**Telecom Customer Churn Prediction**

**Abstract**

Customer churn, or customer attrition, is a critical issue in the telecom industry, leading to significant revenue losses and increased customer acquisition costs. To address this challenge, machine learning (ML) techniques can be leveraged to analyze customer behavior and predict churn probability. This project aims to develop an accurate and efficient ML-based model for telecom customer churn prediction.

The study involves data collection, preprocessing, feature engineering, and model training using multiple algorithms, including Decision Tree, Random Forest, K-Nearest Neighbors (KNN), logistic regression, SVM and XGBoost. The models are evaluated based on performance metrics such as accuracy, precision, recall and F1-score to determine the best-performing approach. Additionally, hyperparameter tuning techniques like Grid Search and Random Search are applied to enhance model efficiency.

A comprehensive literature review highlights the existing research on telecom churn prediction, identifying key challenges such as class imbalance, feature selection complexity, lack of real-time prediction, and model interpretability issues. The study also proposes solutions, including the use of Synthetic Minority Oversampling Technique (SMOTE), automated feature selection, real-time model deployment, and explainable AI techniques like SHAP.

By developing a user-friendly predictive system, this project provides telecom companies with actionable insights to identify at-risk customers and implement data-driven retention strategies. The proposed approach aims to enhance customer satisfaction, minimize churn, and improve overall business profitability in the competitive telecom sector.

**Introduction**

Customer churn, also known as customer attrition, is a significant challenge in the telecom industry. It occurs when customers discontinue using a company’s services, leading to revenue loss and increased customer acquisition costs. With growing competition, telecom companies must implement effective strategies to predict and mitigate customer churn.

To address this issue, Machine Learning (ML) techniques can be utilized to analyze customer behavior and predict churn probability. By leveraging historical data, telecom providers can identify patterns and key indicators that contribute to customer churn. This enables them to take proactive measures such as personalized offers, better service quality, and improved customer engagement to retain at-risk customers.

The primary objective of this project is to develop an accurate and efficient ML-based model that predicts telecom customer churn. By using data science methodologies, including feature engineering, model training, and evaluation, we aim to create a robust tool that helps telecom companies minimize churn and enhance customer satisfaction.

**Literature survey**

**Reference 1**

**Title:** A Review on Machine Learning-Based Customer Churn Prediction in the Telecom Industry

**Authors:** Sawsan Barham, Nowfal Aweisi, Ala’ Khalifeh

**Year:** 2023

**Methodology:** This paper reviews 33 studies on customer churn prediction in telecom from 2019–2022. Techniques analyzed include Random Forest, Logistic Regression, Decision Trees, and XGBoost. Random Forest was found to achieve the highest accuracy, with some studies reporting 97.4%. The paper also discusses feature selection and handling imbalanced datasets.

**Drawback:** Lacks empirical validation and does not discuss real-world deployment challenges such as scalability and computational costs. Model interpretability is also not explored.

**Reference 2**

**Title:** Exploratory Data Analysis and Customer Churn Prediction for the Telecommunication Industry

**Authors:** Kiran Deep Singh, Gaganpreet Kaur, Prabh Deep Singh, Vikas Khullar, Ankit Bansal, Vikas Tripathi

**Year:** 2023

**Methodology:** This study focuses on Exploratory Data Analysis (EDA) and machine learning models for churn prediction in telecom. It examines customer behavior and applies XGBoost, achieving 82.80% accuracy on a real-world dataset. The study emphasizes feature selection and customer retention strategies based on key indicators.

**Drawback:** The study does not address class imbalance handling methods like SMOTE, which may affect prediction accuracy. Additionally, it lacks a comparison with deep learning models and does not explore computational costs for large datasets.

**Reference 3**

**Title:** The Impact of SMOTE and ADASYN on Random Forest and Advanced Gradient Boosting Techniques in Telecom Customer Churn Prediction

**Authors:** Mehdi Imani, Majid Joudaki, Zahra Ghaderpour, Ali Beikmohammadi

**Year:** 2024

**Methodology:** This study examines the effects of SMOTE and ADASYN on Random Forest, XGBoost, LightGBM, and CatBoost for customer churn prediction. The dataset contains 4,250 training and 750 testing samples. After applying SMOTE and ADASYN, LightGBM achieved the highest F1-score (89%) and ROC AUC (95%).

**Drawback:** The study does not evaluate the risk of synthetic noise from oversampling techniques. It also notes that hyperparameter tuning had minimal impact on model performance.

**Reference 4**

**Title:** Customer churn prediction in telecom sector using machine learning techniques

**Authors:** Sharmila K. Wagh, A. A. Andhale, K. S. Wagh, J. R. Pansare, S. P. Ambadekar, S. Gawande  
**Year:** 2024

**Methodology:** The study focuses on predicting customer churn in the telecom industry using machine learning techniques. The authors employed Random Forest (RF), Decision Tree, and K-Nearest Neighbors (KNN) on the IBM Telco dataset. SMOTE was used for class imbalance. Feature selection was done using Pearson correlation, and the model achieved 99% accuracy with Random Forest.

