Marketing Campaign Analysis

Emily Wilkins, Matthew Zlotnik

Background and Introduction

What makes a consumer choose to purchase one good or service over another? By which avenues do most purchases happen in an increasingly digital world? What factors influence the success of a marketing campaign? Are certain customer segments more responsive than others? How do we decide which segments to target? These are some of the questions marketers, salespeople, and business analysis ask themselves everyday. And in a world that is more data saturated than ever, not only do analysts have more information to inform their predictions, but customers also have more information to inform their choices as well. In addition, businesses have more avenues by which to reach customers, whether that be through email, catalogs, pop-up ads and coupon offerings on their website, social media, newspaper, television, and so on. But with so many options, choices must be made. It is too costly to attempt customer outreach through every channel and to try to target every potential customer. The most successful promotional campaigns are those that minimize cost and maximize their positive response rate. So how do businesses decide how to structure their campaigns? Is it more a matter of timing and advertising channel or the demographics of the customers themselves? In this project, we hope to better understand what factors drive the success of promotional campaigns.

To investigate this topic, we are using a data set from Kaggle that contains data from a promotional campaign run by a grocery store. The data set contains information from over 2200 customers, including certain demographic information (age, marital status, income, education, number of young children in their home, ect.), the amount they spend in different product categories, how recently they purchased something, the number of purchases they've made through different avenues (in-store, catalog, and online), their responses being contacted the first through fifth time for a previous

campaign (noted in the data AcceptedCmp1-Accepted Cmp5 in the data), and whether they accepted the offer in the previous campaign. Before doing any EDA, our intuition was that we would attempt to answer one of two questions in this project, depending on the structure of the data. If we found that there were large differences in the campaigns, we'd want to investigate what makes a particular campaign successful. If the campaigns themselves were not revealing, but instead we saw clear trends in the customer that accepted the offers, we'd want to investigate which customers are the most responsive. Since this data set does not include any information on the costs or profits associated with these campaigns, we are only concerned with the response rate of the campaign for measuring its success. From our analyses, we hope to provide insights to marketers of what makes a promotional campaign successful in the grocery industry.

EDA

Before we could begin analysis on our data, we had to run exploratory analyses to ensure we had a grasp on the underlying structure of the data. Looking at the data, we realized that there were several columns titled "AcceptedCmp1" through 5. Given that the goal of our data was to predict which customers would accept a new promotional campaign, we assumed that these metrics would be helpful to show how customers had behaved in the past. According to the Kaggle competition website from which we got our data, these columns consisted of customers' responses to being contacted the first, second, third, etc. time for a previous campaign. When we compared these columns to the "Response" class which we were explicitly instructed to predict, we came across some troubling results. Shown in the table in Figure 1 are the results from the accepted campaign columns compared to the response column. We would expect that the accepted campaign columns would simply be a breakdown of which iteration of contact persuaded each individual to accept. However, as we can see, this is not the case, as customers who responded negatively to Response responded positively to individual campaigns. Now with the meaning of these variables very much in doubt, we were forced to abandon these as predictors.

Response	0	1
WhichCampaign		
0	1631	146
1	65	79
2	9	8
3	78	57
4	85	23
5	38	21

Figure 1: Campaign vs. Response

To see if there were any underlying collinearity or correlation issues, we plotted a correlation matrix of our remaining predictors to ascertain if any underlying collinearity problems exist within the data.

