Project Title: AI-Driven Customer Churn Prediction for SaaS Companies

---

Project Synopsis:

This project focuses on building a machine learning model to predict customer churn for SaaS companies by analyzing customer usage patterns, subscription behavior, and support interaction data. The model will help businesses proactively identify at-risk customers and develop retention strategies to improve customer loyalty and revenue.

Objective:

- Develop a predictive model to estimate the likelihood of customer churn.
- Identify key factors contributing to churn and provide actionable insights for retention.
- Build a user-friendly tool/dashboard for churn prediction and analysis.

Data Sources and Features:

1. **Data Sources**:
    - Customer subscription history (e.g., start date, end date, plan type).
    - Product usage metrics (e.g., login frequency, feature usage, session duration).
    - Customer support data (e.g., number of tickets raised, resolution time).
    - Demographic data (e.g., company size, industry, region).
2. **Features**:
    - Subscription tenure and plan type.
    - Engagement metrics (e.g., usage frequency, inactivity periods).
    - Support interactions (e.g., complaints resolved, unresolved issues).
    - Historical churn status and payment behavior (e.g., late payments).

Risk Factors:

- **Class Imbalance**: Churned customers are typically a small proportion of the dataset, which may skew model performance.
- **Data Quality**: Missing or incomplete customer records may affect accuracy.
- **Feature Relevance**: Some features may not directly contribute to churn and could introduce noise.

Data Preprocessing:

1. Handle missing data with imputation techniques (e.g., mean, median for numerical data, mode for categorical).
2. Normalize numerical features (e.g., MinMaxScaler or StandardScaler).
3. Encode categorical variables using one-hot encoding or label encoding.
4. Balance the dataset using techniques like SMOTE (Synthetic Minority Over-sampling Technique).
5. Split data into training, validation, and test sets (e.g., 70-15-15 split).

Model Selection:

- **Baseline Models**: Logistic Regression, Decision Trees.

- **Advanced Models**: Random Forest, Gradient Boosting (e.g., XGBoost, LightGBM), Neural Networks.
- **Reasoning**: Advanced models can capture complex patterns in customer behavior and feature interactions, crucial for accurate churn prediction.

Exploratory Data Analysis (EDA):

1. Analyze churn rates by customer segment (e.g., plan type, region).
2. Correlation analysis to identify relationships between features and churn.
3. Visualize key trends using histograms, boxplots, and scatterplots.
4. Identify high-risk features like prolonged inactivity or unresolved support issues.

Model Evaluation:

- **Metrics**:
  - Precision, Recall, F1-Score: To balance false positives and false negatives.
  - AUC-ROC Curve: To assess the model's discriminative ability.
  - Log Loss: For probability-based evaluation.
- **Validation**: Use cross-validation techniques to ensure robustness.

Model Deployment:

1. **Platform**: Deploy the model as an API using Flask, FastAPI, or Django.
2. **Interface**: Build a web-based dashboard for business users to input data and view churn predictions.
3. **Monitoring**: Implement logging and model performance tracking to retrain as new data becomes available.

Would you like further details on any of these sections?