**Drawback:** The extremely high accuracy raises concerns about potential overfitting, especially on imbalanced data. The use of SMOTE may introduce unrealistic samples. The study also lacks discussion on computational complexity.

**Reference 5**

**Title:** Predicting Customer Churn in Telecom Industry: A Machine Learning Approach for Improving Customer Retention

**Authors:** Abhikumar Patel, Amit G Kumar

**Year:** 2023

**Methodology:** This study applies supervised ML techniques including Bernoulli Naive Bayes, Gaussian Naive Bayes, SVM, KNN, Decision Tree, Random Forest, and XGBoost. XGBoost achieved the highest accuracy of 94%. Feature influence analysis identified international plans and customer service calls as key indicators.

**Drawback:** Potential overfitting with XGBoost was not analyzed. Issues like computational complexity and scalability were not addressed.

**Reference 6**

**Title:** Telecom Customer Churn Prediction Using Enhanced Machine Learning Classification Techniques

**Authors:** Goldy Verma

**Year:** 2024

**Methodology:** This research compares Decision Tree, Random Forest, and KNN on a Kaggle dataset. Accuracy scores were 79% for Decision Tree and KNN, and 82% for Random Forest. The study emphasizes preprocessing and feature selection.

**Drawback:** The study does not handle class imbalance, affecting reliability. Also lacks interpretability analysis, limiting business usefulness.

**Reference 7**

**Title:** Customer Churn Prediction Using Synthetic Minority Oversampling Technique

**Authors:** Aishwarya H M, Soundarya B, Bindhiya T, C Christlin Shanuja, S Tanisha

**Year:** 2023

**Methodology:** The study applies SMOTE and evaluates Gradient Boosting, Random Forest, Decision Tree, and Logistic Regression. Gradient Boosting achieved 95.13% accuracy. Tenure and monthly charges were key features.

**Drawback:** While accuracy improved, SMOTE may distort data. Computational cost and interpretability were not evaluated.

**Reference 8**

**Title:** Churn Prediction of Customer in Telecom Industry using Machine Learning Algorithms

**Authors:** V. Kavitha, S. V Mohan Kumar, M. Harish, G. Hemanth Kumar

**Year:** 2020

**Methodology:** This study uses Random Forest, XGBoost, and Logistic Regression on a Kaggle dataset. Random Forest performed best with 93% accuracy. Preprocessing steps included normalization and feature selection.

**Drawback:** Logistic Regression underperformed and was unsuitable for imbalanced data. Model interpretability and hyperparameter tuning were not addressed.

**Reference 9**

**Title:** Customer Churn Prediction Based on Interpretable Machine Learning Algorithms in Telecom Industry

**Authors:** Liwen Ou

**Year:** 2022

**Methodology:** Focused on interpretability using Random Forest, Decision Tree, and Extra Tree Classifier. Feature importance highlighted tenure, total charges, and monthly charges.

**Drawback:** Decision Tree showed lower accuracy. Class imbalance was not handled. Computational complexity of ensemble models was not discussed.

**Reference 10**

**Title:** Machine Learning-Based Telecom-Customer Churn Prediction

**Authors:** Pushkar Bhuse, Aayushi Gandhi, Parth Meswani, Riya Muni, Neha Katre

**Year:** 2020

**Methodology:** This study compares Random Forest, SVM, XGBoost, Ridge Classifier, and Deep Neural Networks. Random Forest achieved 90.96% accuracy. Feature selection and grid search tuning were applied.

**Drawback:** Deep learning was not extensively tuned. Class imbalance and computational scalability were not discussed.

**Literature Outcome:**

The reviewed studies show that successful churn prediction involves several important steps such as data preprocessing, feature selection, and handling class imbalance. Many researchers used methods like SMOTE to balance the dataset and improve the performance of models when predicting minority classes. Machine learning models such as Decision Tree, Random Forest, KNN, Logistic Regression, SVM, and XGBoost were commonly used across different studies, with ensemble methods often giving better accuracy and stability. Techniques like hyperparameter tuning and proper evaluation using accuracy, precision, recall, and F1-score helped improve model effectiveness. The literature highlights that using multiple models and combining them can lead to more reliable results. These findings support the use of a complete machine learning pipeline, from data preparation to model combination, for better customer churn prediction in the telecom industry.

**Proposed methodology**

This project follows a comprehensive machine learning pipeline to predict customer churn in the telecom industry. The methodology includes key steps such as data preprocessing, data balancing, model building, evaluation, and ensemble learning.

