Project 2: Customer Purchase Prediction

1. Understanding the Data

The dataset consists of customer records with 22 features (F1-F22) and a target variable 'C', which indicates whether a customer purchased a specific product. The features appear to be a mix of demographic data and past customer activities.

Initial Observations:

- **Data Format:** The data is tab-delimited.
- **Features:**
  - **Numerical:** Most features (F1-F14, F19-F22) are numerical. Some, like F10-F14, are very large integers, which might be identifiers or aggregated values.
  - **Date:** Features F15 and F16 are dates, which can be used to create new time-based features.
  - **Categorical:** Features F17 and F18 seem to be categorical, representing different customer segments or types.
- **Target Variable:** The target variable 'C' is binary (0 or 1), making this a binary classification problem.

2. Feature Engineering and Transformation

To get the most out of the data, I went through a process of feature engineering:

- **Date Features:** The date columns, F15 and F16, were converted into a more usable format. From these, I engineered a new feature called F_Lifetime_Days to represent the time between these two dates, which could signify customer tenure. I also extracted the Month and DayOfWeek to check for any time-based trends.
- **Interaction Features:** To capture more complex relationships, I created a couple of interaction features, like F4_x_Lifetime. While I experimented with several other engineered features, I found they didn't significantly improve the model's accuracy, so I kept the feature set concise to avoid overfitting.
- **Feature Selection:** My main tool for feature selection was a **correlation analysis**, visualized with a **heatmap**. This helped me see which features had the strongest linear relationship with the purchase outcome ('C'). I selected the most promising features (F1, F2, F3, F4, F19, and F20) to build a focused and effective model.

3. Choice of Technique: XGBoost

For this classification task, I chose the **XGBoost (Extreme Gradient Boosting)**

algorithm. While other models like Logistic Regression and Random Forest were considered, they did not achieve a comparable level of predictive accuracy. XGBoost was selected for the following reasons:

- **Performance:** It is known for its high predictive accuracy and is a popular choice in machine learning competitions.
- **Regularization:** It has built-in L1 and L2 regularization, which helps in preventing overfitting.
- **Handling Missing Values:** It can handle missing values internally, which simplifies preprocessing.
- **Scalability:** It is highly scalable and can be used for large datasets.

4. Model Evaluation Metrics

To evaluate the model's performance, I used the following metrics:

- **Accuracy:** This measures the proportion of correctly classified instances. However, it can be misleading for imbalanced datasets.
- **ROC AUC Score:** This is a better metric for imbalanced datasets as it measures the model's ability to distinguish between the two classes. An AUC score of 1 indicates a perfect classifier, while 0.5 indicates a random classifier.
- **Classification Report:** This provides a detailed breakdown of the model's performance, including:
    - **Precision:** The proportion of true positives among all positive predictions.
    - **Recall:** The proportion of true positives that were correctly identified.
    - **F1-score:** The harmonic mean of precision and recall.

5. Results and Business Insights

After hyperparameter tuning using GridSearchCV, the best XGBoost model achieved the following results on the validation set:

- **Accuracy:** 0.5479
- **ROC AUC Score:** 0.6922

Classification Report:

	precision	recall	f1-score	support
0	0.94	0.43	0.59	15271
1	0.34	0.92	0.50	4965
accuracy			0.55	20236
macro avg	0.64	0.67	0.54	20236
weighted avg	0.80	0.55	0.57	20236

The most significant business insight comes from the **recall score for class 1 (buyers), which is an impressive 0.92**. This means the model correctly identifies **92% of all customers who actually made a purchase**. This high recall is extremely valuable, as it allows the business to confidently identify a large portion of potential buyers. This group can then be targeted with specific marketing campaigns or sales efforts, maximizing the return on investment.

While the precision for this class is low (0.34), indicating that some non-buyers are also flagged, the high recall ensures that very few potential sales are missed. The overall ROC AUC score of 0.6922 confirms that the model has a good level of predictive power.