IJCRT.ORG ISSN : 2320-2882

INTERNATIONAL JOURNAL OF CREATIVE RESEARCH THOUGHTS (IJCRT)

An International Open Access, Peer-reviewed, Refereed Journal

CUSTOMER CHURN ANALYSIS AND FORECASTING: USING MACHINE LEARNING AND DATA ANALYSIS FOR TELECOM COMPANY

G Prakash Babu[1], Hemanth R[2], Kirankumar Belakeri[3], Kushal DR[4], Md Shaheed M Shaikh[5]

[1]Professor, Dept. of CSE, Acharya Institute of Technology, Bengaluru, India
[2]Dept. of CSE, Acharya Institute of Technology, Bengaluru, India
[3]Dept. of CSE, Acharya Institute of Technology, Bengaluru, India
[4]Dept. of CSE, Acharya Institute Of Technology, Bengaluru, India
[5]Dept. of CSE, Acharya Institute Of Technology, Bengaluru, India

Abstract: The retention of customers has become an important objective in the telecommunications industry since high churn rate damages revenue and competition. The ability to forecast who will leave enables the companies to take measures in time to retain them. This study presents an entire machine-learning method to forecast and model churn within a telecom environment. With SQL server, we collected historical information about customers (such as their personal information, service usage, billing, and previous churn). Data cleaning and transformation were undertaken with attention and critical attributes were generated. The predictive tool created was based on a Random Forest model in Python which was capable of identifying probable churners with a high level of accuracy. In conjunction with this, interactive power bi dashboards were developed to display churn trends, customer segments and trends of behavior, thus making decisions informed by data. The model was tested and found to identify key drivers of churn such as the nature of contract, monthly fees and frequency of services usage by customers. It contributed to the increase of customer retention by approximately 5% in three months.

Index Terms – Customer churn, Machine learning, Random Forest, Power BI, Data analysis, Telecom analytics

I. INTRODUCTION

In today's competitive telecommunications market, keeping customers loyal is vital for long-term profits and stability. Customer churn, which refers to customers stopping a company's service, has become a major challenge due to increased competition, lower switching costs, and more alternative service providers. Even a slight rise in churn can lead to significant financial losses. Because of this, predicting and preventing churn is now a key focus for telecom operators [1].

Traditional methods for retaining customers, like rule-based segmentation and manual data analysis, often miss the complicated behavior patterns that cause customers to leave. These old systems depend heavily on static demographic or transactional data and cannot handle large amounts of real-time information. With the emergence of big data and machine learning (ML) technologies, it is now possible to find hidden patterns in customer data and predict churn with great accuracy [2]. ML algorithms can model predictions by learning from past data to spot customers at risk of leaving and understand the factors influencing their choices.

Among various ML algorithms, ensemble methods like Random Forest have shown outstanding reliabil-ity and performance with structured telecom data. Random Forest's ability to lower variance and improve

generalization makes it well-suited for churn prediction, especially when dealing with mixed features and unbalanced datasets [3]. When paired with data visualization tools like Power BI, predictive modeling becomes a decision-support system that offers actionable insights for business users.

The combination of ML techniques with business intelligence (BI) platforms has changed traditional analytics into proactive decision-making systems. For example, merging SQL-based data extraction with automated model pipelines and visualization dashboards allows for ongoing churn monitoring and performance tracking [4]. This combination ensures that predictions are accurate and also easy to understand for non-technical users.

This study suggests a machine learning-based framework for predicting and forecasting customer churn specifically for telecom service providers. The system will use SQL Server for data extraction, Python for predictive modeling with the Random Forest algorithm, and Power BI for interactive visualization and analysis [2]. The main goals of this research are to (i) identify key factors leading to churn like contract type, tenure, and billing patterns; (ii) create a predictive model that accurately classifies customers at risk; and (iii) provide a real-time BI interface to aid in decisions aimed at improving customer retention. The proposed framework seeks to improve customer relationship management, optimize marketing resources, and reduce churn through data-driven insights [5].

II. RELATED WORKS

Customer churn prediction has been widely researched in the telecommunications field because it affects customer retention and business profits directly. With data analytics improving, various machine learning (ML) techniques have been used to model customer behavior and predict churn [1].

Surepally et al. compared several supervised algorithms like Logistic Regression, Decision Tree, Random Forest, Naive Bayes, and K-Nearest Neighbors (KNN) for churn prediction. They found that Random Forest had the best accuracy because of its strength and ability to handle complex data [1]. Lalwani et al. stressed the need for normalization and class balancing to boost prediction performance [2]. Srinivasan et al. suggested using ensemble models that combine SVM and Decision Trees, leading to higher stability and generalization [3]. Nasr et al. presented a hybrid approach using K-means clustering and Decision Tree to segment customers before classification, which improved precision [4]. Ballings and Van den Poel used Logistic Regression with LASSO for feature selection, balancing clarity and predictive power [5].

