**Problem Statement: Customer Churn Analysis**

**Background:**

You are working as a Data Scientist for a subscription-based video streaming platform. The company is struggling with customer churn, where customers cancel their subscriptions. Churn directly impacts revenue, and reducing it is a top priority for the business. The dataset contains detailed customer attributes, including engagement, behavioral patterns, financial history, and social interactions.

Your role is to analyze the dataset, build a robust predictive model, and provide actionable insights and recommendations to the company’s management team. The insights should help develop retention strategies and focus on specific customer segments.

## ****Tasks for the Candidate****

### **Phase 1: Exploratory Data Analysis (EDA)**

1. Analyze the dataset to uncover trends, patterns, and anomalies.
2. Identify and visualize relationships between customer characteristics, behavior, and the target variable.
3. Segment the data to find groups that exhibit distinct patterns related to the target variable.
4. Highlight any significant correlations or dependencies within the data.

### **Phase 2: Data Cleaning and Preprocessing**

1. Detect and handle missing values in the dataset using appropriate methods.
2. Identify and address outliers across multiple features.
3. Assess the data for noise or inconsistencies and clean it to ensure accuracy.
4. Apply transformations, scaling, or normalization to features where necessary.

### **Phase 3: Feature Engineering**

1. Evaluate existing features to determine their usefulness in predicting the target variable.
2. Engineer new features based on interactions, aggregations, or domain knowledge.
3. Reduce dimensionality by addressing redundancies or highly correlated features.
4. Prepare the data for modeling by ensuring all features are in the appropriate format.

### **Phase 4: Machine Learning**

1. Split the dataset into training and testing subsets.
2. Train multiple machine learning models to predict the target variable.
3. Address class imbalance using appropriate techniques.
4. Evaluate model performance using a combination of metrics (e.g., classification accuracy, precision, recall, F1-score, and ROC-AUC).
5. Show probabilities along with churn

### **Phase 5: Insights and Recommendations**

* **Churn Drivers**: Identify key factors influencing churn based on the model and EDA.
* **Customer Segmentation**: Highlight customer segments most at risk of churn.
* **Retention Strategies**: Provide actionable recommendations for reducing churn, such as:
  + Incentives for high-risk customers.
  + Changes to subscription plans or content offerings.
* Communicate insights in a clear and structured manner, suitable for non-technical stakeholders.

## ****Dataset Overview****

The dataset includes the following columns:

### **Customer Demographics**

* **CustomerID**: Unique customer identifier.
* **Gender**: Gender of the customer (Male/Female).
* **Age**: Age of the customer.

### **Subscription and Financial Metrics**

* **SubscriptionType**: Subscription plan (Basic, Standard, Premium).
* **MonthlySpend**: Amount spent per month by the customer.
* **DiscountReceived**: Total discounts received in the past 6 months.

### **Engagement and Behavior**

* **TotalWatchTime**: Total hours of content watched in the past month.
* **AvgSessionDuration**: Average session duration (in minutes).
* **NumDevicesUsed**: Number of devices used to access the platform.
* **ContentPreferenceScore**: A score (0–100) indicating how well the platform matches the customer's preferences.
* **ActiveDaysLastMonth**: Number of active days in the past month.
* **BingeWatchRatio**: Ratio of binge-watched hours to total watch time.
* **SkippedContentCount**: Number of pieces of content skipped mid-play.
* **DaysSinceLastLogin**: Days since the customer last logged in.

### **Loyalty and Interaction**

* **ComplaintCount**: Number of complaints raised by the customer.
* **LoyaltyScore**: Loyalty score (0–10) based on historical engagement.
* **PlatformType**: Primary platform used (Mobile, Web, TV).
* **AdInteractionCount**: Number of interactions with ads on the platform.

### **Social and Referrals**

* **ReferredFriendsCount**: Number of friends referred by the customer.
* **SocialMediaMentions**: Number of social media interactions or mentions of the platform.

### **Target Variable**

* **Churn**: Whether the customer has churned (1 = Yes, 0 = No).

**Deliverables**

1. A **Jupyter Notebook** or Python script containing:
   * EDA, data cleaning, feature engineering, and model training.
   * Visualizations and model evaluation results.
2. A **Presentation or Report** summarizing:
   * Key insights and findings.
   * Model performance.
   * Actionable recommendations for reducing churn.