CUSTOMER CHURN PREDICTION

Objective

The goal of this machine learning project is to predict customer churn in a telecom company. Churn occurs when a customer discontinues a service, and predicting it allows companies to implement strategies to retain customers. The model uses customer account information, service details, and demographics to predict the likelihood of churn.

Dataset Description

- **Dataset:** Telco Customer Churn
- **Source:** [Kaggle Dataset](https://www.kaggle.com/datasets/blastchar/telco-customer-churn)
- **Records:** 7043 rows (each representing a customer)
- **Features:**
- **Demographic:** gender, age (senior citizen), partner, dependents
- **Services:** phone service, internet service, online security, streaming services, etc.
- **Account info:** tenure, contract type, paperless billing, payment method, monthly charges, total charges
- **Target:** Churn (Yes/No)

Data Preprocessing

Handling Missing Values:

"Total Charges" had some empty strings. Converted to numeric with error coercion and filled missing values with median.

Encoding:

- Binary categorical columns (e.g., Yes/No) encoded using Label Encoding.

- Multi-class categorical features were one-hot encoded.

Feature Scaling:

- Numerical features like `tenure`, `MonthlyCharges`, and `TotalCharges` were standardized using `StandardScaler`.

Dealing with Class Imbalance:

- The dataset was slightly imbalanced (about 26.5% churned).

- Used **SMOTE** (Synthetic Minority Over-sampling Technique) to balance the classes during model training.

Model Building

Models Trained:

1. Logistic Regression

2. Random Forest Classifier

3. XGBoost Classifier

Each model was trained using a train-test split of 80:20 and evaluated on multiple metrics.

Model Evaluation Metrics

Model	Accuracy	Precision	Recall	F1 Score	ROC AUC
Logistic Regression	0.7388	0.7354	0.7390	0.7372	0.8404
Random Forest	0.7771	0.7713	0.7831	0.7772	0.8233
XG Boost	0.7679	0.7624	0.7750	0.7686	0.8178

- Logistic Regression had the best ROC AUC, making it ideal for distinguishing between churners and non-churners.
- Random Forest offered the best balance of precision, recall, and accuracy.

Model Explainability (SHAP)

To interpret model predictions, SHAP was used with the Logistic Regression model.

Key Influencing Features

Contract: Customers on month-to-month contracts are more likely to churn.

tenure: Customers with shorter tenure have higher chances of churning.

MonthlyCharge: High monthly charges increase churn risk.

TechSupport and OnlineSecurity: Availability of these services reduces the probability of churn.

Visualizations

- Confusion Matrices for each model to observe true/false predictions.

- ROC Curves for all models plotted together for performance comparison.

- SHAP Summary Plot to visualize feature impact across the dataset.

Conclusion

- Logistic Regression was chosen as the best model based on its ROC AUC score and interpretability.

- The model shows strong potential for deployment in customer retention systems.

- By targeting at-risk customers identified by the model, businesses can take proactive steps such as offering discounts, enhancing support, or improving service plans.

Future Improvements

- Hyperparameter tuning using GridSearchCV or Optuna

- Deploying the model with a real-time prediction interface (e.g., Flask or Streamlit app)

- Exploring deep learning models for better feature interactions

- Integration with CRM systems for real-time churn alerts

Dataset Link

https://www.kaggle.com/datasets/blastchar/telco-customer-churn