OPTIMIZING TELECOM OPERATIONS WITH SEGMENTATION AND CHURN PREDICTION

Dandu Neha
B.Tech. Student, Dept. of CSE
Institute of Aeronautical Engineering
Hyderabad, India
21951A05B8@iare.ac.in

Mohammed Ameesha
B.Tech. Student, Dept. of CSE
Institute of Aeronautical Engineering
Hyderabad, India
21951A05A4@iare.ac.in

Mukka Dheeraj
B.Tech. Student, Dept. of CSE
Institute of Aeronautical Engineering
Hyderabad, India
21951A0538@iare.ac.in

K.Sangeeta
Assisstant Professor of CSE
Institute of Aeronautical Engineering
Hyderabad, India
K.sangeeta.iare.ac.in

Abstract— **Client turnover is a major challenge for large companies, especially in the telecom sector, as it directly affects profitability. Telecom providers are increasingly focused on developing methods to predict customer churn to mitigate its negative effects. Understanding the factors that contribute to churn is essential for preventing it. This work's primary contribution is the development of a churn prediction model, which helps telecom providers identify customers at high risk of churning. We explore different data analysis scenarios and display the results using graphical representations. Our model, built using machine learning techniques, leverages large datasets and introduces a novel approach for feature design and selection. The datasets used for training, evaluation, and validation include historical customer data from previous months. We experimented with four machine learning algorithms: Logistic Regression, XGBoost, Gradient Boosted Machine (GBM) Trees, Random Forests, and Decision Trees. To enhance accuracy, we applied parameter tuning during training. The goal of this study is to optimize telecom operations through effective segmentation and churn prediction, ultimately improving operational efficiency and customer retention.**

Keywords— ***Customer churn, churn prediction, telecom operations, machine learning, feature selection, customer segmentation, operational efficiency.***

I. INTRODUCTION

In the highly competitive telecommunication industry, customer churn—where customers discontinue their service subscriptions—poses a significant challenge. The cost of acquiring new customers often exceeds the expenses associated with retaining existing ones, making churn reduction a critical focus for telecom operators [1]. High churn rates can erode profitability and market share, underscoring the need for effective churn prediction and management strategies [2]. Machine learning (ML) offers a transformative approach to address this issue. By leveraging vast amounts of historical customer data, including service usage patterns, demographic details, and interaction histories, ML models can predict the likelihood of a customer discontinuing their service [3]. This predictive capability enables telecom companies to proactively engage with at-risk customers, offering personalized incentives or service improvements to retain them [4].

The development of a churn prediction model involves several steps: data collection and preprocessing, feature selection and engineering, model training, and evaluation [5]. Data from various sources such as call records, billing information, customer service interactions, and social media can be integrated to form a comprehensive view of each customer's behavior [6]. Advanced ML algorithms, including logistic regression, decision trees, random forests, and neural networks, can be employed to analyze these data points and identify patterns indicative of churn [7].

Effective churn prediction not only assists in retaining customers but also provides insights into the underlying causes of churn [8]. These insights can drive strategic decisions in product development, customer service enhancements, and targeted marketing campaigns. Furthermore, reducing churn rates directly increases customer lifetime value and provides a more stable revenue base [9,10]. The project aims to build a scalable and accurate churn prediction system that can be seamlessly integrated into a telecom operator's existing infrastructure [11,12]. By doing so, telecom companies can shift from reactive to proactive customer relationship management, improving customer satisfaction and loyalty in a fiercely competitive market.

II. RESEARCH BACKGROUND

Customer churn—referring to the loss of customers who discontinue their service subscriptions—presents a significant challenge. The expense involved in acquiring new customers often exceeds the cost of retaining existing ones, making the reduction of churn a critical focus for telecom operators [1]. High churn rates can adversely affect both profitability and market share, highlighting the necessity for effective churn prediction and management strategies.

