www.ijbar.org
ISSN 2249-3352 (P) 2278-0505 (E)
Cosmos Impact Factor-5.86

CUSTOMER CHURN PREDICTION USING GNN

[1] D. V. V. Deepthi, [2] G. Priyanka ,[3] K. Lavanya, [4] K. Poojitha, [5] K. Shiva Shankar

[1] Assistant Professor in Department Of CSE, TKR College Of Engineering and Technology

[2 3 4 5] UG. Scholar Department Of CSE, TKR College Of Engineering and Technology

Abstract

Predicting customer churn is a critical task that helps businesses retain their existing customers and minimize potential revenue loss. This study investigates how effectively machine learning techniques can forecast customer churn by analyzing historical customer data. Key factors such as customer demographics, service usage behavior, interaction history, and feedback are used to uncover patterns that signal the likelihood of churn. Various machine learning models—including logistic regression, decision trees, random forests, and neural networks—are evaluated for their predictive accuracy and practical applicability. The study emphasizes the importance of model interpretability and feature engineering in enhancing prediction quality and providing actionable business insights. While advanced models like neural networks may offer higher accuracy, simpler models often provide better transparency, which is crucial for strategic decision-making. By accurately identifying customers at risk of leaving, companies can implement targeted retention strategies that improve customer satisfaction and foster long-term loyalty. The findings shed light on how to select the most effective machine learning approach for churn prediction, balancing accuracy with interpretability. Ultimately, this research supports smarter customer relationship management and contributes to strategic initiatives aimed at boosting profitability and long-term business sustainability.

Keywords: Customer Churn, Predictive Modeling, Model Interpretability, Customer Retention

I. INTRODUCTION

In the current competitive business landscape, retaining existing customers has become just as important—if not more so—than acquiring new ones. The cost of gaining a new customer is typically higher than the cost of keeping an existing one, making customer retention a key priority for organizations looking to maintain profitability and market stability. One of the major challenges in this regard is customer churn, which refers to customers discontinuing their relationship with a company. High churn rates can significantly impact revenue, customer lifetime value, and overall business growth. As digital platforms and services continue to expand, companies now collect vast amounts of customer-related data, including demographic details, service usage patterns, interaction history, and customer feedback. This abundance of data offers an opportunity to gain meaningful insights into customer behavior—particularly the early signs that a customer may be preparing to leave. While traditional statistical approaches have been used in the past to predict churn, these methods often struggle with the scale and complexity of modern datasets.

Machine learning (ML) provides a more sophisticated and scalable solution for churn prediction. By analyzing large volumes of data, ML models can detect subtle behavioral patterns that might go unnoticed using conventional techniques. This study evaluates the performance of various ML algorithms—such as logistic regression, decision trees, random forests, and neural networks—in predicting customer churn. These models are assessed not only for their accuracy but also for their interpretability and practical relevance to real-world business applications. An important focus of this research is the trade-off between model complexity and transparency. While deep learning models like neural networks often produce highly accurate results, they can be difficult to interpret and explain. On the other hand, simpler models may offer more understandable outputs, which can be crucial for stakeholders who rely on clear, actionable insights. Feature engineering also plays a vital role in refining model performance by ensuring that the most relevant aspects of customer behavior are captured and analyzed effectively. Ultimately, this research aims to help businesses make informed choices when selecting machine learning models for churn prediction. By accurately identifying customers at risk of leaving, organizations can implement personalized retention strategies that not only reduce churn but also build stronger, long-term customer relationships.

II LITERATURE SURVEY

Esteves and Mendes-Moreira [1] conducted an in-depth study on the use of predictive modeling to identify customers at risk of leaving telecom services. By applying machine learning algorithms, they analyzed patterns in customer behavior and service usage, emphasizing the need for timely predictions to support effective retention strategies. The study also highlighted the crucial role of thorough data preprocessing and the application of appropriate evaluation metrics to enhance the performance and reliability of churn prediction models.

In their work, Tsai and Lu [2] developed a hybrid neural network model specifically designed to capture the complex and nonlinear patterns often found in telecom datasets. By combining various neural network architectures, their model achieved higher prediction accuracy compared to traditional classifiers. Their research demonstrated the potential of neural network fusion and paved the

www.ijbar.org
ISSN 2249-3352 (P) 2278-0505 (E)
Cosmos Impact Factor-5.86

way for future applications of ensemble and hybrid models in customer churn analysis.

