Problem Statement: Predict Customer Churn in at Camtel

Scenario:

You are tasked with building a machine learning model to predict whether a customer will churn (leave the service) in the next month **at Camtel.** The dataset includes customer behavior data such as usage patterns, customer support interaction, billing information, and demographic details. However, you've noticed that over time, customer behavior is changing due to new service offerings, pricing models, and market competition, leading to both **concept drift** and **data shifts**.

Dataset:

- **Historical customer data** for the past 3 years. It includes:
    - Customer demographic information (age, region, income, etc.).
    - Service usage data (monthly minutes, data usage, etc.).
    - Interaction data with customer service (number of support tickets, response times).
    - Billing information (monthly bills, outstanding balance).

    - Target variable: **Churn** (whether the customer left the company or stayed).
- The dataset has been partitioned into different time periods. You'll notice that patterns of churn behavior have changed over time due to shifts in customer preferences and the introduction of new service offerings.

Challenge:

You need to train a model that can handle:

1. **Concept drift:** The patterns that lead to customer churn have shifted over time due to changes in service quality, market competition, and customer expectations.
2. **Data shifts:** The distribution of customers (e.g., age groups, regions) has shifted due to expansion into new markets.

Tasks:

1. **Data Exploration and Preprocessing:**
    - Explore the dataset for missing values, anomalies, and correlations between features.
    - Split the dataset into training and test sets based on different time periods (e.g., first 2 years for training, the last year for testing).

- Handle class imbalance (if there are more non-churners than churners) by applying techniques such as oversampling, undersampling, or using performance metrics like AUC-ROC.

2. **Initial Model Training:**
   - Train a **logistic regression** or **decision tree** model using the first year's data. Evaluate its performance on the following year's data.
   - Investigate how performance changes over time. Do you see any signs of concept drift or data shifts (e.g., model performance deteriorates as you move to newer data)?
3. **Dealing with Concept Drift and Data Shifts:**
   - Apply **time-weighted learning** where more recent data is weighted higher when training the model.
   - Use **online learning** algorithms like **stochastic gradient descent (SGD)** to continuously update the model as new data comes in.
   - Incorporate **ensemble models** to combine models trained on different time periods (e.g., train one model for each year of data and ensemble their predictions).
4. **Evaluate Model Adaptation:**

    - Compare the performance of your adapted models to a baseline model that doesn't handle concept drift or data shifts.
    - Measure how the model handles both **older** and **newer** customer data.
5. **Drift Detection (Optional but Advanced):**
    - Implement a drift detection mechanism to monitor when concept drift occurs in the model's predictions (e.g., ADWIN or Page-Hinkley test).
    - Retrain or adapt the model dynamically when drift is detected.

Metrics to Monitor:

- **Accuracy, Precision, Recall, F1-Score**: Basic performance measures.
- **AUC-ROC**: For imbalanced datasets, this is useful to measure the trade-off between true positive rate and false positive rate.
- **Model performance over time**: Monitor if model performance declines as you move to more recent data (sign of concept drift or data shifts).

Outcome:

Your goal is to build a predictive model that not only performs well on historical data but can also adapt to future changes in customer behavior, ensuring it remains accurate over time.

DATASET LINK