Machine learning (ML) provides a valuable solution to this issue. By analyzing comprehensive historical customer data, including service usage patterns, demographic information, and interaction histories, ML models can predict the likelihood of a customer terminating their service [2]. This

predictive capability enables telecom companies to proactively engage with at-risk customers, offering personalized incentives or improving service quality to enhance retention [3].

The development of a robust churn prediction model involves several key stages: data collection and preprocessing, feature selection and engineering, model training, and evaluation [4]. By integrating data from diverse sources—such as call records, billing information, customer service interactions, and social media—companies can gain a thorough understanding of customer behavior [5]. Advanced ML algorithms, including logistic regression, decision trees, random forests, and neural networks, are used to analyze these data points and identify patterns that signal potential churn [6][7].

Effective churn prediction not only helps in retaining customers but also offers valuable insights into the underlying reasons for churn [8]. These insights can inform strategic decisions regarding product development, customer service improvements, and targeted marketing efforts. Moreover, reducing churn rates directly contributes to increased customer lifetime value and a more stable revenue stream [9]. This research aims to develop a scalable and precise churn prediction system that can be seamlessly integrated into a telecom operator's existing infrastructure, facilitating a shift from reactive to proactive customer relationship management. This approach enhances customer satisfaction and loyalty in a competitive market [10][11].

Prior research has explored various methodologies for churn prediction. For instance, Wei and Chiu (2002) investigate a data mining approach based on telecommunications call details [2]. Ascarza et al. (2016) examine the complexities of proactive churn prevention through plan recommendations [4]. Burez and Poel (2009) address the challenges of class imbalance in churn prediction [8]. Integrating these methodologies with advanced ML techniques can significantly improve the effectiveness of churn prediction systems [12].

III. METHODOLOGY

A. *Proposed work*

To address the critical issue of client turnover in the telecommunications sector, we propose an advanced churn prediction model. This model is designed to enable telecom providers to proactively identify customers who are at high risk of churning, thereby facilitating effective retention strategies.

Our approach leverages advanced machine learning techniques within a robust data environment to ensure accurate and reliable predictions. The system involves a thorough analysis of customer data and employs an innovative method for feature design and selection. This approach helps to pinpoint the most significant factors influencing churn and incorporates them into the predictive model. By analyzing various data scenarios and visualizing the results, we offer valuable insights into customer behavior and the primary drivers of churn.

The model is trained, evaluated, and validated using a comprehensive dataset that includes detailed customer information from recent months. We assess the effectiveness of different predictive methods by testing four algorithms: Random Forest, Logistic Regression, Extreme Gradient Boosting (XGBoost), and Gradient Boosted Machine Trees (GBM). Parameter tuning is applied to these algorithms to further enhance the model's performance.

The resulting churn prediction model provides telecom providers with a powerful tool to anticipate and mitigate customer churn. This proactive approach allows companies to gain deeper insights into customer behavior, implement targeted retention strategies, and ultimately improve profitability.

B. *Flow chart*

Launch the Application:

Begin by starting the web application, either on a local server or through a cloud-based service. This action makes the app available for use through a web browser.

Authenticate Users:

Users must log in using their credentials (username and password) to access the app's functionalities. This step is crucial for maintaining data security and personalizing user experiences.

Import Prediction Model:

Load a pre-existing churn prediction model into the application. This model, which may have been previously trained with machine learning techniques such as Logistic Regression, Decision Trees, or Neural Networks, enables the app to make predictions without needing to retrain.

Upload Telecom Data:

Provide the application with a dataset containing customer information. This dataset should include essential details like customer demographics, account specifics, service usage, and historical churn information, which are necessary for making accurate predictions.

Process Dataset:

The app will read and prepare the uploaded dataset for analysis. This involves ensuring the data is correctly formatted and checking for any missing or inconsistent values.

Configure Parameters:

Adjust settings for the model, such as classification thresholds, the selection of data columns, or model hyperparameters. This customization ensures that the model

operates effectively with the specific dataset and meets business needs.