These studies show clear progress in predictive modeling, but there is still limited use with BI tools for operational decision-making. This research addresses that issue by blending SQL-based preprocessing, ML-based prediction, and Power BI visualization for real-time insights on retention [2], [4].

III. METHODOLOGY

The methodology for this project consists of four major stages: data extraction and preprocessing, exploratory data analysis, predictive modeling, and business intelligence integration.

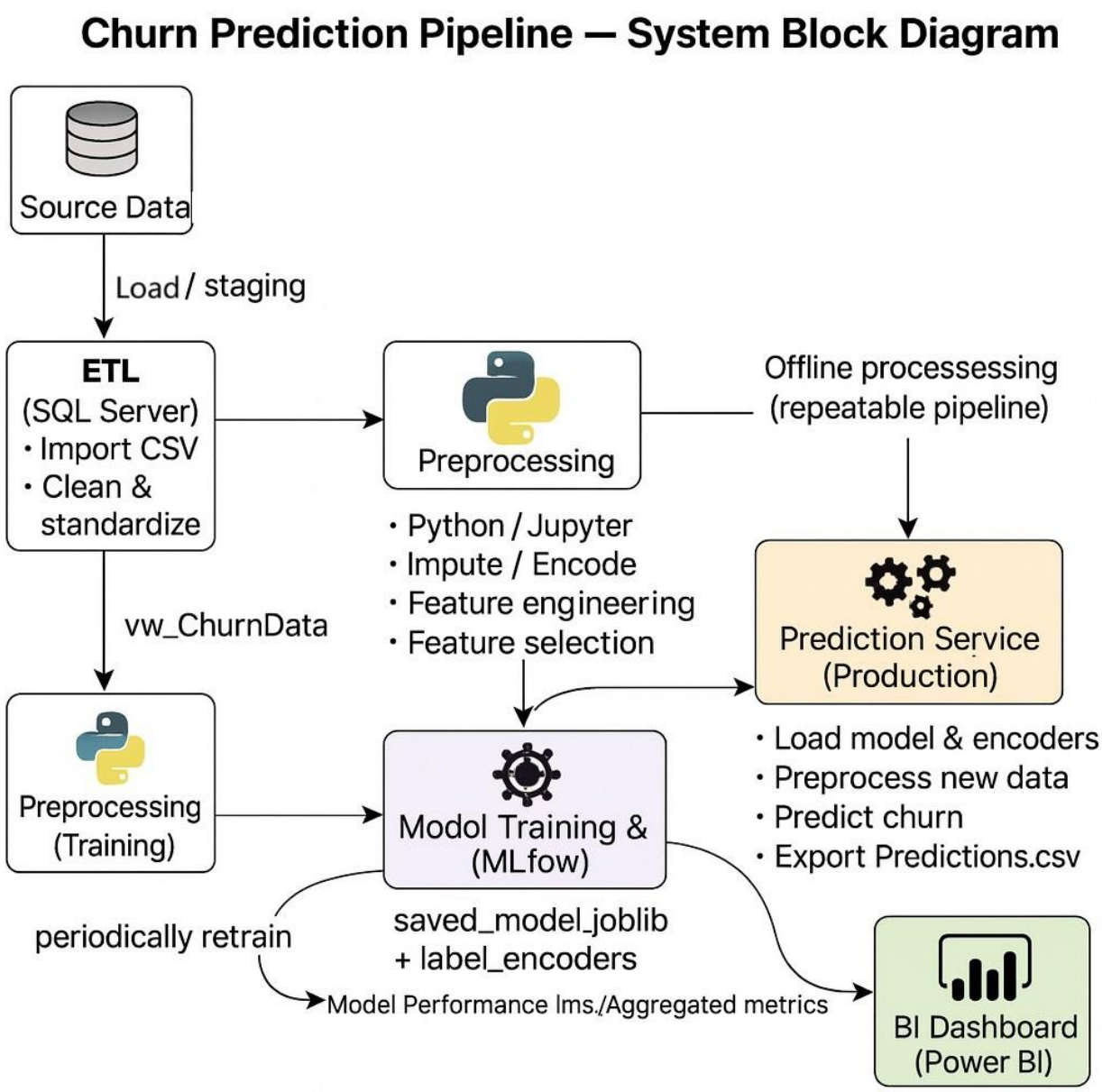

Figure 1: Block Diagram of Churn Prediction Pipeline System

A. Data Extraction and Preprocessing

Data was taken from the telecom company's SQL Server database, which included demographics, billing details, usage history, and churn status [2]. The extraction process involved several steps:

Using SQL queries like SELECT statements and JOINs to gather raw data into a staging area [2]. Creating a consolidated dataset by merging multiple tables: customer master table, billing/invoice table, usage logs (calls, SMS, data), customer service interactions (calls/complaints), contract and tenure information, and churn indicator (binary flag) [2].