* **Data Preprocessing:** Initial preprocessing involved handling missing values, encoding categorical variables, and scaling numerical features to standardize the data for better model performance.
* **Feature Engineering:** Features were selected based on relevance to churn prediction. These included customer demographics, service usage patterns, contract type, and billing-related details. This step helped in improving model interpretability and accuracy.
* **Data Balancing:** The dataset exhibited class imbalance, with a significantly higher number of non-churned customers. To address this, the **Synthetic Minority Over-sampling Technique (SMOTE)** was applied. SMOTE generates synthetic instances of the minority class (churners), helping the model learn more effectively from both classes.
* **Machine Learning Models Implemented**

A diverse set of machine learning classifiers was developed to compare performance:

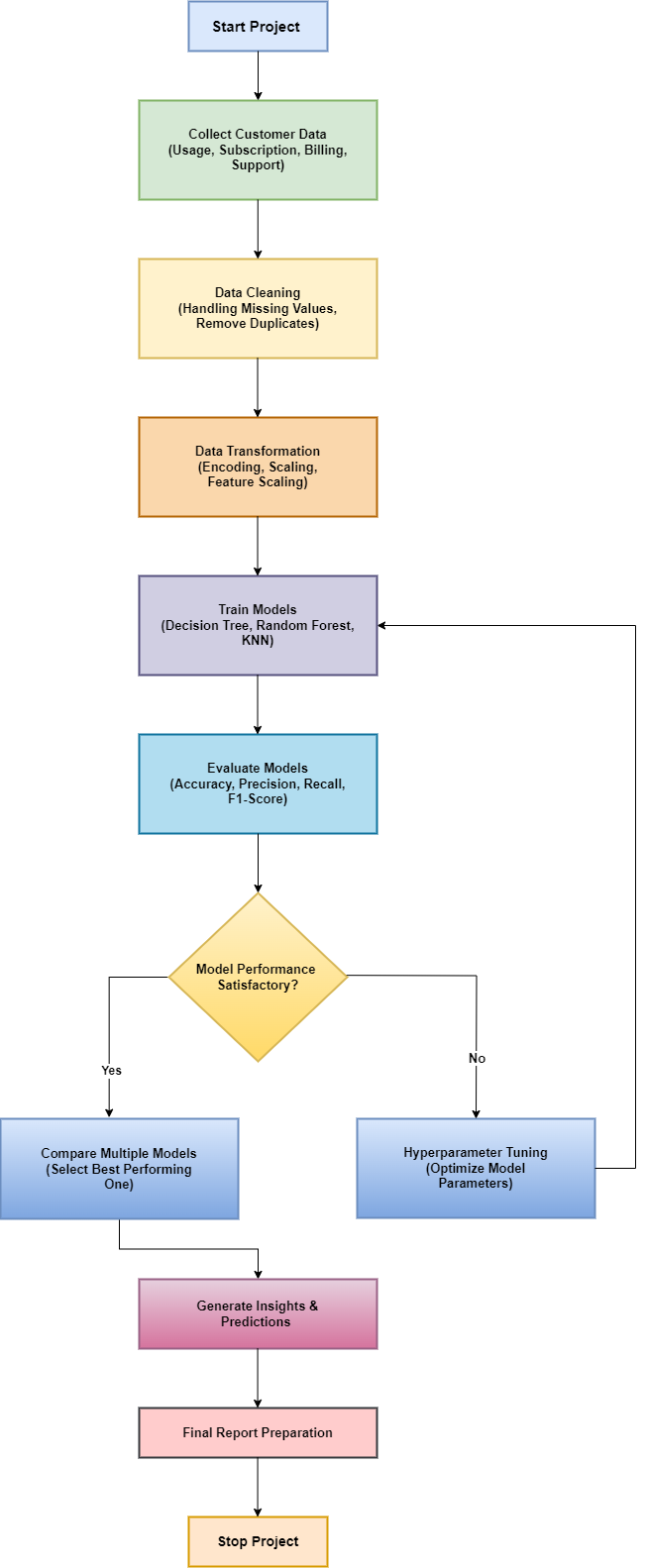
* **Decision Tree Classifier**: A rule-based model that splits customer data into branches based on feature thresholds.
* **Random Forest Classifier:** An ensemble of decision trees that enhances prediction accuracy and controls overfitting.
* **K-Nearest Neighbors (KNN) Classifier:** A distance-based method that predicts churn based on similarities with neighboring data points.
* **Logistic Regression:** A linear model that estimates the probability of churn, useful for interpreting feature impact.
* **Support Vector Machine (SVM):** Constructs an optimal hyperplane in high-dimensional space to separate churn and non-churn classes.
* **XGBoost Classifier:** A gradient boosting algorithm known for its efficiency and performance, especially on structured data.
* **Hyperparameter Tuning: GridSearchCV** was employed to optimize model performance. This involved testing multiple parameter combinations using cross-validation to find the best settings for each model.
* **Model Evaluation:** Model performance was assessed using key classification metrics, particularly focusing on the minority class (churn = 1):
* **Accuracy**
* **Precision**
* **Recall**
* **F1-Score**

The **Random Forest**, **XGBoost**, and **SVM** models demonstrated strong performance across these metrics, with Random Forest achieving the highest balance overall.