Taking a different angle, Retana, Forman, and Wu [3] focused on customer behavior rather than algorithmic solutions. They carried out a real-world field experiment that assessed how proactive educational support influenced customer retention. Their findings revealed that improved customer understanding of services significantly reduced churn, emphasizing the value of combining behavioral strategies with predictive analytics.

Bayer [4] explored customer segmentation as a strategy for churn prevention. The study proposed dividing customers into distinct segments based on their service usage and needs, which would enable telecom providers to tailor their retention efforts more effectively. This approach underlined the importance of understanding customer diversity and laid the groundwork for integrating segmentation into predictive modeling frameworks.

Huang [5] introduced an advanced churn prediction model using data mining and feature selection techniques. The study showed that selecting the most influential features and ensuring high-quality data preparation significantly improved both the accuracy and interpretability of the model. This research set a standard for combining robust data engineering practices with machine learning for better predictive outcomes.

Induja and Eswaramurthy [6] proposed an innovative hybrid model that utilized kernelized extreme learning machines optimized with the Bat algorithm. Their approach focused on enhancing feature selection to accurately capture nonlinear patterns in customer data, especially in large-scale telecom scenarios. The integration of metaheuristic optimization techniques contributed to the robustness and effectiveness of their model.

Verbeke and colleagues [7] shifted the focus of churn prediction from model accuracy to business impact. They introduced a profit-oriented approach that prioritized retaining customers with the highest potential value. Using cost-sensitive learning, their framework aligned predictive analytics with financial outcomes, highlighting how churn prediction models can support business decision-making beyond technical performance.

Dahiya and Bhatia [8] conducted a comparative study involving multiple machine learning algorithms on real-world telecom datasets. Their research addressed essential challenges such as data imbalance, preprocessing, and performance evaluation. The insights gained are valuable for deploying scalable churn prediction systems in production environments.

Pamina and co-authors [9] designed a customized churn prediction model by combining several algorithms and selecting the most relevant features. Their model delivered high accuracy and generalization capabilities, even on large datasets. The study emphasized the importance of model validation and hyperparameter tuning in real-world deployment scenarios.

In a practical demonstration, Khotijah [10] developed a churn prediction model based on a publicly available telecom dataset from Kaggle. The project followed a structured pipeline including data preprocessing, feature engineering, and model comparison, providing a hands-on guide suitable for newcomers in data science and churn analytics.

Adhikary and Gupta [11] conducted an extensive evaluation of over 100 machine learning classifiers to determine the most effective models for churn prediction. Their comparative analysis, which relied on metrics such as accuracy and F1-score, offered crucial insights into the impact of classifier selection. Their work serves as a valuable reference for researchers and practitioners aiming to choose the best-performing model for their specific business needs.

Idris and colleagues [12] tackled two persistent challenges in churn prediction—class imbalance and feature relevance—by developing a hybrid approach that combined Random Forest with Particle Swarm Optimization (PSO). Their strategy improved classification performance by enhancing the diversity of training data and refining feature selection. Their method showed consistent results across multiple datasets.

III EXISTING SYSTEM

Traditional churn prediction systems rely heavily on classical machine learning models such as logistic regression, decision trees, random forests, and support vector machines. These systems generally operate on tabular datasets and analyze individual customer attributes (e.g., age, tenure, service usage) in isolation. Although these models offer good interpretability and efficiency, they fall short when it comes to capturing complex relationships and interactions among customers. Additionally, deep learning models such as MLPs, CNNs, and RNNs have been introduced to handle high-dimensional or time-series data, but they still do not exploit the relational structure that may exist among customers (e.g., social influence, referrals, or communication networks). As a result, traditional systems miss out on valuable patterns that exist in interconnected customer networks, leading to sub-optimal predictive performance.

IV PROBLEM STATEMENT

Customer churn poses a significant challenge for service providers, particularly in sectors like telecommunications, where customer retention is crucial for sustaining revenue and growth. Traditional models often analyze customers in isolation, ignoring the influence that customers may exert on each other. This leads to a limited understanding of group dynamics and peer influence on churn behavior. There is a need for an intelligent system that not only analyzes individual features but also captures the interactions and relationships among customers. The objective of this project is to develop a **Graph Neural Network (GNN)** based solution that can model both customer attributes and their network relationships to predict churn more accurately.