Prepare Data:

Clean and preprocess the data by addressing missing values, converting categorical variables into a suitable format, normalizing numerical data, and, if required, splitting the data into training and test sets. This preparation makes the data suitable for analysis.

Run Churn Predictions:

Use the imported model to evaluate which customers are likely to churn based on the provided data. The model will generate predictions, providing either probability scores or classifications that indicate each customer's risk level.

Update or Retrain Model:

If necessary, the app can train a new model or update the existing one using the provided dataset. This process involves fitting the model to the data, adjusting its parameters, and validating its performance on test data.

Save Model:

After training or updating the model, save it to ensure that the latest version is available for future use. This step prevents the need for retraining each time the model is utilized.

Display Results:

Show the prediction results to the user, which may include detailed reports, customer risk scores, and visual aids such as charts or graphs. These outputs help users interpret and utilize the model's findings effectively.

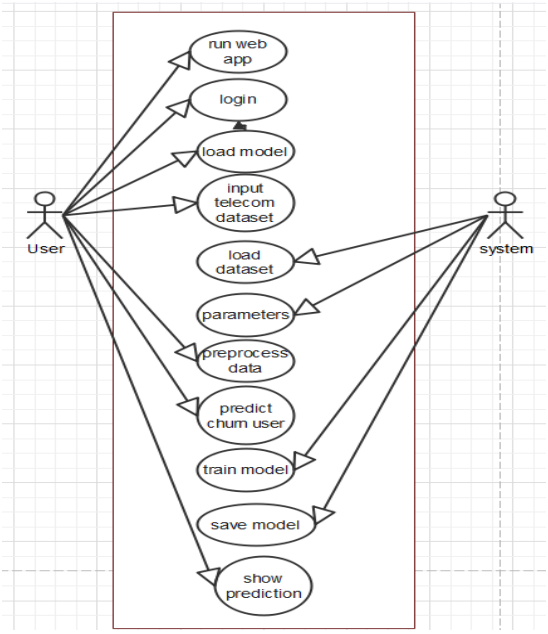

Fig 1. Flow Chart

C. *System Architecture:*

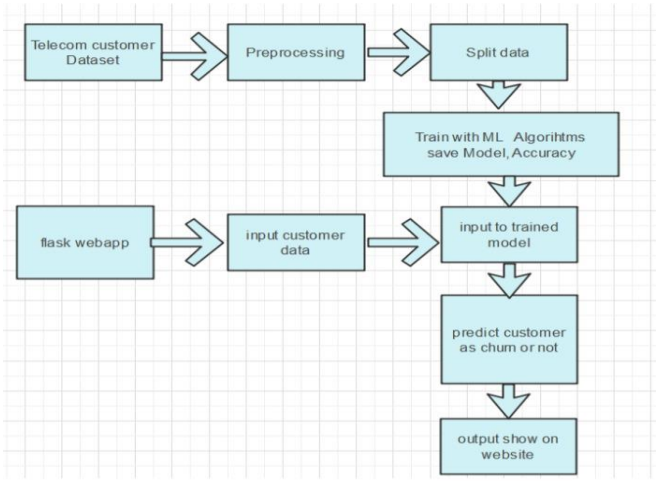

Fig 2. Proposed Architecture

To enhance telecom operations, a comprehensive approach was employed, emphasizing precise churn prediction and operational efficiency. The process starts with gathering an extensive dataset that includes a wide range of customer information, such as demographic details, usage patterns, and historical churn data.

The first step involves preparing the dataset through thorough data preprocessing. This process includes correcting errors, managing missing values, and normalizing the data to ensure consistency. Proper preprocessing is crucial for making the data suitable for analysis and modeling.

Following data preparation, the dataset is split into two subsets: one for training and one for testing. The training subset is used to build and refine predictive models, while the testing subset is used to evaluate the models' performance.