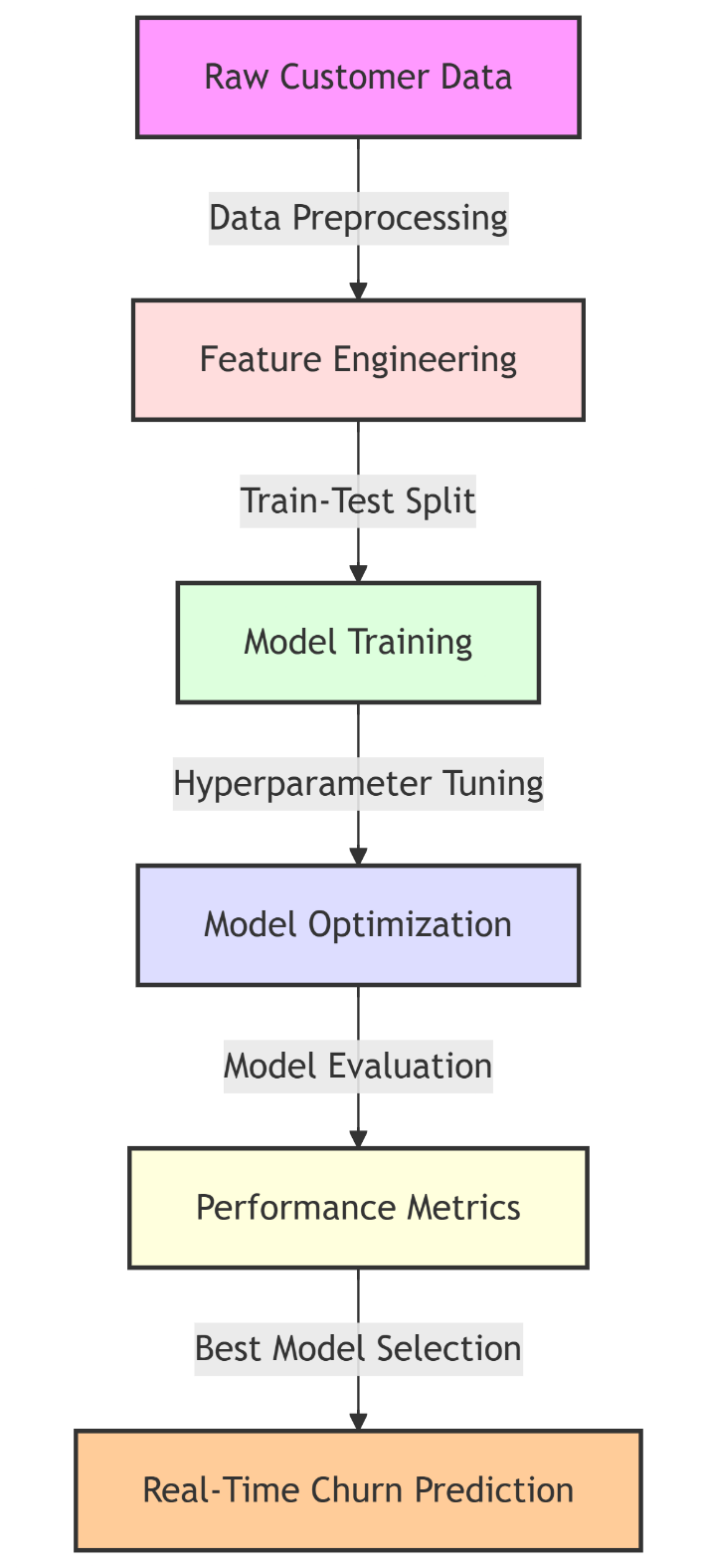
* **Ensemble Learning with Voting Classifier:** To further enhance prediction stability and accuracy, a **soft voting ensemble classifier** was built by combining the predictions of:
* Random Forest
* XGBoost
* SVM
* Logistic Regression

This ensemble approach leveraged the strengths of individual models, resulting in improved generalization on unseen data.

By following this methodology, the project aims to develop an effective churn prediction system that helps telecom companies identify and retain at-risk customers.

**Flowchart**

1. **Start Project**
   * Initiates the process with goal definition.
2. **Collect Customer Data**
   * Data Sources: Usage history, subscription plans, billing records, customer support interaction logs.
3. **Data Cleaning**
   * Techniques: Imputation, duplicate removal, outlier detection.
4. **Data Transformation**
   * Includes: One-hot encoding, label encoding, normalization, feature scaling.
5. **Train Models**
   * Example models: Decision Tree Classifier, Random Forest Classifier, K-Nearest Neighbors (KNN) Classifier, Logistic Regression, Support Vector Machine (SVM), XGBoost Classifier.
6. **Evaluate Models**
   * Using standard metrics (Accuracy, Precision, Recall, F1-Score).
7. **Model Performance Satisfactory?**
   * Decision node:
     + **Yes**: Go to model comparison.
     + **No**: Proceed to hyperparameter tuning.
8. **Hyperparameter Tuning (if needed)**
   * Optimize models for best possible output.
9. **Compare Multiple Models**
   * Finalized top 4 models based on evaluation metrics and combined them using a Voting Classifier for improved prediction accuracy.
10. **Generate Insights & Predictions**
    * Creating business insights and performing churn prediction.
11. **Final Report Preparation**
    * Compiling results, visualizations, and recommendations.
12. **Stop Project**
    * Project closure.

