

TITLE: CUSTOMER CHURN PREDICTION IN TELECOM DATA

Submitted By: Madireddy Bharath Kumar Reddy

Reg No: 12107901

**TABLE OF CONTENTS**

CONTENTS PAGE NO

Title page 1

Table of Contents 2

Introduction 3

Objective & Scope 4

Data Loading & Data Preprocessing 9

Univariate, Bivariate & Multivariate Analysis 12

Feature Engineering 13

Model Building 16

Conclusion 17

Additional Insights from Power BI Dashboard 18

**INTRODUCTION:**

**What is Customer Churn?**

Customer churn refers to the percentage of customers that stop using a company's services during a particular time frame. In the context of the telecommunication industry, churn means the loss of customers due to various reasons such as exiting offers from competitors or network issues. When customers decide to cancel their subscription to a service, it leads to a direct impact on the company's revenue and the length of service.

Churn rate is a critical metric for telecom companies as it affects the lifetime value of a customer. A high churn rate implies that the company is losing customers at a faster rate, which can have negative implications on the company's future revenue and profitability.

To address this issue and minimize churn rate, telecom companies often employ predictive modelling techniques using machine learning algorithms. By predicting which customers are likely to churn, companies can proactively offer better services or incentives to retain them, thereby improving customer satisfaction and increasing profitability.

**Different Churn Scenarios:**

Subscriber churn in the telecom industry manifests in various forms, not limited to customers exiting the service entirely. These different churn scenarios include:

1. **Tariff Plan Churn:** When customers downgrade their plans, such as moving from a $50 monthly plan to a $30 plan, leading to reduced revenue for the provider.
2. **Subscription Churn:** Customers cancel specific subscriptions, such as weekly or monthly add-ons, impacting overall service usage.
3. **Product Churn:** Changes in product types, such as switching from postpaid to prepaid services, which may affect revenue predictability and customer engagement.
4. **Usage Churn:** Instances where customers become inactive or exhibit zero usage, indicating a decline in engagement with the service.
5. **Subscriber Churn:** The most critical form, where customers port out to competitors, leading to a direct loss of market share.

To address this issue and minimize churn rate, telecom companies often employ predictive modeling techniques using machine learning algorithms. By predicting which customers are likely to churn, companies can proactively offer better services or incentives to retain them, thereby improving customer satisfaction and increasing profitability.This report delves into the various aspects of customer churn, exploring the key factors influencing churn, the effectiveness of predictive models in identifying at-risk customers, and the strategies that can be employed to enhance customer retention.

**Objective:**

The objective of this project was to analyze customer data to identify patterns and build a model that predicts churn. This will enable the company to take proactive measures to retain customers.

**Scope:**

The project involved data cleaning, exploratory data analysis (EDA), feature engineering, model building, evaluation, and deriving insights and recommendations.

**Data Loading and Preprocessing:**

**A. Data Source:**

The dataset utilized for this project was obtained from Kaggle. This dataset consisted of 7,043 records and comprised 21 columns, encompassing a wide range of information related to customer demographics, account particulars, and service utilization data.

**B. Loading the Dataset**:

loaded the dataset into environment (Jupyter Notebook), using a programming language Python with library pandas.

**C. Dataset Overview:**

**1.Dataset Dimensions**

The dataset contains 7043 entries (rows) and 21 columns.

**2.Dataset Structure**

An initial inspection using df.info() provides the following details:

* The dataset includes a mix of numeric and categorical variables.
* There are no missing values across the columns, with each column having 7043 non-null entries.

**3.Data Types of Columns**

The dataset includes the following columns with their respective data types:

* 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 (to be converted to numeric)
* Churn: object
* 4.Descriptive Statistics

A summary of the numeric columns is provided below:

SeniorCitizen:

* Mean: 0.162
* Standard Deviation: 0.369
* Most customers are not senior citizens (mean close to 0).

tenure:

* Mean: 32.37 months
* Standard Deviation: 24.56 months
* Tenure varies widely, indicating a diverse customer base in terms of how long they have been with the company.
* 25% of customers have a tenure of 9 months or less, 50% have a tenure of 29 months or less, and 75% have a tenure of 55 months or less.