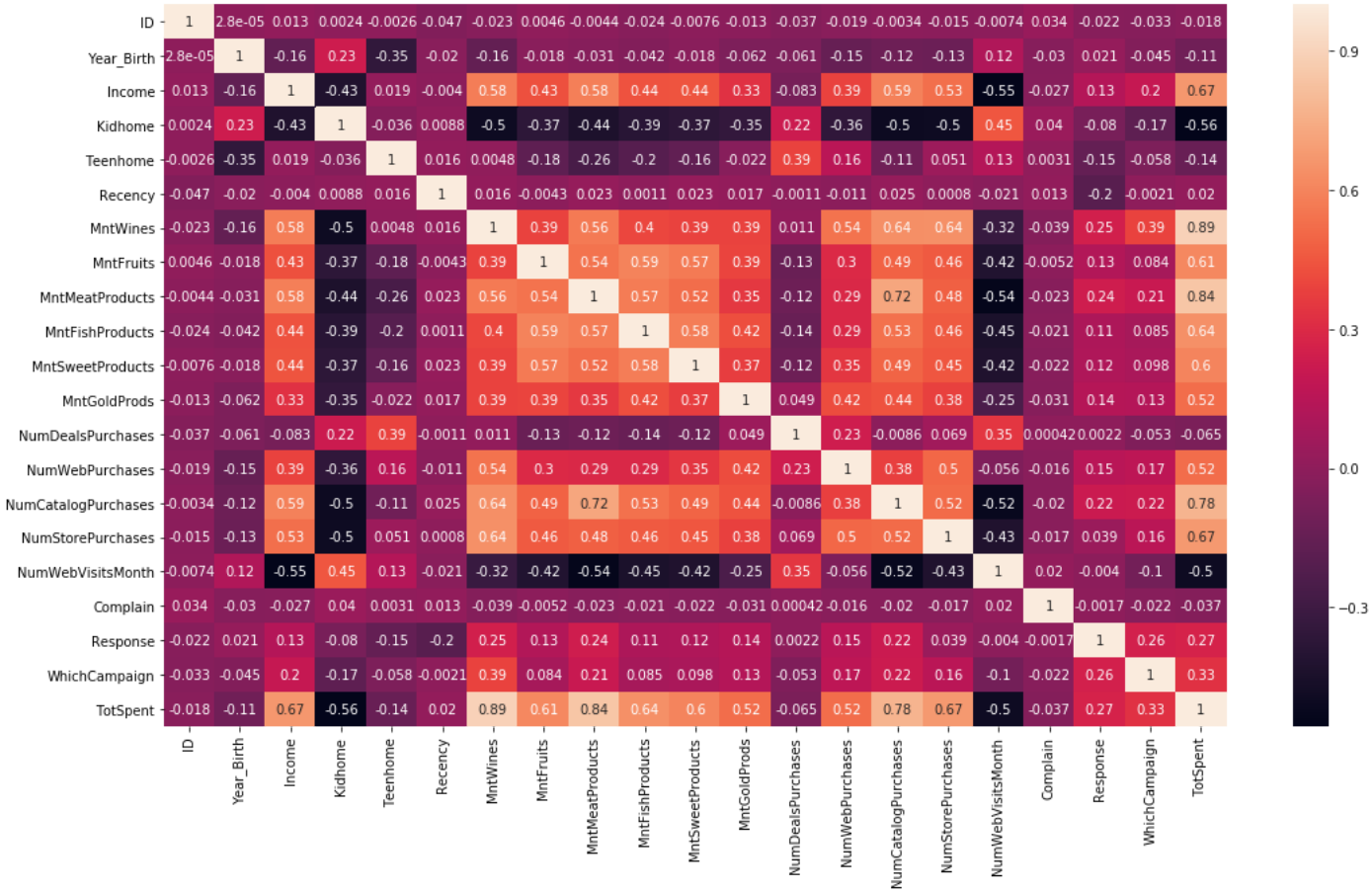

Figure 2: Correlation plot of all variables

There are many interesting correlations within this data to decompose. Firstly, it appears as though, unsurprisingly, a higher income would indicate more being spent on almost all items, though especially wine and meats. Additionally, having a higher income seemingly makes one more likely to have less web visits in a recent month. One way we interpreted this correlation is that higher income families had the means to purchase more of their items in one sale, necessitating less total visits to the website. One of the more confusing predictors in the data is “Kidhome”, a variable which indicates how many children are present in the household. It seems as though having a kid in the house correlates strongly with decreased spending in all categories except for deals purchases. As such, we can likely make the conclusion that for this data, families with small children at home may be single-income households, or in general younger and less wealthy than those with teens or no children left in the house and as such are more likely to shop using discounts.

In the correlations between type of product and avenue of purchase, a few correlations stood out as interesting. Firstly, meat and fish products correlated much less strongly with web purchases, as consumers do not trust buying easily expirable products online. Meat products and wine products tended to correlate more strongly with many other variables than did products of any other kind, possibly indicating that these products are more expensive, and thus more sensitive to changes in circumstance than other products.

Next, we looked at pairwise plots between some of our predictor variables and our response variable. Firstly, we look at the cumulative distribution plot of income for customers who did or did not respond to the last campaign. As these graphs show, customers who responded to the last campaign were of higher average income, and consisted of most of the wealthy (Income $90,000+) customers in the dataset. From this we see that maybe the more price-conscious consumers are not those with low income as initially thought.

