Customer Analytics for Churn Prediction

Rohan Singh Sardar (G01453457), Manohar Babu Katika(G01472869), Rehan Yadali(G01464509), Rohit Chityala (G01441427), Rohit Reddy Kadaru(G01419418), Sagarika Komatireddy(G01467225), Dhanya Sri Vasantha(G01461366), Saaketh Hota(G01461941)

George Mason University

AIT-582-002: Applications of Metadata in Complex Big Data Problems

Prof. Lam Phung

***Project Overview*—This project analyzes customer churn for a business by identifying key factors contributing to customers discontinuing their services. The goal is to develop a predictive model that can accurately classify customers as likely to churn or remain, enabling businesses to take proactive steps to improve retention strategies.**

I. Data Description

[1]The dataset used in this project consists of 7,043 customer records, each containing a combination of numerical and categorical features, which are crucial for understanding and predicting customer churn in the telecommunications industry. Key demographic features include gender, senior citizen status (indicating whether the customer is a senior), and whether the customer has a partner or dependents, providing insights into personal factors that may influence churn behavior. The account details include tenure (the number of months a customer has been with the company), contract type (Month-to-Month, One-Year, or Two-Year), and payment method (such as electronic checks, bank transfers, or credit cards), all of which are important for analyzing customer loyalty and payment patterns. Service-related features describe whether the customer subscribes to phone service, has multiple lines, and the type of internet service they use (DSL, Fiber Optic, or None), along with additional services such as streaming TV and movies. Billing information includes monthly charges and total charges, which help assess the financial aspects of a customer's relationship with the company and their likelihood of leaving. The target variable, Churn, categorizes customers as either 'Yes' (indicating the customer has left) or 'No' (indicating the customer has stayed), and it serves as the primary outcome variable for predicting churn. The dataset, by combining these demographic, account, service usage, and billing features, provides a comprehensive basis for building a predictive model to understand and mitigate customer churn in the telecom sector.

II. Problem Description

The central problem this project addresses is predicting customer churn in a subscription-based service, specifically within the telecommunications industry, where churn is a significant concern. Customer churn refers to the loss of customers who discontinue their services, and it poses a direct threat to businesses by impacting profitability, customer lifetime value, and overall growth. In the highly competitive telecommunications market, where customers have a wide array of alternatives and minimal switching barriers, churn becomes an even more pressing issue. The ability to predict churn accurately enables telecom companies to implement proactive and targeted retention strategies, which can mitigate the negative effects of losing valuable customers. Additionally, reducing churn enhances profitability by retaining existing customers, who are typically more cost-effective to retain than to acquire new ones.

Churn can be influenced by a wide range of factors, including service quality, customer support, pricing, contract type, and personal demographics. For instance, customers who experience poor customer service or service interruptions may be more likely to leave, while others may churn due to high pricing or dissatisfaction with the available plans. Personal factors such as age, whether a customer has dependents, or the length of their tenure with the company can also play a role in their decision to stay or leave. Despite the potential factors influencing churn, many businesses struggle to identify which ones are the most impactful, making it difficult for them to focus their efforts on the right areas for improvement. Moreover, businesses face the challenge of working with vast amounts of customer data, which can be overwhelming and difficult to interpret without advanced analytical tools. This is where a data-driven approach becomes essential. By utilizing machine learning models and advanced analytics, companies can uncover hidden patterns and trends in the data that reveal the primary drivers of churn.

The goal of this project is not only to predict whether a customer will churn but also to uncover and understand the key drivers behind this behavior. By identifying the factors that most strongly contribute to churn, businesses can tailor their strategies more effectively. For example, the model might reveal that older customers with long-term contracts and higher service usage are more likely to churn. Armed with this knowledge, telecom companies can develop personalized retention strategies such as offering discounts, loyalty rewards, or enhanced customer service to these specific customer segments. Similarly, identifying customers who are at risk of leaving due to high monthly charges or poor service quality allows businesses to address these issues before the customer decides to leave.

In addition to identifying customers likely to churn, churn prediction models can be used to segment customers based on

their likelihood of leaving. This segmentation allows businesses to prioritize their retention efforts by focusing resources on the most at-risk customers. High-risk customers could be targeted with special promotions, personalized offers, or dedicated support to improve their experience and reduce the likelihood of churn. On the other hand, customers who are less likely to churn can be rewarded for their loyalty through benefits such as exclusive offers, early access to new services, or discounts, thus fostering customer satisfaction and long-term loyalty. By understanding these customer segments, telecom companies can make smarter, data-driven decisions regarding resource allocation, marketing campaigns, and service improvements.

Moreover, churn prediction models can be used to continuously monitor and track churn patterns over time, providing real-time insights into customer behavior. This dynamic approach enables businesses to adapt to changing market conditions, customer preferences, and service trends. With continuous optimization, retention strategies can be refined and updated to ensure maximum effectiveness. By leveraging these insights, businesses can optimize marketing campaigns, adjust pricing strategies, and improve service offerings to align with customer needs and preferences, ultimately improving customer satisfaction and retention.

The ultimate objective of this project is to build a predictive model that can accurately classify telecom customers as 'likely to churn' or 'not likely to churn' based on various input features. These features may include customer demographics, service usage patterns, billing details, and contract types, among others. The model will provide actionable insights that telecom companies can use to enhance their retention strategies and minimize customer turnover. In doing so, it directly addresses the broader business challenge of reducing churn and increasing profitability in the telecommunications industry. By focusing on retaining customers who are at risk of leaving, telecom companies can reduce the associated revenue loss, stabilize their customer base, and improve customer satisfaction, contributing to long-term growth and a competitive advantage in the market.

III. Project Motivation

In the telecommunications industry, customer churn is a critical business challenge that has far-reaching financial and operational implications. With high levels of competition and minimal switching costs, customers have a plethora of options, making it essential for companies to maintain a strong focus on customer retention. The loss of a customer directly affects revenue, customer lifetime value, and growth potential, with studies suggesting that acquiring a new customer is five to ten times more costly than retaining an existing one. Given the substantial investment required to attract new customers, it becomes imperative for telecom companies to understand and mitigate the factors driving existing customers to leave.

The rapid expansion of data analytics and machine learning offers an unprecedented opportunity for telecom companies to transform raw data into actionable insights. However, the sheer volume of data encompassing demographics, service usage, customer feedback, billing details, and interaction history that presents a significant challenge. Identifying high-risk customers and pinpointing specific drivers of churn in this data-rich environment requires advanced analytical tools and a data-driven approach. Without these insights, companies struggle to deploy effective retention strategies or to focus their efforts on the most impactful areas, such as pricing adjustments, service improvements, or enhanced customer support.