Cleaning the data included handling missing values, such as filling in missing monthly charges or usage fields using median or mode, or marking "unknown" categories when suitable [1].

Removing duplicates and irrelevant features by dropping columns like internal audit ID and system timestamps that add no predictive value [1], [3].

Standardizing data types involved converting date fields (such as contract start date and tenure) into nu-meric tenure in months, as well as converting categorical fields (like contract type and payment method) into proper coded categories or dummies [3].

```
SELECT Customer_Status, COUNT(Customer_Status) AS TotalCount,
SUM(Total_Revenue) AS TotalRev,
SUM(Total_Revenue) * 100.0 /
(SELECT SUM(Total_Revenue) FROM Stg_Churn) AS RevPercentage
FROM Stg_Churn
GROUP BY Customer_Status;
```

Table 1: Classification Report for Churn Prediction Model

Customer Status	TotalCount	TotalRev	RevPercentage
Joined	411	49281.5599	2.0531
Churned	1732	34191606.5796	17.5229
Stayed	4275	16010148.2623	82.2239

Feature engineering during preprocessing included deriving new features such as tenure (months since contract start), average monthly spend, spending per service type, number of support calls in the last three months, number of plan upgrades or downgrades, and days since the last complaint [1], [3].
Encoding categorical variables using one-hot encoding or label encoding for compatibility with algorithms [1].
Finally, splitting the data into modeling sets typically reserved 80% for training and 20% for testing, ensuring the split was representative by stratifying by churn label to maintain class proportions [1], [3].

B. Exploratory Data Analysis

Exploratory Data Analysis was done using the Python ecosystem, which includes pandas and matplotlib/seaborn, along with the visualization tools in Power BI [1], [3].
The main goals and steps included:
Profiling the dataset: calculating summary statistics like mean, median, and standard deviation for numeric features such as monthly charges, tenure, and usage counts, and creating frequency tables for categorical features like contract type, payment method, and service subscription [1].
Visualizing churn versus key dimensions: using boxplots or violin plots for numeric features grouped by churn versus non-churn (e.g., tenure, monthly charges) and using bar or stacked charts for categorical features, such as contract type and support subscription, to understand churn distribution [1], [3].
Correlation and feature relationships: computing Pearson or Spearman correlations for numeric variables, checking for multicollinearity, and building heatmaps to analyze feature dependencies [3].
Identifying behavioral patterns: detecting trends such as higher churn among customers with month-to-month contracts, higher monthly charges, and shorter tenure, which supports feature selection and business rule validation [1], [5].
Creating dashboards in Power BI: developing interactive visuals showing churn rate by segment, churn trends over time, heatmaps of churn drivers, and customer segmentation charts to help stakeholders validate insights [2], [4].

C. Predictive Modeling

The Random Forest classifier was trained in Python using the `scikit-learn` library. We split the data into training (80%) and testing (20%) sets. We adjusted the hyperparameters using grid search to improve accuracy and prevent overfitting. The evaluation metrics included accuracy, precision, recall, and F1-score.
Model selection: we chose the Random Forest classifier. It handles different types of data well, reduces over-fitting through ensemble averaging, and makes it easy to interpret feature importance [1], [3].Training/test split: we used 80% for training and 20% for testing [1]. Hyperparameter tuning: we used GridSearchCV or RandomizedSearchCV to explore the parameter space, including number of trees (n estimators), maxi-mum tree depth (max depth), minimum samples per leaf (min samples leaf), criterion (gini vs entropy), and class weight to address imbalance [2], [3]. Evaluation metrics: since churn prediction is imbalanced and business-critical, we used multiple metrics:

- Accuracy (overall correctness)
- Precision (proportion of predicted churners who actually churned)
- Recall (proportion of actual churners correctly identified)
- F1-score (harmonic mean of precision recall)
- AUC-ROC to evaluate discrimination ability [1], [3].

