Customer Churn Analysis

Dissertation submitted in fulfilment of the requirements for the Degree of

BACHELOR OF TECHNOLOGY

in

COMPUTER SCIENCE AND ENGINEERING

By

Velagana Balaji

Reg.no **:12219897**

Section**: K22UP**

Supervisor

Mr. Himanshu Gajanan Tikle (63982)

(Technical Institutor of EDA-Up Grad)

School of Computer Science and Engineering

Lovely Professional University

Phagwara, Punjab (India)

April, 2025.

DECLARATION STATEMENT

I hereby declare that the research work reported in the dissertation/dissertation proposal entitled **Customer Churn Analysis** in partial fulfilment of the requirement for the award of Degree for Bachelor of Technology in Computer Science and Engineering at Lovely Professional University, Phagwara, Punjab is an authentic work carried out under supervision of my research supervisor **Mr. Himanshu Gajnan Tikle**. I have not submitted this work elsewhere for any degree or diploma.

I understand that the work presented herewith is in direct compliance with Lovely Professional University's Policy on plagiarism, intellectual property rights, and highest standards of moral and ethical conduct. Therefore, to the best of my knowledge, the content of this dissertation represents authentic and honest research effort conducted, in its entirety, by me. I am fully responsible for the contents of my dissertation work.

Signature of Candidate

Velagana Balaji

12219897

SUPERVISOR'S CERTIFICATE

This is to certify that the work reported in the B.Tech Dissertation/dissertation proposal entitled "**Customer Churn Analysis"**, submitted by **Velagana Balaji** at **Lovely Professional University, Phagwara, India** is a Bonafide record of his / her original work carried out under my supervision. This work has not been submitted elsewhere for any other degree.

Signature of Supervisor

Date:

ACKNOWLEDGEMENT

I would like to take this opportunity to express my profound gratitude to **Mr. Himanshu** Sir, whose invaluable guidance, mentorship, and unwavering support made this project on Machine learning on **Customer Churn Analysis**. From the initial conception of the idea to the final stages of this project, his vast knowledge and expertise in data science and machine learning have been instrumental in shaping the direction and depth of my research. His insightful advice and constructive feedback consistently pushed me to explore beyond the surface and helped refine my work to a higher standard. Himanshu Sir's passion for teaching and his dedication to his students have truly been a source of inspiration throughout this journey. His ability to simplify complex concepts, coupled with his approachable nature, created an environment that encouraged learning and curiosity. I am immensely thankful for the time and effort he invested in reviewing my work, providing critical input, and offering continuous encouragement during challenging phases of the project. His feedback was not only crucial for the technical aspects of the project but also helped me build a stronger foundation in the principles of data analysis and machine learning. I would also like to extend my appreciation to my institution for providing the necessary resources and a conducive learning environment that allowed me to pursue this research. My heartfelt thanks to my colleagues and classmates, who have been a constant source of motivation, encouragement, and collaboration. Their shared knowledge and perspectives contributed significantly to my understanding of the subject matter. In conclusion, I would like to express my sincere gratitude to everyone who has contributed to the completion of this project. The learning and experiences I have gained throughout this process will undoubtedly serve as a foundation for my future endeavors in the field of data science and machine learning. This project would not have been possible without the invaluable guidance, support, and encouragement from **Mr. Himanshu** Sir, and I am truly grateful for the opportunity to learn under his mentorship.

TABLE OF CONTENTS

1.ABSTRACT

The investigation of a customer churn dataset forms the foundation of this project, which focuses on understanding how various factors influence customer retention in a subscription-based business. The dataset comprises a broad range of features, including customer demographics, service subscriptions, billing information, contract details, and churn status. These attributes offer a comprehensive view of customer behavior and business interactions, providing valuable insights into churn trends and the operational dynamics that affect customer loyalty. The primary goal of this analysis is to explore the relationships between these variables and their collective impact on churn outcomes. By uncovering these patterns, businesses can develop more effective retention strategies, improve customer satisfaction, and enhance long-term profitability.

In the initial phase of the project, exploratory data analysis (EDA) was performed using descriptive statistics and visualizations to gain a clear understanding of churn distribution across different customer segments. Visualization tools such as bar plots, heatmaps, and boxplots were used to highlight trends related to contract types, payment methods, tenure, service usage, and monthly charges. This allowed for the identification of high-risk customer groups, service-related churn factors, and demographic patterns that influence churn. Furthermore, outlier detection and data normalization techniques ensured the reliability and consistency of the dataset for deeper analysis.

The analysis also addressed anomalies and outliers within the dataset, such as unusual billing behaviors or inconsistencies in service usage, which could potentially skew the results. These findings help identify opportunities for improving service offerings or operational strategies. For example, customers using specific services or payment methods were found to have higher churn rates, prompting further investigation into service quality or payment flexibility.

Building upon these insights, the project implemented machine learning models to predict customer churn with high accuracy. Algorithms such as Logistic Regression, Random Forest, and XGBoost were trained and evaluated using key performance metrics like accuracy, precision, recall, and F1-score. Among these, XGBoost achieved the best performance, effectively capturing complex churn patterns and handling imbalanced class distributions.

The predictive models developed in this project enable businesses to proactively identify customers at risk of leaving and take preventive actions such as personalized offers or support interventions. Moreover, feature importance analysis highlighted key drivers of churn, including contract type, tenure, and monthly charges, guiding businesses in refining their service and pricing strategies.

In conclusion, this customer churn analysis project combines statistical techniques, visualization, and machine learning to deliver actionable insights for improving customer retention. The EDA phase revealed critical churn patterns, while predictive modeling provided a robust framework for future forecasting and strategic decision-making. These insights empower businesses to optimize customer engagement, reduce churn, and drive sustainable growth through data-driven strategies.

2.INTRODUCTION

Customer churn is a critical challenge faced by subscription-based businesses, particularly in highly competitive industries such as telecommunications, where retaining existing customers is significantly more cost-effective than acquiring new ones. In this context, understanding the factors that drive customer churn and predicting churn behavior are essential for developing effective retention strategies, enhancing customer satisfaction, and improving business performance. This project aims to explore customer behavior and identify churn patterns using a structured dataset that includes a variety of features such as customer demographics, service subscription details, payment methods, contract types, and churn status. By investigating the relationships between these attributes, the analysis seeks to uncover insights into why customers leave and how businesses can proactively prevent churn.