This project is motivated by the potential to leverage machine learning for accurate churn prediction, enabling companies to not only identify customers likely to leave but also to understand the underlying reasons for this behavior. By analyzing customer characteristics and engagement patterns, this project aims to develop a predictive model that segments customers based on their likelihood to churn, allowing businesses to prioritize retention efforts more strategically. For instance, if the model identifies that customers with short-term contracts and high monthly charges are at a higher risk of churn, telecom firms could implement targeted retention offers, personalized engagement strategies, or service upgrades to retain these at-risk customers.

Furthermore, predictive insights allow companies to tailor customer experience based on segmentation, with a focus on both at-risk and loyal customers. Customers who are less likely to churn could be rewarded for their loyalty through discounts, exclusive offers, or enhanced services, fostering long-term satisfaction and building a stronger relationship with the brand. On the other hand, high-risk customers can be addressed proactively through targeted interventions, which may include personalized offers, tailored communication, or dedicated support that directly addresses their concerns. This proactive approach not only helps in retaining valuable customers but also prevents revenue loss and contributes to sustainable growth in a saturated market.

Ultimately, this project seeks to empower telecom companies with a robust, data-driven churn prediction model that informs and enhances retention strategies, optimizes resource allocation, and improves overall customer satisfaction. By effectively predicting and mitigating churn, telecom firms can reduce attrition, drive customer loyalty, and secure a competitive advantage, ensuring stability and profitability in an increasingly dynamic market environment.

IV. Proposed Approach

To address the challenge of predicting customer churn in the telecommunications industry, this project will employ a comprehensive, data-driven approach that leverages advanced machine learning models to accurately classify customers based on their likelihood to churn. In an industry where retaining customers is crucial to maintaining profitability and growth, an effective churn prediction model can provide significant value by enabling telecom companies to identify at-risk customers and implement targeted retention strategies.

The proposed approach is structured into several key stages, each meticulously designed to ensure that the data is properly prepared, analyzed, and modeled to deliver actionable insights that can inform and optimize customer retention strategies.

The first phase focuses on data collection and preprocessing, which is essential for ensuring that the dataset is clean, complete, and ready for modeling. This includes handling missing data, removing outliers, and addressing any inconsistencies that could skew the results. Categorical variables, such as contract type, payment methods, and internet service types, will be encoded appropriately to make them compatible with machine learning algorithms, while continuous features like monthly charges and tenure will be scaled to ensure consistent performance across all algorithms. The preprocessing phase ensures that the data is standardized, removing noise and bias that could otherwise affect the accuracy of churn predictions. Once the data is cleaned, the next stage involves exploratory data analysis (EDA), where key trends and patterns within the data are identified.

Following the EDA, the project will move to feature engineering and model selection. Feature engineering involves creating new variables or modifying existing ones to improve model accuracy. For example, customer tenure might be categorized into groups to assess whether customers with long-term commitments are less likely to churn than those with shorter subscriptions. These engineered features will provide more granular insights and allow the model to better understand customer behavior. Once the features are refined, various machine learning algorithms will be evaluated, including decision trees, random forests, logistic regression, and gradient boosting machines. These models will be tested to determine which one best captures the relationships between customer attributes and churn likelihood.

After selecting the most suitable model, the next phase will focus on model training and evaluation. In this stage, the model will be trained using historical customer data, and its performance will be assessed using multiple evaluation metrics such as accuracy, precision, recall, F1-score, and ROC-AUC. These metrics will provide a comprehensive understanding of the model’s ability to correctly identify churned versus non-churned customers. The model will be tested on a separate validation dataset to assess its generalizability, ensuring it can perform well on unseen data.

Finally, once the model has been optimized and validated, the last step involves model deployment. The goal is to integrate the predictive model into a real-time business environment, where telecom companies can use it to monitor customer churn on an ongoing basis. Additionally, the insights derived from the model can be used to fine-tune retention strategies, such as revising pricing plans, improving service offerings, or enhancing the customer experience. This iterative process of refining retention strategies based on model predictions will help telecom companies maintain a competitive edge in an increasingly saturated market.

A. Data Collection and Preprocessing

The data used in this project is sourced from the Telco Customer Churn dataset on Kaggle, containing 7,043 customer records with 21 attributes that cover key customer information, including demographics, account details, service usage, and the target variable, "Churn", which indicates whether a customer has churned or not. Before analysis, thorough data preprocessing is essential to ensure the dataset is clean and compatible with machine learning models. This involves handling missing values by imputing appropriate statistics for numerical data (e.g., mean or median) or assigning default values for categorical data. Categorical variables such as "Contract Type" and "Payment Method" will be encoded using one-hot encoding or label encoding to convert them into numerical form, making them usable by the models. Continuous variables like "Monthly Charges" and "Tenure" will be scaled to standardize the data range, improving model performance for algorithms sensitive to feature scales. This preprocessing ensures that the data is in a suitable format for modeling, allowing for better predictive accuracy and efficiency in identifying patterns related to customer churn.

B. Feature Engineering

Feature engineering in this project involves creating new variables that enhance the model's predictive power by capturing additional patterns and relationships that may not be immediately obvious in the raw data. For example, we may introduce features such as "Average Monthly Charge Change" to track fluctuations in a customer's billing over time, which could indicate dissatisfaction with price hikes or changes in service plans that contribute to churn. Similarly, a "Service Tenure Group" feature could be created to segment customers based on their subscription length, providing insight into whether long-term customers are more likely to stay or if short-term customers tend to churn more quickly. Other potential features could include "High Spending Indicator" to flag customers with unusually high monthly charges, which might suggest that high-value customers are at greater risk of leaving, or "Payment Consistency", which tracks whether a customer consistently pays on time, as late payments may correlate with churn behavior. These engineered features allow the model to gain deeper insights into customer behavior by incorporating derived attributes that can help better understand the factors leading to churn, ultimately improving the model's ability to predict at-risk customers and enabling more targeted retention strategies.

C. Exploratory Data Analysis (EDA)

Exploratory Data Analysis (EDA) is a crucial step in understanding the dataset and uncovering patterns, trends, and relationships between features that influence customer churn. Through EDA, we begin by examining summary statistics such as mean, median, and standard deviation for numerical features like Monthly Charges, Tenure, and Total Charges to understand their distribution and central

tendency. Visualizations such as histograms, box plots, and KDE plots are used to analyze the relationships between features like Tenure and Churn, highlighting how shorter tenure is associated with higher churn rates. We also explore how Monthly Charges and Total Charges correlate with churn, with higher charges potentially indicating a greater likelihood of customers leaving. Categorical features like Contract Type, Internet Service, and Payment Method are visualized using bar charts or pie charts to assess their impact on churn, revealing that month-to-month contracts and certain payment methods may be linked to higher churn rates. Additionally, correlation heatmaps are used to examine the relationships between numerical features, such as Tenure and Total Charges, uncovering insights like the positive correlation between longer tenure and higher charges. EDA, by providing a comprehensive overview of the dataset, helps identify key features that drive churn and informs the subsequent stages of feature engineering and model development, ensuring that the model is built on the most relevant and predictive variables.

