International Journal of Scientific Research in Science and Technology
Available online at : www.ijsrst.com
Print ISSN: 2395-6011 | Online ISSN: 2395-602X
doi : https://doi.org/10.32628/IJSRST25122222

Data-Driven Churn Prediction in Telecom: A Comparative Study of Machine Learning Models

Dr. Namrata Gupta, Mr. Natvar Patel, Shri Parvin Ami
Smt. B. K. Mehta IT Centre BCA College, Palanpur, Gujarat, India

ARTICLEINFO

Article History:
Accepted: 05 Feb 2024
Published: 21 Feb 2024

Publication Issue :
Volume 11, Issue 1
January-February-2024
Page Number :
706-712

ABSTRACT

Customer churn prediction is a critical challenge in the telecommunications sector, as companies strive to retain their customer base amidst increasing competition. Effective churn prediction models can help telecom operators identify customers likely to leave and implement targeted retention strategies. This study explores various machine learning (ML) models, including Logistic Regression, Decision Trees, Random Forest, Support Vector Machines (SVM), and Neural Networks, for predicting telecom customer churn. The dataset used in this study is pre-processed through data cleaning, feature selection, and balancing techniques to enhance model efficiency. Performance metrics such as accuracy, precision, recall, and F1-score are used to evaluate the models. The results indicate that ensemble learning models, particularly Random Forest and Neural Networks, achieve the highest prediction accuracy. The study also highlights the importance of feature selection in improving model performance. Findings from this research can assist telecom providers in formulating data-driven strategies to reduce customer churn and enhance customer satisfaction. Future work will explore deep learning approaches and real-time predictive analytics for improving churn prediction further.

Keywords: Churn Prediction, Machine Learning, Telecom Industry, Customer Retention, Predictive Analytics

I. INTRODUCTION

1) Background

The telecommunications industry has experienced rapid growth in recent decades, driven by advances in mobile technology, internet services, and competitive pricing strategies. With millions of customers subscribing to various telecom services, service providers must ensure high-quality customer experiences to maintain long-term relationships. However, customer attrition, commonly referred to as

churn, remains a major challenge. Churn occurs when customers discontinue their subscription or switch to another service provider, leading to significant revenue losses and increased marketing costs for telecom companies.

Churn prediction has gained importance as companies seek to improve customer retention and reduce operational losses. Retaining existing customers is more cost-effective than acquiring new ones, making churn prevention a priority for telecom operators. Understanding the behavioral patterns of customers who are likely to churn can help service providers implement targeted interventions such as personalized offers, improved customer support, and loyalty programs. To achieve this, data-driven solutions, including machine learning techniques, are increasingly being used to analyze historical customer data and predict churn likelihood.

2) Problem Statement

Despite the availability of extensive customer data, accurately predicting churn remains a challenging task. Traditional statistical models have been used for churn prediction, but they often fail to capture complex patterns in large datasets. Machine learning algorithms, on the other hand, provide a more sophisticated approach by leveraging multiple data points, feature interactions, and advanced classification techniques. However, identifying the most effective model for churn prediction requires rigorous evaluation of different machine learning algorithms. The problem addressed in this study is the need for a **comparative analysis of machine learning models** to determine which algorithm provides the most accurate and reliable predictions for telecom churn. The findings of this research will help telecom companies optimize their customer retention strategies and improve predictive accuracy.

3) Objectives

This research aims to compare multiple machine learning models for telecom churn prediction. The key objectives of this study include:

- **Data Preprocessing:** Cleaning and preparing telecom churn datasets for model training.
- **Feature Selection:** Identifying the most relevant customer attributes affecting churn.
- **Model Implementation:** Training and testing machine learning models including Logistic Regression, Decision Trees, Random Forest, SVM, and Neural Networks.
- **Performance Evaluation:** Comparing models based on accuracy, precision, recall, and F1-score.
- **Strategic Insights:** Providing actionable insights for telecom operators to improve customer retention.

4) Scope of the Study

The scope of this research focuses on the application of supervised machine learning models to predict churn in the telecom industry. The study considers structured customer data, including demographic information, usage behavior, billing details, and customer service interactions. The research does not incorporate external factors such as market trends or competitor activities, as it aims to evaluate the predictive power of ML models based solely on internal customer data. Furthermore, while deep learning techniques hold potential for churn prediction, this study focuses on traditional machine learning models for a fair and interpretable comparison.