The dataset used in this project comprises over 7,000 records with 21 attributes, offering a comprehensive view of each customer's engagement with the service. Key variables include gender, senior citizen status, tenure, monthly charges, total charges, contract duration, payment methods, and value-added services such as internet, tech support, and streaming subscriptions. The project begins with data preprocessing to ensure accuracy and completeness—handling categorical encoding, normalizing numerical values, and checking for missing data and outliers. With a clean dataset, exploratory data analysis (EDA) techniques are employed to reveal patterns and trends related to customer churn. Visualizations such as count plots, boxplots, and heatmaps help highlight significant correlations between churn behavior and key features, such as short contract duration, high monthly charges, low tenure, and payment via electronic checks.

A central component of the analysis involves identifying the key drivers of churn through correlation analysis and statistical hypothesis testing. For example, the relationship between contract type and churn probability is explored to determine whether longer contracts help retain customers. Similarly, the impact of payment methods and service subscriptions on churn likelihood is analyzed to inform decisions about billing flexibility and service offerings. Hypothesis testing ensures that the observed patterns are statistically significant and not the result of random variation, strengthening the reliability of the insights gained.

To further enhance the understanding of churn behavior, the project leverages machine learning models to predict the likelihood of customer churn. Three models—Logistic Regression, Random Forest, and XGBoost—are trained and evaluated using performance metrics such as accuracy, precision, recall, F1-score, and confusion matrices. Among these, XGBoost delivers the highest performance, with an accuracy of over 81%, making it the most effective model for predicting churn. Feature importance analysis using Random Forest and XGBoost reveals that features like tenure, contract type, monthly charges, and payment

method are among the most influential predictors of churn. This information is vital for businesses aiming to focus their retention strategies on the most impactful factors.

Additionally, customer segmentation is performed using clustering techniques to group customers based on their service usage, payment behavior, and engagement levels. Techniques such as K-means clustering help identify distinct customer profiles, such as high-risk churners, loyal long-term users, or customers with specific service preferences. These segments enable more personalized marketing efforts, tailored service recommendations, and targeted interventions designed to reduce churn and increase customer lifetime value.

The findings from this analysis have practical implications for improving business operations and strategy. For instance, companies can prioritize long-term contracts, promote secure and convenient payment methods, and offer loyalty incentives to customers with short tenure or minimal engagement. The analysis also serves as a foundation for deploying real-time churn prediction systems within customer relationship management (CRM) platforms. A Streamlit-based web application was developed to allow real-time churn predictions based on user inputs, and the model was deployed using Docker containers with CI/CD pipelines for seamless updates and scalability.

By applying EDA, clustering, hypothesis testing, and machine learning, this project provides a comprehensive view of customer churn dynamics. The insights gained enable businesses to make data-driven decisions to optimize customer retention strategies, allocate resources more effectively, and improve profitability. Furthermore, the project lays the groundwork for advanced predictive modeling and automation tools that can evolve alongside changing customer behaviors and business needs. As industries increasingly rely on data to drive strategy, the ability to understand and act on churn insights becomes a vital competitive advantage in achieving sustainable growth and customer loyalty.

3.PROBLEM STATEMENT

The increasing availability of customer data in subscription-based industries has created valuable opportunities for businesses to better understand customer behavior, reduce churn, and improve operational efficiency. However, effectively analyzing and leveraging this data remains a challenge, as it involves a complex interplay of variables such as customer demographics, service usage patterns, contract details, billing methods, and engagement metrics. The core problem addressed in this project is how to efficiently identify at-risk customers and understand the factors that drive customer churn, enabling businesses to implement proactive retention strategies and enhance overall performance. Customer churn prediction is a critical aspect of customer relationship management, as it helps businesses recognize behavioral patterns and preferences that lead to customer dissatisfaction and attrition. By identifying distinct segments of customers based on their risk of churn, service preferences, or payment behaviors, businesses can tailor retention efforts, personalize customer interactions, and implement targeted incentives that improve customer loyalty and lifetime value. Likewise, understanding the underlying factors contributing to churn—such as contract type, tenure, payment method, and service issues—is essential for optimizing customer experience and business profitability. Operational inefficiencies like limited support services, inflexible contracts, or inadequate payment options can contribute significantly to churn, negatively impacting revenue and customer satisfaction. Therefore, businesses must analyze customer data to identify patterns that influence churn rates and affect long-term retention.

The primary goal of this project is to perform a comprehensive analysis of a telecom customer churn dataset to uncover meaningful patterns and insights that can inform strategic decision-making. Through data cleaning, preprocessing, exploratory data analysis (EDA), and machine learning modeling, the project aims to highlight the key drivers of churn and provide actionable recommendations for reducing customer attrition. Clustering techniques are also applied to segment customers into distinct behavioral groups, allowing for more personalized

engagement and targeted interventions. Furthermore, predictive models such as Logistic Regression, Random Forest, and XGBoost are used to forecast churn risk, with XGBoost delivering the highest performance in terms of accuracy and reliability. These findings not only empower businesses to take early action against potential churn but also serve as a foundation for deploying real-time churn prediction tools integrated into CRM platforms. Ultimately, this project provides a data-driven approach to enhancing customer retention, optimizing business operations, and gaining a competitive edge in markets where customer loyalty is a key driver of long-term success.

4. LITERATURE REVIEW

A comprehensive understanding of customer churn has become a central concern in subscription-based industries, as businesses aim to leverage data-driven insights to enhance customer retention, reduce attrition, and improve operational efficiency. Customer churn analysis enables companies to identify behavioral patterns and key risk indicators associated with customer loss, facilitating the development of proactive strategies to retain high-value customers. The foundation of churn analysis lies in the recognition that not all customers behave the same; as noted by Smith (1956) in his early work on market segmentation, a one-size-fits-all approach is insufficient for maximizing customer engagement and long-term value. In this context, segmentation based on churn risk—driven by factors such as tenure, contract type, service usage, and billing preferences—allows businesses to tailor retention efforts and optimize customer experience.