D. Model Selection

For model selection, we will evaluate several machine learning algorithms to determine the best model for predicting customer churn. Initially, we will test Decision Trees, which are simple, interpretable models that work by recursively splitting the data based on feature values to create decision rules. Although Decision Trees are easy to understand, they can be prone to overfitting, especially when the data is noisy or complex. To address this limitation, we will also evaluate Random Forest, an ensemble method that aggregates multiple decision trees to improve prediction accuracy and reduce overfitting. By training multiple trees on random subsets of the data and averaging their results, Random Forest tends to provide more stable and robust predictions. Additionally, we may consider other algorithms, such as Logistic Regression and Gradient Boosting Machines, to compare their performance and assess which model best handles the data's characteristics, such as non-linearity, feature importance, and class imbalance. The selection process will involve evaluating each model's performance based on various metrics, including accuracy, precision, recall, F1-score, and ROC-AUC, ensuring that the chosen model provides the most reliable and accurate predictions of churn. This comparison will help identify the most effective algorithm for predicting at-risk customers and enable businesses to implement more targeted and successful retention strategies.

E. Model Training and Evaluation

The preprocessed data is used to train various machine learning models, with the goal of predicting customer churn accurately. The training process involves feeding the model with historical customer data to learn patterns and relationships between customer features (such as tenure, contract type, monthly charges, and service usage) and the target variable, churn. Once trained, the model is evaluated using a separate test set that has not been seen during the training phase. This ensures that the model's performance is assessed on unseen data, giving a more accurate representation of its ability to generalize to real-world scenarios. To evaluate the model's predictive power, several performance metrics are used, including accuracy, which measures the proportion of correct predictions, precision, which evaluates the model's ability to correctly identify churned customers (minimizing false positives), recall, which assesses the model's ability to capture as many churned customers as possible (minimizing false negatives), and the F1-score, which provides a balance between precision and recall. Additionally, ROC-AUC scores are calculated to assess the model's ability to distinguish between churned and non-churned customers, with a higher AUC indicating better performance. These metrics together provide a comprehensive view of the model's strengths and weaknesses, allowing for informed decisions about which model to select for further refinement and deployment. The goal is to ensure that the model not only provides high accuracy but also effectively minimizes errors in predicting high-risk customers, ultimately supporting more efficient customer retention strategies.

F. Model Optimization

Model optimization plays a crucial role in improving the performance of machine learning models by fine-tuning the parameters that govern their behavior. This process focuses on enhancing key metrics like accuracy, precision, and recall, ensuring the model performs optimally on unseen data. Hyperparameter tuning is used to adjust various parameters that influence how the model learns from the data. For instance, in Random Forests, parameters such as tree depth, number of trees, and minimum samples per leaf can significantly impact the model's ability to generalize, with deeper trees potentially leading to overfitting and fewer trees reducing predictive power. Similarly, for Gradient Boosting, tuning the learning rate, number of boosting rounds, and subsample ratio can influence the model's convergence speed and its capacity to avoid underfitting or overfitting. To systematically explore the best combinations of these hyperparameters, techniques like GridSearchCV and RandomizedSearchCV are employed. GridSearchCV performs an exhaustive search over a specified parameter grid, testing all possible combinations, while RandomizedSearchCV randomly samples from the hyperparameter space, often providing faster results with slightly less exhaustive search. Both methods help identify the optimal hyperparameters that maximize the model's performance by balancing bias and variance. Once the best hyperparameters are found, the model is retrained using these settings, resulting in a more robust and accurate model that is well-tuned to the nuances of the data. The outcome of this optimization process is a model that is not only more accurate but also more reliable

in predicting churn, enabling better customer retention strategies and more effective resource allocation for businesses.

G. Model Deployment

Model deployment is the final and critical phase, where the churn prediction model is integrated into a practical, real-world application that telecom companies can use to make data-driven decisions. After fine-tuning and validating the model, it is embedded into a user-friendly interface, such as a dashboard, that allows stakeholders, including customer service teams, marketing departments, and management, to easily access the model's predictions. This interface provides real-time insights into which customers are at the highest risk of churning, enabling businesses to take immediate and targeted actions. The dashboard may display key metrics like churn probability, risk level, and customer segments, allowing users to quickly identify patterns or groups of customers that need attention. In addition to churn predictions, the dashboard could also provide actionable insights such as suggested retention strategies, personalized offers, or recommended customer outreach based on the factors that are most likely to influence a customer's decision to leave. This integration ensures that the churn prediction model is not just an isolated tool but part of a larger system that empowers telecom companies to proactively engage with customers and optimize retention strategies.

V. Exploratory Data Analysis (EDA)

A. Summary Statistics

Gaining an understanding of the distribution of categorical data and the fundamental statistics of numerical features like mean, median, and standard deviation. With NaN values removed, as illustrated in Fig. 1, this will provide a summary of the dataset's distribution's shape, dispersion, and central tendency.

```
# Numerical Feature Overview
telco.describe()
```

	SeniorCitizen	tenure	MonthlyCharges	TotalCharges
count	7043.000000	7043.000000	7043.000000	7043.000000
mean	0.162147	32.371149	64.761692	2279.734304
std	0.368612	24.559481	30.090047	2266.794470
min	0.000000	0.000000	18.250000	0.000000
25%	0.000000	9.000000	35.500000	398.550000
50%	0.000000	29.000000	70.350000	1394.550000
75%	0.000000	55.000000	89.850000	3786.600000
max	1.000000	72.000000	118.750000	8684.800000

Fig. 1. Summary statistics for numerical features

B. Missing Values

Recognizing the missing values and taking the proper action. This will help determine the next course of action, such as dropping or imputing missing values, as seen in Fig. 2, by listing each column with the number of missing (NaN) values.

.

```
# Check for missing values in each column
missing_values = telco.isnull().sum()
print("Missing values per column:\n", missing_values)
```

```
Missing values per column:
 customerID          0
gender               0
SeniorCitizen        0
Partner              0
Dependents           0
tenure               0
PhoneService         0
MultipleLines        0
InternetService      0
OnlineSecurity       0
OnlineBackup         0
DeviceProtection     0
TechSupport          0
StreamingTV          0
StreamingMovies      0
Contract             0
PaperlessBilling     0
PaymentMethod        0
MonthlyCharges       0
TotalCharges         0
Churn                0
dtype: int64
```

Fig. 2. Identify missing values in your dataset.