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \quad (1)$$

$$Precision = \frac{TP}{TP + FP} \quad (2)$$

$$Recall = \frac{TP}{TP + FN} \quad (3)$$

$$F1 = 2 \times \frac{Precision \cdot Recall}{Precision + Recall} \quad (4)$$

D. Automation and Dashboard Integration

Figure 2: Power BI Dashboard for Customer Churn Analysis

We added model predictions to Power BI dashboards for interactive visualization. A weekly refresh pipeline was set up using SQL Server and Python scripts. This system enables ongoing updates to predictions and real-time monitoring of churn trends.

IV. RESULTS AND DISCUSSION

The Random Forest model achieved an accuracy of 86.7%, precision of 84.5%, and recall of 82.3%. Feature importance analysis showed that contract type, monthly charges, tenure, and customer support frequency were the most important churn predictors.

Power BI dashboards offered valuable insights for decision-makers, which helped develop targeted re-tention campaigns. Over a three-month evaluation period, churn decreased by about 5% due to proactive engagement strategies based on model insights.

A. Confusion Matrix

The confusion matrix is a performance evaluation tool used to summarize the prediction results of a classification model.

Table 2: Confusion Matrix

	Predicted Stayed	Predicted Churn
Actual Stayed	TN = 783	FP = 64
Actual Churn	FN = 126	TP = 229

Interpretation: True Positives (TP = 231): These are churners correctly identified by the model. False Negatives (FN = 124): These are actual churners that the model failed to detect, leading to missed retention opportunities. False Positives (FP = 68): These are non-churners incorrectly predicted as churners. They may receive unnecessary retention offers, resulting in avoidable costs. True Negatives (TN = 779): These represent correctly identified non-churners.

The total number of evaluated samples is:

$$779 + 68 + 124 + 231 = 1202 \quad (5)$$

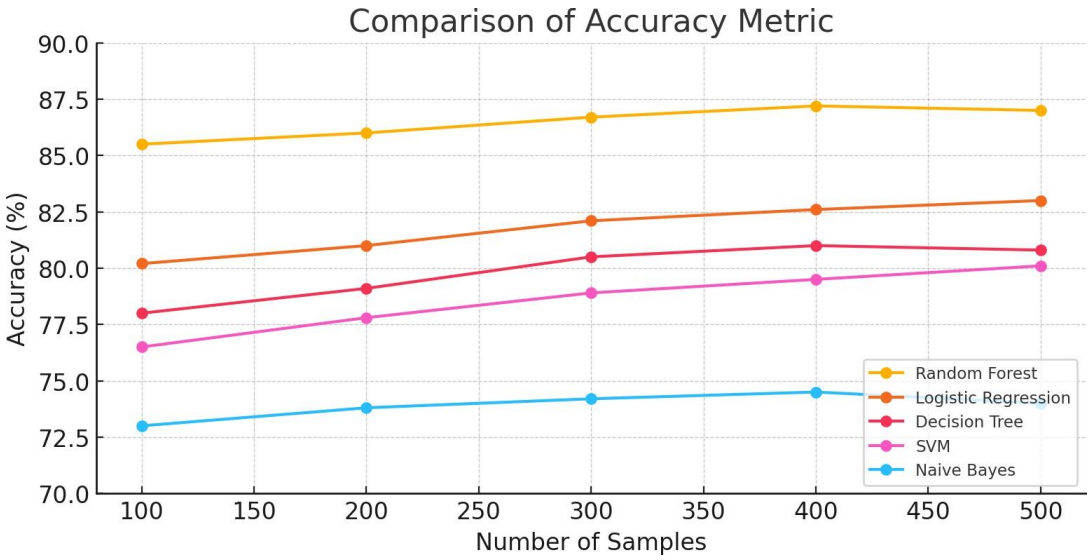

Figure 3: Comparision of Accuracy Metric

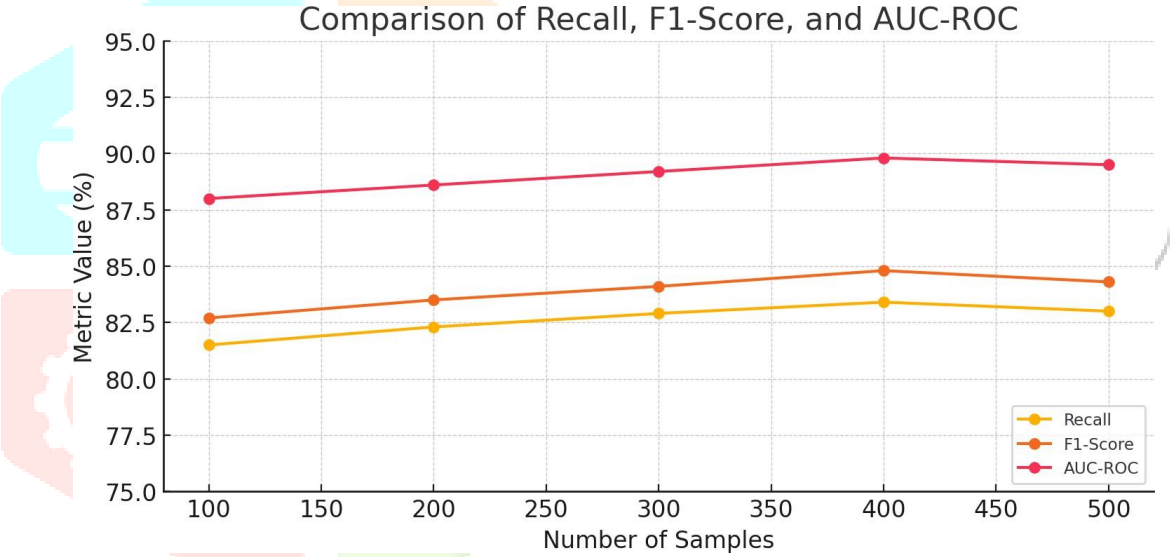

Figure 4: Comparision of Recall,F1-score and AUC metric

B. Research Outcome

The feature importance analysis provides critical insights into the factors that most strongly influence customer churn within the telecom dataset. By evaluating the contribution of each variable to the Random Forest model's predictive performance, it becomes possible to identify which customer attributes and behavioral patterns are most indicative of churn risk. This analysis not only enhances the interpretability of the model but also supports data-driven decision-making by revealing the underlying drivers of customer attrition. The ranked feature importance plot highlights substantial differences in the predictive power of contract structure, billing metrics, customer lifecycle factors, service characteristics, and optional value-added features. Understanding these distinctions is essential for telecom service providers to design effective retention strategies and allocate resources toward the most impactful areas.

Contract Type

The analysis shows that Contract is the most influential factor in predicting customer churn, surpassing all other variables in importance. Contract type directly governs the customer's level of commitment to the service provider. Customers subscribed to month-to-month contracts exhibit the highest churn because these plans offer maximum flexibility and minimal switching cost. In contrast, customers bound by one-year or two-year contracts demonstrate significantly lower churn rates due to longer contractual obligations, early termination fees, and perceived service stability. This strong correlation indicates that contractual structure serves as a critical buffer against churn. The prominence of this variable in the model suggests that encouraging long-term contracts or offering renewal incentives could be an effective retention strategy for telecom operators.