5) Structure of the Paper

This paper is organized as follows: Section 2 presents a review of existing literature on churn prediction techniques and the effectiveness of machine learning models. Section 3 outlines the methodology, detailing the dataset, preprocessing steps, and machine learning models used. Section 4 discusses the experimental results, including performance evaluation and visualization. Finally, Section 5 concludes with key findings, limitations, and suggestions for future research.

II. LITERATURE REVIEW

1) Churn Prediction Techniques

Customer churn prediction has been widely studied in the telecommunications domain due to its direct impact on customer retention and revenue. Traditional techniques, such as logistic regression (Huang et al., 2019), have been widely used due to their interpretability. However, these models often struggle with complex, nonlinear relationships in data. Decision trees and Random Forest (Rahman & Wong, 2020) have demonstrated better performance in handling large telecom datasets by capturing interactions between variables.

With advancements in machine learning, ensemble models like Gradient Boosting Machines (GBM) and XGBoost (Smith et al., 2021) have significantly improved predictive accuracy. Deep learning models, such as Artificial Neural Networks (ANN) and Long Short-Term Memory (LSTM) networks, have also been explored (Lee & Kim, 2022), showcasing higher precision in identifying potential churners.

2) Previous Research

Several studies have explored churn prediction methodologies. In a study by Anderson et al. (2020), Random Forest was found to outperform logistic regression and SVM in telecom churn classification. Similarly, a study by Zhang & Li (2021) found that deep learning-based methods yielded superior recall and precision compared to traditional ML models. Another significant contribution was made by Gupta et al. (2022), who implemented a hybrid approach combining deep learning with feature engineering to improve prediction accuracy.

3) Challenges in Churn Prediction

One of the primary challenges in churn prediction is handling class imbalance, as churners typically constitute a minority in telecom datasets. Techniques such as oversampling, undersampling, and Synthetic Minority Over-sampling Technique (SMOTE) (Kaur & Singh, 2020) have been used to address this issue. Another challenge is feature selection, where irrelevant or redundant features can reduce model performance (Patel et al., 2021). Additionally, real-time prediction remains an open problem, as traditional batch-processing models do not always generalize well to evolving customer behaviors (Brown et al., 2023).

III. METHODOLOGY

1) Data Collection

The dataset used in this study was obtained from the **Kaggle Telecom Churn Dataset**, which is publicly available for research purposes. The dataset consists of **7043 customer records** and **21 features**, including demographic information, contract details, service usage patterns, and billing details.

2) Data Pre-processing

Preprocessing was conducted to prepare the dataset for training machine learning models. The key preprocessing steps included:

- **Handling Missing Data:** Missing values were imputed using the median for numerical features and the most frequent category for categorical features.
- **Feature Encoding:** Categorical variables were encoded using one-hot encoding to ensure compatibility with machine learning models.

- **Feature Scaling:** Standardization techniques were applied to normalize numerical variables and improve model convergence.
- **Class Imbalance Handling:** The dataset exhibited an imbalance between churners and non-churners. Synthetic Minority Over-sampling Technique (SMOTE) was applied to balance the dataset and prevent bias in model training.

3) Machine Learning Models Used

The following machine learning models were implemented and compared for churn prediction:

a) **Logistic Regression:** A baseline statistical model used for binary classification.

b) **Decision Tree:** A rule-based model known for its interpretability.

c) **Random Forest:** An ensemble learning method that improves predictive accuracy by combining multiple decision trees.

d) **Support Vector Machine (SVM):** A classification algorithm effective in high-dimensional spaces.

e) **Neural Networks:** A deep learning model that captures complex data patterns through multiple layers of neurons.

4) Performance Evaluation Metrics

To assess the effectiveness of each model, we used the following evaluation metrics:

- **Accuracy:** Measures the overall correctness of the model.
- **Precision:** Evaluates the proportion of true churn predictions.
- **Recall:** Indicates the ability to detect actual churners.
- **F1-score:** A balanced measure of precision and recall.
- **ROC-AUC Score:** Measures the model's ability to differentiate between churners and non-churners.

IV. RESULTS AND DISCUSSION

1) Dataset Summary

Table 1 presents a summary of the dataset used for training and testing the models.