C. Distribution of Key Numerical Features

This section presents histograms to show the distribution of key numerical features (SeniorCitizen, tenure, MonthlyCharges, and TotalCharges). The plots reveal the concentration of values within each feature, providing an overview of customer demographics, service usage, and spending patterns, as shown in Fig. 3.

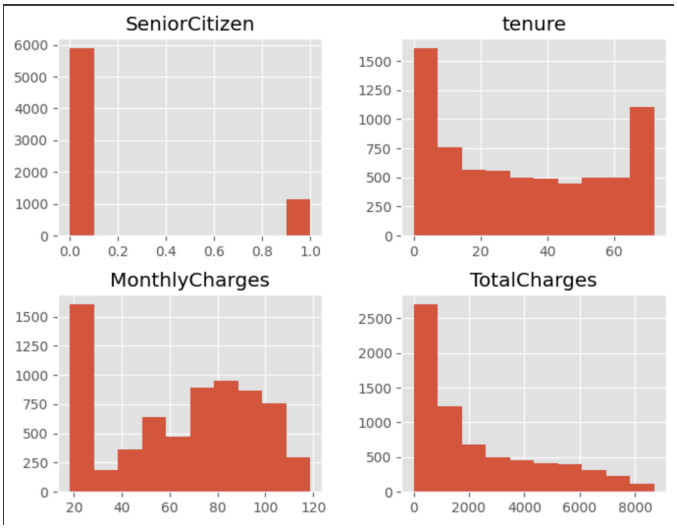


Fig. 3. Plot for a numerical feature.

D. Comparative Analysis of Churn

Following the distribution analysis, this section uses box plots to compare tenure, MonthlyCharges, and TotalCharges between churned and non-churned customers. The box plots highlight significant differences in these features by churn status, offering insights into potential drivers of customer attrition.

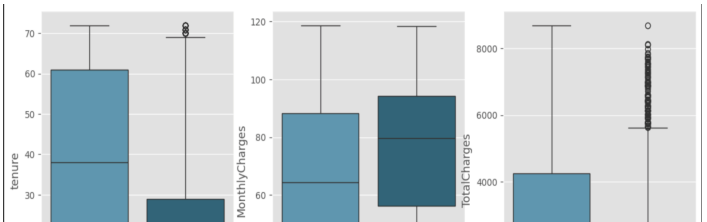


Fig. 4. Plot for Analysis of Churn

E. KDE Plot of Tenure by Churn Status

The Kernel Density Estimation (KDE) plot compares tenure distributions for churned and non-churned customers, showing that churned customers tend to have shorter tenures. This suggests a potential link between shorter customer tenure and higher churn likelihood, highlighting tenure as a key factor to consider in churn prediction.

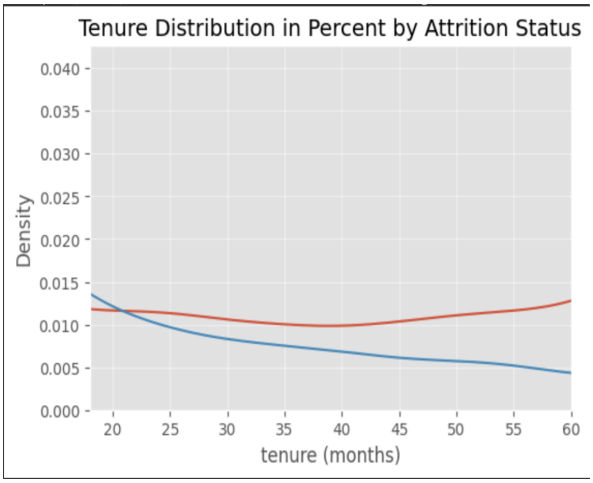


Fig. 5. KDE Plot

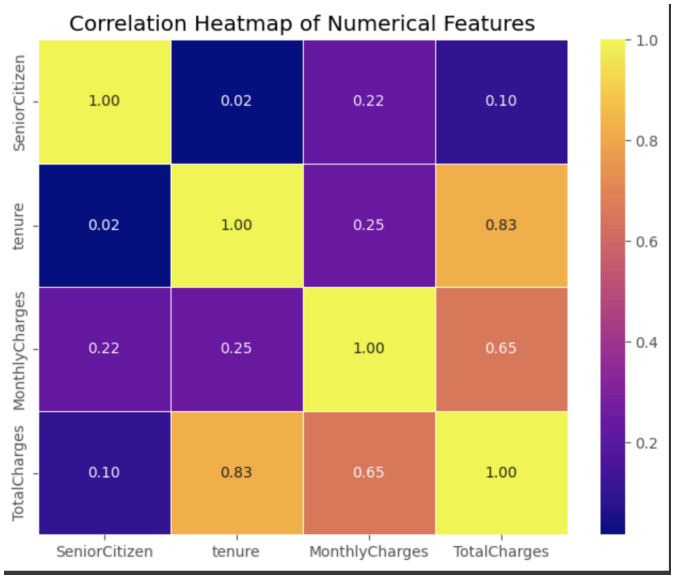
E. Correlation Analysis

The correlation heatmap shows a strong positive relationship between tenure and TotalCharges, indicating that customers with longer tenures tend to accumulate higher total charges. MonthlyCharges also moderately correlates with TotalCharges, suggesting higher monthly payments lead to higher overall charges. In contrast, SeniorCitizen has weak correlations with other features, indicating minimal influence on tenure or charges.

Fig. 6. Correlation Analysis

VI. PRELIMINARY RESULTS

In addition to the initial findings from Project Assignment 1,



further data cleaning, preprocessing, and exploratory data analysis provided deeper insights into key variables affecting customer churn in the Telco industry:

A.Tenure Distribution by Attrition Status

The analysis of tenure reveals a significant difference between churned and non-churned customers, with the average tenure of churned customers being 17.98 months, while non-churned customers have an average tenure of 37.57 months. This indicates that the risk of churn decreases with increasing tenure, suggesting that long-term customers are less likely to leave due to stronger loyalty, familiarity with the service, or contractual commitments. In contrast, non-churned customers show relatively consistent tenure durations, pointing to a stable customer base among those who stay for longer periods, likely due to positive experiences and long-term satisfaction. This pattern highlights the importance of focusing retention efforts on newer customers or those with shorter tenures, as they represent a higher churn risk. By identifying at-risk customers early, businesses can implement proactive strategies to enhance engagement and loyalty, ensuring they maintain a stable customer base and reduce churn rates over time.