Several advanced machine learning algorithms are applied to develop the predictive models. These algorithms include logistic regression, known for its straightforward classification capabilities; extreme gradient boosting (XGBoost), renowned for its high accuracy and ability to handle complex patterns; gradient boosted machine (GBM) trees, which enhance prediction accuracy through boosting methods; random forests, which provide robustness through ensemble learning; and decision trees, valued for their interpretability.

Once the models are trained with the training data, they are tested using the testing data to assess their effectiveness and accuracy. The trained models are then integrated into a web-based system, which allows users to input new customer data and receive real-time churn predictions.

The final stage involves presenting the prediction results on a user-friendly web platform. This interface allows telecom companies to easily access and interpret the predictions,

facilitating data-driven decisions that improve customer retention and operational performance.

1) Specific Steps:

1. Data Collection:

For this project, a telecom customer dataset from Kaggle was utilized. This dataset includes various features related to customer connections and services, with churn labeled as binary values (0 or 1).

2. Preprocessing:

The dataset contains 30 features. The preprocessing phase involves evaluating and selecting the most impactful features. All 30 features are used for training the machine learning models. The data is split into training and testing sets, with 80% of the data used for training and the remaining 20% reserved for testing.

3. Train-Test Split and Model Fitting:

The dataset is divided into training and testing subsets to evaluate model performance on unseen data and to check how well the model generalizes. Model fitting is conducted as part of the model building process.

4. Model Evaluation and Predictions:

The final step involves evaluating model performance using accuracy scores. This includes creating a model instance, fitting the model with training data, and making predictions on the test data. These predictions are compared against actual test values using the accuracy score function to measure the model's effectiveness. This process is repeated across various classification algorithms to obtain their respective test accuracy scores.

This approach ensures that the churn prediction model is robust, reliable, and capable of providing actionable insights for enhancing telecom operations.

D. *Algorithms*

Random Forest

The Random Forest classification model consists of an ensemble of decision trees. It operates by aggregating the results from multiple decision trees to produce a final prediction. Unlike individual decision trees that evaluate all possible feature splits, Random Forest selects a subset of features for consideration in each tree. This approach enhances the model's accuracy by reducing overfitting and improving generalization. Random Forest predictions are derived from the collective outputs of the decision trees within the forest.

Logistic Regression (LR)

Logistic Regression is employed to estimate categorical outcomes based on a set of independent variables. It differs from linear regression, which is used to predict continuous dependent variables. In the context of machine learning, logistic regression is utilized to model the probability of a categorical event occurring. For instance, in studies involving machine data, logistic regression was applied alongside Random Forest (RF) and XGBoost to predict failures. The effectiveness of these models was assessed using various performance metrics.

XGBoost

XGBoost, an ensemble learning technique, leverages the strengths of multiple predictive models to achieve accurate results. It combines several weak learners—models that have high bias and only slightly better performance than random guessing—into a robust learner. Each weak learner contributes valuable information, and the aggregation of these learners reduces both bias and variance, leading to a more accurate final prediction.

Decision Tree (DT)

Decision Trees are widely used in various fields such as character recognition, medical diagnosis, and voice recognition due to their ability to simplify complex decision-making processes. The Decision Tree model works by recursively splitting the data space into smaller subspaces based on feature values. This approach decomposes a complex problem into a series of simpler, more interpretable decisions, making the model both effective and easy to understand.

IV. IMPLEMENTATION

A. Logical Design

Logical design focuses on how users interact with the system and how the system communicates information back to users. It encompasses three main aspects: User Interface Design, Data Design, and Process Design.

- **User Interface Design** addresses how users input data into the system and how the system presents information to them.
- **Data Design** involves the representation and storage of data within the system, ensuring that data is organized and accessible.
- **Process Design** deals with the flow of data through the system, including its validation, security, and transformation at various stages.

B. Physical Design

In this context, Physical Design does not pertain to the tangible layout of hardware components. Instead, it refers to the detailed design of system components, including the structure of user interfaces, product databases, and processing units.

