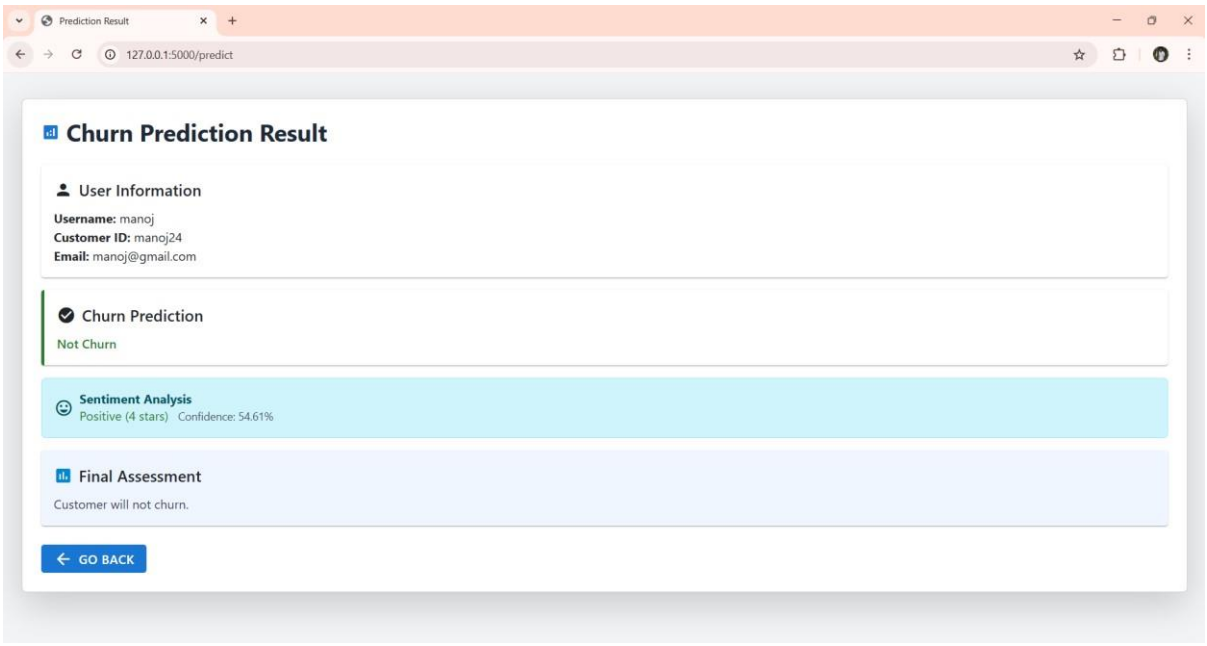


IARJSET ISSN (O) 2393-8021, ISSN (P) 2394-1588

International Advanced Research Journal in Science, Engineering and Technology

Impact Factor 8.311 ⁜ Peer-reviewed & Refereed journal ⁜ Vol. 13, Issue 3, March 2026

DOI: 10.17148/IARJSET.2026.13340



VI. CONCLUSION AND FUTURE WORK

This study presented a comprehensive Customer Churn Prediction system that integrates resampling techniques, BERT-based textual feature extraction, and machine learning classifiers to improve prediction performance. By addressing class imbalance using methods such as SMOTE and ADASYN, the model achieved better detection of churned customers, particularly improving recall and F1-score for the minority class. The use of BERT enabled the system to extract meaningful contextual insights from customer feedback, enhancing the overall predictive capability. Ensemble models such as Random Forest and XGBoost demonstrated superior performance compared to traditional approaches, making the proposed system reliable for real-world business applications.

For future work, the system can be extended to real-time churn prediction using streaming data and deployed on cloud-based platforms for scalability. Further improvements may include integrating deep learning ensemble models, implementing explainable AI techniques for better interpretability, and developing an interactive dashboard for business users. Additionally, incorporating customer behavioral analytics and sentiment trend analysis could further enhance retention strategies and decision-making processes.

Future Scope

The proposed Customer Churn Prediction system can be improved in several ways to increase scalability, accuracy, and practical use. In the future, the system can operate as a real-time prediction model by using streaming customer data to provide instant churn alerts. Deploying it on cloud platforms can increase scalability and support large-scale enterprise applications. We can explore deep learning models like hybrid neural networks or transformer-based approaches to enhance prediction performance. Adding Explainable AI (XAI) techniques will make the model easier to understand. This will help businesses grasp the key factors that influence churn decisions. Additionally, creating a dashboard-based decision support system can help management visualize churn trends and assess customer risk levels. Future research could also focus on adding behavioral analytics, sentiment trend analysis, and personalized recommendation systems to improve customer retention strategies. These enhancements will make the churn prediction system smarter, more flexible, and more suitable for real-world business situations.

REFERENCES

[1]. I. Goodfellow, Y. Bengio, and A. Courville, *Deep Learning*. MIT Press, 2016.
[2]. T. Chen and C. Guestrin, "XGBoost: A Scalable Tree Boosting System," in *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 2016, pp. 785–794.
[3]. L. Breiman, "Random Forests," *Machine Learning*, vol. 45, no. 1, pp. 5–32, 2001.
[4]. N. V. Chawla, K. W. Bowyer, L. O. Hall, and W. P. Kegelmeyer, "SMOTE: Synthetic Minority Over-sampling Technique," *Journal of Artificial Intelligence Research*, vol. 16, pp. 321–357, 2002. [5] H. He, Y. Bai, E. A. Garcia,

DOI: 10.17148/IARJSET.2026.13340

and S. Li, "ADASYN: Adaptive Synthetic Sampling Approach for Imbalanced Learning," in *IEEE International Joint Conference on Neural Networks (IJCNN)*, 2008, pp. 1322–1328.

[5]. J. Devlin, M. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding," in *Proceedings of NAACL-HLT*, 2019, pp. 4171–4186.

[6]. G. E. Batista, R. C. Prati, and M. C. Monard, "A Study of the Behavior of Several Methods for Balancing Machine Learning Training Data," *ACM SIGKDD Explorations Newsletter*, vol. 6, no. 1, pp. 20–29, 2004.

[7]. T. Hastie, R. Tibshirani, and J. Friedman, *The Elements of Statistical Learning*, 2nd ed. Springer, 2009.

[8]. S. Verbeke, D. Martens, and B. Baesens, "Social Network Analysis for Customer Churn Prediction," *Applied Soft Computing*, vol. 14, pp. 431–446, 2014.

[9]. W. Verbeke, D. Martens, C. Mues, and B. Baesens, "Building Comprehensible Customer Churn Prediction Models with Advanced Rule Induction Techniques," *Expert Systems with Applications*, vol. 38, no. 3, pp. 2354–2364, 2011.