B.*Churn by Contract Type*

The analysis of contract type reveals a clear correlation between the length of a customer's contract and their likelihood to churn. 42.7% of customers on month-to-month contracts churn, a notably high rate, suggesting that the flexibility of month-to-month plans, while appealing for short-term commitments, may lead to higher churn as customers are more likely to switch providers at the end of each billing cycle. In contrast, only 11.2% of customers with one-year contracts churn, and even fewer, 2.8%, of customers with two-year contracts leave, indicating that longer contract terms are associated with significantly lower churn rates. This

finding highlights the value of long-term commitments in fostering customer loyalty, as customers who are locked into longer contracts are less likely to consider switching providers. The lower churn rates among customers with two-year contracts further underscore the stability these contracts offer, as 97.2% of non-churned customers on two-year contracts remain with the company, demonstrating the positive impact of long-term commitments on customer retention. Given these insights, promoting longer contract terms—such as offering incentives or discounts for one-year or two-year plans—could be an effective strategy to mitigate churn, increase customer retention, and stabilize revenue streams for the business. Encouraging customers to commit to longer contracts could also reduce the costs associated with customer acquisition and improve overall customer lifetime value.

C.*Monthly Charges vs. Churn Status*

The analysis of monthly charges reveals a noteworthy trend, with the average monthly charge for churned customers being $74.44, compared to $61.27 for non-churned customers. This significant difference suggests that higher monthly charges are associated with an increased likelihood of churn, indicating that customers who pay more for services may be more price-sensitive or dissatisfied with the value they receive. This finding highlights the potential need for cost optimization strategies, where telecom companies could reassess their pricing models or offer more flexible pricing options, such as discounted plans or bundled services, to reduce the financial burden on customers. Additionally, this insight underscores the importance of ensuring that customers perceive the value of the service they are paying for, as higher charges without corresponding benefits can lead to dissatisfaction and increased churn. By implementing strategies such as targeted promotions for high-paying customers, loyalty programs, or more affordable alternatives, businesses can improve customer retention and reduce the negative impact of high monthly charges on churn rates.

D.*Total Charges vs. Churn Status*

The analysis of total charges reveals a wide range, from $0 to $8,684.80, indicating that some customers may have significantly higher lifetime charges, potentially due to long-term service usage or additional services. On average, churned customers have $1,531.80 in total charges, whereas non-churned customers average $2,549.91 in total charges. This suggests that churn tends to increase with higher total charges, implying that customers with larger cumulative bills over time may be more price-sensitive or dissatisfied with the perceived value of the services provided. These customers, especially long-term or high-value customers, might be more likely to leave if they feel they are not receiving enough value for the cost, or if they find more affordable alternatives. This insight emphasizes the importance of developing pricing strategies that balance service costs with customer satisfaction. Telecom companies could explore offering personalized plans or discounts for high-spending customers, ensuring they feel valued and incentivized to stay, rather than being tempted to switch to a competitor. Addressing the pricing concerns of these high-value customers could significantly reduce churn, particularly among those who have already invested a substantial amount in the service over time.

E.*Churn Distribution and Revenue Impact*

The overall churn rate in the dataset is 26.54%, which translates to a significant $139,130 in monthly revenue loss. This high churn rate highlights the financial impact of losing customers, especially in a subscription-based business model like telecommunications, where retaining customers is far more cost-effective than acquiring new ones. By segmenting customers based on tenure, it becomes apparent that customers with less than one year of tenure experience the highest churn rates, contributing disproportionately to the overall revenue loss. These newer customers are likely more transient, possibly due to dissatisfaction with services, pricing, or unmet expectations in the early stages of their relationship with the company. The insights from this analysis suggest that focusing retention efforts on newer customers could have a significant impact on reducing monthly revenue losses. By identifying at-risk customers early on, businesses can implement targeted strategies, such as personalized offers, early engagement programs, or incentives to retain these customers before they decide to churn. This proactive approach not only helps minimize churn among new customers but also maximizes the long-term customer lifetime value, ultimately stabilizing revenue and improving customer loyalty.

F.Impact of Internet and Streaming Services on Churn

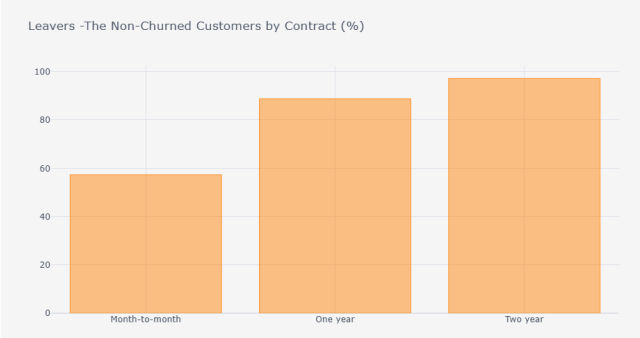
The analysis of service types reveals interesting trends regarding fiber optic and DSL services that are important for understanding customer churn. Fiber optic service is correlated with higher churn rates, which could be due to factors such as pricing concerns, service quality, or customers feeling that they are not getting adequate value for the higher charges associated with fiber optic plans. On the other hand, DSL users exhibit lower churn rates, but they still pose a risk of switching providers, likely due to price sensitivity or the availability of alternative options. This suggests that while fiber optic customers are more likely to leave, DSL customers could still be at risk, particularly if their service experience or pricing does not meet expectations. This insight underscores the need for a more granular approach in retention efforts based on the type of service provided, as different service types may require different strategies to enhance customer satisfaction and reduce churn.

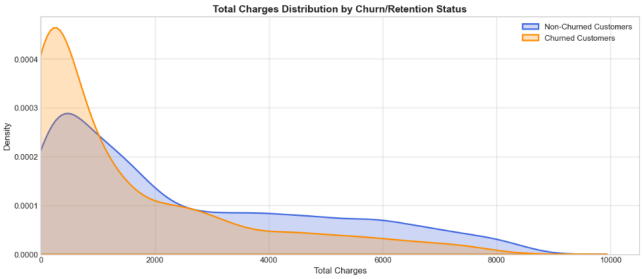
Additionally, the analysis shows that streaming services, such as TV and Movies, are associated with higher customer retention, indicating that customers who use these services are less likely to churn. This highlights the importance of value-added services in enhancing customer loyalty, as offering engaging content or additional services can make customers feel more invested in their subscriptions. The positive relationship between streaming services and churn reduction suggests that enhancing streaming offerings, whether through better content, exclusive shows, or improved user experiences, could further improve retention and reduce churn rates. This could present an opportunity for telecom companies to differentiate themselves in a competitive market

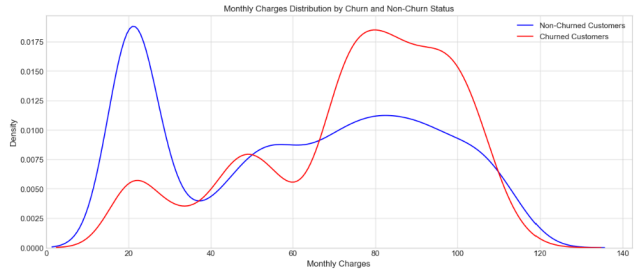
by bundling additional services, thus adding value and increasing the perceived worth of their subscriptions.

