CUSTOMER CHURN PREDICTION

Submitted By:

Aman Vishwakarma
12113404
Lovely Professional University

INTRODUCTION

The goal of this study is to leverage a telecoms company's dataset to anticipate client attrition. Our goal is to develop strong machine learning models through the examination of diverse consumer demographics, account details, and service utilization trends. These models will assist in identifying clients who are likely to leave, allowing the business to create focused retention plans and raise client satisfaction levels all around.

OBJECTIVE:

The primary objective of this project is to develop a predictive model that can identify customers at risk of churning, enabling the company to take proactive measures to retain them.

TABLE OF CONTENTS

DATASET DESCRIPTION

The dataset used in this project comes from a telecommunications company and contains 7,043 rows and 21 columns. Each row represents a unique customer, and each column represents a specific feature related to customer demographics, account information, and service usage. Here is a brief description of each column:

1. **customerID**: A unique identifier for each customer.
2. **gender**: The gender of the customer (Male, Female).
3. **SeniorCitizen**: Indicates if the customer is a senior citizen (0: No, 1: Yes).
4. **Partner**: Indicates if the customer has a partner (Yes, No).
5. **Dependents**: Indicates if the customer has dependents (Yes, No).
6. **tenure**: Number of months the customer has been with the company.
7. **PhoneService**: Indicates if the customer has phone service (Yes, No).
8. **MultipleLines**: Indicates if the customer has multiple lines (Yes, No, No phone service).
9. **InternetService**: Type of internet service (DSL, Fiber optic, No).
10. **OnlineSecurity**: Indicates if the customer has online security service (Yes, No, No internet service).
11. **OnlineBackup**: Indicates if the customer has online backup service (Yes, No, No internet service).
12. **DeviceProtection**: Indicates if the customer has device protection (Yes, No, No internet service).
13. **TechSupport**: Indicates if the customer has tech support (Yes, No, No internet service).

14. **StreamingTV**: Indicates if the customer has streaming TV service (Yes, No, No internet service).
15. **StreamingMovies**: Indicates if the customer has streaming movies service (Yes, No, No internet service).
16. **Contract**: The type of contract the customer has (Month-to-month, One year, Two year).
17. **PaperlessBilling**: Indicates if the customer has paperless billing (Yes, No).
18. **PaymentMethod**: The method of payment (Electronic check, Mailed check, Bank transfer (automatic), Credit card (automatic)).
19. **MonthlyCharges**: The amount charged to the customer monthly.
20. **TotalCharges**: The total amount charged to the customer.
21. **Churn**: Indicates if the customer churned (Yes, No).

This dataset provides a comprehensive view of various factors that can influence customer churn, enabling us to build and evaluate predictive models effectively.

METHODOLOGY

1. IMPORTING ALL THE REQUIRED LIBRARIES

- First, we start by importing all the required libraries for reading, cleaning, visualizing and preprocessing data.
- All the other libraries required to run and test Machine Learning (ML) models are imported later based on requirement.

2. UNDERSTANDING THE DATASET

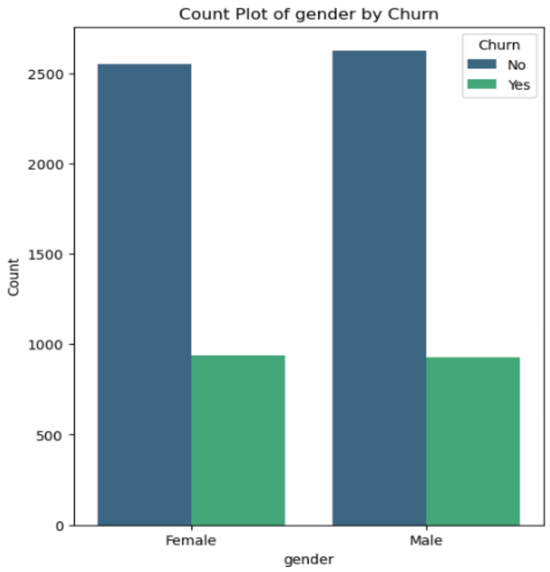
- We then read the data and store it onto a variable.
- We try and understand using info, describe and shape using the available functions.
- We then check the count of the different values in our target feature to understand the distribution of data.