Several studies have emphasized the importance of understanding the drivers of churn. According to Kumar and Shah (2014), long-term customer value is influenced not only by the nature of customer relationships but also by operational and service-level factors such as contract flexibility, payment ease, and support responsiveness. These insights are particularly relevant in churn prediction, where transactional, demographic, and service-related variables must be analyzed collectively to predict customer behavior accurately. Moreover, the relationship between customer dissatisfaction and churn has been well documented, with evidence showing that poor service quality, lack of personalization, and rigid billing models are strong predictors of attrition.

Advancements in data science and machine learning have further strengthened churn analysis. Algorithms such as Logistic Regression, Random Forest, and XGBoost are frequently used to model churn likelihood, offering businesses powerful tools for forecasting and decision-making. These models enable organizations to identify high-risk customers,

prioritize retention strategies, and allocate resources efficiently. Clustering techniques like K-means are also employed to segment customers into behaviorally similar groups, supporting more targeted and personalized interventions. Research also highlights the use of correlation analysis and hypothesis testing to validate assumptions about churn drivers, ensuring that strategies are grounded in statistically significant relationships.

The use of predictive modeling for churn is aligned with broader trends in data-driven business management. As noted by Chintagunta (2002), analytical rigor in customer behavior studies can yield substantial benefits when integrated with strategic planning and operational execution. This includes optimizing outreach timing, refining product and service bundles, and improving customer satisfaction through targeted engagement. The literature consistently supports the view that a predictive, segmented, and personalized approach to customer management leads to increased retention, higher lifetime value, and improved profitability. Therefore, the integration of exploratory data analysis, machine learning, and customer segmentation presents a robust framework for churn analysis, offering businesses a competitive edge in managing customer relationships in an increasingly dynamic and data-intensive environment.

5. METHODOLOGY

1. Data Exploring/Inspection

The first step in the churn analysis process involves exploring and inspecting the dataset to understand its structure, ensure data quality, and identify potential issues. The dataset includes critical customer attributes such as gender, senior citizen status, tenure, contract type, payment method, internet and support services, monthly and total charges, and churn status. During inspection, essential exploratory tasks are performed, including checking for null values, duplicate entries, and verifying correct data types. Summary statistics like mean, median, standard deviation, and frequency counts are calculated for both numerical and categorical variables to get an overview of customer distribution. Visualizations such as count plots, histograms, and box plots are used to detect imbalances, patterns, and outliers that may influence the predictive model.

2. Data Preprocessing

Data preprocessing ensures the dataset is clean, consistent, and suitable for model training. The following steps are undertaken:

- **Handling Missing Values**: The dataset is checked for missing values, particularly in numerical columns like TotalCharges. Missing entries are imputed using appropriate techniques like mean substitution or removed depending on their distribution and impact.
- **Removing Duplicates**: Duplicate customer records are eliminated to maintain data uniqueness and reliability.
- **Type Conversion**: Data types are corrected where necessary, especially for features like TotalCharges, which may initially be interpreted as strings.
- **Outlier Detection**: Outliers in numerical columns (e.g., MonthlyCharges, Tenure) are identified using IQR or Z-score techniques, and treated if they are determined to distort model performance.

3. Feature Engineering

Feature engineering is performed to enhance the dataset by creating meaningful variables that help improve model accuracy and interpretability:

- **Tenure Binning**: Customer tenure is grouped into categories (e.g., short-term, mid-term, long-term) to analyze churn tendencies across different lifecycle stages.
- **Service Count**: A new feature is created to represent the total number of services each customer subscribes to, indicating engagement level.
- **Average Monthly Spend**: Calculated as TotalCharges / Tenure to assess spending habits and financial commitment.
- **IsEngaged**: A binary feature indicating whether a customer uses multiple services (e.g., internet + tech support), helping to identify loyalty indicators.

4. Feature Scaling

To ensure that machine learning models, especially distance-based algorithms, treat all features fairly, numerical features are scaled:

- **Standardization**: Features like Tenure, MonthlyCharges, and TotalCharges are standardized using z-score scaling to have zero mean and unit variance.
- **Normalization**: When required (e.g., for K-Means clustering or PCA), min-max normalization is applied to compress values into the [0, 1] range.

5. Target Variable Approach

In this supervised classification project, the **target variable** is the Churn column, which indicates whether a customer has left the service. This binary variable (Yes/No) is encoded appropriately (e.g., 1 for churned, 0 for retained) for machine learning models to predict churn likelihood. The goal is to build a classifier that can accurately distinguish between churned and non-churned customers.

6. Data Visualization

Visualization plays a vital role in understanding patterns and distributions related to churn:

- **Univariate Analysis**: Count plots and pie charts are used to analyze churn rates across individual variables like Contract, PaymentMethod, and SeniorCitizen.
- **Bivariate Analysis**: Bar plots and box plots help study the impact of features like Tenure and MonthlyCharges on churn.
- **Multivariate Analysis**: Heatmaps are used to show correlations between numerical features. Pair plots and grouped bar charts help reveal complex interactions between variables like service usage and churn.
- **PCA Visualization**: After clustering or dimensionality reduction, scatter plots are used to visualize group separation and churn tendencies.

7. Customer Segmentation

Though churn prediction is a supervised task, unsupervised clustering techniques like **K-Means** or **Hierarchical Clustering** are employed to segment customers based on behavioral attributes like ServiceCount, Tenure, MonthlyCharges, and SupportUsage. This segmentation allows for better profiling of at-risk customers and personalization of retention strategies. The **Elbow Method** is used to determine the optimal number of customer clusters.

8. Churn Pattern Analysis

The analysis investigates how various factors influence churn:

- **Contract and Tenure**: Month-to-month contracts and lower tenure are strongly associated with higher churn rates.
- **Payment Methods**: Customers paying via electronic checks exhibit a higher tendency to churn.
- **Service Engagement**: Customers who subscribe to fewer services or lack tech support/security are more likely to leave.
- **Monthly Charges**: A trend of increasing churn with rising monthly charges is observed, potentially indicating dissatisfaction with pricing or value.

