# Telecom Customer Churn Predictor with LLM Explanation

## Abstract

This project addresses the problem of customer churn in the telecom industry by building a predictive machine learning model using the XGBoost algorithm. The model is deployed using Streamlit to create an interactive user interface where users can input customer details and receive real-time predictions. A unique feature of the application is the integration of a Large Language Model (LLM) — DeepSeek Qwen3-8B (via OpenRouter API) — to provide natural language explanations of model predictions. The system is trained on a publicly available dataset and achieves strong classification performance, with test accuracy of 77.10% and AUC of 0.8268. This integration bridges the gap between black-box ML models and interpretability, enhancing both user trust and practical value.

## 1. Introduction

Customer churn is a major concern in subscription-based industries, particularly in telecom services, where retaining existing users is more cost-effective than acquiring new ones. This project aims to build a data-driven solution to predict which customers are likely to leave the service and to explain these predictions in human-understandable terms using a language model. Enhancing model interpretability can assist business stakeholders in taking timely and targeted retention actions.

## 2. Dataset Description

The dataset used is a standard telecom customer churn dataset with 7043 entries and 21 features, including:  
  
- Demographics: gender, senior citizen status, partnership, dependents   
- Account Info: tenure, contract type, payment method, billing method, monthly and total charges   
- Services Signed Up: internet, phone, online security, backup, tech support, streaming services   
- Target Variable: Churn (Yes/No)  
  
During preprocessing:  
- TotalCharges was converted from object to float.  
- customerID column was removed.  
- Null values (11 rows with tenure = 0) were dropped.  
- 22 duplicate entries were also removed.

## 3. Methodology

3.1 Preprocessing and Feature Engineering:  
- Categorical variables were encoded using Label Encoding and One-Hot Encoding.  
- Numeric variables were standardized.  
- SMOTE (Synthetic Minority Over-sampling Technique) was applied to handle class imbalance (~26% churn).  
  
3.2 Model Training:  
Several models were evaluated using cross-validation:  
- XGBoost (selected): 0.8410 mean accuracy  
- Random Forest: 0.8386  
- Gradient Boosting: 0.8242  
Other models tested: Logistic Regression, KNN, SVM, Decision Tree, and Naive Bayes.  
  
3.3 Evaluation Metrics:  
- Test Accuracy: 77.10%  
- AUC: 0.8268  
- Confusion Matrix: TN: 857, FP: 174, FN: 147, TP: 224  
- Precision (Churn): 0.56, Recall (Churn): 0.60, F1-score (Churn): 0.58  
  
3.4 Feature Importance (XGBoost):  
Top predictive features included:  
- PaymentMethod\_Electronic check  
- InternetService\_Fiber optic  
- Contract\_Two year

## 4. Deployment Using Streamlit

The model is deployed as an interactive web application built using Streamlit. Users can:  
  
- Enter customer attributes via dropdowns and number inputs.  
- Submit the data to receive:  
 - Churn prediction (Yes/No)  
 - Prediction probability  
 - Model-extracted explanation via LLM  
- View encoded data used for prediction

## 5. LLM Integration for Explainability

To enhance interpretability, an LLM is integrated through OpenRouter API (model: gpt-4o). The application sends a dynamically generated prompt that includes the customer’s input profile and the model’s prediction.  
  
This improves user understanding and trust in the system, particularly for non-technical stakeholders.

## 6. Sample Output

Example Input:  
- Gender: Male   
- Tenure: 1 month   
- Contract: Month-to-month   
- Payment Method: Electronic check   
- Internet Service: Fiber optic   
- No online security or tech support   
  
Model Prediction: Churn = Yes   
LLM Explanation: “The customer is likely to churn due to a short tenure, use of a month-to-month contract, and reliance on fiber optic service with no security or tech support. Electronic check payment is also associated with higher churn risk...”

## 7. Conclusion

The project successfully integrates machine learning with modern NLP techniques to predict and explain telecom customer churn. The XGBoost model shows strong performance on an imbalanced dataset, and the Streamlit-based web app enhances usability. LLM-based explanations make the system more transparent and user-friendly.

## 8. Key Learnings and Skills Utilized

Technical Skills:  
- Python (Pandas, NumPy, Scikit-learn, XGBoost)  
- Data preprocessing and EDA  
- Handling imbalanced data (SMOTE)  
- Model selection, cross-validation, and hyperparameter tuning  
- Streamlit for app deployment  
- Joblib for model serialization  
- Prompt engineering for LLMs  
- Integration with OpenRouter API (LLM access)  
  
Key Learnings:  
- Importance of feature encoding and column alignment for deployment  
- Model explainability improves user adoption  
- LLMs can bridge the gap between technical insights and user comprehension  
- A robust UI can make a technical system accessible to non-experts