Group 9: Anjana Khabir, Kush Desai, Shruthi Suresh, Nadifa Hossain

<u>Problem Statement</u>: Identify and target persuadable customers from the Wave-1 non-responder group to optimize the Wave-2 marketing upsell campaign.

Introduction

In the Wave-1 campaign, 801,821 business customers were targeted, but not all upgraded. From a random sample of 75,000, we aim to identify those most likely to upgrade in Wave-2. Instead of mass emailing, we seek a targeted approach to reduce costs, minimize annoyance, and maximize profitability. Using logistic regression and neural networks (with and without interaction terms, and new features) we'll segment persuadable customers and compare targeted marketing to mass outreach. The final recommendation will be based on predictive performance, response rates, and profitability to optimize customer selection for Wave-2.

Exploratory Data Analysis:

Our dataset consists of 75,000 records, with 22,500 initially designated as test data.

Key findings are as follows:

a. Data Integrity: No missing values, ensuring a complete dataset.
b. Distributions & Visualizations: The purchase amount (dollars) is right-skewed, indicating most customers buy less value. Recent purchases increase upgrade likelihood, and engagement is declining over time
c. Correlation Analysis: No extreme multicollinearity; tenure influences upgrade decisions, with long-term customers either staying or upgrading.
d. Customer Segmentation: High-value customers respond at higher rates and should be prioritized. TurboTax users show stronger engagement, making them an ideal target.
e. Gender & Response Rates: No significant difference, so gender shouldn't influence targeting.
f. Outlier Analysis: High-spending customers (4.58%) will be retained for insights into premium behavior.

In short, we should prioritize recent and high-value customers in Wave-2 marketing. Gender is not a key factor. Segmentation based on customer tenure and spending behavior can enhance targeting. Re-engagement strategies are crucial as overall customer engagement declines.

Feature Engineering

During feature engineering, we experimented with adding multiple new columns that logically aligned with customer behavior. However, after testing these features, we observed the following:

a. Multicollinearity became an issue, rather than improving model performance, they introduced redundancy.
b. For Instance, the new feature "average order value" was strongly correlated with the existing dollar column, making additional engineered features unnecessary.
c. Removing the newly engineered features had no impact on model performance, indicating they were not significant enough to justify inclusion.

Some examples of dropped Features Due to multicollinearity and redundancy:

a. Avg_Order_Value (dollars/numords) – Hrs).
b. Is_Recent_buyers (customers who purchase within the last 3 months) – Strong overlap with numords.
c. Loyalty Score / Recency Frequency Ratio – Overlapping with last.highly correlated with total dollars spent
d. Rea1_Engagement_Score – Provided no additional value beyond the last.
e. Version1_and_Upgraded – Constant value leading to no variance.
f. Has_Owned_Tax_Prod_and_Upgraded (customers who previously bought a tax product and are more likely to upgrade) – Identifies customers who previously owned a tax product and upgraded.
g. High_Spending_Customer, Customer Lifetime Value (CLV), and Is_Male / Is_Unknown_Gender

Ultimately, we **removed the newly created columns** and focused on retaining only the most meaningful variables, ensuring the model remained interpretable and effective without introducing unnecessary complexity.

Interactions

The interaction plots were analyzed to determine whether adding interaction terms to the logistic regression model would improve predictive performance. However, after multiple trials, we observed that the interaction between 'bizflag' and 'upgraded' showed a moderate effect but was insignificant in terms of performance and p-value. Therefore, the base model without interactions appears to be sufficient for our analysis.

Modeling

1. Overview of Predictive Modeling Approach

The goal of this analysis was to predict which businesses that did not respond to the Wave-1 mailing are most likely to respond in Wave-2. To achieve this, we developed predictive models using Logistic Regression and Neural Networks, leveraging historical purchase behavior and business attributes to optimize our marketing strategy.

2. Development of Predictive Models

Logistic Regression Model

We developed a logistic regression model to predict response probability (res1_yes). The model included key independent variables such as:

- Purchase Behavior: numords (number of orders), dollars (total spending), last (time since last purchase)
- Product Indicators: version1, owntaxprod, upgraded
- Geographic Attributes: zip_bins
- Interaction Effects: dollars:bizflag and owntaxprod:upgraded

We assessed multicollinearity using Variance Inflation Factor (VIF) and found no severe correlations. The final logistic model achieved an AUC of 0.755, indicating good predictive power.

Neural Network Model

We also implemented a Multi-Layer Perceptron (MLP) to capture non-linear relationships. The best-performing network had one hidden layer (5 neurons) with tanh activation. This model slightly outperformed logistic regression, achieving an AUC of 0.775.

3. Model Comparison and Evaluation

We compared models using:

- AUC (Area Under Curve): Neural Network (0.775) vs. Logistic Regression (0.755)
- Interpretability: Logistic regression was easier to explain, making it useful for understanding key drivers of response.
- Feature Importance: zip_bins, upgraded, and last were the strongest predictors, while bizflag and dollars had minimal impact.

Based on the performance we decided to move forward, with neural network model for our final targeting.

Selection Criteria for Wave-2 Mailing

To determine whom to target in Wave-2, we followed these steps:

1. Filtered test set non-responders (res1_yes = 0).
2. Computed predicted response probabilities using logistic regression.
3. Applied the breakeven threshold (0.0235) to filter businesses where expected profit exceeds mailing cost.
4. Selected 5,830 businesses for the Wave-2 mailing based on predicted response rates.

Expected Profit Estimation for Wave-2

Based on our best model i.e Neural Networks below is the final profit and ROME Analysis:

- Response Rate (scaled): 3.55%
- Buyers in Wave-2: ~206 expected responders
- Total Cost: $8,220
- Total Revenue: $12,402
- Projected Profit: $4,182
- Return on Marketing Expenditure (ROME): 50.8%

Key Learnings About Business Upgrade Behavior

- **Business vs. Individual:** bizflag had **minimal impact**, meaning businesses and individuals exhibited similar response behaviors.

- **Past Upgraders Respond More:** Businesses that previously upgraded had a significantly higher likelihood of responding.
- **Spending History Alone is Weak:** dollars had a **low correlation with response**, indicating that spending alone does not predict upgrade likelihood.

Conclusion

Our analysis identified the most persuadable customers from the Wave-1 non-responders to optimize the Wave-2 marketing campaign. Instead of mass emailing, we used predictive modeling (Logistic Regression & Neural Networks) to target high-potential customers, reducing costs and improving efficiency. The Neural Network model (AUC 0.775) outperformed logistic regression and helped us select 5,830 businesses most likely to upgrade, maximizing response rates and profitability. Key insights show that past upgraders are more likely to respond, while spending history alone is not a strong predictor. By implementing a data-driven approach, we expect a 50.8% return on marketing expenditure (ROME), ensuring a smarter, cost-effective strategy for Wave-2.