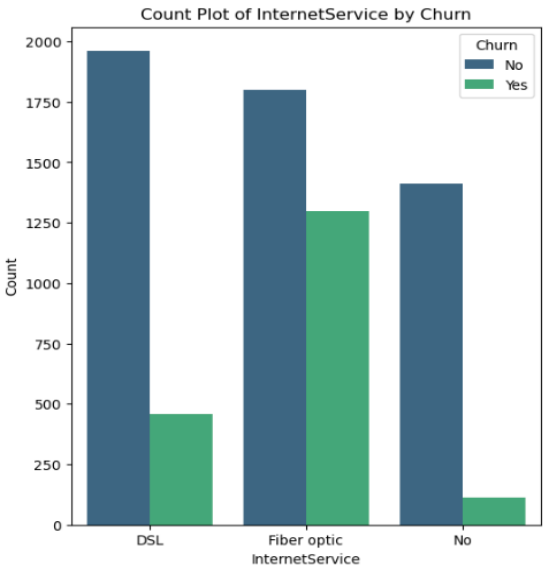
3. DATA PREPROCESSING

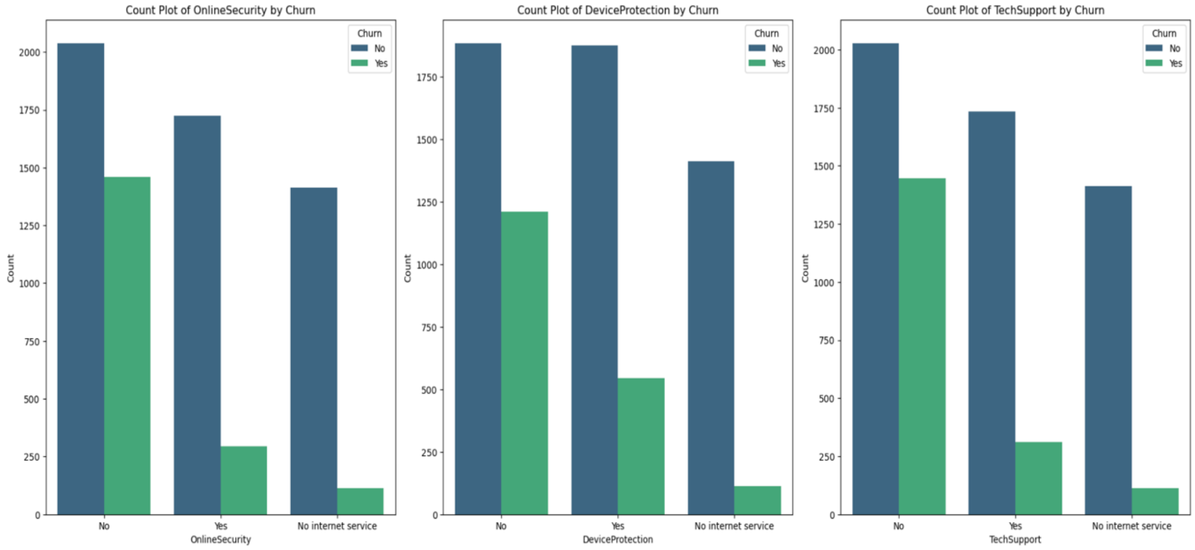
- We remove the unnecessary columns
- We proceed to convert all the binary values 1's and 0's.
- We then encode the categorical data into numerical form for further processing the data.