Statistical tests and correlation analysis confirm these trends, validating assumptions about churn drivers.

9. Insights and Recommendations

Based on the analysis, several key insights and actionable recommendations are provided:

- **Target Short-Term Contract Customers:** Offer incentives or discounts to encourage longer-term commitments.
- **Improve Support Services:** Promote value-added services like online support and security to boost engagement**.**
- **Adjust Billing Models:** Provide flexible, transparent billing options to reduce churn among electronic check users.

- **Identify High-Risk Segments:** Use segmentation results to create personalized outreach strategies for high-risk customers.
- **Predictive Intervention:** Deploy the churn prediction model within a CRM to trigger retention workflows in real-time**.**

6. RESULT

```
[7]: telco_base_data.head()
```

[7]:

	customerID	gender	SeniorCitizen	Partner	Dependents	tenure	PhoneService	MultipleLines	InternetService	OnlineSecurity	...	DeviceProtection	TechSupport
0	7590-VHVEG	Female	0	Yes	No	1	No	No phone service	DSL	No	...	No	No
1	5575-GNVDE	Male	0	No	No	34	Yes	No	DSL	Yes	...	Yes	No
2	3668-QPYBK	Male	0	No	No	2	Yes	No	DSL	Yes	...	No	No
3	7795-CFOCW	Male	0	No	No	45	No	No phone service	DSL	Yes	...	Yes	Yes
4	9237-HQITU	Female	0	No	No	2	Yes	No	Fiber optic	No	...	No	No

Fig-1: head() of the Dataset.

Explanation: Displays the top 5 rows of the Dataset.

```
[10]: telco_base_data.columns.values
```

```
[10]: array(['customerID', 'gender', 'SeniorCitizen', 'Partner', 'Dependents',
       'tenure', 'PhoneService', 'MultipleLines', 'InternetService',
       'OnlineSecurity', 'OnlineBackup', 'DeviceProtection',
       'TechSupport', 'StreamingTV', 'StreamingMovies', 'Contract',
       'PaperlessBilling', 'PaymentMethod', 'MonthlyCharges',
       'TotalCharges', 'Churn'], dtype=object)
```

Fig-2: Columns of the Dataset.

Explanation: Displays the list of all the columns**.**

```
[40]: telco_base_data.shape
      print("No.of rows: ",telco_base_data.shape[0])
      print("No.of columns: ",telco_base_data.shape[1])

      No.of rows:  7043
      No.of columns:  21
```

Fig-3: Shape of the Dataset.

Explanation: Displays the number of rows and columns in the Dataset**.**

```
[18]: # Concise Summary of the dataframe, as we have too many columns, we are using the verbose = True mode
      telco_base_data.info(verbose = True)

      <class 'pandas.core.frame.DataFrame'>
      RangeIndex: 7043 entries, 0 to 7042
      Data columns (total 21 columns):
       #   Column            Non-Null Count  Dtype
      ---  ------            --------------  -----
       0   customerID        7043 non-null   object
       1   gender            7043 non-null   object
       2   SeniorCitizen     7043 non-null   int64
       3   Partner           7043 non-null   object
       4   Dependents        7043 non-null   object
       5   tenure            7043 non-null   int64
       6   PhoneService      7043 non-null   object
       7   MultipleLines     7043 non-null   object
       8   InternetService   7043 non-null   object
       9   OnlineSecurity    7043 non-null   object
       10  OnlineBackup      7043 non-null   object
       11  DeviceProtection  7043 non-null   object
       12  TechSupport       7043 non-null   object
       13  StreamingTV       7043 non-null   object
       14  StreamingMovies   7043 non-null   object
       15  Contract          7043 non-null   object
       16  PaperlessBilling  7043 non-null   object
       17  PaymentMethod     7043 non-null   object
       18  MonthlyCharges    7043 non-null   float64
       19  TotalCharges      7043 non-null   object
       20  Churn             7043 non-null   object
      dtypes: float64(1), int64(2), object(18)
      memory usage: 1.1+ MB
```

Fig-4: Info() of the Dataset.

Explanation: Displays the number of Non-Null rows, Data types of columns of the Dataset**.**

```
[11]: # Checking the data types of all the columns
      telco_base_data.dtypes
```

```
[11]: customerID          object
      gender              object
      SeniorCitizen        int64
      Partner             object
      Dependents          object
      tenure               int64
      PhoneService        object
      MultipleLines       object
      InternetService     object
      OnlineSecurity      object
      OnlineBackup        object
      DeviceProtection    object
      TechSupport         object
      StreamingTV         object
      StreamingMovies     object
      Contract            object
      PaperlessBilling    object
      PaymentMethod       object
      MonthlyCharges     float64
      TotalCharges        object
      Churn               object
      dtype: object
```

Fig-5: Columns of the Dataset.

Explanation: Displays the Data types of columns of the Dataset**.**

```
# Check the descriptive statistics of numeric variables
telco_base_data.describe()
```

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

SeniorCitizen is actually a categorical hence the 25%-50%-75% distribution is not propoer

75% customers have tenure less than 55 months

Average Monthly charges are USD 64.76 whereas 25% customers pay more than USD 89.85 per month

Fig-6: describe() of the Dataset.