MonthlyCharges:

* Mean: $64.76
* Standard Deviation: $30.09
* Monthly charges range from $18.25 to $118.75.
* 25% of customers pay $35.50 or less per month, 50% pay $70.35 or less, and 75% pay $89.85 or less.

**5.Unique Values in Categorical Columns:**

* customerID: Unique identifiers for each customer.
* gender: ['Female', 'Male']
* Partner: ['Yes', 'No']
* Dependents: ['No', 'Yes']
* PhoneService: ['No', 'Yes']
* MultipleLines: ['No phone service', 'No', 'Yes']
* InternetService: ['DSL', 'Fiber optic', 'No']
* OnlineSecurity: ['No', 'Yes', 'No internet service']
* OnlineBackup: ['Yes', 'No', 'No internet service']
* DeviceProtection: ['No', 'Yes', 'No internet service']
* TechSupport: ['No', 'Yes', 'No internet service']
* StreamingTV: ['No', 'Yes', 'No internet service']
* StreamingMovies: ['No', 'Yes', 'No internet service']
* Contract: ['Month-to-month', 'One year', 'Two year']
* PaperlessBilling: ['Yes', 'No']
* PaymentMethod: ['Electronic check', 'Mailed check', 'Bank transfer (automatic)', 'Credit card (automatic)']
* TotalCharges: ['29.85', '1889.5', '108.15', ..., '346.45', '306.6', '6844.5']
* Churn: ['No', 'Yes']

**D. Checking for Null Values:**

After loading the dataset, inspected it to identify any missing values (null values). Initially there are no null values in the dataset.

**E. Changing Data Type of a Column:**

Converted the 'totalcharges' column from an object data type to numeric using pd.to\_numeric(). As it has to be numeric value.

**G. Checking for Null Values Again**:

After the data type conversion, rechecked for null values and found 11 null values in the 'totalcharges' column, which is a very small percentage (0.15%) of the total dataset. Since the percentage of null values is low, it is safe to ignore them from further processing.

**H. Creating Tenure Group Column**:

Created a new column called 'tenure\_group' based on the 'tenure' column. This new column categorizes customers based on their tenure (length of time they've been a customer). This helps in understanding customer behaviour and identifying patterns related to churn across different tenure periods.

1 – 12, 13 – 24, 25 – 36, 37 – 48, 49 – 60, 61 – 72

**I. Removing Unnecessary Columns**:

Removed the 'tenure' and 'customer id' columns from the dataset, because they were no longer needed for analysis or modelling.

**J. Churn Distribution**

Firstly, examined the distribution of the target variable, Churn, to understand its balance within the dataset.

Churn Count:

* No: 5163
* Yes: 1869

Churn Percentage:

* No: 73.46%
* Yes: 26.54%

The data is highly imbalanced with a ratio of approximately 73:27.

**K. Churn Analysis by Various Features**

To get deeper insights, analysed the churn distribution across different features.

**Senior Citizen:**

Churned:

* Not Senior Citizens: 1393
* Senior Citizens: 476

Not Churned:

* Not Senior Citizens: 4508
* Senior Citizens: 666

Insights:

* A higher number of non-senior citizens tend to churn compared to senior citizens. However, the ratio of senior citizens who churn (476 out of 1142) is relatively high compared to non-senior citizens.
* Senior citizens seem to be more loyal, with a lower absolute count of churn despite their smaller representation in the dataset.

**Gender:**

Churned:

* Female: 939
* Male: 930

Not Churned:

* Female: 2549
* Male: 2625

Insights:

* Churn rates are almost equal between genders, suggesting that gender is not a significant predictor of churn.
* Both male and female customers have nearly equal representation in terms of churn, indicating similar satisfaction or dissatisfaction levels.

**Partner:**

Churned:

* No: 1200
* Yes: 669

Not Churned:

* No: 2441
* Yes: 2733

Insights:

* Customers without partners are more likely to churn compared to those with partners.
* Having a partner seems to be a stabilizing factor, reducing the likelihood of churn.

