Telco Customer Churn Prediction using XGBoost

Introduction:

Customer churn is a critical challenge for telecom companies, as retaining existing customers is often more cost-effective than acquiring new ones. Predicting churn accurately helps businesses design personalized marketing and retention campaigns. This project utilizes the Telco Customer Churn dataset and applies advanced machine learning techniques to predict churn behavior. The focus is on using the XGBoost algorithm, known for its robustness and superior performance in handling imbalanced datasets and non-linear relationships.

Abstract:

The Telco Customer Churn Prediction project aims to identify customers who are likely to discontinue services, enabling the company to take proactive retention measures. The dataset used in this study contains customer demographic information, account details, and service usage patterns. An XGBoost classifier model was trained to predict customer churn based on encoded and preprocessed features. The model achieved an Area Under the Precision-Recall Curve (AUC-PR) score of 0.859, demonstrating strong classification performance and the ability to distinguish between churn and non-churn customers effectively.

Tools Used:

1. Python: Used for data preprocessing, feature engineering, model training, and evaluation.
2. Pandas & NumPy: For data manipulation, transformation, and numerical computations.
3. XGBoost: The primary model used for binary classification and prediction of churn.
4. Matplotlib & Seaborn: Used to visualize results such as confusion matrices and feature importance plots.
5. SHAP: For explainable AI to interpret the contribution of each feature to the model predictions.
6. Scikit-learn: For metrics evaluation, train-test splitting, and confusion matrix generation.

Steps Followed:

1. Data Import and Cleaning: The Telco Customer Churn dataset was imported and cleaned by removing irrelevant columns such as CustomerID, Country, State, and Lat Long. Missing values were handled appropriately to ensure data consistency.

2. Feature Engineering: Categorical columns including City, Gender, Partner, Internet Service, Contract, and others were encoded using one-hot encoding to convert them into machine-readable numerical formats.

Train-Test Split: The dataset was divided into training and testing sets using a stratified samplin approach to preserve the class balance between churned and non-churned customers.

3. Model Training: The XGBoost model was trained using DMatrix for optimized performance. Hyperparameters such as max_depth=4, learning_rate=0.1, subsample=0.9, and colsample_bytree=0.5 were tuned to prevent overfitting and achieve generalization.

4. Model Evaluation: The evaluation metric chosen was AUC-PR, as the dataset is imbalanced. The final model achieved an AUC-PR score of 0.859, showing strong predictive capability. The confusion matrix highlighted accurate predictions for both churn and non-churn customers.

5. Feature Importance and Explainability: SHAP analysis and XGBoost's built-in feature importance were used to interpret the model. Features like Contract Type, Monthly Charges, Tenure, and Internet Service were identified as key drivers of churn.

6. Result Export: The model predictions, including actual labels, predicted probabilities, and classes, were saved to an Excel file for submission and further analysis.

Conclusion:

This project successfully demonstrated how machine learning techniques can be leveraged to predict customer churn with high accuracy. By applying XGBoost and using AUC-PR as the performance metric, the model provided actionable insights that can help telecom companies retain valuable customers. Additionally, SHAP visualizations improved interpretability by revealing the most influential factors behind churn decisions. Future work can focus on integrating this predictive model into a real-time dashboard for business stakeholders and enhancing model accuracy through hyperparameter optimization or deep learning approaches.