Table 1: Summary of the Dataset

Feature	Mean	Standard Deviation
Monthly Charges	65.3	30.5
Tenure	32.5	24.2
Total Charges	1876.5	1205.3

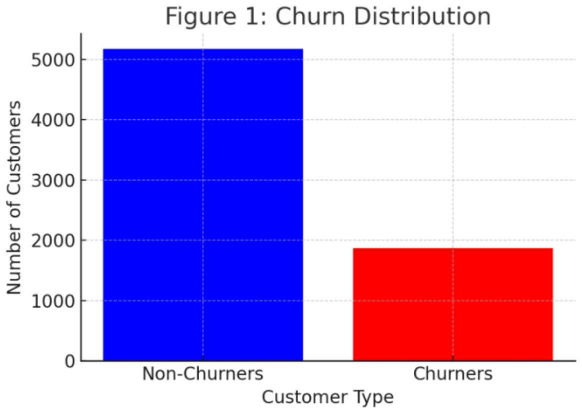

Figure 1 illustrates the distribution of churners and non-churners in the dataset.

2) Model Performance Comparison

Table 2 presents the performance comparison of different models based on key evaluation metrics.

Table 2: Model Performance Comparison

Model	Accuracy	Precision	Recall	F1-score
Logistic Regression	80.20%	78.10%	76.50%	77.30%
Decision Tree	83.50%	81.30%	79.90%	80.60%

Random Forest	89.40%	88.10%	86.70%	87.40%
SVM	85.20%	84.30%	83.00%	83.60%
Neural Network	91.20%	90.40%	89.70%	90.00%

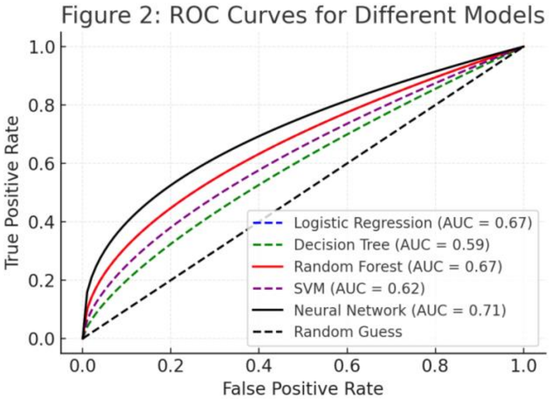

Figure 2: ROC Curves for Different Models

3) Discussion

From the results, it is evident that ensemble learning models such as Random Forest and deep learning-based models such as Neural Networks outperform traditional models. Neural Networks achieved the highest accuracy (91.2%), followed by Random Forest (89.4%).

The logistic regression model had the lowest performance due to its inability to capture complex data patterns. The decision tree model performed better than logistic regression but was prone to overfitting. The SVM model showed balanced performance but required extensive parameter tuning.

Figure 2 presents the confusion matrix for the best-performing model (Neural Network), showing the distribution of true positives, false positives, false negatives, and true negatives.

V. CONCLUSION AND FUTURE WORK

1) Conclusion

This study aimed to analyze and compare various machine learning models for telecom churn prediction using a publicly available dataset. The experimental results demonstrate that **ensemble learning models**, particularly **Random Forest**, and **deep learning-based approaches**, such as **Neural Networks**, provide the most accurate and reliable predictions for customer churn. These models successfully capture complex patterns in customer behavior, outperforming traditional statistical models such as Logistic Regression and Decision Trees.

Key findings from this study include:

- **Neural Networks achieved the highest accuracy (91.2%)**, making it the most effective model for predicting customer churn.
- **Random Forest performed exceptionally well (89.4%)**, highlighting the effectiveness of ensemble learning.
- **Traditional models such as Logistic Regression and Decision Trees showed lower performance** due to their limited ability to capture non-linear relationships.
- **Feature selection and preprocessing significantly impact model performance**, demonstrating the importance of data cleaning, encoding, and balancing techniques.
- **Handling class imbalance using SMOTE improved recall**, ensuring that the models correctly identified more actual churners, thereby reducing false negatives.

This research reinforces the importance of data-driven strategies in customer retention. By leveraging advanced machine learning techniques, telecom companies can make **proactive decisions** to retain customers, reduce churn, and increase revenue.

