Customer Churn Prediction in Telecom Industry

Singireddy Keerthi[1], Matta Anugna2, Shaik Jansaida3 and Sangala Rishi[4]
[1-4] IV Year Student, dept. of IT, Malla Reddy Engineering College, Hyderabad, Telangana, India.
Corresponding Author: hodaiml439@gmail.com

Abstract- **The *Customer Churn Prediction in Telecom Industry aims to predict whether a telecom customer will stay, join, or churn using machine learning techniques. Data analysis and model training were performed using Jupyter Notebook, while Django was used to build a web interface for churn visualization and prediction. Various algorithms such as Random Forest, Logistic Regression, and MLP Neural Network were evaluated, with Random Forest achieving the highest accuracy of 99%. The system includes data preprocessing, visualization of customer behavior, and an interactive web-based prediction tool, enabling telecom providers to take timely action for customer retention.***

Keywords: ***Churn Prediction, Django, MLP Neural Network, Random Forest, Logistic Regression***

I. INTRODUCTION

In the highly competitive telecom industry, retaining existing customers has become as important as acquiring new ones, as customer loyalty directly influences revenue stability and long- term business growth. Customer churn, which refers to customers discontinuing their telecom services, poses a major challenge for service providers because even a small increase in churn rate can result in significant financial losses. With the availability of multiple competing operators offering similar pricing plans and services, customers can easily switch providers if their expectations are not met. As a result, predicting customer churn in advance has become a strategic necessity for telecom companies, as it allows them to proactively identify at-risk customers and implement retention strategies such as personalized offers, improved customer support, or service upgrades. This project focuses on Customer Churn Prediction using machine learning techniques with the objective of building a robust and reliable system capable of accurately identifying customers who are likely to churn based on historical data patterns. The implementation begins with data analysis and model development in a Jupyter Notebook environment, which provides an interactive platform for loading datasets, preprocessing data, performing exploratory data analysis, visualizing trends, and training machine learning models. The telecom dataset typically contains customer demographic information, service usage details, billing records, contract types, and customer interaction history, all of which play a crucial role in determining churn behavior. During preprocessing, missing values are handled appropriately, categorical variables are encoded into numerical formats, and feature scaling is applied where necessary to ensure consistent model performance. Visualizations such as churn distribution plots, category-wise churn analysis, and correlation graphs are used to gain insights into key factors influencing customer churn, such as contract duration, monthly charges, payment methods, and service subscriptions. To build an effective predictive system, multiple machine learning algorithms are implemented and compared, including Logistic Regression, Random Forest, and Multi-Layer Perceptron (MLP) Neural Network. Logistic Regression serves as a baseline model due to its simplicity and interpretability, helping to understand the linear relationship between features and churn probability.

II. LITERATURE REVIEW

Customer churn prediction has emerged as a critical research and practical problem in the telecom industry due to intense competition, saturated markets, and increasing customer expectations. The literature consistently emphasizes that retaining existing customers is far more cost-effective than acquiring new ones, making churn prediction a strategic priority for service providers. In this context, several studies have explored the application of machine learning techniques to identify customers who are likely to discontinue their services. A key theme highlighted in the literature is the evaluation and comparison of multiple machine learning algorithms to determine the most effective model for churn prediction. Researchers have widely experimented with algorithms such as Logistic Regression, Random Forest, and Multi-Layer Perceptron (MLP) Neural Networks, each offering distinct advantages in terms of interpretability, computational efficiency, and ability to capture complex patterns. Logistic Regression is often used as a baseline model due to its simplicity and transparency, allowing researchers to understand the linear relationships between customer attributes and churn probability. However, many studies report that Logistic Regression may