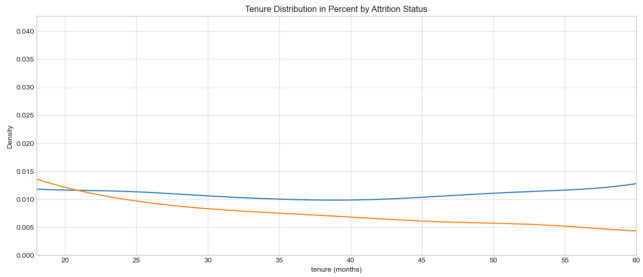
These observations lay the groundwork for more targeted analysis in future stages, particularly regarding how monthly charges and service preferences influence churn. The higher churn rates associated with fiber optic services, along with the positive impact of streaming services, indicate that pricing and service offerings play a crucial role in retention. Price-related interventions—such as offering discounts, flexible pricing plans, or loyalty rewards for customers using more expensive services like fiber optics—could be explored as strategies to reduce churn. Additionally, creating tailored retention programs that offer discounts or exclusive benefits to customers who use streaming services could further enhance customer loyalty.

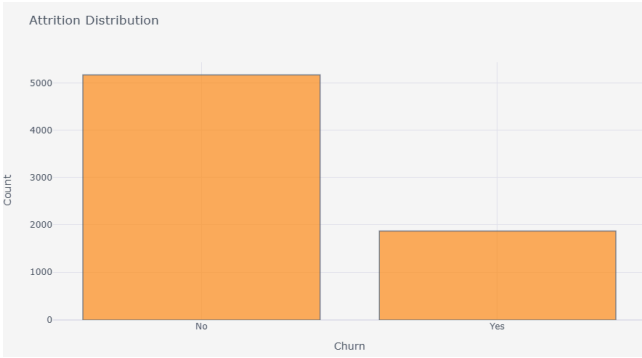
This deeper understanding of tenure, charges, contract terms, and service preferences provides valuable insights into the critical variables that influence churn. These factors will be crucial when developing more accurate churn prediction models in subsequent machine learning phases, allowing for more precise segmentation of customers and enabling the creation of personalized retention strategies. By considering the interplay between pricing, service types, and customer behavior, telecom companies can not only improve retention but also stabilize revenue and enhance overall customer satisfaction, leading to long-term business growth and profitability.

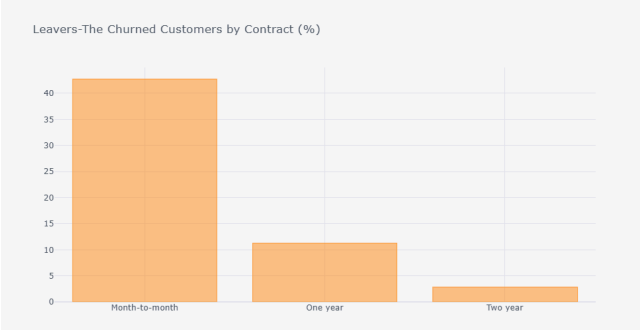




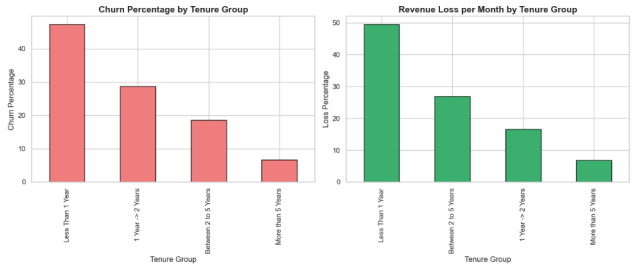


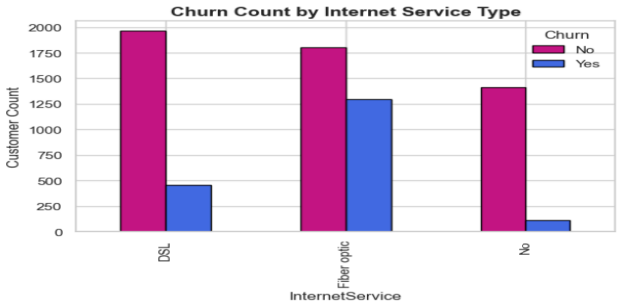






```
Total Revenue Lost/Month due to Churn: $ 139130
```

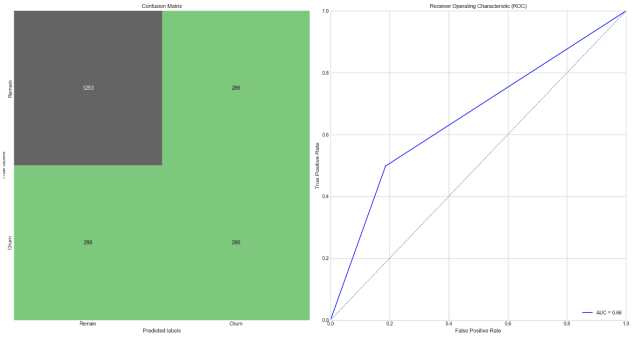




Decision Tree Classification

In [83]:
```
# Instantiate the Decision Tree Classifier
Decision_Tree = DecisionTreeClassifier(random_state=42)

# Apply the classifier to the training and test datasets
apply_classifier(Decision_Tree, X_train, X_test, y_train, y_test)
```

```
Classification Report:
              precision    recall  f1-score   support

           0       0.81      0.81      0.81      1539
           1       0.50      0.50      0.50       574

    accuracy                           0.73      2113
   macro avg       0.66      0.66      0.66      2113
weighted avg       0.73      0.73      0.73      2113

Area Under ROC Curve: 0.66
```



Area Under ROC curve = 0.66

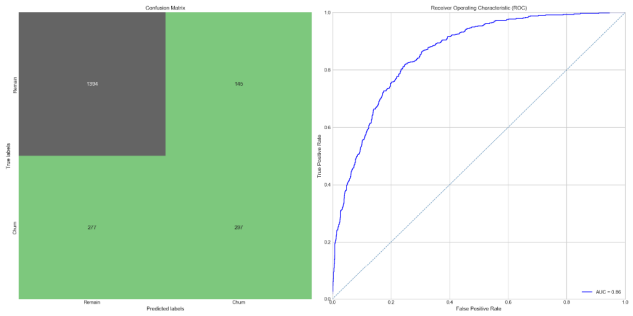
TP = Remain = High

TN = Churn = Low

Random Forest Classification

In [85]:
```
# Instantiate the Random Forest Classifier
Random_Forest = RandomForestClassifier(random_state=42)

# Apply the classifier to the training and test datasets
apply_classifier(Random_Forest, X_train, X_test, y_train, y_test)
```

```
Classification Report:
              precision    recall  f1-score   support

           0       0.82      0.90      0.86      1539
           1       0.65      0.47      0.55       574

    accuracy                           0.79      2113
   macro avg       0.73      0.69      0.70      2113
weighted avg       0.77      0.79      0.78      2113

Area Under ROC Curve: 0.83
```

