Oscar Li, Anara Myrzabekova, Suhang Xiang, Diego Gonzalez Donamaria

Group Project #1: Quicken Quickbooks Upgrade Assignment

Professor: Olivia Natan

Class: UGBA 167.2

Date: 21 March 2023

Q: Describe **how you developed** your response model and discuss its **expected predictive performance**. If you created **new variables** to include in the response model, please describe these as well.

A: To develop our response model, we first analyzed our dataset by tabulating the relationship between response rate and total dollars ordered. Out of 20,000 people, only 4.80% responded to wave one mailing. Our initial approach was to develop the RFM model, which categorized individuals based on their response rate into RFM index bins. We examined the mean probability of the model's prediction against actual data and observed patterns of average dollars by recency, frequency, and monetary decile. However, we concluded that the RFM model didn't take into account other individual characteristics that could impact the response rate. This resulted in a mailing size of 7,184 people with a 3.8% response rate and a profit of $6,387. We determined that the individuals predicted by the RFM model have low customer quality in responsiveness, and will use this as our baseline model to make comparisons. As a result, we plan to construct machine learning models to identify previously unseen patterns in our dataset.

To develop our second model, we initially constructed a naive logistic regression using all variables, including categorical variables such as gender and zip code. However, we anticipated that modeling on all variables without filtering would result in a poorer model fit. After analyzing the logistic regression model using summary statistics and odds ratio, we eliminated variables that were insignificant, such as gender (F, M, U), bizflag, sincepurch, and owntaxprod.

Our odds ratio indicated that all zip_bins were significant, but they decreased the odds of a second-wave mailing response by a factor of 1.29% within a range of 0.7% - 0.18%. This decrease was reasonable because each zip bin had 1000 customers and a lot of variability in a zip

code. We believe it can helped the model avoid overfitting the data, resulting in better generalization.

To evaluate our model's performance, we used gains, decile response means summary, and predicted response probability decile graphics, specifically focusing on decile 0, as it strongly predicted the bins of people likely to respond to the second wave mailing. We initially considered self-selecting zip_bins based on wave 1 respondent characteristics, grouping them by their response rate, and even interacting with bizflag with zipbins since QuickBooks products are geared towards small and medium-sized companies. However, we believe including all zip_bins is helpful in increasing response rate and higher model fit.

There was also no need to apply regularizaation because there were no signs of overfiting, more of model underfitting by Random Forest. Overall, we generated a function to built 15 different models based on different variable combinations. We selected the best model based on decile graphics and attempt to find a balance between all the numeric values such as gross profit, wave 2 mailing size, mean response rate from model prediction, gain's decile 0.

Here are the main **new variable details**: (xx = {logit, nn, random_forest}, # = range(1, 21)

- `response_prob_xx`: from the model predictions to calculate the response probability of each individual.
- `prob_dec_xx`: based on the response_prob_xx,' we built ten groups that had similar probability to conduct decile analysis.
- `adj_prob_dec_xx`: adjust the probability dec from high probability to low.
- 'adj_response_prob_xx': Assignment Note 3 "assume that the response probabilities in wave 2 are only 50% of the response probabilities you predict based on wave one responses."
- `target_xx`: mark all individuals which 'adj_response_prob' >= break_even
- `decile_response_xx` average decile response bins based
- `zip_bus_bin_#`: interaction base on bizflag * zip_bins

Q: What **criteria** did you use to decide who should receive the wave 2 mailing?

A: Our criteria is based on our understanding of the variables, how we interact them with other variable, and filtering. We also try to recognize patterns through decile analysis visualizations. Lastly, we let machine learning models to help us detect underlying nonlinear patterns across zipbins, assign individuals their appreciate response rate accordingly.

After experimenting with different variable combinations and interacting them, we decide to keep zip_bins 1 - 20 and all the other significant variables when filtering through odds ratio and p-value significance.

Here are the details of criterias that we used to adjust the variables in the Variable Filter Part and Part 2:

1. **RFM model**: RFM considers the response rate a variable to create response rate for each individuals and doesn't refer to other criteria. Thus, we built Logistic, Neural Network, and Random Forest to use other user characteristics.
2. **Odd ratio and p-value:** We threw out some variables which are not significant. Also, we refer to the contribution of the variables from the odd ratio to evaluate the effect of variables.
3. **New variables analysis**: Observe the interaction and effectiveness of new variables we created.
4. **Predicted response probability decile visualization** : Plots comparing model predicted probability decile with actual response decile.
5. **Gains Metric Visualization**: Plot of Model Gains by Response Decile and Response Probability. We also specifically looked at Decile 0.
6. **Profit and Modeling**: Gross profit and value of gains decile based on different models (**Logistic, NN, Random Forest**) with the same predictor variables.
7. **Characteristics of ('res1' = 1)**: we observe the customers who responded to wave one mailing.

Q: How much **profit** do you anticipate with your wave two mailing to a subset of 20,000 customers?

A: Wave 2 mailing to a subset of 20,000 customers, we target **5078** customers. And the net profit we predict is **$7548**, which is based on our chosen neural network model.

Detailed Calculation:
Response rate mean of choose model (after 50% reduce): 4.827%

Cost of mail: $1.41* 5078 = $7159

Gross profit: $60 * 5078 * 0.04827 = $14707

Net Profit: $14706.9 - $7159.98 = $7548

Q: What did you **learn** about the type of consumers who are likely to **upgrade**?

A: To determine which individuals are more likely to upgrade, we analyzed the characteristics of those who have previously upgraded and targeted those who have not. We found that consumers who are likely to upgrade have similar characteristics, so our decile analysis will group them into same bins. We observed significant differences in the 'numorder,' 'dollars,' and 'last' variables across different deciles. The deciles with high response rates were concentrated in specific regions, such as 'zip_bins_18', `zip_bins_12`, etc. Additionally, we examined if they had not responded and how long it has been since their last purchase to determine if they require an upgrade. Here is our visualization of the people who responded in wave 1. Customers who are likely to upgrade will have similar characteristics as wave 1.