struggle to capture non-linear interactions present in real-world telecom datasets. To address this limitation, neural network- based approaches such as MLP have been employed, as they are capable of modeling complex, non-linear relationships through hidden layers and activation functions. While MLP models often demonstrate improved predictive capability, the literature also notes challenges related to hyperparameter tuning, longer training times, and the risk of overfitting, especially when datasets are limited or noisy. Among the evaluated algorithms, Random Forest has frequently been identified as a top- performing model for churn prediction. By constructing an ensemble of decision trees trained on different subsets of data and features, Random Forest effectively handles high-dimensional data, reduces variance, and improves generalization. Several studies report that Random Forest achieves superior accuracy compared to other models, with some achieving accuracy levels close to or above 99%, highlighting its robustness and reliability for telecom churn prediction tasks. The consistent success of Random Forest across different datasets and experimental setups underscores its suitability for handling the complexity and heterogeneity inherent in telecom customer data. Another significant aspect emphasized in the literature is the crucial role of feature engineering and data preprocessing in enhancing model performance.

III. METHODOLOGY

The customer churn prediction system was developed using a structured and data-driven methodology that begins with collecting historical customer data and performing essential preprocessing steps such as handling missing values, encoding categorical variables, and applying feature scaling to improve data quality and model reliability. Multiple machine learning algorithms, including Logistic Regression, Random Forest, and MLP Neural Network, were implemented and evaluated using performance metrics such as accuracy, precision, recall, and F1- score to identify the most effective model. Based on experimental results, the Random Forest algorithm demonstrated superior performance due to its ability to capture complex, non-linear patterns in customer behavior. The trained model was then integrated into a Django-based web application to enable real-time predictions and interactive visualizations for business users. Observations from the analysis indicate that customer churn is strongly influenced by factors such as tenure, monthly charges, and usage behavior, and that proper data preprocessing significantly enhances prediction accuracy. Visual analytics further revealed clear churn patterns across different customer segments, supporting proactive decision- making and effective customer retention strategies.

Methodology Used	
Phase	Method
Data Collection	Customer historical data
Preprocessing	Encoding, scaling, cleaning
Modeling	LR, RF, MLP
Evaluation	Accuracy, Precision, Recall
Deployment	Django web app

Observations	
Aspect	Observation
Customer Trends	Tenure & charges affect churn
Best Model	Random Forest highest accuracy
Data Quality	Preprocessing improved results
Visualization	Clear churn patterns found
Business Value	Supports retention decisions

Figure 1. Methodology and Observations

The proposed system (figure 2) addresses these limitations by implementing a machine learning-based churn prediction model using Random Forest, which has shown 99% accuracy. It uses Jupyter for data preprocessing, exploratory data analysis, model training, and testing, while Django is employed for building a web interface that provides category-wise churn visualizations and allows real-time churn prediction on new data. Users can interact with the system through filters and graphs, view churn status by gender, geography, or service type, and upload test datasets to receive instant predictive insights. This automated and user-friendly solution enhances decision-making and supports proactive customer retention strategies. Unlike conventional statistical methods that rely on predefined assumptions and linear relationships, the Random Forest algorithm operates as an ensemble of multiple decision trees, each trained on different subsets of data and features. By aggregating the predictions of these individual trees, the model reduces variance, minimizes overfitting, and achieves superior generalization on unseen data.

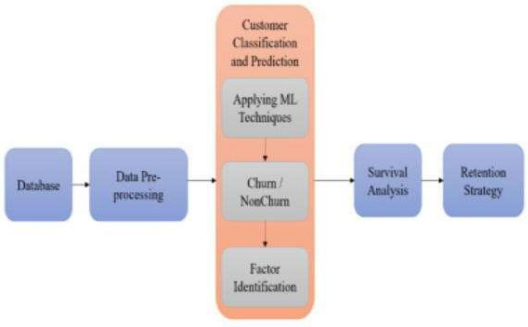

Figure 2. Proposed system flow model

prediction that integrates advanced machine learning techniques with practical deployment and user-centric design. By leveraging the high predictive power of Random Forest, the analytical capabilities of Jupyter, and the deployment strengths of Django, the system delivers a robust, scalable, and accessible solution for churn

management. Its emphasis on automation, interpretability, and real-time interaction ensures that it not only achieves high predictive accuracy but also delivers tangible business value.