For example, while the physical design of a personal computer includes hardware elements like the monitor, CPU, and hard drive, Physical Design in information systems focuses on creating detailed specifications for the system's user interfaces, data structures, and processing

logic. This phase involves designing how these elements will be implemented and integrated, including the development of hardware and software specifications to support the proposed system's requirements.

V. Result and Discussion

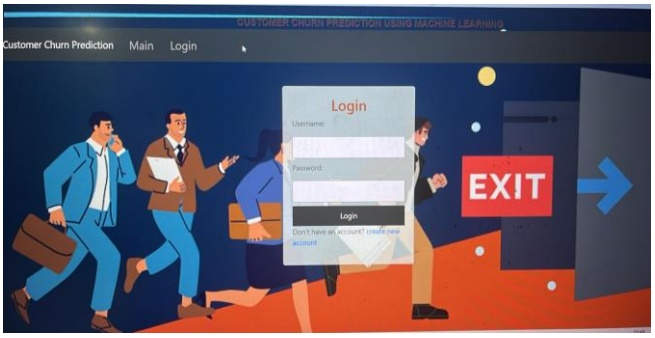

Fig 3 Login Page

The login page serves as the entry point for users to access the customer churn prediction system. It features fields for entering a username or email address and a password. To ensure proper authentication, the page includes validation mechanisms to verify that all required fields are filled out correctly and that the email address adheres to a valid format when applicable. This validation process helps prevent errors and ensures that only authorized users can access the application.

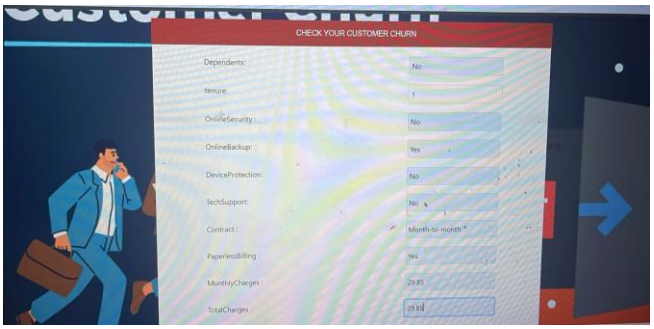

Fig 4 Customer churn checking Page

The customer churn checking page is designed for users to input customer data in order to evaluate the likelihood of churn. It includes fields for entering various customer attributes, such as demographic information, transaction history, and usage patterns. The page interfaces with a backend machine learning model that has been trained to predict churn based on the provided data. Upon submission, the model processes the inputs and generates a prediction indicating the likelihood of the customer leaving. The results are presented in an easily interpretable format, which may include a probability score, risk category (e.g., high, medium, low), or visual indicators like color-coded status.

Fig 5 Churn Prediction Page

The churn prediction result page provides users with detailed insights into the likelihood of customer churn based on the data they have submitted. Key features of this page include:

1. **Prediction Outcome**: Displays a clear result, such as "Likely to Churn" or "Not Likely to Churn," offering a straightforward overview of the customer's status.

2. **Probability Score**: Presents a numerical probability or confidence level, such as “Churn Probability: 92.85%,” which helps users gauge the certainty of the prediction.

This page is designed to facilitate quick understanding and interpretation of the churn predictions, aiding in data-driven decision-making for customer retention strategies.

VI. Conclusion

The primary objective of this study is to enhance profitability for telecom companies by improving the accuracy of customer churn predictions. Accurate forecasting of customer attrition is crucial, as it directly impacts a key revenue stream for telecom firms. This research focuses on developing a predictive system that achieves high performance, as indicated by elevated Area Under the Curve (AUC) values.

To build and validate the model, the dataset was divided into an 80/20 split, allocating 80% for training and 20% for testing. This approach was employed to optimize the model and evaluate its performance. The process involved hyperparameter tuning and validation to refine the predictive accuracy of the models. Additionally, feature engineering, which includes transforming and selecting relevant features, was conducted to prepare the data for machine learning algorithms.