**Block Diagram**

1. **Raw Customer Data**
   * Input: Collected from the telecom company (e.g., demographics, usage, service features).
   * Format: CSV, Excel, or database.

* **Data Preprocessing**
  + Tasks: Handling missing values, encoding categorical variables, removing duplicates, scaling numeric features.

1. **Feature Engineering**
   * Tasks: Creating new relevant features, selecting or transforming existing ones to improve model accuracy.

* **Train-Test Split**
  + Splitting the dataset into training and testing sets (e.g., 80/20) to evaluate model performance.

1. **Model Training**
   * Training machine learning algorithms like Decision Tree Classifier, Random Forest Classifier, K-Nearest Neighbors (KNN) Classifier, Logistic Regression, Support Vector Machine (SVM), XGBoost Classifier.

* **Hyperparameter Tuning**
  + Finding the best combination of parameters for each model using techniques like GridSearchCV.

1. **Model Optimization**
   * Further improvements to boost model efficiency or accuracy (e.g., using pipelines, ensemble methods).

* **Model Evaluation**
  + Metrics: Accuracy, Precision, Recall, F1-score.

1. **Performance Metrics**
   * Comparing models using evaluation metrics.

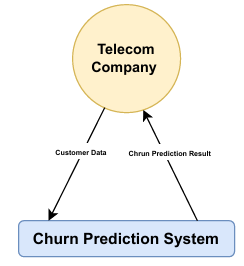
* **Best Model Selection**
* Selecting the most accurate model for deployment.
* Compared the performance of all models and selected the top 4 performing models and combined them using a Voting Classifier for improved prediction accuracy.

1. **Real-Time Churn Prediction**

* Final deployed model used in a real-time system (via Streamlit) to make predictions based on new data.

**Data Flow Diagrams**

* **Level 0 DFD**



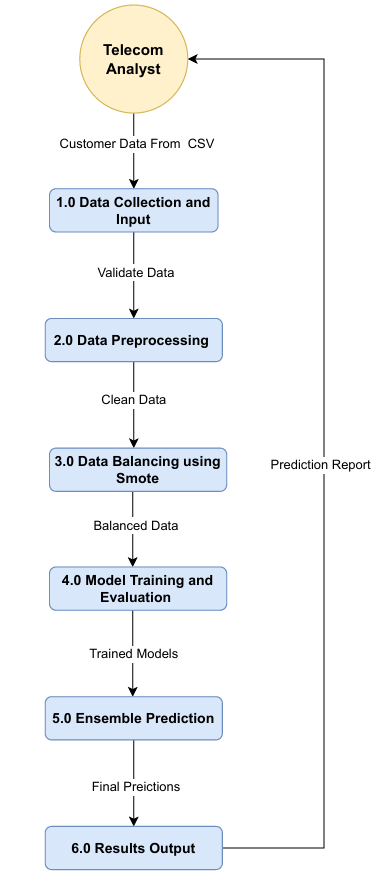
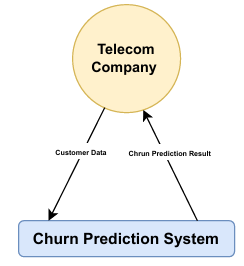
**Entities:**

* **Telecom Company:**
  + This is the external entity that interacts with the churn prediction system.
  + It acts as both the data provider and the consumer of results**.**

**Main Process:**

* **Churn Prediction System:**
  + This is the core process that takes customer data as input, processes it through a machine learning pipeline, and generates churn predictions.
  + Internally, it includes data preprocessing, model training, ensemble prediction, and result generation (which will be detailed in Level 1 DFD).

**Data Flows:**

* **Customer Data (→):**
  + This flow represents the **input** from the telecom company. It includes all customer details like demographics, service usage, billing info, etc.
* **Churn Prediction Results (←):**
  + This flow represents the **output** of the system — whether a customer is likely to churn or not, including churn probability and risk level.
* **Level 1 DFD**

**Entities:**

* **Telecom Company:**
* This represents the end-user or company analyst who interacts with the system.
* Provides input data (manually or via CSV upload).
* Receives the churn prediction report for analysis and customer retention efforts.