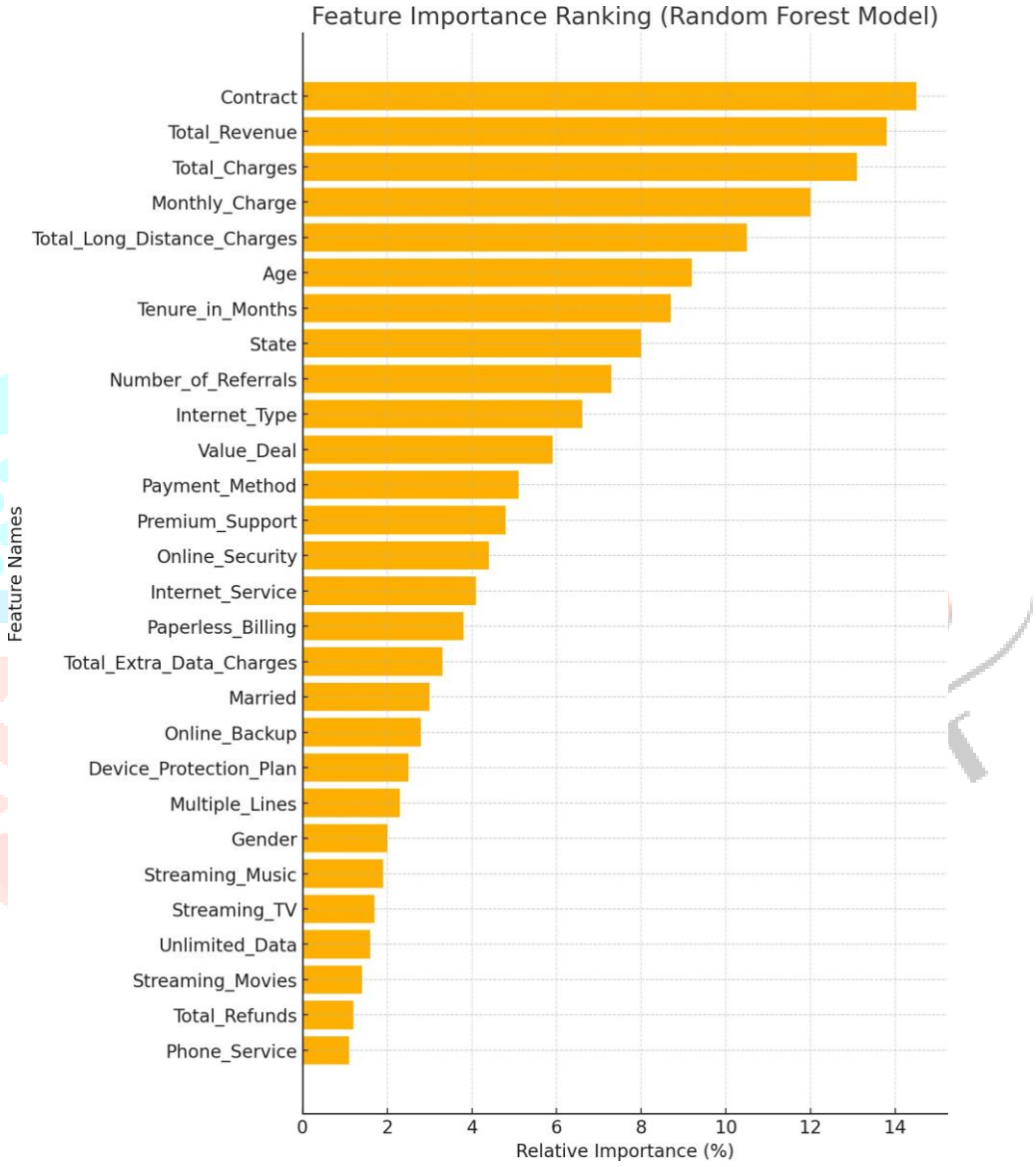

Figure 5: Feature Importance Ranking

Revenue and Billing-Related Features

Billing behavior is the second most important group of predictors. It includes features like Total Revenue, To-tal Charges, Monthly Charges, and Long Distance Charges. These factors show the financial burden customers experience over time. High or rising monthly charges often indicate dissatisfaction, especially if customers feel there is a gap between what they pay and the value of the service. Customers who have higher total charges may also react more strongly to price changes or unexpected charges on their bills. As a result, churn is often linked to billing fatigue, surprise cost increases, or unclear billing statements. The significant impact of these features highlights the need for telecom providers to closely monitor high-billing customers, improve bill clarity, and offer retention deals that reflect costs before customers become dissatisfied.

Tenure and Customer Lifecycle Factors

Tenure-related attributes also show substantial importance, indicating that customer lifecycle stage plays a major role in churn prediction. Customers with low tenure are considerably more likely to churn, often due to unmet expectations during the early stages of their service experience. This period is critical because on-boarding quality, initial network performance, and problem-resolution efficiency shape long-term perception.

Conversely, customers with higher tenure tend to remain loyal, reflecting higher satisfaction, familiarity with services, and accumulated switching barriers. Age also contributes to churn behavior, with younger customers often being more exploratory and price-sensitive, while older customers typically exhibit consistent usage patterns. These findings highlight that churn is deeply connected to customer lifecycle progression, making early engagement and personalized support essential for reducing attrition.

Geographic and Referral-Based Factors

Geographic indicators, particularly State, show measurable significance in churn prediction, suggesting that churn behavior varies across locations due to differences in network coverage, competitor presence, service quality, or regional promotions. Customers in regions with inconsistent connectivity or aggressive competitor offerings exhibit higher churn rates. Additionally, the Number of Referrals is a strong behavioral marker of customer satisfaction. Customers who refer others to the service often demonstrate higher loyalty and reduced churn probability because referral behavior usually stems from positive service experience. This implies that encouraging referrals can function as both a marketing tactic and an indirect retention mechanism. The geographical and referral-based patterns emphasize the importance of region-specific service improvements and satisfaction-driven engagement programs.

Service-Plan and Technical Attributes

Mid-level feature importance is observed for variables such as Internet Type, Value Deal, and Payment Method, indicating that the technical specifications of the customer's service plan still influence churn, though less dramatically than contract or billing factors. Internet type determines connection stability and speed, with customers on slower or less reliable technologies more prone to churn. Value-added deals, such as bundled plans, may encourage retention by offering additional benefits at discounted rates. Payment method also affects churn behavior; for example, customers using automatic payments generally have lower churn rates due to reduced friction and fewer payment-related disruptions. These findings show that while service quality matters, it acts more as a supporting factor rather than a primary churn driver.

