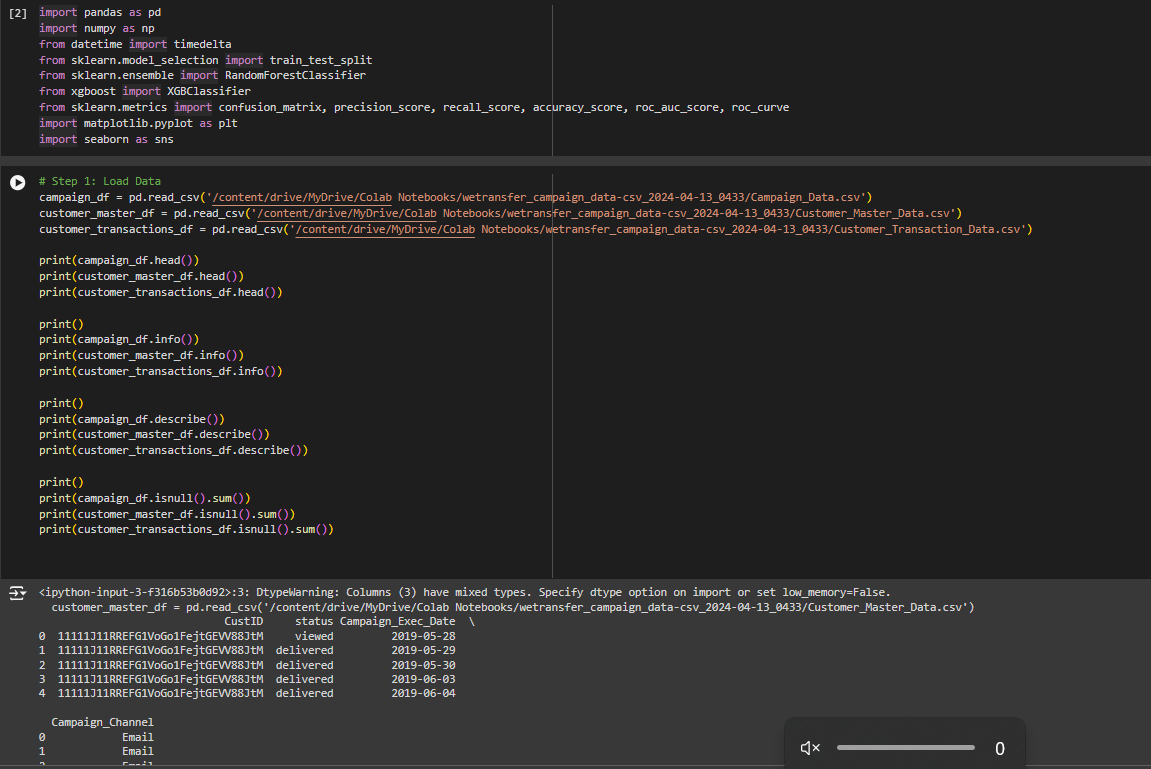
**Capstone Project - Case Study 2**

**Case Study 2 [Campaign Effectiveness]:**

* Campaigns are run by Croma to boost marketing and sales of their different products and categories. There can be different kinds of campaigns that are run – like cashbacks on specific credit cards, discounts during a certain duration, etc.
* Conduct analysis to help determine how effective campaigns that have been run historically are. The following information should be collated:
  + Campaign Details:
    - Each record should correspond to a particular Email or SMS campaign run
    - Details of the campaign
    - Date time when the campaign was run
  + Customer Details:
    - Customer ID (who was sent Email or SMS)
    - Age, Gender, Marital Status, Residential Pincode, Corporate, Tata Employee, Ethnicity
    - Number of Transactions (qty and value) in the Last 3/6/12 Months
    - % of Premium/Mainstream/Value Transactions
  + Campaign Outcome:
    - Whether there was any transaction within 1 month of campaign – accordingly create an outcome variable – 0 for no transactions and 1 for transactions (the transaction should have connection to the campaign)
    - Exclude records which were not sent any campaigns
    - Transactions which were not related to the campaign should not be considered either 0 or 1 but should be excluded
* Using the above data, develop a campaign effectiveness model using Random Forest and XGBoost – use a 80:20 train:test split
* Create a confusion matrix for both the models and report the Precision, Recall and Accuracy
* Also, report the KS statistic and the ROC area of the models
* Report Model Results appropriately using charts generated in Python – leveraging matplotlib/seaborn packages

### Step 1: Data Preparation

1. Load the Data: Import the dataset into a pandas DataFrame.
2. Convert Data Types: Ensure that the Campaign\_Exec\_Date column is in datetime format.
3. Merge Additional Customer Data: Assuming customer details and transaction data are available in separate datasets, merge them with the campaign data.



