Model Architecture

The proposed model architecture for churn prediction integrates systematic preprocessing, feature transformation, multiple machine learning algorithms, and a user-friendly deployment interface. The overall pipeline is structured into four key stages: data preprocessing, model training, evaluation, and deployment.

1. Data Preprocessing

The raw dataset is first cleaned and prepared for modeling. The *Total Charges* column is converted to numeric format, with missing values imputed using the median. For the *Competitive Influence* feature, missing entries are replaced with zero, assuming no external competitive impact. Redundant and non-predictive columns such as customer identifiers and geospatial attributes are removed to reduce noise.

Categorical features are encoded using **Label Encoding**, while numerical features are standardized with **StandardScaler** to ensure uniform scaling. Additionally, one-hot encoding is applied to categorical variables for correlation analysis. This preprocessing ensures that the data is well-structured for training both linear and tree-based models.

2. Model Training

Three different machine learning algorithms are employed to capture diverse decision boundaries and patterns:

- **Logistic Regression**: Used as a baseline linear model, trained with class balancing and feature scaling.
- **Random Forest Classifier**: An ensemble method leveraging bagging and feature importance, robust against overfitting.
- **XGBoost Classifier**: A gradient boosting algorithm optimized for classification tasks, tuned with log-loss as the evaluation metric.

Data is split into training and testing subsets (80:20) using stratified sampling to maintain churn class distribution. Class imbalance is addressed using the class_weight=balanced parameter in logistic regression and the inherent resampling mechanisms of ensemble models.

3. Model Evaluation

Each model is evaluated using accuracy, ROC AUC score, classification report, and confusion matrix visualization. A **5-fold cross-validation** strategy is applied to assess model generalizability, with Random Forest selected as the benchmark due to superior performance.

Feature importance analysis is conducted on the Random Forest model to identify the most influential attributes driving churn. Key drivers include *Monthly Charges, Total Charges, Contract Type, Technical Support, and Competitive Influence*. Correlation heatmaps are used to visualize interdependencies among features.

4. Model Deployment

The final **XGBoost model** is serialized and deployed using a **Gradio interface**. The interface allows real-time input of customer attributes such as demographics, service usage, billing patterns, and competitor influence. Predictions are expressed as a churn probability percentage, with weighted adjustments applied to critical features for business interpretability.

The deployed tool not only provides an interpretable churn score but also highlights the impact of specific business drivers, making it suitable for strategic decision-making in customer retention initiatives.