**Dependents:**

Churned:

* No: 1543
* Yes: 326

Not Churned:

* No: 3390
* Yes: 1784

Insights:

* Customers without dependents are more likely to churn compared to those with dependents.
* The presence of dependents appears to be associated with customer retention, possibly due to greater service needs.

**Contract:**

Churned:

* Month-to-month: 1543
* One year: 326

Not Churned:

* Month-to-month: 3390
* One year: 1784

Insights:

* Customers with month-to-month contracts are more likely to churn compared to those with longer-term contracts.
* Longer-term contracts (one year or more) are associated with higher customer retention, suggesting that commitment reduces churn.

**Tenure Group:**

Churned:

* 1 - 12: 1037
* 13 - 24: 294
* 25 - 36: 180
* 37 - 48: 145
* 49 - 60: 120
* 61 - 72: 93

Not Churned:

* 1 - 12: 1138
* 13 - 24: 730
* 25 - 36: 652
* 37 - 48: 617
* 49 - 60: 712
* 61 - 72: 1314

Insights:

* Customers with shorter tenures (1 - 12 months) have the highest churn rate, indicating that new customers are at a higher risk of churn.
* As tenure increases, the likelihood of churn decreases significantly, suggesting that longer-tenured customers are more loyal and satisfied with the service.

**Univariate Analysis**

In the univariate analysis, explored individual variables to understand their distribution and characteristics without considering the relationships with other variables.

1. Count of Churn

Started by examining the count of churn within the dataset. The countplot of churn revealed that the number of churned customers ('Yes') is lower compared to non-churned customers ('No'). This indicates an imbalanced dataset with more loyal customers than churned ones.

2. Count of Gender

Next, investigated the count of customers based on their gender. The countplot illustrated that the number of male and female customers is approximately equal, suggesting a balanced gender distribution in the dataset.

3. Count of Contract

We then analyzed the count of customers based on their contract type. The countplot of contract types revealed that most customers have month-to-month contracts, followed by two-year contracts, and then one-year contracts.

4. Count of Senior Citizen

The countplot of senior citizen status showed that the majority of customers are not senior citizens (SeniorCitizen = 0) in the dataset, with a smaller count of senior citizen customers.

5. Count of Tenure Group

Further, explored the count of customers grouped by their tenure. The analysis revealed the distribution of customers across different tenure groups, providing insights into the length of time customers have been with the company.

6. Box Plot for Numerical Columns

Lastly, utilized box plots to visualize the distribution of numerical columns such as monthly charges and total charges. These box plots provided information about the central tendency, variability, and potential outliers in the numerical data.

For MonthlyCharges:

* The lower whisker indicates a minimum value of $18.25, and the upper whisker extends to $118.75.
* The interquartile range (IQR), which represents the middle 50% of the data, spans from $35.59 (Q1) to $89.86 (Q3), with a width of $54.28.
* The median monthly charge is $70.35, indicating that half of the customers have a monthly charge below this value.
* The distribution of monthly charges shows no extreme outliers beyond the upper whisker, suggesting a relatively symmetrical spread of data.

For TotalCharges:

* The lower whisker indicates a minimum value of $18.80, and the upper whisker extends to $8684.80.
* The interquartile range (IQR) spans from $401.45 (Q1) to $3794.74 (Q3), with a width of $3393.29, indicating a wide spread of data.
* The median total charge is $1397.48, suggesting that half of the customers have a total charge below this value.
* The distribution of total charges exhibits positive skewness, with a longer tail towards higher values, as evidenced by the difference between the median and the upper quartile.

Overall, the univariate analysis offered valuable insights into the distribution and characteristics of individual variables within the dataset,

**Bivariate Analysis:**

Payment Medium and Churn:

* Insight: Customers using electronic checks as their payment method exhibited the highest churn rates.
* Implication: This suggests that the payment method may significantly influence customer retention, with electronic check users being more prone to churn compared to other payment methods.
* Visualization: Countplot showing churn rates across different payment mediums.

