**Project Goal, Simplified**

**Our Simple Goal: Get More Bang for Our Buck**

The entire point of this project is to figure out exactly **where to spend our advertising money** to get the most customers.

Right now, we're spending a lot of money (high Customer Acquisition Cost, or **CAC**) on ads that don't always work. We're going to use this data to build a smart system that helps us stop wasting money on bad clicks and focus all our effort on the users who are most likely to buy our product.

**The Business Problem (What We're Trying to Solve)**

**The Big Challenge: Wasting Money on the Wrong People**

We're running many ads—on Facebook, Google, other websites—but we don't truly know **which ad and which person is worth the most investment**. It’s like throwing darts in the dark.

* **The Cost Problem:** Our current ads are expensive because too many people click but never buy anything. This makes our overall customer cost too high.
* **The Intelligence Gap:** We can see *how many* people bought something (*reporting*), but we can't tell the marketing team *which new person* is about to buy (*predicting*).

**Assumptions**

### The Trend of Future Data Will Also Be Same.

**The Technical Term:** We call this **Data Stability** or **Non-Drift**.

**Simplified Takeaway:** **"What worked yesterday will keep working tomorrow."** If our target audience dramatically changes their online habits, the model will need to be re-trained.

### We Must Keep All Features on Which Model Will Be Trained.

**The Technical Term:** This is a **Deployment Dependency**.

**Simplified Takeaway:** **"If the model needs a piece of information, we must guarantee we can collect it for every new person."** No missing information is allowed for making a prediction.

* 1. **The Data Accurately Represents the Real Customer.**

### We Have the Ability to Act on the Prediction

**The Assumption:** The marketing team *can* adjust their ad spend or targeting in real-time (or near real-time) based on the model’s prediction.

## Marketing and Conversion Hypotheses

* **Demographic Hypothesis:** Clients who successfully convert (Conversion = Yes) will predominantly belong to the **Age** group of less than 18 and will be mostly **female**.
* **Conversion Efficiency Hypothesis:** The overall **Conversion Rate** (Conversions divided by Clicks) is expected to be greater than 50% of the **Click-Through Rate** (Clicks divided by Views).
* **Channel Performance Hypothesis:** The **Social Media** channel is expected to account for the majority of successful conversions, either in raw count or as a percentage of clicks for that channel.
* **Email Engagement Hypothesis:** Customers whose volume of **Previous Purchases** is above the dataset's average are the same group of customers who actively register **Email Clicks**.
* **Loyalty Hypothesis:** There is a **direct, positive correlation** between accumulating a high number of **Loyalty Points** and achieving a positive conversion outcome (Conversion = Yes).
* **Ad Quality Hypothesis (Inverse Relationship):** Campaigns that exhibit an extremely high **Click-Through Rate** are expected to correlate with a subsequent **low Conversion Rate**, suggesting that the audience clicking the ad is of poor quality.
* **Engagement Hypothesis (Behavioral):** Converting customers (Conversion = Yes) will display significantly higher engagement metrics, specifically in their **average Time on Site** and **Pages Per Visit**, compared to customers who do not convert.

## Research Questions

### I. Customer Segmentation & Performance

1. Which **Gender** segment exhibits the **highest Conversion Rate**?
2. Which **Gender** segment exhibits the **highest Click-Through Rate**?
3. Which **Age Group** exhibits the **highest Conversion Rate**? (Age Groups Teenager(18 to less than 21), Young (21 to less than 50), Old (50 and above)).
4. Which **Age Group** exhibits the **highest Click-Through Rate**?
5. Which **Income Group** exhibits the **highest Conversion Rate**? (Income Groups: Lower-Class (≤25% percentile), Middle-Class (≤50% percentile), Upper-Class (≤75% percentile), Elite-Class (>75% percentile)).
6. Which **Income Group** exhibits the **highest Click-Through Rate**?

### II. Campaign and Channel Effectiveness

1. Which **Campaign Channel** has the **top highest Click-Through Rates**?
2. Which **Campaign Channel** has the **top highest Conversion Rates**?
3. Which **Campaign Channel** has the **lowest Click-Through Rates**?
4. Which **Campaign Channel** has the **lowest Conversion Rates**?
5. Which **Campaign Type** has the **top highest Click-Through Rates**?
6. Which **Campaign Type** has the **top highest Conversion Rates**?
7. Which **Campaign Type** has the **lowest Click-Through Rates**?
8. Which **Campaign Type** has the **lowest Conversion Rates**?

### III. Top/Lowest Customer Behavioral Metrics

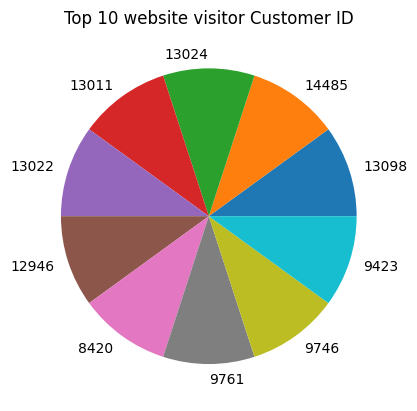
1. Who are the **top 10 customers** with the **highest** values for each of the following behavioral metrics: **Website Visits**, **Pages Per Visit**, **Time On Site**, **Social Shares**, **Email Opens**, **Email Clicks**, **Previous Purchases**, and **Loyalty Points**?
2. Who are the **top 10 customers** with the **lowest** values for each of the following behavioral metrics: **Website Visits**, **Pages Per Visit**, **Time On Site**, **Social Shares**, **Email Opens**, **Email Clicks**, **Previous Purchases**, and **Loyalty Points**?
3. Which are the **top 10 customers** with the **highest volume of Previous Purchases**?
4. Which are the **top 10 customers** with the **lowest volume of Previous Purchases**?