Explanation: Summarize the Statistics of the Dataset**.**

```
telco_base_data['Churn'].value_counts().plot(kind='barh', figsize=(6, 4))
plt.xlabel("Count", labelpad=14)
plt.ylabel("Target Variable", labelpad=14)
plt.title("Count of TARGET Variable per category", y=1.02);
```

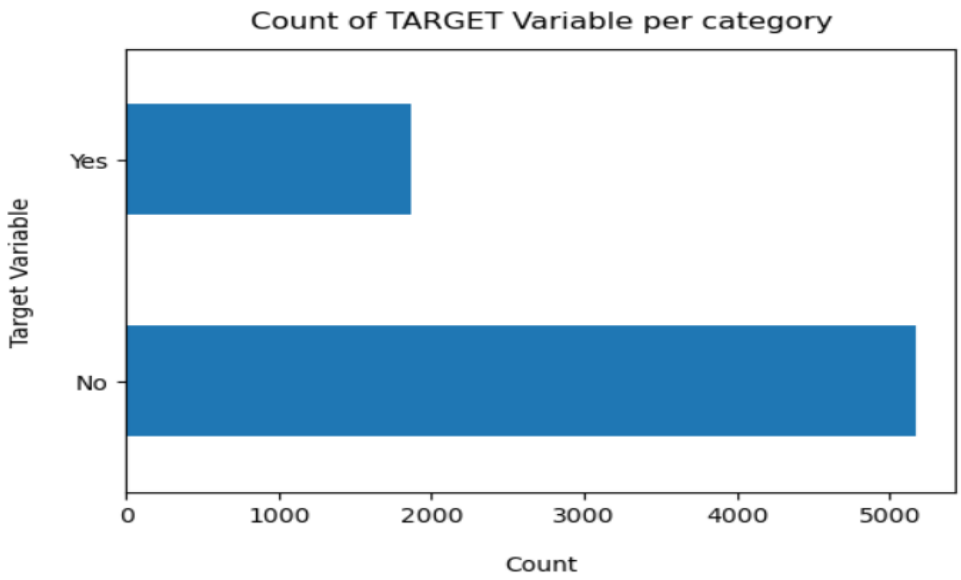


Fig-7: Bar plot.

Explanation: Visualizing the count with bar plot**.**

```
telco_data.TotalCharges = pd.to_numeric(telco_data.TotalCharges, errors='coerce')
telco_data.isnull().sum()
```

```
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       11
Churn               0
dtype: int64
```

Fig-7: Detecting the missing/null values.

Explanation: Displays the percentage of the null values in each column**.**

```
In [22]: # Drop rows where 'Postal code' is null, but keep other rows with nulls in other columns
         df = df.dropna(subset=['Postal Code'])

In [23]: #check whether null values are deleted or not
         print("Null values in Postal Code :",df['Postal Code'].isna().sum())
         print("No.of rows reduced to:",df.shape[0])

         Null values in Postal Code : 0
         No.of rows reduced to: 9983
```

Fig-8: Dropping the rows with missing/null values.

Explanation: Displays the non-null values after dropping it**.**

```
121]: plt.figure(figsize=(16,8))
      telco_data_dummies.corr()['Churn'].sort_values(ascending = False).plot(kind='bar')

121]: <Axes: >
```

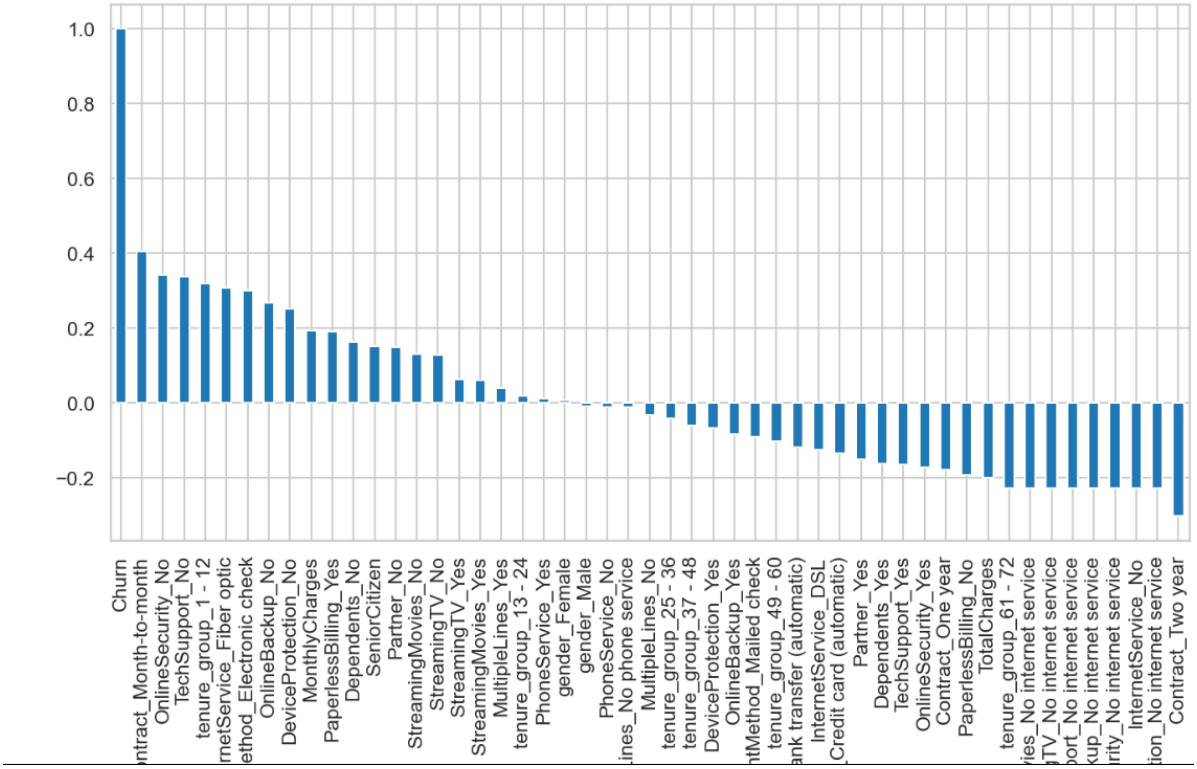


Fig-9: correlation between the variables.

```
plt.figure(figsize=(6,6))
sns.heatmap(telco_data_dummies.corr(), cmap="Paired")
```

```
<Axes: >
```