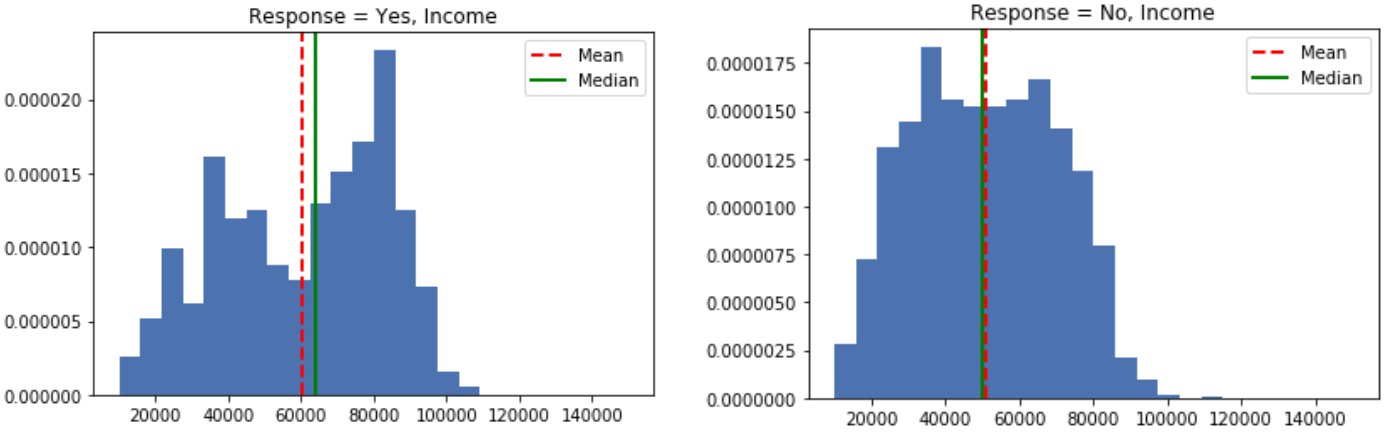

Figure 3: PDF of Income Stratified by Response

Next, we looked at the relationship between kids in the household and responding to the marketing campaign. The table in Figure 4 below indicates that parents with kids home are slightly less likely to have responded to the most recent campaign, with 5% young parents responding compared to households without children.

Response	0	1	Perc
Kidhome			
0	1071	222	17.0
1	789	110	12.0
2	46	2	4.0

Figure 4: Number of kids in household vs. Response

To compare with kids at home, we looked at the response rate amongst parents with teens at home. As shown in the table in Figure 5, households without teens are considerably more likely to respond than households with teens, showing almost twice as likely in a similarly sized sample.

Response	0	1	Perc
Teenhome			
0	921	237	20.0
1	938	92	9.0
2	47	5	10.0

Figure 5:

Finally, we looked at education level and its relationship to our response variable. In the table in Figure 6, we can see that educated individuals (those who have completed “graduation” or higher) are considerably more likely to respond to the last campaign, showing a 16.9% likelihood compared to only 10.3% for those with lower levels of completed education. Having now deciphered many of the basic relationships between our variables, we embarked upon modeling and predicting which customers would respond to the newest marketing campaign.

Response	0	1	Perc
Education			
2n Cycle	181	22	11.0
Basic	52	2	4.0
Graduation	975	152	13.0
Master	313	57	15.0
PhD	385	101	21.0

Figure 6: Education Level vs. Response

Analyses, Findings, Conclusions

To begin our analyses, we decided that we would run several different types of models, and compare their AIC, BIC, and parsimony to determine which would ultimately be selected for our recommendation. As our data is binary, we elected to use logit and probit regression models on the entirety of our data. Additionally, as our data contains many predictors, we decided to include a principal components analysis and subsequent logit and probit regression models to see if the models could be nearly as effective with far fewer parameters. To begin, we first ran a PCA algorithm to focus on areas of high variance in the data.