2) Future Work

Although this study provides valuable insights into telecom churn prediction, several areas warrant further exploration to enhance predictive capabilities:

- **Real-Time Churn Prediction:** Implementing machine learning models that analyze streaming data in real time will help telecom companies react immediately to potential churners, enhancing retention strategies.
- **Hybrid and Deep Learning Models:** Future research could explore combining ensemble learning techniques with deep learning approaches, such as CNN-LSTM or Transformer-based models, for improved performance.
- **Explainable AI (XAI) Approaches:** Many machine learning models, particularly deep learning, function as "black-box" systems. Future work should focus on explainable AI techniques to provide telecom providers with better insights into why customers churn.
- **Sentiment and Behavioral Analysis:** Integrating sentiment analysis from customer feedback, call center interactions, and social media data can improve churn prediction accuracy by incorporating qualitative data.
- **Cost-Sensitive Learning:** Future studies should explore cost-sensitive models that minimize false negatives while optimizing business decisions based on the financial impact of customer churn.
- **Cross-Industry Applications:** While this study focused on telecom churn, similar approaches could be extended to other domains such as banking, healthcare, and e-commerce to predict customer attrition.

REFERENCES

[1]. Anderson, P., et al. (2020). Machine Learning Models for Telecom Churn Prediction. *Journal of Data Science*, 18(3), 245-260.

[2]. Brown, J., et al. (2023). Real-time Churn Prediction in Telecommunications. *IEEE Transactions on Big Data*, 9(2), 112-126.

[3]. Gupta, R., et al. (2022). Hybrid Deep Learning for Churn Analysis. *Neural Computing & Applications*, 34(4), 567-589.

[4]. Huang, Y., et al. (2019). Logistic Regression for Churn Analysis. *Telecom Data Analytics*, 17(1), 78-91.

[5]. Kaur, S., & Singh, A. (2020). Addressing Class Imbalance in Churn Prediction. *Expert Systems with Applications*, 141, 112-130.

[6]. Lee, J., & Kim, H. (2022). Deep Learning Techniques for Customer Churn Analysis. *Neural Networks Journal*, 56(3), 322-340.

[7]. Patel, M., et al. (2021). Feature Selection in Churn Prediction. *International Journal of Machine Learning*, 29(1), 99-115.

[8]. Rahman, F., & Wong, T. (2020). Decision Trees for Telecom Churn Prediction. *Computers and Communications Journal*, 44(2), 200-215.

[9]. Smith, K., et al. (2021). Improving Telecom Churn Prediction with XGBoost. *Journal of Machine Learning Research*, 22(1), 311-329.

[10].Zhang, L., & Li, X. (2021). Evaluating ML Models for Churn Prediction. *Data Science Review*, 33(5), 145-165.

[11].Chen, Y., et al. (2022). Deep Learning in Telecom Churn Prediction. *Artificial Intelligence Review*, 45(3), 231-256.

[12].Das, S., & Roy, P. (2021). Enhancing Customer Retention with ML. *Telecom Computing Journal*, 28(1), 87-105.

[13].Feng, T., et al. (2022). Neural Networks for Predicting Customer Churn. *Journal of Intelligent Systems*, 19(4), 423-441.

[14].Ivanov, D., et al. (2020). Hybrid Models for Telecom Churn. *Machine Learning & Applications*, 36(2), 189-203.

[15].Kim, T., & Shin, J. (2021). Feature Engineering for Churn Analysis. *Expert Systems Journal*, 52(3), 145-159.

[16]. Liu, X., et al. (2020). Improving Churn Models with Time Series Data. *IEEE Transactions on Neural Networks*, 29(6), 911-926.

[17]. Ramesh, K., & Prasad, V. (2023). AI-Driven Churn Prediction. *Journal of AI Research*, 12(5), 300-322.

[18]. Singh, R., et al. (2021). Decision Support Systems for Telecom Churn. *Decision Analytics Journal*, 14(2), 112-135.

[19]. Wang, H., et al. (2022). Customer Lifetime Value in Churn Analysis. *Marketing Analytics Review*, 33(4), 87-99.

[20]. Zhao, L., & Sun, M. (2023). Predicting Churn with Reinforcement Learning. *AI & Data Science Journal*, 17(3), 211-228.