Tartan Data Science Cup

Team: Heteroskedasticity Terminator

Introduction

This research is to analyze the likelihood of customers who will purchase eggs at Kroger in the following week. we have a dataset collected from Kroger database which records 967 unique households who made transactions for days 500 - 704 (spanning 205 days). Throughout our research, we apply the Recency, Frequency, and Monetary analysis (RFM) in order to make evaluation on the probability of purchasing eggs in the following weeks. By analyzing the potential predictors and their effects on the purchasing likelihood, Kroger can make better decision on customer service.

Methods

The dataset includes the demographics informations for each household and their purchasing information, which is categorized by money they spent, name of commodity and department. We first made dataset randomized and then used cross validation procedure to split the dataset into two segments, 80% of original data for training (for model building) and 20% for testing (to test accuracy). In both subsets, we conducted data cleaning procedure. Specifically, we aggregated predictor variables by household in days 500 - 697 and calculated the response variable (dev_eggs) based on data in days 698 - 704 (most recent week).

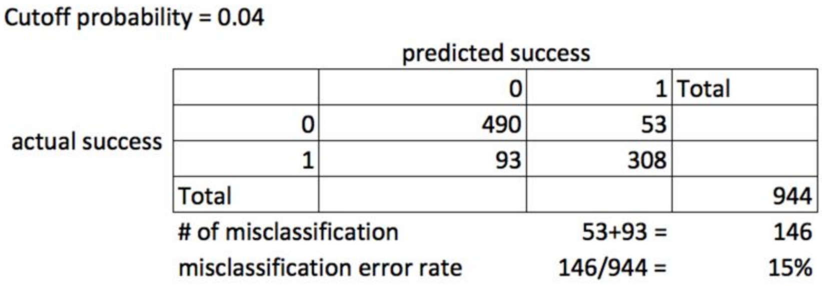
We chose predictor variables based on RFM (Recency, Frequency, and Monetary) analysis which is used to evaluate customer value. First, for recency, we recorded customer recency *visited_in*, tracing back from most recent day 697 to day 500, so the smaller the value is, the more recently that household visits. Second, for frequency, we measured the frequency of visits per household within specific periods in order to make prediction for day 698 - 704, such that *freq_14* means the number of visits for each household during the past 7-14 days, *freq_21* means the number of visits during the past 14-21 days, and *freq_total* is frequency of visits from the beginning of time. Moreover, we specifically calculated the frequency of visiting grocery department from the beginning of time (*Grocery_total)*. Third, for monetary, we focused on eggs, so we calculated total amount of eggs purchased from the beginning of time (eggs_total), average price paid on eggs per household (price_paid_eggs). Plus, we included two demographics information household's income *(income_desc)* and household composition *(hh_comp)* as predictor variables.

Based on our goal of predicting the likelihood of purchasing eggs in the following weeks, we determined the response variable as a binary classification variable (*dv_eggs*), which is equal to 1 if household actually bought eggs in day 698 - 704, 0 otherwise. Accordingly, we conducted our analysis by constructing logistic regression model and use testing subset to testify the validation of our model.

Result

After applying logistic regression model, we have dv_eggs = -3.2521 + (-0.0255)*visited_in + (-0.2868)*freq_14 + (0.2475)*freq_21 + (-0.001)*freq_total + (0.4870)*price_paid_eggs + (0.0019)*Grocery_total + (-0.6563)*hh_compSingle + (-0.2894)*hh_compNoKids + (-0.3120)*hh_compWithKids + (0.0990)*eggs_total.

Then we use testing subset as our new data frame to predict the likelihood of purchasing eggs in days 698 - 704. We specify the cutoff probability as 4%, which means we assume the household would buy eggs (success = 1) if its predicted probability is above 4% and household won't buy eggs (success = 0) if predicted probability is below 4%. According to Table 1, we calculated that among 944 households in testing subset, there are 308 households who are expected to buy eggs during the days 698 - 704 actually bought eggs in days 698 - 704. In addition, the misclassification error rate is 15% which indicates that only 15% of data are predicted wrong based on our logistics regression model.



Cutoff probability = 0.04

		predicted success		
		0	1	Total
actual success	0	490	53	
	1	93	308	
	Total			944
	# of misclassification		53+93 =	146
	misclassification error rate		146/944 =	15%

Table 1. Predicted Success (purchase eggs) vs Actual Success(purchase eggs)

Discussion

To sum up, we applied cross validation procedure and RMF analysis to choose predictor variables, then constructed logistic regression model for the likelihood of purchasing eggs the following weeks after day 704. Based on the training subset that is used to build model and testing subset that is used to evaluate prediction value, we have only 15% misclassification error rate. However, even though the misclassification error is quite low, there are some factors which jeopardize our prediction. First, the selection bias might exist because we removed some independent variables if they were not statistically significant in our model. Second, we only applied logistic model to predict the likelihood of purchasing eggs and choose predictor variables based on logistics regression analysis. Third, there are shortcomings in cross validation procedure with respect to objectivity. For more comprehensive cross validation, we need to split dataset into three subsets, training, validation and testing. In the future, we should apply other regression model like probit regression onto the dataset and pay more attention on potential influential points.