**SyriaTel Customer Churn**

**1. Business Understanding**

**Objective:**

The primary objective of this project is to develop a predictive model that can accurately forecast customer churn for SyriaTel. By identifying customers who are likely to churn, the company can take preemptive actions to retain them. This is crucial for maintaining a stable revenue stream and enhancing customer satisfaction.

**Significance:**

Customer churn is a significant issue for telecommunication companies, leading to lost revenue and increased costs for acquiring new customers. By predicting churn, SyriaTel can focus on retaining existing customers, thereby reducing financial losses and maintaining a positive market presence. This proactive approach not only helps in retaining valuable customers but also enhances the overall customer experience.

**2. Data Understanding**

**Dataset Overview:**

The dataset, sourced from Kaggle, contains 3,333 records and 21 features. These features encompass various aspects of customer demographics, service usage patterns, and a churn label indicating whether a customer has left the service. The dataset provides a comprehensive view of customer behavior, which is essential for developing an accurate predictive model.

**Key Features:**

**Demographics**: Information such as state, whether the customer has an international plan, and whether they have a voicemail plan. These features help in understanding the customer profile and identifying patterns that might be indicative of churn.

**Service Usage**: Metrics including total day minutes, total evening charge, total night minutes, and total international calls. These usage patterns provide insights into customer engagement with the service and can highlight changes in behavior that precede churn.

**Churn Label**: A binary indicator of whether the customer has churned or not. This is the target variable that the model will predict.

**Data Exploration:**

The dataset's initial exploration involves understanding the distribution of each feature, identifying missing values, and assessing the relationships between different features. Exploratory Data Analysis (EDA) techniques such as visualizations (e.g., histograms, box plots, and scatter plots) are employed to uncover patterns and anomalies. This step is critical for informing subsequent data preprocessing and modeling steps.

**3. Environment Setup**

**Tools Used:**

**Pandas**: For data manipulation and analysis.

**NumPy**: For numerical operations and calculations.

**Scikit-learn**: For machine learning algorithms and model evaluation.

**Matplotlib and Seaborn**: For data visualization.

**Imbalanced-learn**: For handling imbalanced datasets.

**XGBoost or LightGBM**: For advanced gradient boosting techniques.

**SHAP**: For model interpretability and understanding feature importance.**Development Environment:**

The project was conducted in a Jupyter Notebook environment, which provides an interactive platform for data analysis and model development. Version control was facilitated through GitHub, ensuring an organized and collaborative workflow. The combination of these tools and platforms allows for efficient data processing, model training, and result visualization.

**4. Steps in the Notebook**

**1. Data Import and Initial Exploration:**

**Loading the Dataset**: The dataset was loaded using Pandas, and initial exploration was conducted using .info() and .head() to understand the structure and preview the data.

**Exploratory Data Analysis (EDA)**: EDA was performed to identify any patterns, anomalies, and relationships within the data. This includes visualizing the distribution of key features, checking for missing values, and exploring correlations between features.

**2. Data Preprocessing:**

**Encoding Categorical Variables**: Categorical variables such as "international plan" and "voice mail plan" were encoded using label encoding to convert them into numerical format suitable for machine learning algorithms.

**Scaling Numerical Features**: Numerical features were standardized using StandardScaler to ensure they have a mean of 0 and a standard deviation of 1. This step is crucial for algorithms that rely on distance metrics.

**Splitting the Data**: The data was divided into training and test sets with an 80-20 split, ensuring the model is trained on a majority of the data and evaluated on a separate subset.

**3. Handling Class Imbalance:**

**SMOTE-Tomek**: The class imbalance was addressed using Synthetic Minority Over-sampling Technique (SMOTE) combined with Tomek links. This technique oversamples the minority class (churned customers) and removes overlapping examples, resulting in a balanced dataset and improved model performance. SMOTE generates synthetic examples for the minority class, while Tomek links identify and remove instances that are difficult to classify, enhancing the dataset's quality.

**4. Model Training:**

**Models Used**:

**Logistic Regression**: A baseline model to start with, providing a good balance between interpretability and performance.

**Decision Trees**: A model that captures non-linear relationships and is easily interpretable.

**Random Forest Classifier**: An ensemble method that combines multiple decision trees to improve accuracy and robustness.

**Hyperparameter Tuning**: GridSearchCV was used to systematically explore a range of hyperparameters for each model. This process involved defining a grid of possible values for parameters such as regularization strength, maximum tree depth, and minimum samples per leaf, and then finding the combination that yields the best performance based on cross-validation.

**5. Model Evaluation:**

**Metrics Computed**: Accuracy, Precision, Recall, F1-Score, and ROC-AUC were used to assess model performance. These metrics provide a comprehensive view of the model's ability to predict churn accurately.

**Confusion Matrix Visualization**: The confusion matrix was visualized to understand the distribution of true positives, false positives, true negatives, and false negatives. This helps in diagnosing the types of errors the model is making.

**6. Advanced Techniques:**

**Gradient Boosting Techniques**: Advanced algorithms like XGBoost were explored to further enhance the model's prediction accuracy and ROC-AUC score. These techniques use boosting to sequentially improve the model by focusing on the hardest-to-predict cases, resulting in a more robust and accurate model.

**5. Observations and Outcomes**

**Initial Results:**

Logistic Regression faced convergence warnings due to feature scaling or imbalanced data. Despite being a simple and interpretable model, its performance was limited in handling the complexity of the data.

**Improvements:**

**SMOTE-Tomek**: Effectively balanced the dataset, leading to improved model robustness and accuracy. This technique enhanced the model's ability to identify churned customers by providing a more representative training set.

**Decision Trees and Random Forests**: Performed well after hyperparameter tuning, capturing complex patterns in the data. These models showed significant improvement in accuracy, precision, recall, and F1-Score compared to the baseline logistic regression model.

**Evaluation Highlights:**

**Precision and Recall for Churn Class**: These were key focus areas to ensure the model not only predicts churn accurately but also identifies the majority of churned customers. High precision reduces false positives, while high recall ensures most churned customers are correctly identified.

**ROC Curve**: Plotted to assess the classifier's ability to distinguish between churned and non-churned customers. A higher ROC-AUC score indicates better model performance.

**6. Recommendations**

**Adopt the Model with the Highest ROC-AUC and Balanced Metrics:**

The decision tree model, after hyperparameter tuning, demonstrated the highest ROC-AUC and balanced precision-recall metrics. It is recommended for deployment due to its robust performance and interpretability.

**Deploy the Model in Production:**

Implement the decision tree model in SyriaTel's production environment to enable real-time churn prediction. This involves integrating the model into existing customer relationship management (CRM) systems for seamless operation.

**Periodically Retrain the Model:**

As customer behavior evolves, it is essential to periodically retrain the model with new data. This ensures that the model remains accurate and relevant, capturing any changes in customer patterns and improving its predictive power over time.

**Strategic Impact**

**Revenue Retention:**

By accurately predicting and reducing churn, SyriaTel can maintain a stable revenue stream and improve overall profitability. Retaining existing customers is more cost-effective than acquiring new ones, making churn prediction a valuable investment.

**Customer Satisfaction:**

Proactive engagement with at-risk customers enhances their loyalty and satisfaction, leading to long-term customer relationships. By addressing issues before customers decide to leave, SyriaTel can improve its service quality and customer experience.

**Operational Efficiency:**

Focused retention efforts allow for more efficient use of resources, targeting interventions where they are most needed. This improves the return on investment (ROI) of retention strategies and optimizes resource allocation.