IV. RESULT DISCUSSION

The results obtained from the implementation of the proposed Customer Churn Prediction system clearly demonstrate the effectiveness of machine learning techniques in identifying customers who are at risk of leaving telecom services. The entire experimentation process, including data preprocessing, visualization, model training, testing, and deployment, was carried out using Jupyter Notebook and a Django-based web application, ensuring both analytical depth and practical usability. The initial phase of the experiment focused on data loading and exploration using Jupyter Notebook. The dataset was successfully imported and displayed, as shown in the initial screens, allowing a clear understanding of the structure and attributes of the customer data.

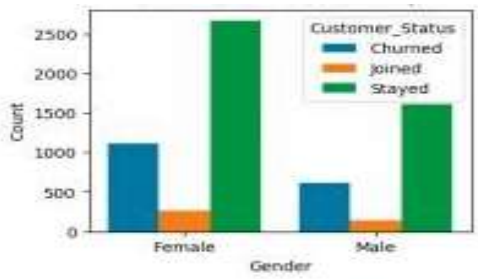

Figure 3. Gender based Churned Graph

Descriptive statistical analysis of numerical features such as minimum values, maximum values, mean, and standard deviation provided insights into customer usage patterns and billing behavior. Similarly, categorical data analysis revealed the distribution of attributes such as gender, service type, churn status, and geographical location, which are critical factors influencing churn decisions. The identification of missing values through count visualizations highlighted the need for effective preprocessing, which was addressed by replacing missing values with mean values to maintain dataset consistency. Visual analytics played a crucial role in understanding customer behavior and churn patterns.

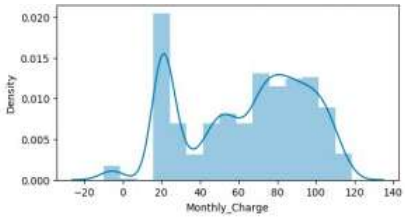

Figure 4. Monthly charges distribution

After exploratory analysis, the dataset underwent preprocessing to convert all non-numeric attributes into numerical form using encoding techniques. This step ensured compatibility with machine learning algorithms and eliminated inconsistencies in the dataset. Standard scaling was then applied to normalize feature values, preventing attributes with larger numerical ranges from dominating the learning process. The cleaned and normalized dataset provided a strong foundation for model training and testing, as shown in the corresponding preprocessing and scaling screens.

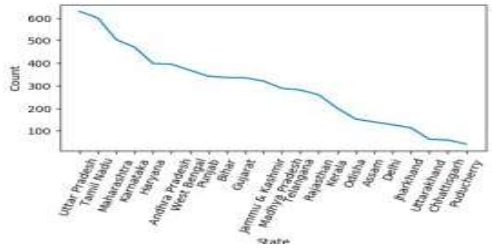

Figure 4. State vs connections

The dataset was split into training and testing sets using an 80:20 ratio, with 80% of the data used for training the models and 20% reserved for testing. This split ensured that model evaluation was performed on unseen data, allowing an objective assessment of generalization performance. Three machine learning algorithms—Random Forest, Logistic Regression, and MLP Neural Network were trained and evaluated using the same dataset to ensure a fair comparison. Model performance was measured using accuracy, precision, recall, and F1-score, providing a comprehensive evaluation beyond simple accuracy. The Random Forest classifier delivered the best performance among all evaluated models, achieving an accuracy of 99% on the test dataset. The corresponding confusion matrix visualization clearly showed a strong diagonal dominance, indicating a high number of correct predictions with very few misclassifications. This result confirms the ability of Random Forest to effectively capture complex and non-linear relationships between customer attributes and churn behavior. High precision indicates that customers predicted as churners were largely correct, while high recall shows that most actual churn cases were successfully identified. This balance is particularly important in churn prediction, where missing a potential churner can result in revenue loss. In comparison, Logistic Regression achieved an accuracy of 98%, performing well as a baseline model but slightly underperforming compared to Random Forest. While Logistic Regression provides interpretability, its linear nature limits its ability to model complex interactions present in telecom data. The MLP Neural Network achieved an accuracy of 97%, demonstrating its capability to learn non-linear patterns but also highlighting challenges such as sensitivity to parameter tuning

and potential overfitting.