### IV. Correlation Analysis

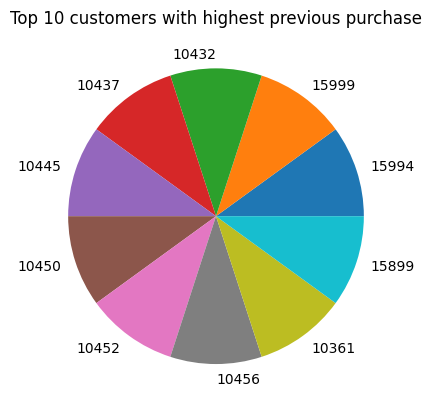
1. What is the correlation between **Income** and **Conversion Rate**?
2. What is the correlation between **Income** and **Click-Through Rate**?
3. What is the correlation between **Ad Spend** and **Conversion Rate**?
4. What is the correlation between **Ad Spend** and **Click-Through Rate**?
5. What is the correlation between **Email Opens** and **Email Clicks**?
6. What is the correlation between **Time On Site** and **Loyalty Points**?

### Essential ML/Classification Questions

1. **Model Performance:** What is the performance of the final classification model (e.g., Precision, Recall, F1-Score, and AUC) on unseen data, and does it meet the business requirement for accuracy?
2. **Misclassification Analysis:** Where does the model fail? Which customer segments or behavioral profiles are most frequently **misclassified** (e.g., predicted to convert but do not, or vice versa)?

****

This pie chart shows the CustomerIDs of top customers. I have to mention customer ids as you don’t provided me the names of customers due to security of data.

****

This pie chart shows top 10 customers with highest previous purchase where each top 10 has equal 9 purchases

**Model Insights and Findings**

**Low Linear Correlation:** The analysis confirms a weak or non-existent *linear* relationship between most numerical features. This justifies using a non-linear model (like Decision Tree/Random Forest) over simpler linear methods (like Logistic Regression).

The convergence rate is independent of age, gender and income group.

**Channel Homogeneity:** All primary campaign channels and types exhibit remarkably similar performance metrics (approx 10.4% Conversion Rate and approx 15% Click-Through Rate). This suggests a need for deeper segmentation, as no single channel currently offers a superior efficiency edge.

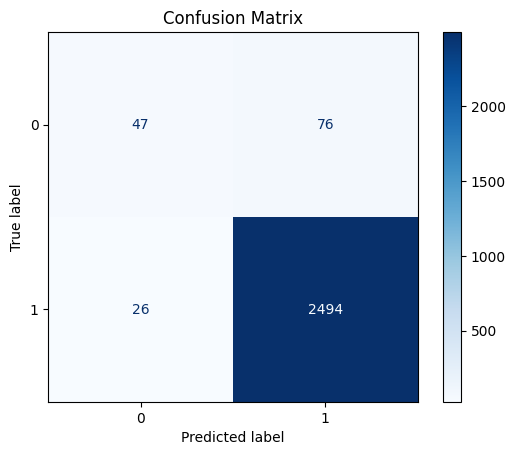
**Data Imbalance Management:** The primary modeling challenge was the severe class imbalance in the target variable. Future work requires more advanced techniques (e.g., SMOTE, ADASYN) or dedicated cost-sensitive learning to ensure robust performance on the minority (Conversion = Yes) class.

**High Accuracy but Misleading:** The Decision Tree classifier achieved $97\%$ overall accuracy on the test set. However, due to severe data imbalance, this high score is potentially misleading and requires validation using Precision, Recall, and F1-Score on the minority class.

**Model Insights and Information**

The model has accuracy of 97% on unseen data.





These graphs prove that while the classification model demonstrates a strong 97% accuracy, further analysis is required, as performance metrics are heavily influenced by the underlying data imbalance.

## Model Development Challenges

* **Model Selection Difficulty:** Initial attempts using Logistic Regression, Random Forest, and Decision Tree all returned high overall accuracy but suffered from unsatisfactory Precision, Recall, and F1-Scores due to the imbalanced data. The Decision Tree was selected as the best performing among the initial set.
* **Data Volume:** The small dataset size ($\approx 8,000$ rows) limits the complexity of the models and the depth of feature engineering, making it challenging to identify complex, non-obvious correlations.
* **Feature Understanding:** A lack of clarity regarding the correlations between features hindered preliminary data analysis, necessitating a reliance on automated model selection to reveal feature importance.

**Hypothesis Testing**

**"None of our original beliefs about our customers or ad performance were proven right.** We need to throw out our old assumptions and let the machine learning model show us the *real* customer profile."