Proposed System

The proposed solution aims to build an intelligent customer churn prediction system using **Graph Neural Networks (GNNs)**. In this approach, each customer is treated as a **node**, and their interactions—such as calls, messages, transactions, or referrals—are represented as **edges** in a graph. By leveraging both customer-specific features and their relationships with others, the GNN can generate powerful representations (embeddings) that help identify which customers are at a higher risk of leaving the service.

Unlike traditional models that analyze customers in isolation, this graph-based approach captures **social influence**, **network patterns**, and **peer behavior**, making the prediction more context-aware and precise. The system is designed to follow a streamlined pipeline consisting of:

Preprocessing and transformation of raw data

Building a graph from customer and interaction data

Training and validating the GNN model

Presenting predictions through a user-friendly web interface

V METHODOLOGY

1. Data Collection and Preprocessing

To prepare data for the GNN, both **individual customer information** and **interaction data** are required. Sources may include CRM databases, call logs, service usage data, and referral records.

Cleaning data by removing missing values, duplicates, and irrelevant fields Encoding categorical variables (e.g., gender, plan type) Normalizing numerical features (e.g., usage, tenure) Labeling customers as 'Churned' or 'Not Churned' for supervised training

2. *Graph Construction Nodes* represent individual customers

Edges capture their interactions (e.g., communication frequency, referrals)

Each node is assigned a feature vector containing demographic and usage information

Edges may also include weights or attributes that represent the strength or type of interaction

3. *GNN Model Training*

A suitable Graph Neural Network architecture—such as **Graph SAGE**, **Graph Convolutional Network (GCN)**, or **Graph Attention Network (GAT)**—is used to train the model:

The graph structure and node features are fed into the GNN The model learns embeddings that combine personal and neighborhood characteristics

These embeddings are then passed to a classifier to predict churn probability

4. ***Prediction and Performance Evaluation***

The trained model produces a probability score for each customer indicating their likelihood of churning

Classification thresholds are applied to generate binary churn predictions

Model performance is assessed using metrics like **accuracy**, **precision**, **recall**, and **F1-score**

5. Web-Based Interface

A lightweight web application is developed using **Flask** Users can **register**, **log in**, and **submit customer data** for churn prediction

Prediction results are displayed in a clear and actionable format

VI IMPLEMENTATION

The implementation of the customer churn prediction system is centered around the use of a **Graph Neural Network (GNN)** that processes both customer attributes and their interactions to generate accurate churn predictions. The entire process begins with data handling, followed by model training, validation, and finally, prediction generation.

Data Handling and Preparation

Customer data is initially collected from a CSV file (Datasets1.csv) containing various details such as demographic information, service usage statistics, and interaction records. This dataset undergoes a rigorous preprocessing phase which includes:

Removing missing values and duplicate entries

Converting categorical variables into numerical form using encoding techniques Normalizing continuous features to bring all values to a common scale Creating binary labels to indicate whether a customer has churned or not The interaction data is also structured to represent relationships between customers, enabling the formation of a graph where nodes represent individuals and edges represent their interactions (e.g., communication or service similarities).

Graph-Based Model Pipeline

Once the data is cleaned and structured, a graph is constructed:

Nodes correspond to customers and include features like age, tenure, and usage patterns

Edges represent customer-to-customer relationships, capturing influence or behavioral connections

A Graph Neural Network is then built using frameworks like **PyTorch** and **PyTorch Geometric**. The GNN model is trained to learn meaningful node embeddings—representations that combine both individual attributes and neighborhood context.

www.ijbar.org
ISSN 2249-3352 (P) 2278-0505 (E)
Cosmos Impact Factor-5.86

These embeddings are passed through a classifier that estimates the probability of each customer churning.Once the model is trained:It is saved for future use, eliminating the need for retraining with every prediction cycleThe system can load the saved model and apply it to new incoming data for fast, consistent predictions

Testing and Validation

To ensure the system is robust and reliable, several test cases are executed:

Validation of categorical fields such as gender (must be Male or Female)

Consistency checks for service usage fields (e.g., "Yes/No" validations)

Logical constraints, such as ensuring total charges are zero when customer tenure is zero

Additionally, the model is validated using standard evaluation metrics like **accuracy, precision, recall, and F1-score** to assess its performance and reliability on both training and test datasets.