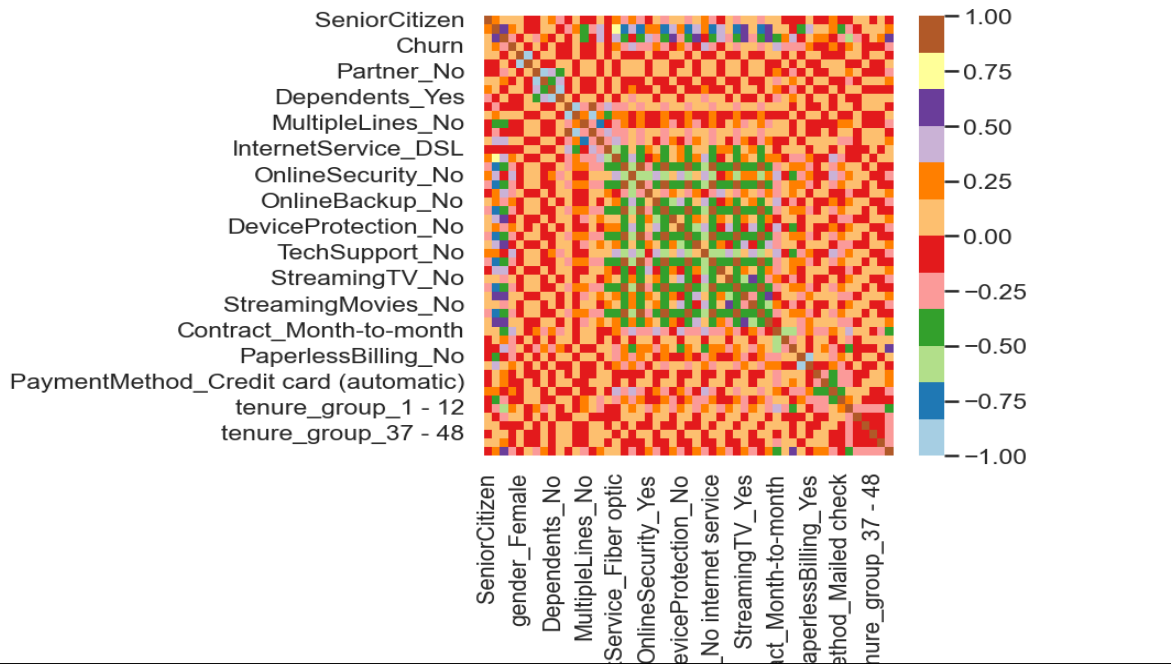
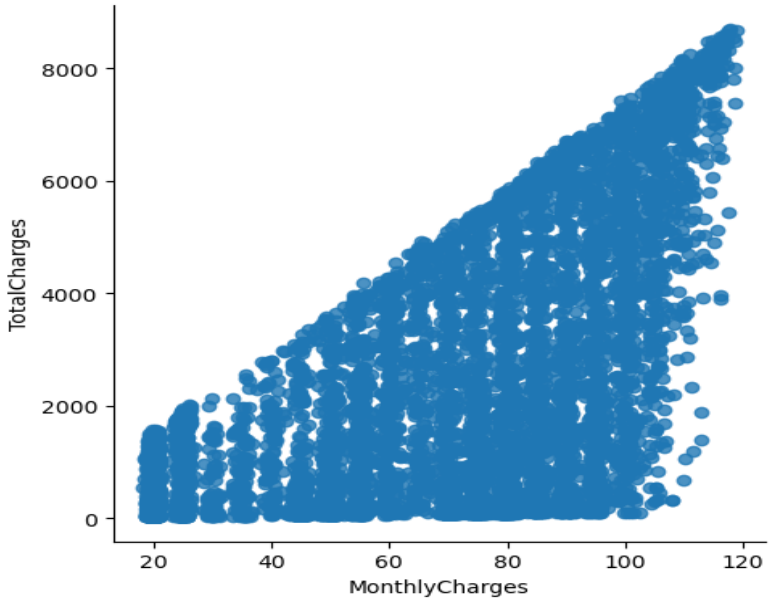


Fig-10: HeatMap between the variables.

```
sns.lmplot(data=telco_data_dummies, x='MonthlyCharges', y='TotalCharges', fit_reg=False)
```

```
<seaborn.axisgrid.FacetGrid at 0x24e9c7de450>
```



Total Charges increase as Monthly Charges increase - as expected.

Fig-11 : Scatter plot of Total charge vs Monthly charge

```
for i, predictor in enumerate(telco_data.drop(columns=['Churn', 'TotalCharges', 'MonthlyCharges'])):
    plt.figure(i)
    sns.countplot(data=telco_data, x=predictor, hue='Churn')
```

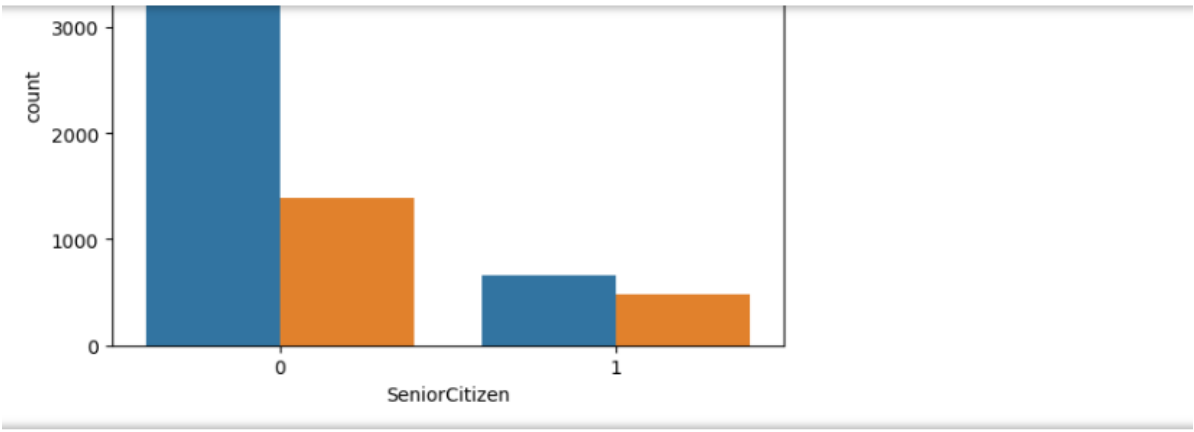
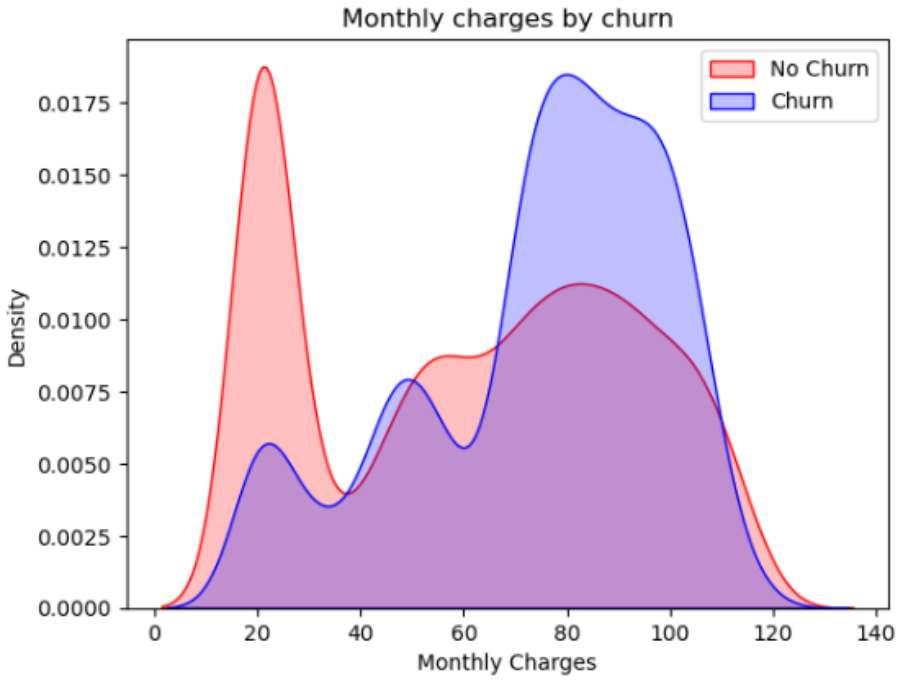