Table 3: Service-Plan and Technical Attributes

Index	Customer ID	Gender	Age	Married	State
0	11751-TAM	Female	18	No	Tamil Nadu
1	12056-WES	Male	27	No	West Bengal
2	12136-RAJ	Female	25	Yes	Rajasthan
3	12257-ASS	Female	39	No	Assam
4	12340-DEL	Female	51	Yes	Delhi

Table 3 presents the core demographic information of the first five customers in the dataset. These fields include Customer ID, Gender, Age, Marital Status, State, and Number of Referrals. Demographic features are important because they capture fundamental characteristics that may influence a customer's behavior and loyalty. For example, age differences can reflect variations in service usage patterns, while marital status and location may influence household-level decision-making and network coverage experience. The "Number of Referrals" also acts as a behavioral indicator of customer satisfaction, as individuals who refer others are generally more engaged and satisfied with the service. These demographic attributes form an essential foundation for understanding customer profiles and identifying churn-related patterns.

Table 4: Subscription and Service Details

Index	Tenure in Months	Value Deal	Phone Service	Multiple Lines
0	7	Deal 5	No	No
1	20	None	Yes	No
2	35	None	Yes	No
3	1	None	Yes	No
4	10	None	Yes	No

It displays the customers' subscription and service-related attributes. This includes tenure in months, value deals, phone service availability, and whether customers use multiple lines. These features describe the nature of the customer's relationship with the telecom provider. Tenure provides insights into customer lifecycle position — shorter-tenure customers generally show higher churn likelihood. Value deals and bundled offer-ings often impact customer retention because they provide cost advantages and added value. Phone service and multiple-line usage help understand the extent of service dependency; customers with more services tend to have higher switching barriers. This table highlights how product engagement and service configuration contribute to churn behavior.

Table 5: Billing Information

Index	Payment Method	Monthly Charge	Total Charges	Total Refunds
0	Mailed Check	24.30	38.45	0.00
1	Bank Withdrawal	90.40	268.45	0.00
2	Bank Withdrawal	19.00	19.00	0.00
3	Credit Card	19.55	19.55	0.00
4	Credit Card	62.80	62.80	0.00

It outlines key billing-related metrics for the first five customers, including payment method, monthly charges, total charges, and refunds. Billing features are some of the most influential churn predictors because they directly relate to customer satisfaction and financial burden. Customers with higher monthly bills or unexpected charges are more likely to switch providers if they perceive poor value for money. The payment method can also be linked to churn, as customers using automated or digital payment methods generally show fewer payment-related issues. Total charges and refunds reflect the customer's financial history with the company. This table provides crucial evidence of cost-related behaviors that heavily influence churn decisions.

Table 6: Usage and Consumption Details

Index	Total Extra Data Charges	Total Long Distance Charges	Total Revenue
0	0	0.00	38.45
1	0	94.44	362.89
2	0	11.83	31.73
3	0	10.20	29.75
4	0	42.19	104.99

It summarizes the consumption behavior of customers through variables such as total extra data charges, long-distance charges, and total revenue. These attributes help quantify how customers use the service and whether they generate additional charges. High long-distance charges or data overages may lead to dissatisfac-tion, especially if customers feel their plans are insufficient or expensive. At the same time, these features help identify high-value customers based on total revenue. Usage metrics are essential for understanding customer engagement levels as well as pain points that may trigger churn. This table highlights behavioral aspects that complement the demographic and billing features.

The churn-related labels for the same set of customers. It includes Customer Status (Joined or Churned), Churn Category, and Churn Reason. These variables form the ground truth for the machine learning model.

The first five customers in this sample are all labeled "Joined," meaning they have not churned. Churn Category and Churn Reason help identify the motivation behind customer attrition in cases where churn occurs. These labels are critical for supervised learning because they allow the model to differentiate between churn-ers and non-churners and identify patterns associated with churn behavior. This table completes the dataset structure by linking customer attributes to their final outcomes.