**Processes:**

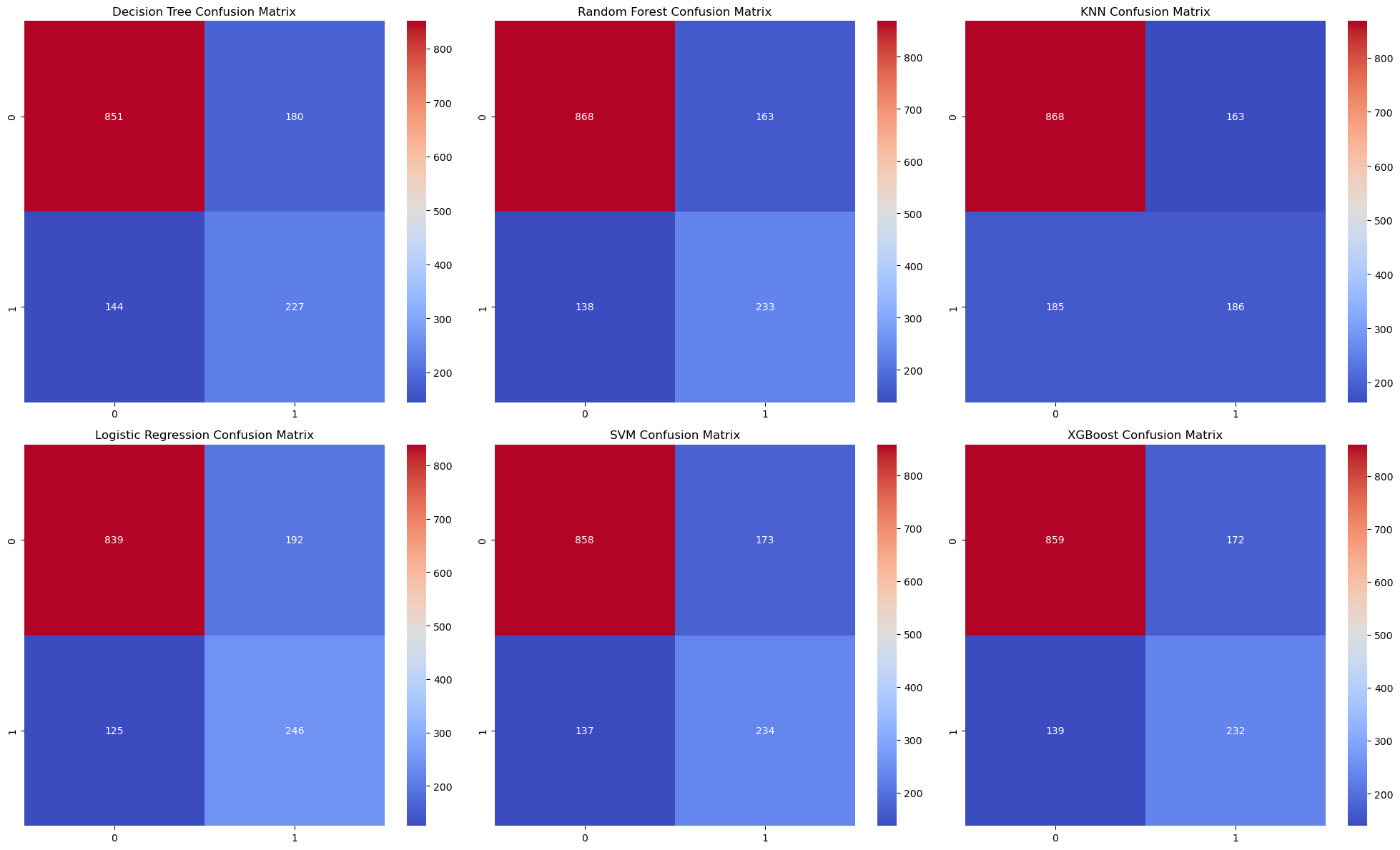
* **Data Collection and Input:**
* Receives customer data from the Telecom Analyst (either via form or CSV).
* Performs initial validation to ensure data format and completeness.
* **Data Preprocessing:**
* Handles missing values.
* Encodes categorical variables (like gender, contract type).
* Scales numerical features (like tenure, charges).
* **Data Balancing using SMOTE:**
* Applies the SMOTE algorithm to balance the dataset by generating synthetic samples for the minority class (churned customers).
* Helps prevent model bias towards non-churners.
* **Model Training and Evaluation:**
* Trains multiple machine learning models like Decision Tree Classifier, Random Forest Classifier, K-Nearest Neighbors (KNN) Classifier, Logistic Regression, Support Vector Machine (SVM), XGBoost Classifier.
* Evaluates them using metrics like accuracy, precision, recall, and F1-score.
* **Ensemble Prediction:**
* Combines top-performing models using Voting Classifier (soft voting).
* Generates the final churn prediction (Yes/No), probability, and risk level.
* **Results Output and Logging:**
* Displays the prediction result to the user.
* Logs predictions (for both single and batch mode) for review and audit.

**Data Flows:**

* Customer Data (Form/CSV) flows from **Telecom Company**.
* Each process passes its output to the next stage:
  + Validated Data → Clean Data → Balanced Data → Trained Models → Final Prediction.
* Prediction Report is finally sent back to the **Telecom Company**.