PCA

Because our data had so many predictors, we decided to try and add a measure of parsimony to our models by employing Principal Components Analysis in an attempt to reduce the dimensionality of our data. To do this, we first standardized all of our numeric data, and, because PCA is an unsupervised learning technique, dropped our response variable. Then we ran a PCA algorithm on the remaining data and analyzed the results.

```
 [1] 6.0893257 1.9513271 1.2312407 1.0193981 0.9893498 0.8255720 0.7637761 0.6424229 0.6146012
[10] 0.4880786 0.4493156 0.4235026 0.3897503 0.3561338 0.3052336 0.2448718 0.2161000
```

Figure 7: Eigenvalues of the Principal Components

Shown in Figure 7 are the eigenvalues of each of the principal components created by the algorithm. Abiding by the Kaiser Rule, we elected to retain only the PCs which had eigenvalues greater than 1, leaving only 4 principal components. However, the fifth principal component is nearly 1, so we examine that component more closely to determine whether or not to include it in our analysis. For our PCA to be meaningful in any way, these components had to be interpretable enough to make a regression on

their scores meaningful. Shown below in Figure 8 are the correlations between the initial standardized independent variables and the five computed principal components.

	Dim.1	Dim.2	Dim.3	Dim.4	Dim.5
Income	-0.74684649	0.070661432	-0.17341710	0.034560230	-0.003751780
Kidhome	0.67017994	-0.083319315	0.32493815	-0.106330308	0.007776559
Teenhome	0.11951333	0.761475688	-0.28775098	0.008092134	-0.019876244
Recency	-0.01413496	0.009252004	-0.04281370	-0.712926321	-0.693679363
MntWines	-0.75175855	0.254931393	0.03812593	0.038498313	-0.028910517
MntFruits	-0.70497157	-0.191297875	0.15919989	-0.051022572	0.068544976
MntMeatProducts	-0.80262802	-0.195527098	0.05085269	-0.034549549	-0.016577042
MntFishProducts	-0.72864965	-0.207401403	0.14025275	-0.042920190	0.038254059
MntSweetProducts	-0.70430712	-0.167605050	0.14984161	-0.064203356	0.010095987
MntGoldProds	-0.58036485	0.150790929	0.26593865	-0.027085627	-0.016254387
NumDealsPurchases	0.12599079	0.691664023	0.44252312	-0.066075597	0.003588332
NumWebPurchases	-0.55315737	0.524765484	0.28667751	0.016139658	0.034719826
NumCatalogPurchases	-0.81684254	0.010045573	0.01143736	-0.028097066	-0.023359418
NumStorePurchases	-0.74716998	0.242338729	0.04277239	0.021563339	0.023789424
NumWebVisitsMonth	0.64998668	0.313240150	0.42150010	-0.013031986	0.028317765
Complain	0.03867918	0.004213761	-0.07884152	-0.692600023	0.704303990
Age	-0.16263968	0.463742040	-0.63629284	-0.029081211	0.029573839

Figure 8: Correlations between features and the first five principal components

To interpret each of the PCs (in the chart called Dims), we must look at the variables which correlate highly with each dim. For dim1, we see that it has a strong positive correlation with kids home and web visits last month, while having a strong negative correlation with income, and amount spent on all five product categories as well as web, store, and catalog purchases. A customer/household which scores highly on Principal Component 1 is likely a low-income family with new kids at home and is only buying the essentials as they are needed. Dim2 shares a strong positive correlation with teens at home, deals purchases, web purchases, age, and amount spent on wine, while having no particularly strong negative correlations and weak negative correlations with the other four product categories. This likely indicates that a household which scores highly on Dim2 has slightly older parents with teenagers at home who enjoy drinking wine and purchase their goods with discounts online. Dim 3 shows strong positive correlations with deals purchases and number of web visits per month, while showing a strong negative correlation with age. Customers scoring highly in Dim 3 are likely younger, price conscious consumers who check the website frequently to find the newest deals.

They have no particular affinity towards any of the five product categories. Dim 4 shows strong negative correlation with recency and complain and weak correlations with all other features. The only product it has a positive correlation with is wine, but that correlation is still weak. Customers who score highly on dim 4 are likely customers who shop more often, as demonstrated by the strong negative correlation with recency, but they don't buy much when they shop. Out of all the product groups, they spend the most on wine. But while these customers like wine, they don't like whining based on their strong negative correlation with complain. These are not frequent shoppers, but as wine provides the highest margins in many retail spaces, these customers could be highly profitable to target. Finally, Dim 5 shows a very high positive correlation with complain and a strong negative correlation with recency, with no other particularly strong or weak positive or negative correlations. A customer who scores highly on Dim 5 likely shopped recently and was not happy with their purchase or service received. Since Dim 5 has an interesting interpretation, especially as the emotional opposite to Dim 4, we decide to include it in our analysis.

