Customer Churn Prediction

1. Introduction

Customer churn prediction is a crucial aspect for businesses, especially in the telecom industry, where retaining customers is vital for maintaining a steady revenue stream. This project aims to build a machine learning model to predict customer churn using a given dataset. This is my third time working on a Data Visualization project and second time learning ML Prediction algorithms.

2. Dataset Description

Numerous characteristics of telecom users are included in the dataset, such as account information, demographic information, and services utilized. The main attributes along with their explanations are:

- **customerID**: Unique identifier for each customer.
- **gender**: Gender of the customer (Male/Female).
- **SeniorCitizen**: Indicates if the customer is a senior citizen (1) or not (0).
- **Partner**: Indicates if the customer has a partner (Yes/No).
- **Dependents**: Indicates if the customer has dependents (Yes/No).
- **tenure**: Number of months the customer has stayed with the company.
- **PhoneService**: Indicates if the customer has phone service (Yes/No).
- **MultipleLines**: Indicates if the customer has multiple lines (Yes/No/No phone service).
- **InternetService**: Type of internet service (DSL/Fiber optic/No).
- **OnlineSecurity**: Indicates if the customer has online security (Yes/No/No internet service).
- **OnlineBackup**: Indicates if the customer has online backup (Yes/No/No internet service).
- **DeviceProtection**: Indicates if the customer has device protection (Yes/No/No internet service).
- **TechSupport**: Indicates if the customer has tech support (Yes/No/No internet service).
- **StreamingTV**: Indicates if the customer has streaming TV (Yes/No/No internet service).
- **StreamingMovies**: Indicates if the customer has streaming movies (Yes/No/No internet service).
- **Contract**: Type of contract (Month-to-month/One year/Two year).
- **PaperlessBilling**: Indicates if the customer uses paperless billing (Yes/No).
- **PaymentMethod**: Payment method used by the customer (Electronic check/Mailed check/Bank transfer/Credit card).
- **MonthlyCharges**: Monthly charges for the customer.
- **TotalCharges**: Total charges incurred by the customer.

- **Churn**: Indicates if the customer churned (Yes/No).

3. Data Preprocessing

3.1 Handling Missing Values

The TotalCharges column had some missing values. These were handled by converting the column to numeric type and filling missing values with the mean of the column.

Code:

```
data['TotalCharges'] = pd.to_numeric(data['TotalCharges'], errors='coerce')

data['TotalCharges'].fillna(data['TotalCharges'].mean(), inplace=True)
```

3.2 Encoding Categorical Variables

Categorical variables were encoded using LabelEncoder to convert them into numeric format required for machine learning models.

Code:

```
categorical_columns = data.select_dtypes(include=['object']).columns

label_encoders = {}

for column in categorical_columns:
    le = LabelEncoder()
    data[column] = le.fit_transform(data[column])
    label_encoders[column] = le
```

3.3 Standardizing Numerical Features

Numerical features were standardized using StandardScaler to ensure they are on a similar scale.

Code:

```
numerical_columns = ['tenure', 'MonthlyCharges', 'TotalCharges']

scaler = StandardScaler()

data[numerical_columns] = scaler.fit_transform(data[numerical_columns])
```

4. Model Training

4.1 Data Splitting

The dataset was split into training and testing sets with an 80-20 ratio to evaluate the model's performance on unseen data.

Code:

```
X = data.drop('Churn', axis=1)
```

```
y = data['Churn']

X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.2, random_state=42)
```

I have made two models one using Logistic regression and one using XGBoost which comes under gradient boosting. These prediction algorithms were a new learning for me. Eventhough I wasnt capable of create bivariate visualizations and complex analysis, I tried to implement these models to my knowledge.

4.2 Logistic Regression Model

A statistical model for binary classification issues is called logistic regression. Logistic regression forecasts the probability of a binary event (such as yes/no, true/false, or churn/no churn), as opposed to linear regression, which predicts continuous outcomes.

What Makes Logistic Regression Useful?
Due to its ease of implementation, interpretability, and computing efficiency, logistic regression is a frequently used method for binary classification applications. It is especially helpful in situations where the logistic function can be used to model a non-linear connection between the dependent and independent variables.

Churn Prediction Using Logistic Regression
Logistic regression is used in the customer churn prediction task to estimate the likelihood that a particular customer will leave based on characteristics including tenure, monthly fees, and service usage.

4.3 XGBoost Model

Gradient Boosting is a machine learning technique used for regression and classification tasks. It builds models sequentially, where each new model aims to correct the errors made by the previous models.

An enhanced kind of gradient boosting is called XGBoost (Extreme Gradient Boosting). It has several improvements to increase performance and speed. Here's a quick rundown:

Efficiency: XGBoost handles enormous datasets and high-dimensional data with ease because it is built with speed and efficiency in mind.
Regularization: To lessen overfitting and increase the model's robustness, it contains regularization parameters (L1 and L2).
XGBoost facilitates parallel processing, which makes use of several cores to accelerate computing.
Handling Missing Data: It is more flexible since it contains built-in techniques to handle missing data.
Tree Pruning: Removes superfluous branches to increase efficiency through the use of an advanced tree pruning algorithm.

An XGBoost classifier was used for training. Overall this model is known for its high performance and scalability.

Code:

```
model = xgb.XGBClassifier(use_label_encoder=False, eval_metric='logloss')

model.fit(X_train, y_train)
```

Model Parameters:

- **use_label_encoder=False**: Disables the internal label encoder.
- **eval_metric='logloss'**: Sets the evaluation metric to logarithmic loss, suitable for binary classification tasks.

Insights from the Churn Prediction Project :

1. Demographics of the Customer

Gender: The data reveals that gender has no discernible effect on customer attrition, meaning that the rates of male and female churn are comparable.

Senior Citizens: Senior citizens have a lesser tendency to churn compared to non-senior citizens. This insight can help in designing targeted retention strategies for young customers.

2. Account Information

Compared to customers with one-year or two-year contracts, those with month-to-month contracts have a higher turnover rate. This implies that long-term agreements may lessen attrition rates.
Paperless Billing: There is a somewhat increased chance of customer attrition among those who choose paperless billing. This may be due to the fact that clients on month-to-month contracts are more likely to receive paperless billing.

3. Charges and Tenure
Tenure: Shorter-tenured customers are more prone to leave. This highlights how crucial it is to get to know new clients right away in order to increase their satisfaction and lower attrition.
Monthly Charges: Higher turnover rates are correlated with higher monthly charges. This implies that clients who spend more can believe they aren't getting enough value, which could cause them to become dissatisfied and to leave.

4. Services Utilized

Internet Service: Those who use fiber optic internet service typically have higher attrition rates than DSL users or those who do not use internet service at all. This may be the result of increased expenses or problems with the quality of the fiber optic service.
Extra Services: Users are more likely to leave if they do not have access to extra services like tech

assistance, device protection, online security, online backup, streaming TV, and streaming movies. This emphasizes how crucial it is to package services in order to keep customers longer.

5. Modes of Payment

Payment Method: Compared to customers who use credit cards, bank transfers, or postal checks, those who use electronic checks have greater turnover rates. This could suggest a relationship between client satisfaction and the ease of payment methods.

6. Performance of Predictive Models

Logistic Regression: This model offered a baseline with reasonable accuracy, but it was not very good at identifying intricate patterns in the data.

XGBoost Model: By capturing more intricate correlations between variables and churn, the XGBoost model fared better than logistic regression. It produced improved precision, recall, F1-scores, and accuracy.

These are the Insights I could deliver from the Data visualization and Predictive analysis of given dataset.