Fig-12: Count plot between Churn, Total Charges, Monthly Charges.

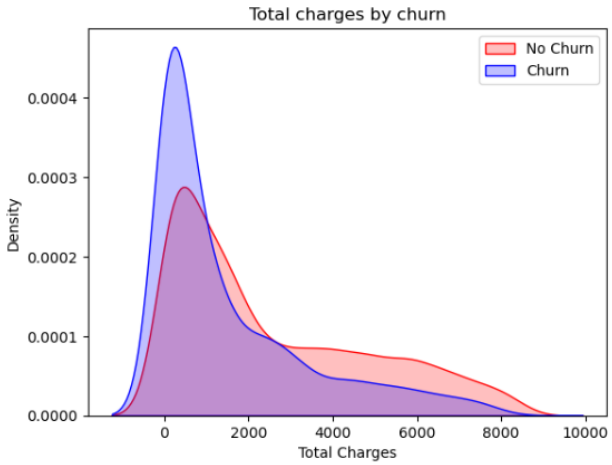
```
[78]: Text(0.5, 1.0, 'Monthly charges by churn')
```



Insight: Churn is high when Monthly Charges ar high

Fig-13: KDE plot

[80]: Text(0.5, 1.0, 'Total charges by churn')



**Surprising insight ** as higher Churn at lower Total Charges

However if we combine the insights of 3 parameters i.e. Tenure, Monthly Charges & Total Charges then the picture is bit clear :- Higher Monthly Charge at lower tenure results into lower Total Charge. Hence, all these 3 factors viz **Higher Monthly Charge**, **Lower tenure** and **Lower Total Charge** are linkd to **High Churn**.

Fig-14: KDE plot

7.ANALYSIS

The analysis of the customer churn dataset is centered on understanding customer behavior and identifying patterns that contribute to churn, using exploratory data analysis (EDA), statistical techniques, clustering, and machine learning models. By examining variables such as tenure, monthly charges, contract type, service subscriptions, and payment methods alongside demographic features like senior citizen status and gender, key insights were derived to support proactive retention strategies and improve business performance.

The dataset reveals significant trends in churn behavior across different customer profiles. Customer segmentation, achieved through clustering techniques such as K-Means, highlighted distinct patterns in engagement and churn risk. High-risk customer clusters were characterized by short tenure, month-to-month contracts, and reliance on electronic check payments—traits commonly linked to dissatisfaction and churn. These segments were further analyzed to understand defining attributes such as service usage patterns, support subscriptions, and billing preferences. This segmentation enables businesses to develop personalized retention strategies tailored to the needs and vulnerabilities of each segment.

The relationship between customer churn and service features was a critical area of focus. Correlation and feature importance analyses revealed that customers lacking value-added services such as tech support or online security were more likely to churn. For example, the absence of internet service or streaming features was a common characteristic of churned customers, suggesting that increasing customer engagement through bundled services could help reduce attrition. Similarly, churn rates were highest among those on month-to-month contracts, reinforcing the value of promoting longer-term commitments through incentives and loyalty programs.

Temporal patterns in churn behavior provided further insights into customer lifecycles. Tenure analysis showed that churn was most frequent within the first few months of

service, indicating a need for improved onboarding experiences and early engagement. Customers with longer tenure exhibited greater loyalty, suggesting that investments in customer relationship management during early phases could yield long-term retention benefits. This analysis supports the development of lifecycle-based interventions designed to reduce early-stage churn.

The impact of billing preferences and contract types on churn was also examined. Analysis revealed that customers using electronic checks as a payment method had the highest churn rates, likely due to dissatisfaction with payment flexibility or service reliability. In contrast, those using credit cards or bank transfers showed higher retention. These findings point to the importance of offering flexible, modern billing options as part of customer experience improvements. Contract type also played a crucial role, with one- and two-year contracts associated with lower churn, highlighting opportunities to offer value-based long-term plans.

Feature importance analysis from machine learning models—specifically Random Forest and XGBoost—identified tenure, contract type, monthly charges, and service count as the strongest predictors of churn. This aligns with observed behavioral trends and enables businesses to prioritize these variables in retention-focused strategies. For example, customers with high monthly charges but limited services are more likely to churn, suggesting a mismatch between pricing and perceived value.