Regression on PCA

Once we had our PCA components selected, we then assigned scores to each of the rows of our initial dataset and replaced all of the independent variables with the PC scores for each of the five principal components we kept. We decided to run both logit and probit regressions on our PCA scores so that they could be compared to the logit and probit models of our given dataset. The results of the logit regression are shown below in Figure 9. As you can see, PC1, 3, 4, and 5 were all statistically significant. PC1 being negative, if you will recall from our earlier analyses of each PC, indicates that low-income families with young children at home who are only typically buying essentials are less likely to respond to the marketing campaign. While all of these coefficients will be moderately difficult to interpret, we can loosely determine that if the customer perfectly matches PC1, then they are 21.8% less likely, on average, to accept

the marketing campaign. PC3, 4, and 5 all have approximately equal positive coefficients in the regression, which indicates that younger price-conscious consumers, casual wine purchasers, and frequent shoppers that like to complain all are about 40% more likely, on average, to respond to the marketing campaign than people who do not relate strongly to these groups. The AIC of our PCA Logit model was 1671.381 and the BIC of the model was 1705.602. We will use these statistics to compare our five models.

```
Call:
glm(formula = Response ~ ., family = binomial(link = "logit"),
    data = PC_data)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-1.6035  -0.5913  -0.4217  -0.2781   2.6271

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept) -2.01526    0.07387 -27.281  < 2e-16 ***
PC1         -0.21750    0.02476  -8.785  < 2e-16 ***
PC2         -0.03850    0.04270  -0.902    0.367
PC3          0.40976    0.05860   6.993 2.70e-12 ***
PC4          0.42712    0.06801   6.280 3.39e-10 ***
PC5          0.44056    0.06935   6.353 2.12e-10 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 1875.5  on 2215  degrees of freedom
Residual deviance: 1659.4  on 2210  degrees of freedom
AIC: 1671.4

Number of Fisher Scoring iterations: 5
```

Figure 9: Logit Model using PCA data

The next model we ran was a probit regression on our PCA scores. The results of this regression are shown in the output in Figure 10. As is typical in the relationship between logit and probit regressions, the coefficients of the five principal components go the same direction as in the logit model, but have slightly smaller magnitudes across the board. This model would be interpreted almost entirely the same as the logit regression

on our principal components, and has an AIC of 1666.94 and a BIC of 1701.161, which are slightly better than the logit.

```
Call:
glm(formula = Response ~ ., family = binomial(link = "probit"),
    data = PC_data)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-1.5380  -0.5984  -0.4201  -0.2537   2.6962

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept) -1.17271    0.03787 -30.970  < 2e-16 ***
PC1         -0.12492    0.01369  -9.127  < 2e-16 ***
PC2         -0.01685    0.02376  -0.709    0.478
PC3          0.23532    0.03220   7.308 2.71e-13 ***
PC4          0.23283    0.03696   6.299 2.99e-10 ***
PC5          0.23976    0.03745   6.403 1.53e-10 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 1875.5  on 2215  degrees of freedom
Residual deviance: 1654.9  on 2210  degrees of freedom
AIC: 1666.9

Number of Fisher Scoring iterations: 5
```

Figure 10: Probit Model using PCA data

Regression on Given Variables

Next, we ran logit and probit regressions on the variables provided in the data to see if our PCA-logit mixture could be more predictive while also being more easily interpretable and more parsimonious. The results of our logit regression model are shown below in Figure 11. As you can tell, this model is considerably more complex, fitting a coefficient to each of our 17 variables, as well as each of the different possible values of our two factor variables (marital status and education). The variables and coefficients which emerged as significant from this regression were presence of a PhD,