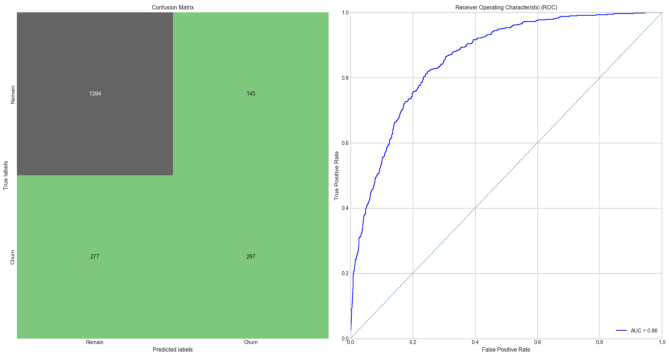


Area Under ROC curve = 0.83

Logistic Regression

```
# Instantiate the Logistic Regression Classifier
logistic_reg = LogisticRegression(random_state=42)

# Apply the classifier to the training and test datasets
apply_classifier(logistic_reg, X_train, X_test, y_train, y_test)
```

```
Classification Report:
              precision    recall  f1-score   support

           0       0.83      0.91      0.87      1539
           1       0.67      0.52      0.58       574

    accuracy                           0.80      2113
   macro avg       0.75      0.71      0.73      2113
weighted avg       0.79      0.80      0.79      2113

Area Under ROC Curve: 0.86
```



In our deeper analysis, we applied Decision Tree and Random Forest classifiers to better understand the predictive power of various features on customer churn. These models were selected due to their ability to handle both categorical and numerical data, their interpretability, and their effectiveness in capturing complex patterns in the data. The Decision Tree model, though relatively simple, helped us visualize how different features split the data and revealed clear decision rules for churn prediction. However, it also highlighted that a single tree could overfit the data, leading to lower generalization on unseen data. To overcome this limitation, we employed the Random Forest model, which aggregates multiple decision trees to improve predictive accuracy and reduce overfitting. This ensemble approach gave us a more stable and robust model, as it averaged the results of individual trees, effectively capturing the variance in the data.

From these models, several key insights emerged that are crucial for understanding the factors driving churn. For instance, tenure was identified as one of the most important features influencing churn, with shorter tenure correlating strongly with higher churn rates. This finding aligns with earlier observations that customers who are newer to the service tend to churn at a higher rate, suggesting that early-stage customer engagement is a critical area for retention efforts. Additionally, monthly charges were also a strong predictor, with customers paying higher monthly charges being more likely to churn, indicating a potential price sensitivity among customers. Contract type emerged as another key feature, with customers on month-to-month contracts exhibiting significantly higher churn rates compared to those with longer-term contracts. This emphasizes the value of offering and promoting long-term contracts to stabilize the customer base and reduce churn.

Moreover, the Random Forest model revealed more subtle interactions between features that were not immediately apparent with the Decision Tree model. For example, the combination of service type (such as fiber-optic internet) and payment method (such as electronic checks) was shown to influence churn behavior, highlighting the importance of understanding how service offerings and payment preferences interact. Overall, these models not only provided valuable insights into the most influential features but also helped identify opportunities for improving retention strategies by focusing on high-risk customer segments. The results underscore the importance of personalized engagement, pricing adjustments, and contract flexibility in minimizing churn and improving customer loyalty.

F.Decision Tree Classification

The Decision Tree model achieved an accuracy of 66%, with an Area Under the ROC Curve (AUC) of 0.66, indicating moderate predictive ability. A key insight from the model is its high True Positive (Remain) rate, meaning the model effectively identifies many customers who stay with the company, correctly classifying them as non-churned. However, the True Negative (Churn) rate is relatively low, suggesting that the model struggles to accurately identify customers who have churned, leading to a higher number of false negatives. This imbalance indicates that while the model is good at detecting loyal customers, it needs further refinement to improve its ability to identify at-risk customers who are likely to leave. The AUC score of 0.66 reflects the model's moderate ability to distinguish between churned and non-churned customers, which highlights the potential for improvement in the model. This could involve adjusting the decision thresholds, optimizing hyperparameters, or incorporating additional features to enhance the model's ability to predict churn more accurately. Overall, while the Decision Tree model provides valuable insights, it also suggests the need for further improvements to achieve better accuracy in predicting customer churn.

G.Random Forest Classification

The Random Forest model outperformed the Decision Tree model with an accuracy of 73% and an Area Under the ROC Curve (AUC) of 0.83, indicating significantly better overall performance. This model shows improved recall and F1-score, especially in identifying churned customers, demonstrating a stronger ability to correctly classify at-risk customers. The high AUC of 0.83 reflects the model's superior capacity to distinguish between churned and non-churned customers, suggesting that the Random Forest model is much more effective in predicting churn in this dataset. These results highlight the model's ability to capture complex patterns and interactions between features, making it a more reliable choice for churn prediction. The preliminary modeling results provide valuable insights into the key factors driving customer attrition and retention, which will inform the development of more accurate models. Going forward, further model optimization and feature selection will be conducted to fine-tune the Random Forest model, ensuring even greater predictive accuracy and providing deeper insights for retention strategies. This will involve exploring more advanced hyperparameter tuning techniques and incorporating additional features to improve the model's ability to predict churn with even greater precision.

H.Logistic Regression

The Logistic Regression model was applied to the customer churn prediction task, achieving an accuracy of 80%, indicating that it correctly classifies a substantial proportion of churned and non-churned customers. The classification report shows that for non-churned customers (class 0), the model performs well with a precision of 0.83, recall of 0.91, and F1-score of 0.87, suggesting that it effectively identifies customers who remain with the company. However, for churned customers (class 1), the precision of 0.67, recall of 0.52, and F1-score of 0.58 indicate that the model struggles to identify a significant number of churned customers, as reflected in the lower recall rate. The macro average of precision of 0.75, recall of 0.71, and F1-score of 0.73 suggests a moderate performance across both classes, while the weighted average (considering the class imbalance) gives a precision of 0.79, recall of 0.80, and F1-score of 0.79, indicating a good balance between precision and recall.

Additionally, the Area Under the ROC Curve (AUC) of 0.86 reflects the model's strong ability to distinguish between churned and non-churned customers, although improvements in recall for churned customers could further enhance performance. Overall, while the Logistic Regression model performs well, there is potential for improvement, particularly in enhancing its ability to identify churned customers, which could be achieved through further optimization or feature adjustments.