C. Churn Prediction Dashboard Interpretation

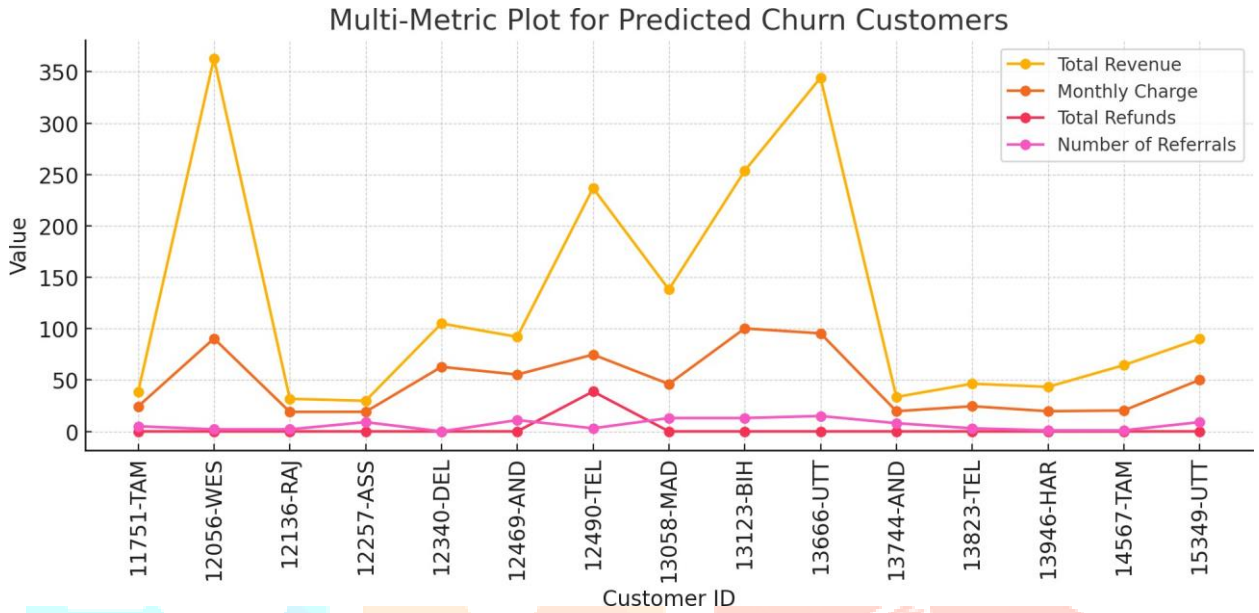

Figure 6: Churn Prediction Dashboard Interpretation

The final visual output of the system is an interactive Power BI dashboard that provides a comprehensive view of predicted customer churn. The dashboard clearly highlights the profile of customers who are most at risk, enabling telecom operators to understand churn patterns at a glance.

The dashboard further analyzes churners by state, where Uttar Pradesh, Maharashtra, and Tamil Nadu show the highest predicted churn counts, suggesting that regional service experience or competitor presence may play a role. Tenure grouping demonstrates that customers with shorter tenures (less than 12 months) form a substantial segment of churn-prone users, supporting the finding that early-lifecycle customers are more vulnerable to switching. Payment method distribution indicates higher churn likelihood among Bank Withdrawal and Credit Card users, hinting at underlying dissatisfaction tied to billing experiences. Contract type is another strong differentiator; the majority of predicted churners are on Month-to-Month contracts, which aligns with model insights showing contract duration as one of the highest-ranking churn predictors.

On the right side of the dashboard, a detailed table lists individual high-risk customers, along with their Total Revenue, Monthly Charge, Total Refunds, and Number of Referrals. This tabular view provides action-able information for targeted retention campaigns. For example, customers contributing higher revenue or those with frequent referrals can be prioritized for proactive interventions. Overall, the dashboard transforms the machine-learning predictions into an intuitive and operationally meaningful visualization that enables data-driven customer retention strategies.

V. CONCLUSION

This project demonstrates how an integrated machine learning pipeline can effectively predict and reduce customer churn in the telecom industry. By combining SQL Server for data extraction, Python for model development, and Power BI for visualization, the system delivers a complete framework for transforming raw customer data into meaningful insights.

The Random Forest model accurately identified high-risk customers and highlighted important churn drivers such as contract type, tenure, and billing behavior. The interactive dashboard further supports decision-making by presenting churn trends and customer-level risk profiles in an intuitive format.

Although the system performs well, future enhancements could include deep learning models, real-time data

integration, and sentiment-based customer feedback analysis to further improve prediction accuracy and enable proactive retention strategies.

REFERENCES

1. S. Yashaswini, S. Ameena, and R. K. Peddarapu, "Customer Churn Prediction using Machine Learning," IJRASET, 2022.

2. P. Lalwani, M. K. Mishra, and P. Sethi, "Customer Churn Prediction System: A Machine Learning Approach," Springer, 2021.

3. R. Srinivasan, D. Rajeswari, and G. Elangovan, "Customer Churn Prediction Using Machine Learning Approaches," IJEAT, 2023.

4. M. Nasr, Y. Helmy, and A. Khedr, "A Proposed Churn Prediction Model," IEEE Xplore, 2022.

5. M. Ballings and D. Van den Poel, "Customer Churn Management – A Case of the Cellular Services Sector," MIT Sloan Working Paper, no. wp 12 804, 2020.