Teen in the house, recency of last purchase, the amount spent on wines, amount spent on meat, amount spent on gold products, number of purchases using deals, number of purchases on the web, catalog, and store, and the number of web visits in the last month. Presence of a PhD has the strongest positive coefficient among all significant variables, indicating that the company should try as much as possible to target highly educated individuals with their marketing campaign. Amount spent on wines, meats, and gold products also had positive and significant coefficients in this regression, indicating that customers who spend more on these product categories are more likely to respond to the marketing campaign and thus should be prioritized in targeting. Additionally, the number of web visits in the last month and the quantity of web purchases had significant positive coefficients, suggesting that customers who enjoy shopping online may be more likely to respond positively to the marketing campaign. This model had an AIC of 1448.481 and a BIC of 1613.881, indicating that it performed considerably better than the PCA models listed before.

```
Call:
glm(formula = Response ~ ., family = binomial(link = "logit"),
    data = campaign_data2)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-2.7273  -0.5058  -0.3082  -0.1538   3.0568

Coefficients:
                        Estimate Std. Error z value Pr(>|z|)
(Intercept)            -0.974131   1.479501  -0.658  0.51027
EducationBasic         -1.038183   0.785225  -1.322  0.18612
EducationGraduation     0.175389   0.277876   0.631  0.52792
EducationMaster         0.360272   0.314901   1.144  0.25259
EducationPhD            0.803756   0.301479   2.666  0.00767 **
Marital_StatusAlone    -0.604754   1.930888  -0.313  0.75413
Marital_StatusDivorced -1.087996   1.462932  -0.744  0.45705
Marital_StatusMarried  -2.089670   1.455928  -1.435  0.15121
Marital_StatusSingle   -1.086474   1.456694  -0.746  0.45576
Marital_StatusTogether -2.115114   1.458801  -1.450  0.14709
Marital_StatusWidow    -0.923722   1.486488  -0.621  0.53433
Marital_StatusYOLO     -0.885959   2.043514  -0.434  0.66462
Income                  0.042435   0.080574   0.527  0.59843
Kidhome                 0.150692   0.102008   1.477  0.13960
Teenhome               -0.583799   0.098147  -5.948 2.71e-09 ***
Recency                -0.772765   0.076414 -10.113  < 2e-16 ***
MntWines                0.583069   0.095537   6.103 1.04e-09 ***
MntFruits               0.008081   0.082004   0.099  0.92150
MntMeatProducts         0.394516   0.099362   3.970 7.17e-05 ***
MntFishProducts        -0.086666   0.088922  -0.975  0.32974
MntSweetProducts        0.099052   0.083097   1.192  0.23326
MntGoldProds            0.225741   0.073511   3.071  0.00213 **
NumDealsPurchases       0.072833   0.082769   0.880  0.37888
NumWebPurchases         0.220247   0.082214   2.679  0.00739 **
NumCatalogPurchases     0.318577   0.103758   3.070  0.00214 **
NumStorePurchases      -0.569933   0.098310  -5.797 6.74e-09 ***
NumWebVisitsMonth       0.503035   0.098693   5.097 3.45e-07 ***
Complain                0.022916   0.074875   0.306  0.75956
Age                    -0.031775   0.074654  -0.426  0.67038
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 1875.5  on 2215  degrees of freedom
Residual deviance: 1390.5  on 2187  degrees of freedom
AIC: 1448.5

Number of Fisher Scoring iterations: 6
```

Figure 11: Logit model with given variables

The next model we ran was a probit regression on the independent variables given. Again, as is common with logit and probit models, this regression was extremely similar, with no coefficient sign changes or major differences in interpretation. The AIC and BIC of the probit regression were 1443.391 and 1608.791, indicating that this link function performed slightly better than the logit link.