An issue of data imbalance was noted, with only 5% of the dataset representing instances of customer churn. This imbalance was addressed through under-sampling techniques and the use of tree-based algorithms, which are less sensitive to such imbalances. Four tree-based algorithms were selected for their effectiveness in handling this type of prediction: Decision Trees, Random Forests, and Gradient Boosting Machines (GBM).

FUTURE SCOPE

The proposed churn prediction system establishes a robust foundation for future improvements that can greatly enhance telecom providers' capabilities in managing and reducing customer churn. A significant opportunity for advancement is the integration of real-time data streams, which would facilitate immediate analysis and prediction of churn. This capability would enable telecom companies to respond promptly to churn signals and implement effective measures as they arise.

Another promising area for development is the adoption of advanced algorithms. As machine learning and artificial intelligence technologies progress, incorporating cutting-edge techniques such as deep learning, reinforcement learning, and hybrid models could further refine prediction accuracy and better address complex customer behaviors.

Enhancing personalization is also crucial for future enhancements. By integrating customer segmentation and profiling, the system could deliver more tailored retention strategies based on individual preferences, usage patterns, and service needs. This would enable telecom providers to design interventions that are more relevant and effective for each customer.

Moreover, broadening the system to include data from multiple channels would enrich the dataset and offer deeper insights into customer behavior. By incorporating information from sources such as social media, customer service interactions, and IoT devices, the system could provide a more comprehensive view of customer patterns, ultimately improving the accuracy of churn predictions.

VII.Reference

[1] Gerpott, T.J., Rams, W., & Schindler, A. (2001). Analysis of customer retention, loyalty, and satisfaction within the German mobile cellular telecommunications market. *Telecommunications Policy*, 25, 249–269.

[2] Wei, C.P., & Chiu, I.T. (2002). Utilizing telecommunications call data for churn prediction through data mining techniques. *Expert Systems with Applications*, 23(2), 103–112.

[3] Qureshi, S.A., Rehman, A.S., Qamar, A.M., Kamal, A., & Rehman, A. (2013). A machine learning approach for predicting telecommunications subscribers' churn. In *Proceedings of the Eighth International Conference on Digital Information Management*, 131–136.

[4] Ascarza, E., Iyengar, R., & Schleicher, M. (2016). Risks associated with proactive churn prevention via plan recommendations: Insights from a field experiment. *Journal of Marketing Research*, 53(1), 46–60.

[5] Bott, R. (2014). Using multilayer perceptron neural networks for customer churn prediction in the telecommunications industry: A modeling and analysis approach. *IGARSS*, 11(1), 1–5.

[6] Umayaparvathi, V., & Iyakutti, K. (2016). A review of customer churn prediction in the telecommunications sector: Datasets, methodologies, and metrics. *International Research Journal of Engineering and Technology*, 3(4), 1065–1070.

[7] Yu, W., Jutla, D.N., & Sivakumar, S.C. (2005). A model for aligning churn strategies with management in mobile telecom. In *Proceedings of the Communication Networks and Services Research Conference* (Vol. 3, pp. 48–53).

[8] Burez, D., & den Poel, D.V. (2009). Addressing class imbalance in customer churn prediction. *Expert Systems with Applications*, 36(3), 4626–4636.

[9] Zhan, J., Guidibande, V., & Parsa, S.P.K. (2016). Identification of influential communities in large networks. *Journal of Big Data*, 3(1), 16. https://doi.org/10.1186/s40537-016-0050-7.

[10] Barthelemy, M. (2004). Analysis of betweenness centrality in complex networks. *European Physical Journal B*, 38(2), 163–168. https://doi.org/10.1140/epjb/e2004-00111-4.

[11] Elisabetta, E., Meyerhenke, H., & Staudt, C.L. (2014). Approximation techniques for betweenness centrality in evolving large networks. *CoRR*. arXiv:1409.6241.

[12] Brandusoiu, I., Toderean, G., & Ha, B. (2016). Approaches for predicting churn in the prepaid mobile telecommunications sector. In *Proceedings of the International Conference on Communications*, 97–100.