Intuit Report

EDA

Zip Bins vs. Res1 Analysis: This analysis explores the relationship between geographic distribution (zip code bins) and the response to the first wave of mailing (Res1). It was found that customers in the first zip bin had a significantly higher response rate compared to other bins. This suggests that geographic areas represented by the first zip bin may be more receptive to marketing efforts and could be targeted again for better response rates.

Numords (Number of Orders) Impact: We see a gradual increase in the positive responses with the recent records having the highest response and the older ones having lower positive response. This also goes well with our previous findings where we observe steady increase in the number of recent orders.

Last Order Timing (Last): Analyzes the effect of the recency of the last order on marketing response. Customers who made purchases more recently were more likely to respond positively, highlighting the importance of recent engagement in predicting marketing success.

QuickBooks Version (Version1): Evaluates the influence of using QuickBooks version 1 on response rates. Users of version 1 showed a slightly higher likelihood to respond, which might indicate specific product satisfaction or a tendency towards loyalty among this group.

Owntaxprod (Tax Software Purchase): Here we look at how purchasing tax software influences response rates to marketing. Customers who purchased tax software were more slightly more likely to respond, suggesting that this product purchase might be a predictor of engagement.

Upgraded from Version 1 to 2: Customers who upgraded their QuickBooks software from version 1 to 2 were significantly more likely to respond to marketing efforts, indicating that customers who engage with product updates are also more responsive to marketing.

Additionally, we observed that Sex, bizflag, dollars and sincepurch didn't have any significant importance when it comes to the response in wave 1.

Model Comparison

1. Initial Logistic Regression Model: This model incorporated a wide array of variables, including demographic information, purchase history, and product interactions. Notable variables included sex, business flag, number of orders, total dollars spent, time since last order, QuickBooks version, and more. The model achieved modest predictive performance, with a Pseudo R-squared of 0.114 and an AUC of 0.755, identifying significant predictors such as number of orders and dollars spent, version1, owntaxprod, upgraded and zip_bins_factor.

2. Refined Logistic Regression Model: The second iteration focused on simplifying the model by removing variables with high p-values, indicating a lack of significant impact on the outcome. This refinement maintained the model's performance while streamlining the variable set for efficiency.

3. Introduction of Zip Code Variables: The third model iteration significantly improved predictive power by incorporating new zip code variables for granular geographic insights. This is because from our data visualizations, EDA with res1 we had observed that one of the zip bins (i.e zip bin 1) had significantly higher response indicating zips inside this bin might be more responsive for the next mail wave. Exploring even deeper, we see there is a huge spike in the number of positive responses for the zip 00801 followed by 00804 and 00000, which have been then introduced as new variables/labels to the model.This addition raised the Pseudo R-squared to 0.148 and the AUC to 0.768, highlighting the importance of geographic data in predicting customer behavior.

4. Further Refinement with Geographic Insights: The fourth model refined the geographic variables by removing non-significant ones, like the '00000' zip code variable, streamlining the model without compromising its predictive capabilities.

5. Final Model Adjustments: The final iteration removed the zip_bins_factor variable due to its non-significant bins, slightly adjusting performance metrics but achieving the highest expected profits for both the test set and the total population. This model demonstrated a balance between simplicity and predictive power, with a Pseudo R-squared of 0.146 and an AUC of 0.766.

Comparison Metrics

1. Pseudo R-squared: A measure of the model's explanatory power, indicating how well the independent variables predict the dependent variable.

2. AUC (Area Under the ROC Curve): A metric used to evaluate the model's ability to distinguish between the classes. A higher AUC indicates a model with better prediction quality.

3. Log-likelihood: A measure of the likelihood that the model's parameters are correct given the observed data.

4. Permutation Importance: This technique was used to assess the importance of each variable in the model. By randomly shuffling values of each predictor and measuring the decrease in model performance (specifically, AUC decrease), the method identifies variables that significantly contribute to the model's predictive power.

5. Expected Profit Calculation: Beyond statistical metrics, the models were also evaluated based on their ability to generate expected profit. This practical measure considered the cost of marketing efforts and the potential revenue from positive responses, thereby providing a financial assessment of each model's utility.

Variable Creation

Zip Code Variables Creation:

Three new zip code variables were introduced to capture more granular geographic effects on customer behavior. These variables were derived from the 'zip5' field, representing different geographical areas or regions. The inclusion of these variables aimed to leverage geographic diversity, assuming that customers from different areas might exhibit varying response patterns due to regional differences in market dynamics, economic factors, or cultural preferences.

Wave-2 Mailing Criteria & Expected Profit

Given that the conversion rate from response to actual purchase or profitable action is 50%, the formula to calculate the Breakeven Response Rate (BRR) would adjust to account for this conversion probability:

BRR = Cost per Mail Piece / (Margin × Conversion Rate)

The breakeven response rate for the campaign is approximately 4.7%. This means that to break even, at least 4.7% of the recipients need to respond and go through the conversion process successfully, considering the cost per mail piece and the margin generated from each conversion.

Profit Calculation:

1. mailing_cost_lr1 calculates the total cost of mailing. It multiplies the mailing cost by the proportion of individuals targeted by the model and the total count of the population targeted in the test set (22500).
2. margin_lr1 calculates the total margin from responders. It considers the margin per responder, the conversion rate (assumed to be 0.5 or 50%), the proportion targeted by the model, and the total population in the test set.
3. Profit_lr1 calculates the net profit by subtracting the total mailing cost from the total margin gained from responders.

Scaling to Target Population:

1. The adjusted population targeted by the model (target_pop) is calculated by applying the model's targeting proportion to the difference between the total population and the population already contacted.
2. mailing_trgt_pop_cost and margin_trgt_pop calculates the total mailing cost and margin for this adjusted target population, respectively.
3. profit_trgt_pop calculates the expected profit for the target population by subtracting the total mailing cost from the total margin.

The expected profit was calculated for both the test set and the total population, allowing for a comprehensive understanding of the campaign's potential financial impact. Specifically, for the test set, the expected profit was calculated to be $13,112.70. This figure represents the net profit anticipated from the subset of customers included in the test set, after accounting for the costs associated with the mailing campaign and the expected revenue from positive responses.

For the total population, which encompasses the broader customer base, the expected profit was estimated to be $444,861. This larger figure reflects the scaled-up potential of the campaign when applied to the entire customer base, highlighting the significant financial benefit that could be realized from a well-targeted wave-2 mailing campaign.

Type of Businesses likely to upgrade

Purchase History: Businesses with higher numbers of orders (numords) or higher spending (dollars) might be more inclined to upgrade, as these variables often indicate a greater reliance on or engagement with the service.

Engagement Metrics: Variables like last (time since last order) and version1 (usage of a specific version) might reveal how recent and type of engagement affects the likelihood of upgrading. Frequent and recent use might indicate higher satisfaction or need, leading to a higher propensity to upgrade.

Geographic Factors: With zip code variables (e.g., zip_bins_factor, zip5_rc_00801), the model might identify geographic areas with businesses more likely to upgrade, possibly due to regional economic conditions, market saturation, or the presence of industries that heavily rely on the product.

Previous Upgrades: The upgraded variable directly indicates businesses that have upgraded in the past, which could be a strong predictor of future upgrades due to demonstrated willingness or perceived value in enhanced services.

‘sex’,’bizflag’ and ‘owntaxprod’ didn’t have any significant impact on the prediction and have shown a similar districution which targeting the population for wave-2 mailing.

*graphs present in the RMD and notebook.