Contract Type and Churn:

* Insight: Monthly contract customers are more likely to churn due to the absence of contract terms, providing them with more flexibility.
* Implication: Implementing longer-term contracts or offering incentives for contract renewal may help reduce churn among monthly contract customers.
* Visualization: Countplot illustrating churn rates based on contract type.

Security and Tech Support Services:

* Insight: Customers without online security or tech support services experienced higher churn rates.
* Implication: Enhancing these services may improve customer satisfaction and retention, as lack of support may lead to dissatisfaction and prompt customers to switch providers.
* Visualization: Countplot depicting churn rates for customers with and without online security and tech support services.

Senior Citizens and Churn:

* Insight: Senior citizens were found to have higher churn rates compared to non-senior citizens.
* Implication: Tailoring retention strategies to address the unique needs and preferences of senior citizens may help mitigate churn within this demographic.
* Visualization: Countplot displaying churn rates among senior citizens and non-senior citizens.

Monthly Charges Analysis:

* Insight: Customers who churned had a higher median monthly charge compared to those who did not churn.
* Implication: Higher monthly charges may contribute to churn, indicating a need to review pricing strategies or offer additional value to justify costs and retain customers.
* Visualization: Box plot and KDE plot demonstrating the distribution of monthly charges for churned and non-churned customers.

By analyzing the relationship between churn and various factors, we can identify key drivers of customer attrition and develop targeted strategies to improve retention and enhance overall customer satisfaction.

**Multivariate Analysis:**

Monthly Charges by Contract Type and Churn:

* Insight: The bar plot illustrates the relationship between monthly charges, contract type, and churn. Customers with month-to-month contracts and higher monthly charges are more likely to churn.
* Implication: It suggests that customers on month-to-month contracts may be sensitive to pricing, and offering incentives or discounts on longer-term contracts could potentially reduce churn among this group.

Churn by Tenure Group and Senior Citizen Status:

* Insight: The catplot displays the count of churn across different tenure groups, segmented by senior citizen status. It shows that senior citizens tend to churn more across various tenure groups compared to non-senior citizens.
* Implication: Implementing targeted retention strategies for senior citizens, such as personalized support or loyalty programs, may be beneficial in reducing churn within this demographic.

Contracts Distribution by Tenure Group and Churn:

* Insight: The count plot illustrates the distribution of contract types across different tenure groups and churn status. It reveals that customers with longer tenure are more likely to have one-year contracts.
* Implication: Understanding the contract preferences of customers with different tenure lengths can help in tailoring contract renewal offers or promotions to improve retention.

Correlation Analysis:

* Insight: The correlation heatmap shows the correlation coefficients between senior citizen status, monthly charges, and total charges. Monthly charges have a moderate positive correlation with total charges, while senior citizen status has a weak positive correlation with both monthly and total charges.
* Implication: This indicates that there is some association between monthly charges, total charges, and senior citizen status, although the correlations are not very strong. Further analysis may be needed to explore these relationships in more detail.

Overall Insights:

* High Churn Factors: Month-to-month contracts, lack of online security and tech support, first-year subscribers, and fiber optic internet users are associated with higher churn rates.
* Low Churn Factors: Long-term contracts, subscriptions without internet service, and customers engaged for 5+ years are associated with lower churn rates.
* Neutral Factors: Gender, availability of phone service, and the number of multiple lines have minimal impact on churn.

By analyzing these multivariate relationships, we gain deeper insights into the interplay between different variables and their impact on churn, allowing us to develop more targeted strategies for customer retention and satisfaction.

**Feature Engineering:**

Conversion of Target Variable:

* Action: The target variable 'Churn' was converted into a binary numeric variable where 'Yes' is represented as 1 and 'No' as 0.
* Outcome: The binary representation facilitates modelling as machine learning algorithms typically require numeric inputs.

Creation of Dummy Variables:

* Action: All categorical variables were transformed into dummy variables using one-hot encoding.
* Outcome: This process expanded the dataset, creating binary columns for each category of the original categorical variables. These binary columns are more suitable for modelling purposes.