**Result**

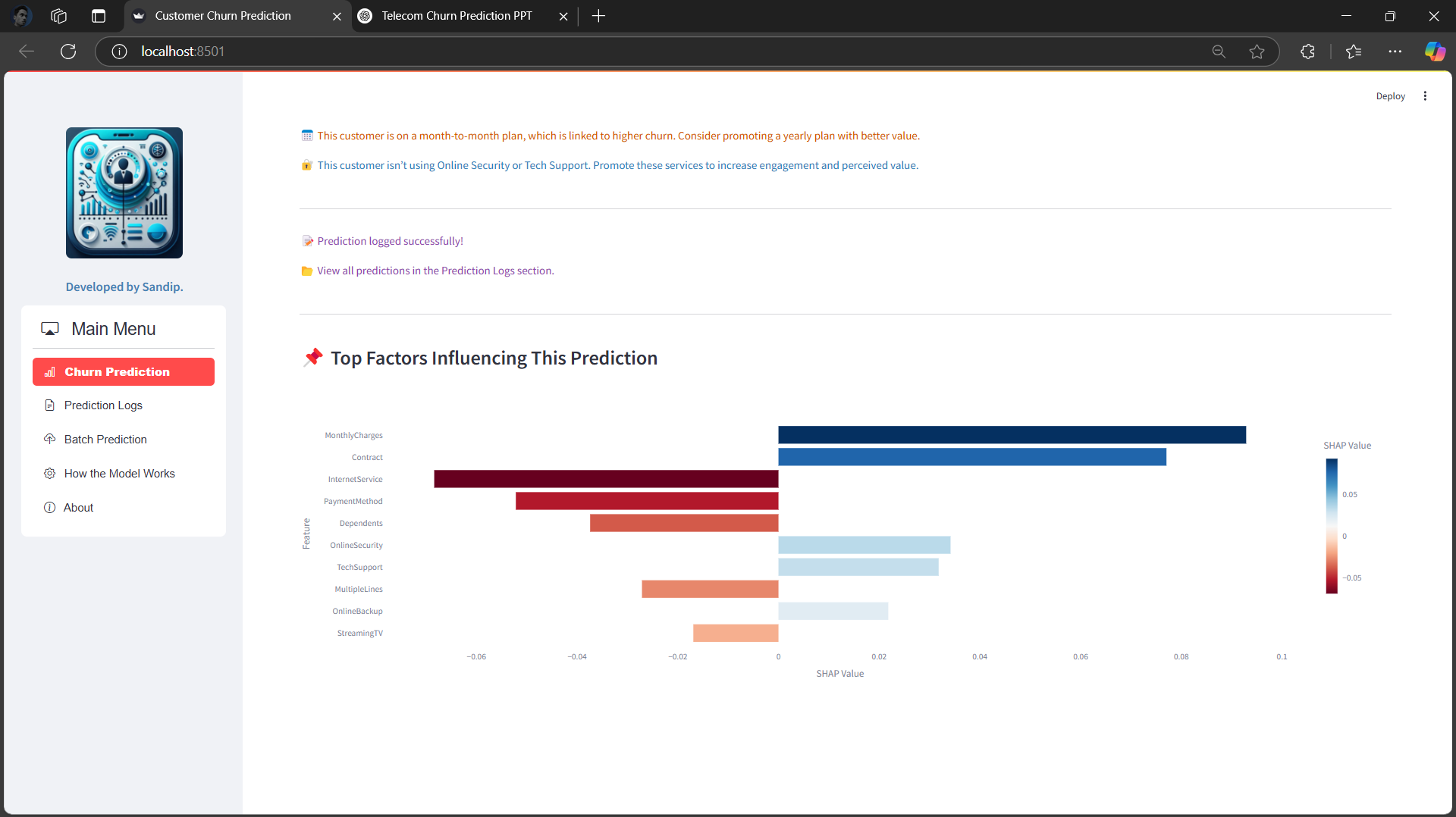
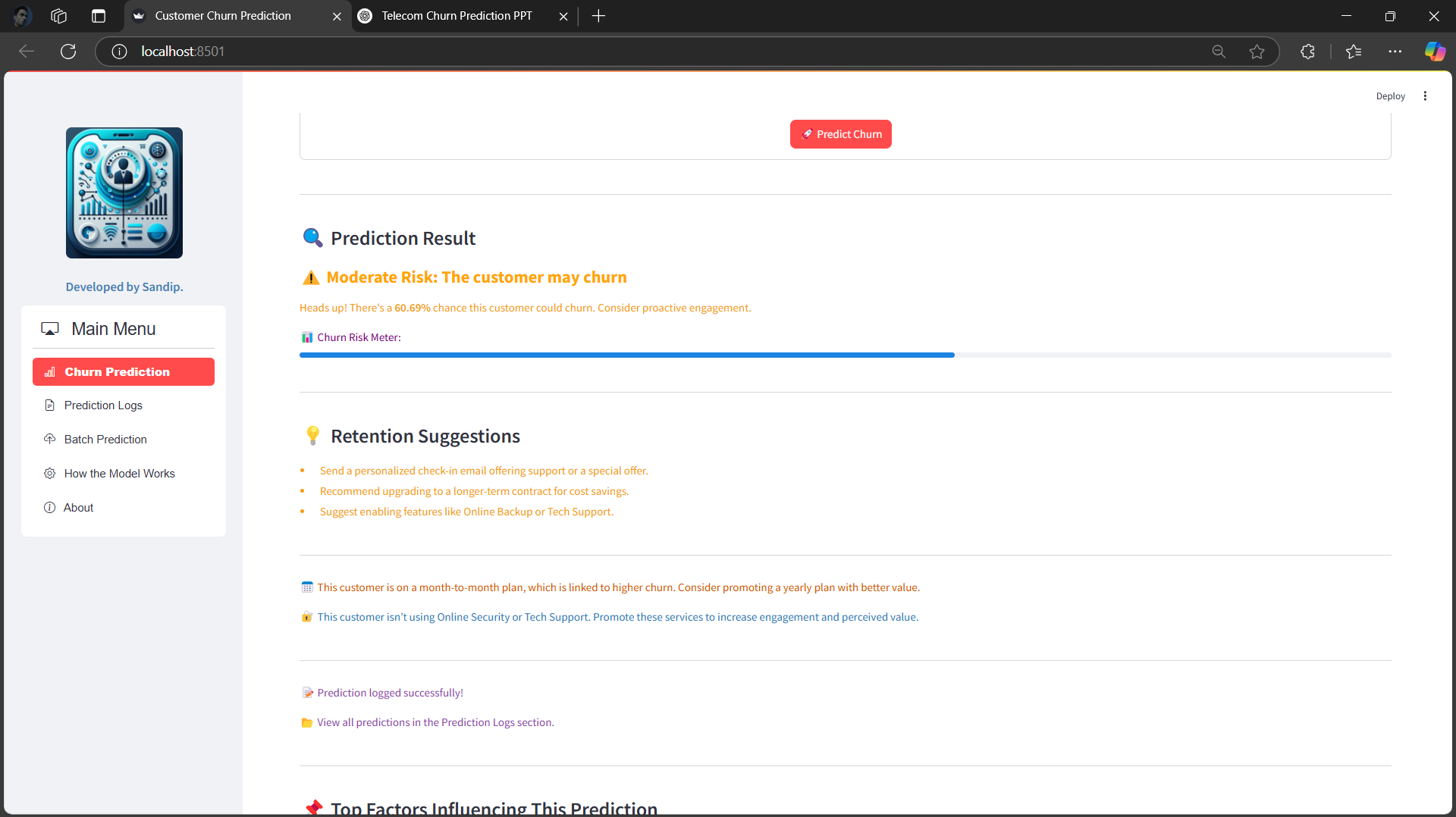
**Model Comparison**

