**Name: Khurram Shahin**

**Registration Number:12112093**

**Customer Churn Prediction in a Telecommunications Company Project Report:**

**Table Of Contents**

* **Introduction**
* **Domain knowledge**
* **Libraries used and Approach.**
* **Data Description**
* **Data Cleaning**
* **Data Exploration**
* **Univariate Analysis**
* **Bivariate Analysis**
* **Findings and Insights**
* **Limitations**
* **Recommendations**
* **Model Prediction**
* **Model Evaluation**
* **Conclusion**

**1.Introduction**

The Telco Customer Churn dataset is a detailed compilation of data that meticulously records customer behaviors and interactions within a telecommunications company. This dataset includes a wealth of information such as customer demographics, service usage patterns, billing details, and churn indicators. It serves as an invaluable resource for analyzing customer retention and attrition trends, enabling profound insights that can guide targeted marketing strategies, improve customer satisfaction, and enhance predictive models for churn prevention.

**Problem Description**:

The analysis is cantered around a comprehensive dataset from a telecommunications company, capturing detailed customer information and service usage patterns. The primary objective is to gain insights into customer churn behaviour and its intricate relationships with various factors such as demographics, service plans, and billing details. This analysis has substantial implications for customer retention strategies, marketing efforts, and improving overall customer satisfaction within the telecommunications sector.

**Objectives of EDA:**

**Data Familiarization:** Begin by familiarizing yourself with the dataset, understanding its composition, and gaining a clear overview of the variables it contains.

**Data Cleaning**: Identify and address any data quality issues, including missing values, inconsistencies, or outliers, to ensure the dataset's integrity and reliability.

**Churn Distribution Analysis:** Investigate the distribution of churn within the customer base and determine any underlying patterns or trends that may influence customer churn rates.

**Service Usage Analysis:** Explore the relationship between different service features (such as Internet Service, Phone Service, Streaming TV) and churn, identifying which services are associated with higher or lower churn rates.

**Contract Type Analysis:** Examine the impact of different contract types (Month-to-month, One year, Two year) on customer churn. Identify any trends that suggest certain contract types are more stable.

**Demographic Disparities:** Analyses the churn rates across different demographic groups, such as gender, senior citizen status, and dependents, to understand how demographic factors influence churn.

**Billing and Payment Analysis:** Investigate the relationship between billing methods, payment methods, and churn. Assess how different billing strategies and payment methods affect customer retention.

**Segmentation Opportunities:** Identify opportunities for customer segmentation based on demographics, service usage, or other relevant factors. Use these segments to develop targeted marketing and retention strategies.

**Data Visualization**: Create data visualizations, such as bar plots, histograms, and stacked bar charts, to present your findings effectively and facilitate data communication.

**Domain knowledge**

**Dataset Description:**

The dataset provides a comprehensive overview of customer data for a telecommunications company, detailing various attributes that influence customer behavior and churn rates. Key attributes include:

**Customer ID:** A unique identifier for each customer.

Gender: Gender of the customer, allowing for analysis of gender-based patterns in churn and service usage.

**Senior Citizen**: Indicates whether the customer is a senior citizen, enabling the study of age-related trends.

Partner: Whether the customer has a partner, useful for demographic segmentation.

**Dependents:** Indicates if the customer has dependents, providing insights into family status and its impact on service preferences.

**Tenure:** The number of months the customer has been with the company, critical for understanding customer loyalty and retention.

**Phone Service**: Whether the customer has a phone service, aiding in the analysis of service adoption.

**Multiple Lines:** Indicates if the customer has multiple lines, relevant for understanding the complexity of service usage.

**Internet Service**: The type of internet service the customer has (DSL, Fiber optic, etc.), essential for analyzing service preferences.

**Online Security:** Whether the customer has online security service, helping to identify the adoption of additional services.

**Online Backup:** Indicates if the customer has online backup service, useful for assessing service penetration.

**Device Protection:** Whether the customer has device protection service, providing insights into value-added service adoption.

**Tech Support:** Indicates if the customer has tech support service, relevant for customer support analysis.

**Streaming TV:** Whether the customer uses streaming TV services, critical for understanding entertainment service usage.

**Streaming Movies:** Indicates if the customer uses streaming movie services, providing insights into media consumption.

**Contract:** The type of contract the customer has (Month-to-month, One year, Two year), crucial for analyzing contract-related churn patterns.

**Paperless Billing:** Whether the customer uses paperless billing, relevant for digital service adoption analysis.

**Payment Method**: The payment method used by the customer, aiding in the study of payment preferences and their impact on churn.