4. EXPLORATORY DATA ANALYSIS

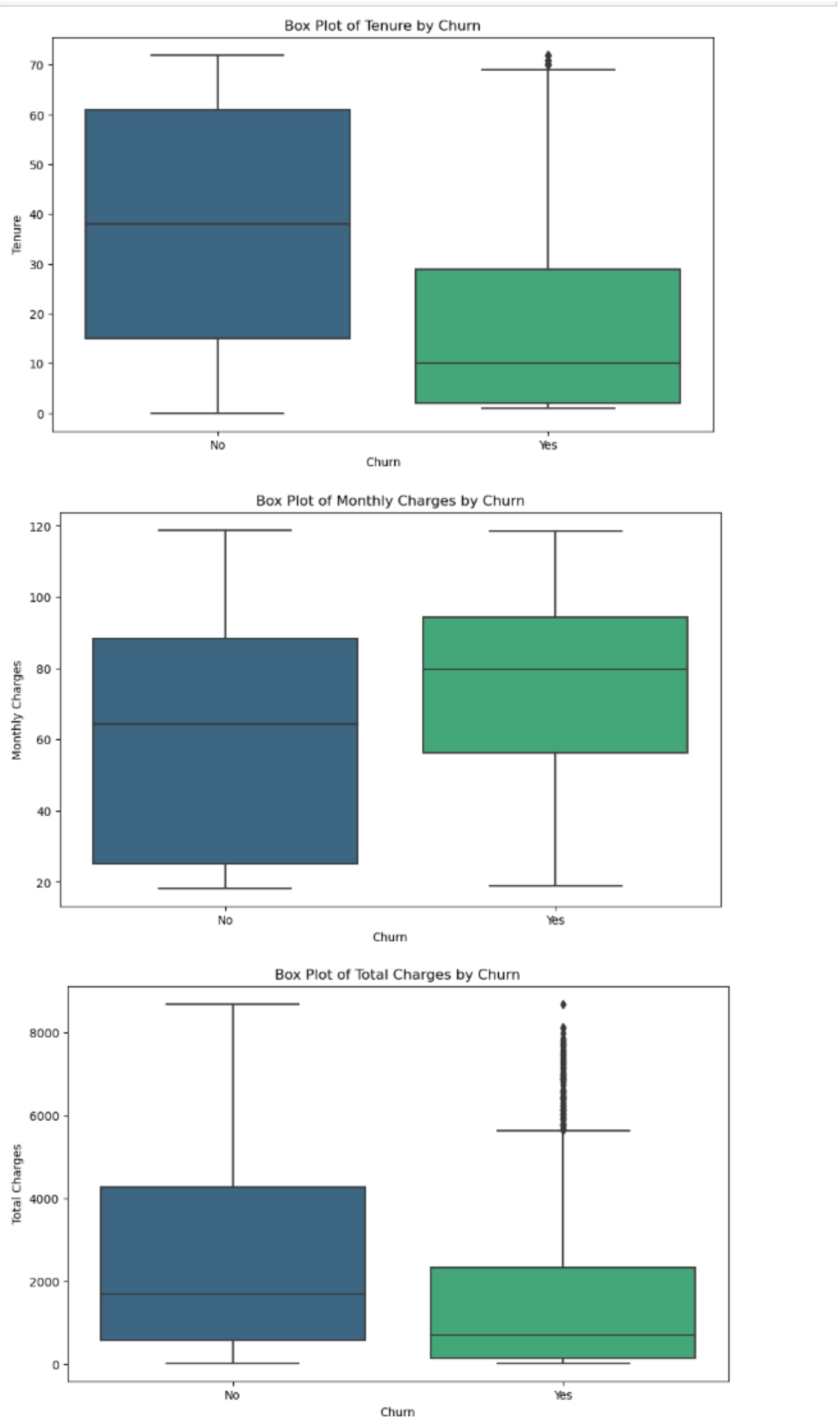
- We plot various graphs to better understand the relationship between various features and the target variable.
- We plotted count plots between different variables to understand how each feature is responsible for customer churn.


Count Plot of gender by Churn


Count Plot of InternetService by Churn



- From the graphs we can understand that that more people who have opted Fiber Optic service have left compared to others. The customers who have not taken any service, not leaving as much.
- We cannot find any noticeable difference between gender and whether or not the customer will churn or not.
- Also in the count plot of Tech support to Churn, customers with no techsupport are leaving more than others. This could be due to the inability to solve the problems they might be facing and they are unable to find the help they require.
- Then, boxplots were plotted to understand the distribution better.