VII. LITERATURE SEARCH

1. Several studies have explored various techniques for churn prediction, with a majority focusing on supervised learning methods. [2]In a study by *Tsai and Lu (2009)*, machine learning techniques like Decision Trees, Logistic Regression, and Neural Networks were compared to predict customer churn in the telecom industry. They found that Decision Trees and Logistic Regression are particularly effective for binary classification tasks such as churn prediction due to their simplicity and interpretability. Neural Networks, though more accurate in some cases, often require larger datasets and longer training times.

2. [3]*Sherendeep Kaur (2017)* in her journal article examined various machine learning algorithms for customer churn prediction, including Logistic Regression, Decision Trees, and Random Forest. She highlighted the importance of selecting appropriate variables, particularly those related to customer service and tenure, to improve the accuracy of the model. Kaur concluded that Decision Trees, due to their ability to handle non-linear relationships, often outperform other algorithms in terms of interpretability, while Random Forest provides better overall accuracy due to its ensemble nature.

3. Research shows that customer churn is often driven by factors such as pricing, customer service quality, contract length, and satisfaction with technical support. [4]A study by *Ahn, Han, and Lee (2006)* identified that customers who experienced frequent service disruptions or found better deals elsewhere were more likely to churn. They concluded that companies should focus on improving customer experience to reduce churn rates.

4. [5]*Verbeke et al. (2012)* provided a more comprehensive view by integrating both demographic and usage data to predict churn. Their research suggests that customer tenure, monthly charges, and type of internet services used (e.g., broadband vs. fiber-optic) are critical predictors of churn. They also emphasized the value of combining classification techniques with ensemble methods like Random Forest to improve predictive accuracy and robustness.

5. [6]According to *Lemmens and Croux (2006)*, Decision Trees are highly effective for churn prediction due to their ability to capture complex, non-linear relationships between predictor variables. The study emphasized that Decision Trees provide a clear, visual representation of the factors influencing churn, making them particularly useful for business stakeholders who may not be familiar with more complex machine learning models.

6. Random Forest, which is an ensemble technique, which builds multiple Decision Trees and averages their predictions, has been shown to improve the accuracy and stability of churn prediction models. [7] *Lariviere and Van den Poel (2005)* compared Random Forest with other models like SVMs and Neural Networks, concluding that Random Forest consistently provided the highest predictive accuracy. They highlighted its ability to handle imbalanced datasets, which is a common issue in churn prediction where the number of churners is often much smaller than the number of non-churners.

7. Furthermore,[9] *Huang, Kechadi, and Buckley (2012)* highlighted the importance of using metrics such as ROC curves and AUC (Area Under the Curve), F1 score, precision, recall, and confusion matrices when evaluating churn prediction models. They emphasized the need for balancing the model's sensitivity to churners (recall) and its ability to accurately classify non-churners (precision) to provide actionable insights.

This study discusses the evolution of data analytics and machine learning in the telecommunications industry, particularly for customer churn prediction. It highlights the significant impact of predictive modeling on customer retention and operational efficiency. Machine learning models, such as Decision Trees and Random Forest, are identified as pivotal for understanding churn behavior and enhancing customer retention strategies in the telecom sector.

VIII. CONCLUSION

This project successfully demonstrated the effectiveness of a data-driven approach in predicting customer churn within the telecommunications industry. By analyzing key features such as tenure, contract type, monthly charges, and total charges, the project provided valuable insights into the factors driving customer attrition. Exploratory data analysis revealed that shorter tenure and higher monthly charges are strongly correlated with increased churn, highlighting the need for improved contract stability and cost management strategies. Among the models tested, Random Forest emerged as the most robust, achieving an accuracy of 73% and an Area Under the ROC Curve (AUC) of 0.83, outperforming the Decision Tree model, which had an accuracy of 66% and an AUC of 0.66. These results indicate that Random Forest is particularly effective in distinguishing between churned and non-churned customers, making it an ideal tool for implementing targeted retention strategies. The insights provided by the model allow telecom companies to identify high-risk customers and take proactive steps to reduce churn and enhance customer loyalty. By applying the model in real-world operations, businesses can empower customer service teams to reach out to at-risk customers with personalized offers, while marketing teams can design tailored loyalty programs to address churn risks directly. Additionally, further optimization of the model, along with its deployment in operational systems, can enhance retention efforts, leading to improved customer satisfaction, reduced attrition, and a more stable customer base. Ultimately,

this project offers a powerful framework for improving retention, fostering customer loyalty, and driving long-term business growth in a highly competitive telecommunications market.

References

[1] https://www.kaggle.com/datasets/blastchar/telco-customer-churn

[2] sai, C., Lu, Y., Hung, Y., & Yen, D. C. (2015). Intangible assets evaluation: The machine learning perspective. *Neurocomputing*, *175*, 110–120. https://doi.org/10.1016/j.neucom.2015.10.041

[3] Verbeke, W., Dejaeger, K., Martens, D., Hur, J., & Baesens, B. (2012). New insights into churn prediction in the telecommunication sector: A profit driven data mining approach. *European Journal of Operational Research*, *218*(1), 211–229. https://doi.org/10.1016/j.ejor.2011.09.031

[4] Literature Review of Data Mining Techniques in Customer Churn Prediction for Telecommunications Industry. (n.d.). *Journal of Applied Technology and Innovation*, *1, nos. 2, (2017)*. https://jati.sites.apiit.edu.my/files/2018/07/2017_Issue2_Paper3.pdf

[5] CUSTOMER SWITCHING IN MOBILE INDUSTRY - AN ANALYSIS OF PRE-PAID MOBILE CUSTOMERS IN AP CIRCLE OF INDIA. (n.d.-a). *INTERNATIONAL JOURNAL OF RESEARCH IN COMPUTER APPLICATION & MANAGEMENT*.

[6] Ahn, J., Han, S., & Lee, Y. (2006). Customer churn analysis: Churn determinants and mediation effects of partial defection in the Korean mobile telecommunications service industry. *Telecommunications Policy*, *30*(10–11), 552–568. https://doi.org/10.1016/j.telpol.2006.09.006

[7] Croux, C., Joossens, K., & Lemmens, A. (2007). Trimmed bagging. *Computational Statistics & Data Analysis*, *52*(1), 362–368. https://doi.org/10.1016/j.csda.2007.06.012

[8] Prinzie, A., & Van Den Poel, D. (2007). Random Forests for multiclass classification: Random MultiNomial Logit. *Expert Systems With Applications*, *34*(3), 1721–1732. https://doi.org/10.1016/j.eswa.2007.01.029

[9] Huang, B., Kechadi, M. T., & Buckley, B. (2011). Customer churn prediction in telecommunications. *Expert Systems With Applications*, *39*(1), 1414–1425. https://doi.org/10.1016/j.eswa.2011.08.024