**Monthly Charges:** The monthly charges incurred by the customer, essential for revenue analysis.

**Total Charges:** The total charges incurred by the customer to date, providing a cumulative financial perspective.

**Churn:** Whether the customer has churned, the primary variable of interest for retention analysis.

**Libraries used and approaches.**

**Libraries Used:**

* Pandas: For data manipulation and analysis.
* NumPy: For numerical operations and array handling.
* Matplotlib: For data visualization.
* Seaborn: For advanced data visualization.
* Scikit-learn: For machine learning algorithms and preprocessing.
* Stats models: For statistical modeling**.**

**Data Familiarization:**

Load the dataset and understand its structure using pandas.

Inspect data types, summary statistics, and initial data points using functions like info(), describe(), and head().

**Data Cleaning:**

Check for and handle missing values using isnull(), fillna(), and dropna().

Identify and address inconsistencies or outliers using statistical methods or visual inspection.

Convert data types if necessary to ensure compatibility for analysis.

**Exploratory Data Analysis (EDA):**

Churn Distribution Analysis: Calculate churn rates using value counts().

Visualize churn distribution using bar plots.

Contract Type and Churn Analysis: Analyze the relationship between contract types and churn using group by operations.

Visualize with stacked bar charts to show the proportion of churn across different contract types.

Seniority Level and Churn Analysis: Compare churn rates between senior citizens and non-senior citizens using group by and visualization techniques.

Service Usage and Churn Analysis: Examine how different services impact churn using cross-tabulations and visualizations.

Payment Methods and Churn Analysis: Analyze the impact of different payment methods on churn.

Visualize using bar charts or pie charts.

Feature Scaling:

Normalize the numerical features to ensure all variables are within a range of 0 to 1 using Min Max Scaler or Standard Scaler from scikit-learn.

**Machine Learning Models**

**Logistic Regression:**

Fit a logistic regression model using Logistic Regression from scikit-learn.

Evaluate feature importance and model accuracy.

**Random Forest:**

Train a Random Forest classifier using Random Forest Classifier.

Determine feature importance and compare with logistic regression results.

**Support Vector Machine (SVM):**

Implement an SVM model using SVC.

Analyze accuracy, true positive, and true negative rates, and the AUC (Area Under the Curve).

Evaluation and Metrics:

Use confusion matrix, accuracy score, precision, recall, F1-score, and ROC-AUC for model evaluation.

Compare models to select the best-performing one based on these metrics.

**Visualization:**

Create various plots using matplotlib and seaborn to present the findings effectively.

Visualize distributions, relationships, and model performance to facilitate data communication and insights extraction.

**Data Exploration**

**Feature Engineering:**

Feature engineering is a critical step in the data analysis process, where new features are created from the existing data to improve model performance and gain deeper insights. In this section, we have engineered features to understand the demographics of senior citizens and analyze the relationships between partners and dependents.

**1. Percentage of Male and Female Senior Citizens**

We start by calculating the percentage of male and female senior citizens within the dataset. This helps in understanding the demographic distribution and can also highlight any potential biases or trends in the data.

**2. Customers with and without Dependents and Partners**

Next, we analyze the number of customers who have dependents and partners. This analysis helps in understanding the support structure of the customers, which can be crucial for targeted marketing strategies and customer support.

**Univariate Analysis:**

I have Categorized Univariate Analysis In two types First For Categorical Data And other For Numerical Data

Let’s start with Categorical Data

1. Univariate Analysis on Multiple Lines column:

After Analyzing we see most of the users uses single line type Distribution and some users Does not use any kind of phone services

2. Univariate analysis on contract with customer Column

After Analysing it discovered that most of the customers are in the month-to-month contract. While there are equal number of customers in the 1 year and 2-year contracts.

3. Univariate analysis on Payment Method

After Analysing it is discovered that most of the customer uses electronic check to pay for their services and remaining three categories of customers are approx. Identical in numbers

4. Univariate analysis on Gender Distribution

After Analysing it is discovered that 49.5 percent of customers are female, and 50.5 percent of customers are male

5.Univariate Analysis on Senior Citizen Distribution

After Analysing it is discovered that about 16 percent of customer population are senior citizen and reast are young ones

I have done some feature engineering and finds out the number of percentages of male and female senior citizen and finds out the number of male and female senior citizens are same in number

6. Univariate Analysis on customers

After Analysing it is discovered that About 50% of the customers have a partner, while only 30% of the total customers have dependents

It also shows that among the customers who have a partner, only about half of them also have a dependent, while other half do not have any independents. Additionally, as expected, among the customers who do not have any partner, a majority (80%) of them do not have any dependents.

**Bivariate Analysis**

As we see Our Target variable in to predict the churn variable let’s find relationship of different columns with Churn

Let’s First check the distribution of churn rate

As we can See In our data, approx 74% of the customers do not churn. Clearly the data is skewed as we would expect a large majority of the customers to not churn. This is important to keep in mind for our modelling as skewness could lead to a lot of false negatives. We will see in the modelling section on how to avoid skewness in the data.