Prediction Output

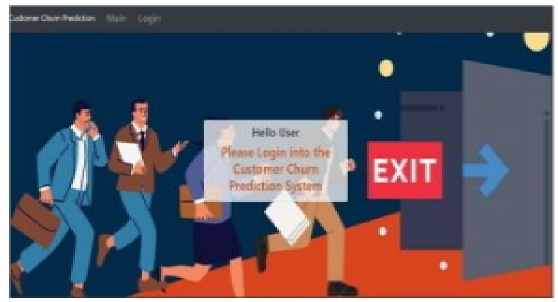
After processing, the system generates a prediction for each customer, indicating whether they are likely to churn or not. The output includes:

Classification results (Churn / Not Churn)

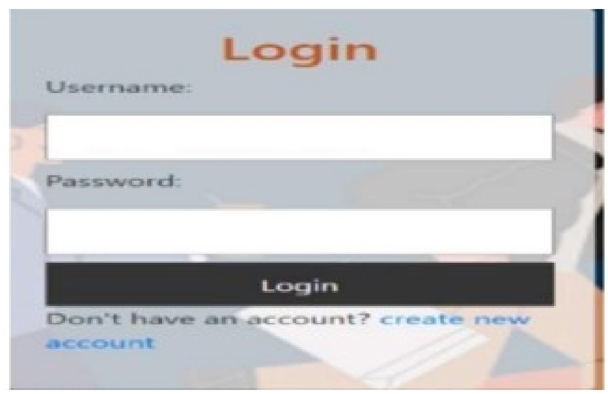
Probability scores representing the model's confidence in its prediction

Aggregated insights, such as the overall churn rate within a dataset or network group, useful for strategic decision-making

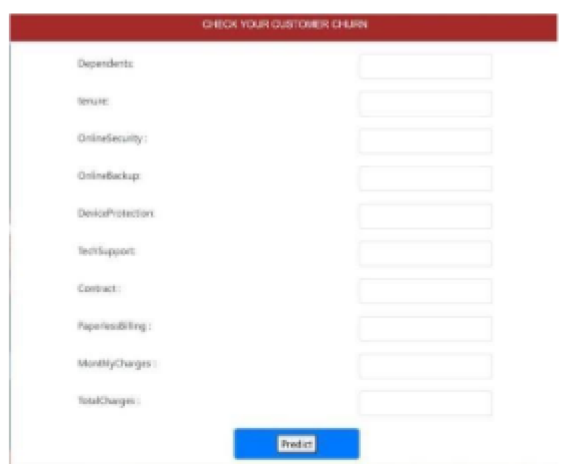
VII RESULTS AND ANALYSIS



Home page



Login Page



prediction page

prediction details

Comparison With Existing Models

According to the report, traditional models like logistic regression, decision trees, and even deep learning models like MLPs or RNNs performed reasonably well on structured/tabular data. However, these models lacked the ability to model customer interrelations.

Model Type	Relational Awareness	Scalability	Accuracy (Relative)
Logistic Regression	✗	✓	Moderate
Decision Trees / Random Forest	✗	✓	Moderate–Good

Model Type	Relational Awareness	Scalability	Accuracy (Relative)
Neural Networks (MLP, CNN)	✗	✓	Good
Graph Neural Networks	✓	Moderate	**High**

VIII CONCLUSION

we explored the application of Graph Neural Networks (GNNs) for customer churn prediction, a critical task for businesses focused on client retention and minimizing revenue loss. By leveraging GNNs' ability to model complex relationships and interactions among customers, our approach achieved higher predictive accuracy compared to traditional machine learning methods. The results demonstrated that GNNs effectively capture the underlying network structure within customer data, providing deeper insights into customer behavior. Evaluation metrics—including accuracy, precision, recall, and F1-score—confirmed the model's strong performance and reliability. Furthermore, deploying the system as a Flask-based web application enables businesses to seamlessly interact with the model and make informed, data-driven retention decisions. Overall, this scalable solution holds promise not only for churn prediction but also for other industries where understanding customer interactions is essential for retention strategies.