Geographic and demographic factors also influenced churn behavior. While gender showed no significant correlation, senior citizens were slightly more prone to churn, possibly due to usability or support challenges. Businesses targeting this segment could consider tailored communication, simplified interfaces, or dedicated support services to improve satisfaction and retention.

Clustering and dimensionality reduction using Principal Component Analysis (PCA) provided a clear visualization of customer groups in 2D space. These clusters exhibited minimal overlap, validating the segmentation approach. Each group showed unique churn tendencies—some highly sensitive to pricing and others influenced by contract flexibility—enabling targeted communication and intervention strategies.

The insights derived from this analysis culminated in several actionable recommendations. Businesses should focus on high-risk customer segments identified through clustering and modeling, introduce personalized engagement programs in the early tenure phase, and offer bundled service plans to increase customer value perception. Additionally, billing systems should support modern payment methods, and contract incentives should be strategically deployed to encourage long-term commitments. Improved technical support and service quality for disengaged or underserved customers can further reduce churn and enhance loyalty.

This comprehensive analysis not only uncovers the key drivers of customer churn but also provides a strong foundation for predictive modeling and data-driven decision-making. By integrating machine learning insights with customer segmentation and behavioral trends, businesses can proactively mitigate churn, enhance customer satisfaction, and ensure sustainable growth in a competitive, subscription-driven market

8.CONCLUSION

This project provided an in-depth analysis of customer churn behavior in a subscription-based telecom environment, utilizing a comprehensive dataset comprising customer demographics, account information, service usage, contract types, billing methods, and churn status. The core objective was to uncover key factors influencing customer churn and to develop a predictive framework that enables businesses to proactively identify at-risk customers and implement effective retention strategies.

The analytical process began with thorough data exploration and cleaning, ensuring data integrity by addressing missing values, correcting data types, and detecting outliers. Exploratory Data Analysis (EDA) revealed several trends—such as higher churn among customers on month-to-month contracts, short tenure periods, and those using electronic check payments. These initial findings provided foundational insights into customer dissatisfaction drivers and guided subsequent modeling efforts.

Feature engineering played a pivotal role in enriching the dataset with meaningful variables such as ServiceCount, AverageMonthlySpend, and tenure-based categories, which helped to capture the behavioral and financial attributes of customers. Standardization and normalization techniques were applied to prepare the data for machine learning algorithms, ensuring fair contribution from all numerical features.

Customer segmentation was performed using clustering techniques like K-Means, uncovering distinct groups of customers based on their engagement levels, usage patterns, and churn likelihood. High-risk segments were defined by low tenure, minimal services, and unfavorable billing methods, while low-risk, high-value customers exhibited long-term contracts and broad service utilization. These insights empower businesses to create tailored outreach strategies, design loyalty programs, and optimize resource allocation across customer groups.

To predict churn with precision, multiple machine learning models—Logistic Regression, Random Forest, and XGBoost—were trained and evaluated. Among these, XGBoost achieved the highest performance with an accuracy of 81.6% and the best balance between precision, recall, and F1-score. Feature importance analysis reinforced earlier observations from EDA, confirming that tenure, contract type, monthly charges, and support service availability were the most influential churn predictors.

The study also incorporated dimensionality reduction techniques like PCA to visualize customer segmentation and validate the separation of clusters. These visual insights further demonstrated the effectiveness of the segmentation approach and provided intuitive understanding of customer behavior patterns across different risk profiles.

Based on the insights drawn from the analysis, several strategic recommendations were proposed:

- **Targeted Retention Campaigns:** Focus on customers in short-term contracts and early-tenure phases with personalized incentives to extend contract duration and improve satisfaction.
- **Service Bundling & Upselling:** Promote value-added services like tech support or online security to increase engagement and reduce churn.
- **Billing Optimization:** Offer modern, flexible billing options and encourage electronic payment adoption through incentives to reduce churn related to inconvenient payment methods.
- **Lifecycle-based Interventions:** Develop onboarding programs and early engagement strategies to support new customers during the critical first few months.
- **Automated Churn Monitoring:** Integrate predictive models into CRM systems to enable real-time churn prediction and trigger timely retention actions.

The outcomes of this project not only offer immediate, actionable strategies to reduce churn but also establish a solid framework for further development. Future work may involve integrating real-time data pipelines for continuous model updating, expanding feature sets with behavioral and sentiment analysis, and deploying the solution to cloud-based platforms for scalable enterprise use.

In conclusion, this project successfully demonstrated how a data-driven approach combining EDA, clustering, and machine learning can yield deep insights into customer churn dynamics. By identifying high-risk customers early and understanding the key factors influencing their decisions, businesses can enhance customer satisfaction, optimize operational strategies, and sustain long-term profitability in an increasingly competitive, subscription-driven marketplace.

.

9. REFERENCES:

1. **Kaggle** - A platform for datasets and data science projects. *https://www.kaggle.com*
2. **Towards Data Science** - Articles on data science techniques and tools. *https://towardsdatascience.com*
3. **Medium**: Analytics Vidhya - Insights on data science, machine learning, AI. *https://medium.com/analytics-vidhya*
4. **GeeksforGeeks** - Tutorials on data science, Python, and machine learning. *https://www.geeksforgeeks.org*
5. **DataCamp** - Learn data science with online courses and projects. *https://www.datacamp.com*
6. **Coursera** - Data science courses from top universities. *https://www.coursera.org*
7. **Statista** - Insights and statistics for e-commerce trends. *https://www.statista.com*
8. **Stack Overflow** - Community discussions on coding and data analysis. *https://stackoverflow.com*
9. **GitHub** - A repository for data science projects and open-source code. *https://github.com*
10. **Investopedia** - Insights into financial and sales data analysis. *https://www.investopedia.com*
11. **edX** - Data science and analytics courses from universities worldwide. *https://www.edx.org*
12. **UCI Machine Learning Repository** - Datasets for data science and machine learning. *https://archive.ics.uci.edu/ml/index.php*
14. **Analytics Vidhya** - Comprehensive guides on data science projects and tools. *https://www.analyticsvidhya.com*

15 **KDNuggets** - Articles and tutorials on big data, analytics, and machine learning. *https://www.kdnuggets.com*

10. GITHUB REPOSITORY LINK

Git hub link