Box Plot of Tenure by Churn

Box Plot of Monthly Charges by Churn

Box Plot of Total Charges by Churn

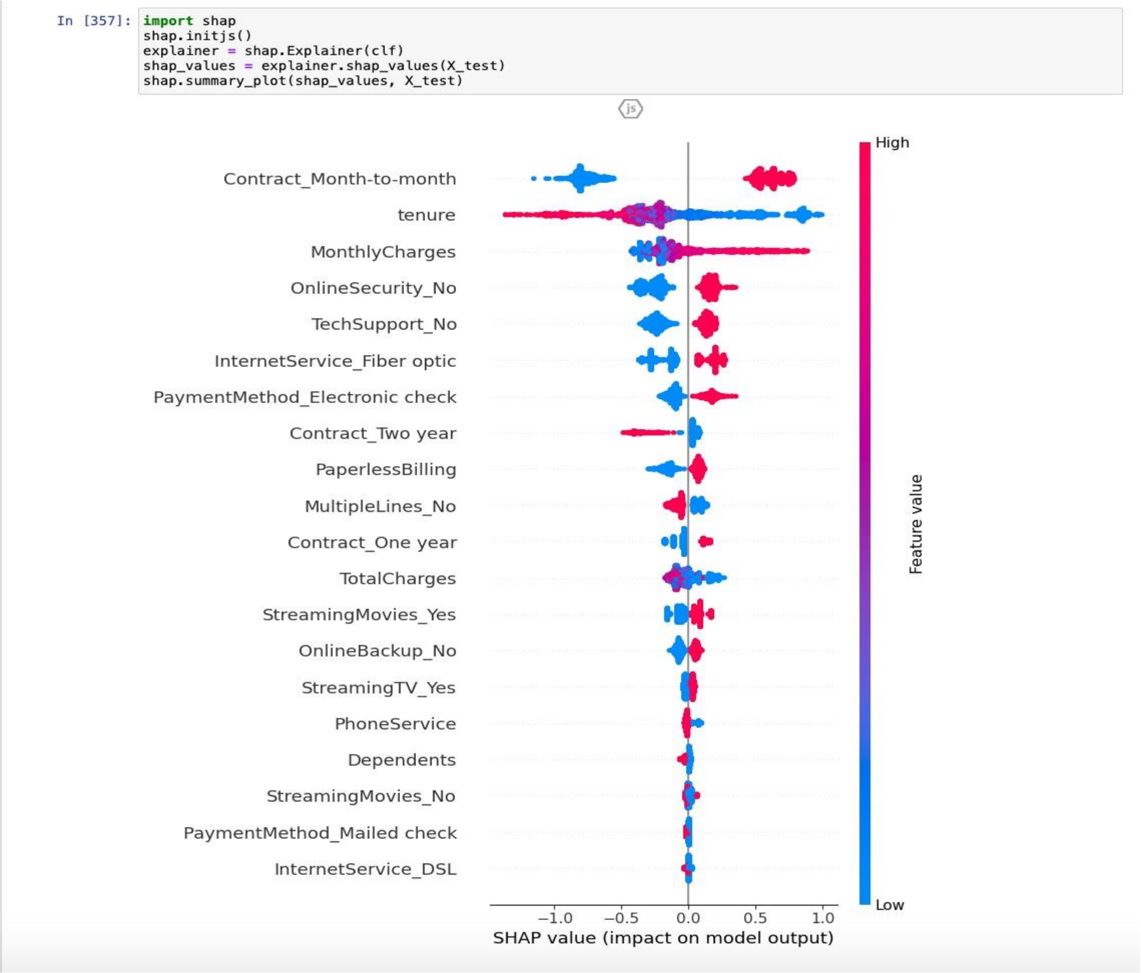
- We can see from the box plots that people who have been present longer are less likely to leave and those people who are newer are more likely.
- Also from the other boxplots we can find that people with little monthly charges are not leaving as much

but this is not the case for Total charges as people with minimal total charges are leaving more than others.