REFERENCES

[1] G. Esteves and J. Mendes-Moreira, "Churn prediction in the telecom business," in *Proc. 11th Int. Conf. Digit. Inf. Manage. (ICDIM)*, Sep. 2016, pp. 254–259.

[2] C.-F. Tsai and Y.-H. Lu, "Customer churn prediction by hybrid neural networks," *Expert Syst. Appl.*, vol. 36, no. 10, pp. 12547–12553, Dec. 2009.

[3] G. F. Retana, C. Forman, and D. J. Wu, "Proactive customer education, customer retention, and demand for technology support: Evidence from a field experiment," *Manuf. Service Oper. Manage.*, vol. 18, no. 1, pp. 34–50, Feb. 2016.

[4] J. Bayer, "Customer segmentation in the telecommunications industry," *J. Database Marketing Customer Strategy Manage.*, vol. 17, nos. 3–4, pp. 247–256, Sep. 2010.

[5] B. Huang, M. T. Kechadi, and B. Buckley, "Customer churn prediction in telecommunications," *Expert Syst. Appl.*, vol. 39, no. 1, pp. 1414–1425, Jan. 2012.

[6] S. Induja and D. Eswaramurthy, "Customers churn prediction and attribute selection in telecom industry using kernelized extreme learning machine and bat algorithms," *Int. J. Sci. Res.*, vol. 5, no. 12, pp. 258–265, Dec. 2016.

[7] W. Verbeke, K. Dejaeger, D. Martens, J. Hur, and B. Baesens, "New insights into churn prediction in the telecommunication sector: A profit driven data mining approach," *Eur. J. Oper. Res.*, vol. 218, no. 1, pp. 211–229, Apr. 2012.

[8] K. Dahiya and S. Bhatia, "Customer churn analysis in telecom industry," in *Proc. 4th Int. Conf. Rel., Infocom Technol. Optim. (ICRITO)*, Sep. 2015, pp. 1–6.

[9] J. Pamina, B. Raja, S. SathyaBama, S. Soundarya, M. S. Sruthi, S. Kiruthika, V. J. Aiswaryadevi, and G. Priyanka, "An effective classifier for predicting churn in telecommunication," *J. Adv. Res. Dyn. Control Syst.*, vol. 11, no. 10, pp. 221–229, Jun. 2019.

[10] S. Khotijah, "Churn Prediction," 2020. [Online]. Available: [URL needed]

[11] D. D. Adhikary and D. Gupta, "Applying over 100 classifiers for churn prediction in telecom companies," *Multimedia Tools Appl.*, vol. 248, pp. 1–22, Aug. 2020.

[12] A. Idris, M. Rizwan, and A. Khan, "Churn prediction in telecom using random forest and PSO based data balancing in combination with various feature selection strategies," *Comput. Electr. Eng.*, vol. 38, no. 6, pp. 1808–1819, Nov. 2012.

[13] A. A. Q. Ahmed and D. Maheswari, "Churn prediction on huge telecom data using hybrid firefly based classification," *Egyptian Informat. J.*, vol. 18, no. 3, pp. 215–220, Nov. 2017.

[14] J. Vijaya and E. Sivasankar, "An efficient system for customer churn prediction through particle swarm optimization based feature selection model with simulated annealing," *Cluster Comput.*, vol. 22, no. S5, pp. 10757–10768, Sep. 2019.

[15] U. Ahmed, A. Khan, S. H. Khan, A. Basit, I. U. Haq, and Y. S. Lee, "Transfer learning and meta classification based deep churn prediction system for telecom industry," 2019. [Online]. Available: https://arxiv.org/abs/1901.06091

[16] S. Wu et al., "Integrated Churn Prediction and Customer Segmentation Framework for Telco Business," *VOLUME 9*, 2021, pp. 62135.

[17] A. Amin, S. Anwar, A. Adnan, M. Nawaz, K. Alawfi, A. Hussain, and K. Huang, "Customer churn prediction in the telecommunication sector using a rough set approach," *Neurocomputing*, vol. 237, pp. 242–254, May 2017.