### Step 2: Data Cleaning and Feature Engineering

1. Filter Campaign Data: Exclude records which were not sent to any campaigns.
2. Create Outcome Variable: Identify transactions related to the campaign within 1 month and create an outcome variable.
3. Aggregate Transaction Data: Calculate the number of transactions and their value over the last 3/6/12 months, and the percentage of premium/mainstream/value transactions.

### Step 3: Model Building

1. Prepare Training and Test Sets: Use an 80:20 train-test split.
2. Build Random Forest and XGBoost Models: Train both models on the training data.

**Random Forest:**

1. \*Accuracy\*: 0.9814211878131304

2. \*Precision\*: 0.05405405405405406

3. \*Recall\*: 0.0003908317580345225

4. \*ROC AUC\*: 0.5001131354998049

The confusion matrix is as follows:

- True Negatives (TN): 271,887

- False Positives (FP): 35

- False Negatives (FN): 5,112

- True Positives (TP): 2

### Interpretation

- Accuracy is quite high at 98.14%, indicating that the model is correctly predicting the class labels the majority of the time.

- \*Precision\* is very low at 5.41%, suggesting that when the model predicts a positive class, it is correct only about 5.41% of the time.

- \*Recall\* is extremely low at 0.039%, indicating that the model is missing almost all of the positive cases.

- \*ROC AUC\* is around 0.5, which is equivalent to random guessing.

**Case Study Context:**

* You are likely trying to predict customer engagement (e.g., clicking on a campaign, making a purchase).
* The dataset appears to be highly imbalanced, with very few positive cases (customers engaging) compared to negative cases (customers not engaging).

**Model Performance Issues:**

* Despite high accuracy, the extremely low precision and recall indicate that the model is not effectively identifying true positive cases.
* The ROC AUC close to 0.5 confirms that the model is not able to distinguish between engaged and unengaged customers.

### Step 4: Model Evaluation

1. Confusion Matrix: Create confusion matrices for both models.
2. Performance Metrics: Calculate Precision, Recall, Accuracy, KS statistic, and ROC area.
3. Visualizations: Use matplotlib and seaborn to create visualizations for model results.

Yes, the KS statistic and ROC curve analysis are effective tools for evaluating how well models predict the effectiveness of historically run campaigns. Here’s how they align with the objective of determining campaign effectiveness:

### KS Statistic

* Purpose: The KS (Kolmogorov-Smirnov) statistic measures the maximum distance between the cumulative distributions of the positive and negative classes (responders and non-responders) predicted by the model.
* Relevance: By computing the KS statistic, you assess how well the model discriminates between customers who respond positively to campaigns (buy products) and those who do not. A higher KS statistic indicates better model performance in separating these two groups.

### ROC Curve

* Purpose: The ROC (Receiver Operating Characteristic) curve plots the true positive rate (Sensitivity) against the false positive rate (1 - Specificity) for different threshold values.
* Relevance: The ROC curve visually represents the trade-off between sensitivity and specificity of the model. The AUC (Area Under the Curve) quantifies the overall discriminative ability of the model. A higher AUC suggests that the model has better overall predictive performance.

### Application to Campaign Effectiveness Analysis

* Evaluation: When evaluating campaign effectiveness historically, you aim to determine how well the models predict whether a customer will make a transaction (purchase) in response to a campaign.
* Metrics Interpretation:
  + A high KS statistic indicates that the model can effectively rank customers based on their likelihood to respond positively to campaigns.
  + A high AUC in the ROC curve suggests that the model has good discriminatory power in distinguishing between positive and negative outcomes (transactions related to the campaign versus no transactions).

### Conclusion

By leveraging the KS statistic and ROC curve analysis, you can quantitatively assess and compare the predictive performance of models (such as Random Forest and XGBoost) in the context of historical campaign effectiveness. These metrics provide insights into how well the models capture the relationship between campaign variables and customer responses, aiding in strategic decision-making and optimization of future marketing efforts.