Poisson Model with Random Effect

The final model we tested was a log-Poisson model with a random effect. Because teens at home seemed to be a major factor in each of our previous regressions, we wondered how the data would react to this feature being held as a random effect. The results of this model are shown in the output of Figure 12 below. Holding the number of teenagers in the household constant, the significance of having a PhD decreases. However, the coefficient is still large, positive, and somewhat significant, indicating that targeting PhDs in the marketing campaign will likely yield good results. Additionally, kids at home and the amount spent on wine, meats, and gold products all still remained positive and significant. However, in this model, the only coefficients to remain significant at $p < .01$ were recency (-.522), amount spent on wines (.273), and number of web visits last month (.367). This means that the optimal customers to target based on this model are those that have not made a purchase recently, but tend to purchase wines and have visited the website many times in the past month. The AIC and BIC of this model are 1649.393 and 1809.09, respectively, meaning that it performed slightly worse than the probit regression model.

```
Generalized linear mixed model fit by maximum likelihood (Laplace Approximation) ['glmerMod']
 Family: poisson  ( log )
Formula: Response ~ Education + Marital_Status + Income + Kidhome + Recency +      MntWines + MntFruits +
MntMeatProducts + MntFishProducts +      MntSweetProducts + MntGoldProds + NumDealsPurchases + NumWebPurchases +
NumCatalogPurchases + NumWebVisitsMonth + Complain + Age +
    (1 | Teenhome)
   Data: campaign_data2_re

     AIC      BIC   logLik deviance df.resid
  1649.4   1809.1   -796.7   1593.4     2188

Scaled residuals:
    Min      1Q  Median      3Q     Max
-1.4291 -0.3476 -0.2483 -0.1570  7.2467

Random effects:
 Groups   Name        Variance Std.Dev.
 Teenhome (Intercept) 0.1299   0.3604
Number of obs: 2216, groups:  Teenhome, 3

Fixed effects:
                            Estimate Std. Error z value Pr(>|z|)
(Intercept)               -1.3056167  1.0835266  -1.205  0.22822
EducationBasic            -0.8034703  0.7456320  -1.078  0.28123
EducationGraduation        0.1738979  0.2306567   0.754  0.45089
EducationMaster            0.2212999  0.2586612   0.856  0.39224
EducationPhD               0.5515819  0.2458390   2.244  0.02485 *
Marital_StatusAlone       -0.4050558  1.4401755  -0.281  0.77852
Marital_StatusDivorced    -0.9029896  1.0399725  -0.868  0.38524
Marital_StatusMarried     -1.6202754  1.0350976  -1.565  0.11750
Marital_StatusSingle      -0.9592670  1.0346588  -0.927  0.35386
Marital_StatusTogether    -1.6418734  1.0389981  -1.580  0.11405
Marital_StatusWidow       -0.9170395  1.0558300  -0.869  0.38509
Marital_StatusYOLO        -0.5945721  1.4495741  -0.410  0.68168
Income                    -0.0042227  0.0864484  -0.049  0.96104
Kidhome                    0.1732149  0.0817507   2.119  0.03411 *
Recency                   -0.5220887  0.0601856  -8.675  < 2e-16 ***
MntWines                   0.2730563  0.0614378   4.444 8.81e-06 ***
MntFruits                 -0.0076698  0.0604487  -0.127  0.89903
MntMeatProducts            0.2342101  0.0723819   3.236  0.00121 **
MntFishProducts           -0.0364575  0.0655902  -0.556  0.57832
MntSweetProducts           0.0567225  0.0599119   0.947  0.34376
MntGoldProds               0.1304703  0.0537026   2.429  0.01512 *
NumDealsPurchases          0.0050175  0.0583511   0.086  0.93148
NumWebPurchases            0.1316187  0.0611317   2.153  0.03132 *
NumCatalogPurchases        0.2092936  0.0663443   3.155  0.00161 **
NumWebVisitsMonth          0.3678048  0.0691141   5.322 1.03e-07 ***
Complain                  -0.0008945  0.0574090  -0.016  0.98757
Age                       -0.0008745  0.0585571  -0.015  0.98809
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Figure 12: Poisson Model with TeenHome as a random effect

Marketing Strategy, Recommendations, Limitations, and Future Work

By comparing the AIC, BIC, interpretability and parsimony of all of our models, we came to the conclusion that our probit regression on all of our given variables was the most appropriate to use for predicting whether or not a customer will respond positively to the next marketing campaign. If the company has exact data on each possible target for the next marketing campaign, our model can give an approximate probability of success for

any given individual. If the company only has general information, then they can target, as the probit model recommends, highly educated households without teens at home and who have not made purchases recently, but tend to purchase meats and wines when they do make purchases at the store.

While we believe that this will increase the profitability and effectiveness of the company's next marketing campaign, we also accept that our analysis has limitations. First and foremost, the data does not specify if any continuous variables are averages for a household, total for a household, or the maximum/minimum variables for members of a household. Likewise with the discrete variables, it was difficult to make interpretations like exactly what having a PhD indicated for the total or average education level within a household. Also, we could not be sure if the next marketing campaign that the company will create would be the same or similar at all to the ones which were run in the past, which could heavily skew the effectiveness of our recommendations.