- This could be due to not opting for different features provided by the company and this would align with our previous findings.
- Correlation matrix is also used to understand the correlation between different features.

5. **FEATURE SELECTION**
    - Feature selection is done to extract all the relevant features that would be the most apt for running machine learning models.
    - We considered all the relevant features for training and testing our model.
6. **RUNNING ML MODELS**
    - Data was split into train and test sets.
    - Various ML models were trained with the train sets and further tested.
    - The models are further tested using various performance metrics like accuracy, precision, recall,f1 score and auc score.
7. **SHAP ANALYSIS**
    - We employed SHAP (SHapley Additive exPlanations) values to interpret the models, particularly focusing on the logistic regression model. SHAP values helped in understanding the contribution of each feature to the model's predictions, providing valuable insights into which factors most significantly influence customer churn.
    - We analyze how each and every feature is affecting the model and what can be conveyed from the plotted graph.

- We can see that there are trends from which we can understand what reasons are inclining the customers to reject further services from the telecommunications company.
- We can see that customers who have not opted for online security and techsupport etc are more inclined to leave than others.
- Customers who have been present longer are very less likely to leave than others.
- Also customers who have opted for shorter contracts are very intended on leaving than others. If the monthly charges of customers are high, then the customer is being more inclined to leave.



In [357]:
```
import shap
shap.initjs()
explainer = shap.Explainer(clf)
shap_values = explainer.shap_values(X_test)
shap.summary_plot(shap_values, X_test)
```

8. LISTING PEOPLE LIKELY TO LEAVE IN THE FUTURE

- We then identify all the customers who are likely to reject further services from the telecommunications company in the future. This enables us to take preventive measures so as not to lose these customers.
- We can provide offers or other benefits, or convince them to go with any other mode of service which is more likely to retain the customer based on their interests.
- Through SHAP analysis, we are able to find out what features are pushing the customer to retain the services more and focus more on those features and push other customers to enable or switch to those features which would make these customers more likely to stay with the company.
- This analysis would help prevent more and more customers from leaving the company as we can take the required measures based on the patterns we have found.

RESULTS

NAME	**ACCURACY**	**PRECISION**	**RECALL**	**F1**	**AUC**	**ACCURACY (**after hyperparameter tuning)
Logistic Regression	82.25	69.64	58.44	63.55	86.13	82.25

Decision Tree	71.75	54.90	52.54	53.69	64.63	76.01
Random Forest	79.48	68.27	53.08	59.72	86.09	81.05
Naïve Bayes	69.55	46.21	86.86	60.33	83.64	69.76
Adaboost Classifier	81.33	67.29	57.37	61.93	86.34	NA
XGBoost Classifier	79.48	69.23	53.08	60.09	86.34	81.33

FURTHER INSIGHTS

We have understood that people who have not been taking the services of the company for a long period of time are very inclined to leave. People who are not opting for security and other beneficial services are also leaving at a high rate. Also it is also helpful in tying the customer down to longer contracts.

We can take best measures based on these findings as we can provide offers at the beginning of the customers usage and also provide security services and tech support services at a smaller rate at the beginning so that they understand the importance and how useful it is for them.

CONCLUSION

In this project, we aimed to predict customer churn using a dataset from a telecommunications company. We began with data cleaning and preprocessing, including encoding categorical variables and standardizing numerical features. Exploratory data analysis helped us understand the relationships and distributions within the dataset.

We implemented and evaluated several machine learning models, including logistic regression, decision tree, random forest, and more. Each model was assessed using metrics such as accuracy, precision, recall, F1-score, and AUC to determine their effectiveness in predicting customer churn.

Our comprehensive approach allowed us to identify key drivers of customer churn, providing valuable insights for improving customer retention strategies and enhancing business profitability.