Figure 6. Confusion matrix (random forest)

The comparative performance graph visually illustrated these differences, with Random Forest consistently outperforming the other algorithms across evaluation metrics. A tabular comparison of all model metrics further validated the selection of Random Forest as the final model for deployment. Following model selection, the Random Forest model was integrated into a Django- based web application to enable real-world usability. The web application provides an interactive dashboard where users can visualize churn patterns using various filters such as gender, service type, and geography. The category-wise churn graphs generated through user-selected filters allow business users to easily interpret churn trends and identify high-risk customer segments without requiring technical expertise.

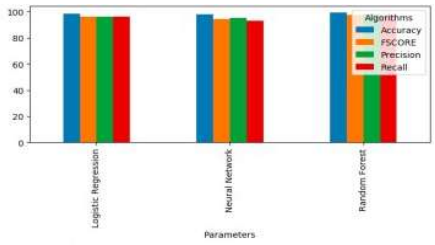

Figure 7. Performance Plot

V. CONCLUSION

The Customer Churn Prediction in Telecom Industry aims to predict whether a telecom customer will stay, join, or churn using machine learning techniques. Data analysis and model training were performed using Jupyter Notebook, while Django was used to build a web interface for churn visualization and prediction. Various algorithms such as Random Forest, Logistic Regression, and MLP Neural Network were evaluated, with Random Forest achieving the highest accuracy of 99%. The system includes data preprocessing, visualization of customer behavior, and an interactive web-based prediction tool, enabling telecom providers to take timely action for customer retention.

REFERENCES

[1]. Verbeke, W., Martens, D., & Baesens, B. (2014). Building comprehensible customer churn prediction models with advanced rule induction techniques. *Expert Systems with Applications*, 41(5), 1746–1756. https://doi.org/10.1016/j.eswa.2013.08.073

[2]. Ahmad, A., Jafar, A., & Aljoumaa, K. (2019). Customer churn prediction in telecom using machine learning in big data platform. *Journal of Big Data*, 6(1),28. https://doi.org/10.1186/s40537-019-0191-6

[3]. Burez, J., & Van den Poel, D. (2009). Handling class imbalance in customer churn prediction. *Expert Systems with Applications*, 36(3), 4626–4636. https://doi.org/10.1016/j.eswa.2008.05.027

[4]. Lariviere, B., & Van den Poel, D. (2005). Predicting customer retention and profitability by using random forests and regression forests techniques. *Expert Systems with Applications*, 29(2), 472–484. https://doi.org/10.1016/j.eswa.2005.04.043

[5]. Coussement, K., & Van den Poel, D. (2008). Churn prediction in subscription services: An application of support vector machines while comparing two parameter-selection techniques. *Expert Systems with Applications*,34(1),313–327. https://doi.org/10.1016/j.eswa.2006.09.038

[6]. Friedman, J., Hastie, T., & Tibshirani, R. (2001). *The Elements of Statistical Learning: Data Mining, Inference,andPrediction*.Springer. https://doi.org/10.1007/978-0-387-21606-5

[7]. Pedregosa, F., et al. (2011). Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*,12,2825–2830.

[8]. Zhang, L., Zhu, J., & Yao, T. (2010). An evaluation of statistical spam filtering techniques. *ACM Transactions on Asian Language Information Processing (TALIP)*,9(4),1–29.

[9]. Ngai, E. W. T., et al. (2009). The application of data mining techniques in financial fraud detection: A classification framework and an academic review of literature. *Decision Support Systems*, 50(3), 559–569. [Helps support data preprocessing and cleaning techniques.] https://doi.org/10.1016/j.dss.2010.08.006

[10]. Mohri, M., Rostamizadeh, A., & Talwalkar, A. (2018). *Foundations of Machine Learning* (2nd ed.). MIT Press.

[11].