# **2. Analysis on Churn vs Tenure:**

After Analysing form the below plot, the customers who do not churn, they tend to stay for a longer tenure with the telecom company.

**3.Bivariate Analysis on Churn vs Contract Type:**

After Analysing we saw in the correlation plot, the customers who have a month-to-month contract have a very high churn rate

**4.Bivariate Analysis on Churn vs seniority:**

After Analysing we saw Senior Citizens have almost double the churn rate than younger population.

**5.Bivariate Analysis on Churn vs Monthly Charges:**

After Analysing we saw the chances of customer churn is high when the monthly charges are high

# Now we Have COMPLETED THE EDA PART AND GAINED THE INSIGHTS ITS TIME TO DEVELOP SOME PREDICTIVE MODELS AND COMPARE ITS ACCURACY

**Machine Learning Model**

**1.Logistic Regression**

The reason behind choosing logistics regression for churn prediction is due to its suitability for binary classification, interpretability, efficiency, and ability to provide probability outputs, making it a practical and insightful model for understanding and predicting customer churn.

After applying Logistic Regression, We found

**Results Summary:**

The logistic regression model's performance on the test set is summarized below:

Accuracy: 80%

Precision: 0.70

Recall: 0.60

F1 Score: 0.64

ROC-AUC Score: 0.76

**Observations:**

Accuracy: The model correctly predicted 80% of the instances, indicating a strong overall performance.

Precision and Recall:

For class 0 (No Churn), the precision and recall are relatively high at 0.86 and 0.88, respectively. This indicates the model is effective at identifying non-churners.

For class 1 (Churn), the precision and recall are lower at 0.64 and 0.60, respectively. This suggests there is room for improvement in identifying churners.

F1 Score: The F1 score for churners is 0.62, reflecting a balanced trade-off between precision and recall. The F1 score for non-churners is higher at 0.87.

ROC-AUC Score: The ROC-AUC score of 0.76 indicates that the model has a good ability to discriminate between churners and non-churners.

Scaling the variables in logistic regression is a pivotal step that can significantly enhance model accuracy. By ensuring that all variables are within a standardized range of 0 to 1, the model becomes more robust and better equipped to handle diverse data inputs. In my experience, this transformation led to a noticeable improvement, elevating accuracy from 79.7% to 80.7%.

Additionally, the scaling process offers a coherent alignment between variable importance as observed in Random Forest algorithms and the insights gleaned from exploratory data analysis (EDA). This synchronization underscores the effectiveness of scaling in logistic regression, as it enables the model to leverage the most influential features accurately, ultimately contributing to enhanced predictive performance

**2.Random Forest Classifier**

The insights extracted from the random forest algorithm shed light on key predictors of churn: monthly contract, tenure, and total charges emerge as highly influential variables. These findings closely mirror the outcomes of logistic regression and are in perfect harmony with the expectations we derived from our exploratory data analysis (EDA). This consistency across different methodologies underscores the reliability and robustness of our predictive models, reinforcing our confidence in their ability to accurately forecast churn patterns.

After applying Random Forest , We found

**Results Summary:**

The Random Forest model's performance on the test set is summarized below:

Results Summary:

Accuracy: 81%

Macro Average Precision: 0.76

Macro Average Recall: 0.72

Macro Average F1 Score: 0.73

Weighted Average Precision: 0.80

Weighted Average Recall: 0.81

Weighted Average F1 Score: 0.80

3.Support Vector Machine

After applying Support Vector Machine Algorithm, We found

**Results Summary:**

Support Vector Machine Algorithm performance on the test set is summarized below

Accuracy: 82%

Macro Average Precision: 0.78

Macro Average Recall: 0.74

Macro Average F1 Score: 0.76

Weighted Average Precision: 0.81

Weighted Average Recall: 0.81

Weighted Average F1 Score: 0.80

In this journey of refining our churn prediction model, we've witnessed a significant enhancement in accuracy, particularly with the SVM algorithm, achieving an impressive 82% accuracy rate. However, accuracy alone doesn't provide a complete picture of model performance. Therefore, our next steps involve a thorough examination of metrics such as true positive and true negative rates, as well as calculating the Area Under the Curve (AUC) to gain a deeper understanding of our model's predictive power.

By delving into these metrics, we aim to uncover insights that go beyond simple accuracy figures. This comprehensive analysis will allow us to make informed decisions about our model's strengths and areas for improvement, ultimately leading to more robust and reliable churn predictions. Stay tuned for updates as we continue to refine and optimize our predictive model for even better performance!

​