Exporting Updated Dataset:

* Action: The preprocessed dataset with dummy variables was saved as 'updated\_tele\_data.csv'.
* Outcome: The updated dataset is ready for further model building and analysis, ensuring seamless integration into subsequent stages of the project.

By preprocessing the data and engineering relevant features, the dataset is now optimized for modeling, allowing for the development of predictive models to forecast churn and inform strategic decision-making.

**Model Building for Churn Prediction:**

**Introduction:**

The objective of this analysis is to develop predictive models for churn prediction in a telecommunications dataset. Churn prediction is crucial for businesses to identify customers who are likely to discontinue their services, allowing proactive retention strategies to be implemented.

**Data Preprocessing:**

We conducted preprocessing tasks on a telecommunications dataset to prepare it for model building. Here are the key steps involved:

Handling Missing Values:

* Checked for missing values across all features.
* No missing values were found in the dataset, eliminating the need for imputation.

Encoding Categorical Variables:

* Converted the target variable 'Churn' into a binary numeric variable, where 'Yes' was encoded as 1 and 'No' as 0.
* Utilized one-hot encoding to convert categorical variables into binary numeric variables.

Splitting the Data:

* Split the dataset into features (X) and the target variable (y).
* Segregated the data into training and testing sets using the train\_test\_split function from sklearn.

These preprocessing steps ensured that the dataset was clean, properly formatted, and ready for further analysis and model building.

**Model Selection:**

For churn prediction, we opted for three distinct machine learning models, each with its own characteristics and advantages:

Decision Tree Classifier:

* Decision trees are powerful and interpretable models that recursively partition the feature space based on the most discriminative features.
* We trained a Decision Tree Classifier with hyperparameters such as maximum depth and minimum samples per leaf to control model complexity and prevent overfitting.

Random Forest Classifier:

* Random Forest is an ensemble learning method that constructs multiple decision trees during training and combines their predictions through voting or averaging.
* By aggregating predictions from multiple trees, Random Forest reduces overfitting and provides robust performance on unseen data.
* We utilized a Random Forest Classifier with parameters like the number of trees (n\_estimators), maximum depth of each tree, and minimum samples per leaf to fine-tune the model.

K Nearest Neighbors (KNN) Classifier:

* KNN is a simple yet effective algorithm that classifies data points based on the majority class of their nearest neighbors in the feature space.
* It is a non-parametric method that does not make strong assumptions about the underlying distribution of the data.
* We applied KNN with default hyperparameters and evaluated its performance on the dataset.

These models were chosen due to their suitability for binary classification tasks and their ability to handle non-linear relationships between features and the target variable. Additionally, each model offers different trade-offs in terms of interpretability, computational efficiency, and predictive accuracy, allowing us to explore various aspects of the dataset and select the most suitable model for churn prediction.

**Model Building:**

**Before Balancing the Dataset:**

Decision Tree Classifier:

* Trained the Decision Tree model with hyperparameters: max\_depth=6, min\_samples\_leaf=8.
* Achieved an accuracy of 79.96% on the test data.

Random Forest Classifier:

* Trained the Random Forest model with hyperparameters: n\_estimators=100, max\_depth=6, min\_samples\_leaf=8.
* Achieved an accuracy of 80.03% on the test data.

K Nearest Neighbors (KNN) Classifier:

* Trained the KNN model with default hyperparameters.
* Achieved an accuracy of 77.83% on the test data.

**After Balancing the Dataset Using SMOTEENN:**

Decision Tree Classifier (Balanced):

* Trained the Decision Tree model with hyperparameters: max\_depth=10, min\_samples\_leaf=1, min\_samples\_split=10.
* Achieved an accuracy of 93.76% on the test data.

Random Forest Classifier (Balanced):

* Trained the Random Forest model with hyperparameters: n\_estimators=300, max\_depth=None, min\_samples\_leaf=1, min\_samples\_split=2.
* Achieved an accuracy of 95.13% on the test data.

K Nearest Neighbors (KNN) Classifier (Balanced):

* Trained the KNN model with hyperparameters: n\_neighbors=3, metric='euclidean', weights='distance'.
* Achieved an accuracy of 97.95% on the test data.