****

**Model Ranking**

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| Rank | Model | Accuracy | Precision (1) | Recall (1) | F1-Score (1) | Reason |
| 🥇 1 | Random Forest | 0.785 | 0.59 | 0.63 | 0.61 | High accuracy, balanced performance |
| 🥈 2 | XGBoost | 0.778 | 0.57 | 0.63 | 0.60 | Balanced, powerful, avoids overfitting |
| 🥉 3 | SVM | 0.779 | 0.57 | 0.63 | 0.60 | Stable, reliable, and consistent |
| 4 | Logistic Regression | 0.774 | 0.56 | 0.66 | 0.61 | High recall, linear model |
| 5 | Decision Tree | 0.769 | 0.56 | 0.61 | 0.58 | Overfitting, lower overall performance |
| 6 | KNN | 0.752 | 0.53 | 0.50 | 0.52 | Weak on class 1 (churners), poor generalization |

* The **Voting Classifier**, combining **Random Forest, XGBoost, SVM**, and **Logistic Regression**, was selected as the final model for deployment.
* The model provides:
  + **Binary churn prediction (Yes/No)**
  + **Churn probability (%)**
  + **Churn risk level** categorized as **Low**, **Moderate**, or **High**
* The system was deployed using **Streamlit**, with features such as:
  + **Real-time churn prediction**
  + **Batch predictions via CSV upload**
  + **Prediction logs for record tracking**
* This system enables telecom companies to **identify at-risk customers** and take **proactive retention measures**, potentially reducing customer churn significantly.



**Conclusion**

**Key Findings:**

* + Customers with shorter tenure, higher monthly charges, and no tech support, Online Security, Device Protection are more likely to leave.
  + Customers with long-term contracts and lower monthly charges are less likely to leave.

**Modeling Outcome**

* + Multiple machine learning models were evaluated.
  + Random Forest delivered the best overall performance, while Logistic Regression showed the highest recall, useful for detecting churners.
  + An Ensemble Voting Classifier combining top models improved robustness and generalization.
  + A full-featured Streamlit app was deployed to enable real-time churn prediction and business usability.

**Recommendations for Telecom Companies:**

* + Offer discounts on monthly charges and long-term contracts to reduce churn.
  + Improve customer support and engagement to keep customers happy.
  + Introduce or improve value-added services (e.g., Tech support, device protection).