**Results and Discussion:**

* Before balancing the dataset, all three models achieved reasonable accuracy scores: Decision Tree (79.96%), Random Forest (80.03%), and K Nearest Neighbors (77.83%).
* However, the models showed a notable disparity in recall scores, particularly for identifying churned customers. This indicates the influence of class imbalance on model performance.
* After balancing the dataset using SMOTEENN:

1. Decision Tree (Balanced) achieved an accuracy of 93.76% with improved recall (95.90%).
2. Random Forest (Balanced) further increased accuracy to 95.13%, with significantly enhanced recall (98.11%).
3. KNN (Balanced) also exhibited substantial improvement, achieving an accuracy of 97.95% with a recall of 98.74%.

* Balancing the dataset significantly improved model performance, particularly in correctly identifying churned customers (higher recall scores).
* These results highlight the importance of addressing class imbalance in predictive modeling tasks, as it can greatly impact the model's ability to identify minority classes.

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Model | Accuracy (Before) | Precision (Before) | Recall (Before) | F1-Score (Before) | Accuracy (After) | Precision (After) | Recall (After) | F1-Score (After) |
| Decision Tree Classifier | 79.96% | 68.09% | 46.67% | 55.38% | 93.76% | 92.82% | 95.90% | 94.34% |
| Random Forest Classifier | 80.03% | 69.11% | 45.33% | 54.75% | 95.13% | 93.25% | 98.11% | 95.62% |
| K Nearest Neighbors (KNN) | 77.83% | 61.54% | 44.80% | 51.85% | 97.95% | 97.51% | 98.74% | 98.12% |

**Conclusion:**

In conclusion, this project successfully addressed class imbalance in predicting customer churn using machine learning models. By implementing SMOTEENN and Hyperparameter Tuning, achieved significant improvements in model performance, particularly with the Random Forest Classifier. These results underscore the importance of addressing class imbalance for accurate churn prediction in telecommunications datasets, ultimately enhancing customer retention strategies and business outcomes.

**Additional Insights from Power BI Dashboard:**

**Recommendations for Customer Churn Reduction:**

After conducting thorough analysis and visualizing insights through the Power BI dashboard, several patterns and trends have emerged, providing valuable recommendations for reducing customer churn.

* **Month-to-Month Contract Type:** Customers subscribed to month-to-month contracts exhibit a higher likelihood of churning. It's imperative to focus on strategies to retain these customers, such as offering incentives for long-term contracts or personalized retention offers.
* **Tenure of 0-10 Months:** Customers with a tenure of 0-10 months are more prone to churning. Implementing targeted retention campaigns or providing exceptional service during the early stages of the customer lifecycle can mitigate churn risk.
* **Fiber Optic Internet Service:** Customers using fiber optic internet service show a higher churn rate. Investigating the reasons behind dissatisfaction with this service and addressing any underlying issues can help in retaining these customers.
* **Electronic Check Payment Method:** Customers using electronic check as their payment method are more likely to churn. Offering alternative payment options or streamlining the payment process can improve customer satisfaction and reduce churn.
* **Lack of Online Backup and Tech Support:** Customers without online backup and tech support services are more prone to churning. Highlighting the benefits of these services and providing proactive support can enhance customer loyalty and reduce churn.
* Absence of Device Protection: Customers without device protection are at a higher risk of churning. Offering device protection plans or bundling these services with existing offerings can add value to the customer experience and decrease churn likelihood.

Conclusion: Addressing these key areas highlighted in the Power BI dashboard can significantly reduce customer churn. By focusing on improving customer satisfaction, providing additional services, and implementing targeted retention strategies, the company can enhance customer retention rates and ensure long-term business success.

*“Additionally, as part of this project, I created a comprehensive Power BI dashboard to visualize and analyze the data, facilitating deeper insights into customer churn patterns and trends.”*

**Power BI Dashboard Link:**

<https://lpuin-my.sharepoint.com/:u:/g/personal/bharath_12107901_lpu_in/Ed0-I8zBtSxPnFGxf28Y8SQBGlHIQ3Wx1xO4DM0hoXJbFQ>

**----Thank you----**