Online Retail Promotions

* Start by exploring the "spend" variable, examining its distribution, central tendency, and spread to understand its characteristics.
* Next, identify and gather the relevant variables that may influence customer spend.
* Look at the correlation between "spend" and other variables using a correlation matrix or scatter plots to see which variables have a significant impact on spend.
* Perform any Feature Engineering required on the data like converting required variables into factors and creating new variables to answer the objectives.
* Select a regression model that is appropriate for the data and the business problem. In this case, since the "spend" variable is continuous and has a skewed distribution with many zero values, we could consider using quassipoisson model because there seems to be overdispersion in the data. We could also build zero-inflated model such as a zero-inflated poisson regression which account for excess zeros in the data or a hurdle model which the probability of non-zero values for the response variable ("spend"), while the count component models the actual counts.

Chart, histogram, scatter chart

Description automatically generatedChart, histogram

Description automatically generated

The spend variables has only 579 non-zero values. Even the log-transformation doesn’t show any significant change in the distribution.

Predictor table for DV = spend (Actual dollars spent in the following two weeks of campaign)

|  |  |  |  |
| --- | --- | --- | --- |
| S.no | Variable | Expected sign of effect | Rationale |
| 1 | recency | + | Recent customers might spend more. |
|  | historysegment | excluded | Taking its effect from history. |
|  | history | +/- | Converted to historycategory and the rationale is explained below. |
|  | Mens/womens | +/- | Promoting a mens merch to someone who got mens, increases spend and vice-versa. |
|  | zipcode | +/- | Urban customers may spend more compared to suburban and rural. |
|  | newcustomer | + | New customers are likely to be more receptive to marketing campaigns and promotions. |
|  | channel | + | Some customers might like to use phone over web. |
|  | campaign | + | Either a Men's or Women's promotional email could be associated with an increase in spend compared to not receiving an email. |
|  | visit | +/- | The chances of spending increase but how much of spend could vary. |
|  | conversion | +/- | Conversion 1 indicates that customer spent money but how much of spend could vary. |

Diagram, schematic

Description automatically generated

Most of the variables except Recency and history are categorical in nature. The corrplot shows there are no significant correlations among the variables of the predictor table except for mens and womens which are not considering. Hence, we should be good on Multicollinearity of the models we build.

* Converted historysegment,channel,campaign,conversion,newcustomer,zipcode into factors.
* Created a new variable called historycategory to answer 4.c considering customers who spent more than the median history as higher category and lower than median history as lower category.

Models-

1. We notice that we have an overdispersion ( Var(spend) > Mean (spend) ) and to address it, we created a quassipoission model.

**qpoisson <- glm(spend ~ recency + zipcode + campaign + newcustomer + visit + conversion , data = cust\_data, family = quasipoisson (link=log))**

1. We created a hurdle poission model. The hurdle model is divided into two parts: the zero hurdle model and the count model. The zero hurdle model estimates the probability of a customer not spending any money, while the count model estimates the amount spent by customers who do spend money.

**hpoisson <- hurdle(spend ~ recency + zipcode + newcustomer + channel**

**+ campaign | visit + conversion, data=cust\_data2, link= 'logit',dist="poisson")**

1. The 3rd model we build is a zero-inflated poisson model which is like the hurdle model. But lets add interactions to this.

**zip <- zeroinfl(spend ~ recency + history + zipcode + campaign \* mens + campaign \* womens + campaign \* newcustomer+ campaign\*channelphone + campaign\* channelweb +**

**+ campaign\*history | visit , data=cust\_data2, dist="negbin", link = "logit")**Interactions are added to answer question number 4. visit could be the logit predictor because spend = 0 when visit =0.

**Models summary**

|  |  |  |  |
| --- | --- | --- | --- |
|  | | | |
|  | *Dependent variable:* | | |
|  |  | | |
|  | spend | | |
|  | *glm: quasipoisson* | *hurdle* | *zero-inflated* |
|  | *link = log* |  | *count data* |
|  | (1) | (2) | (3) |
|  | | | |
| recency | 0.003\*\*\* (0.001) | 0.007\*\*\* (0.001) | 0.004 (0.010) |
| factor(historycategory)1 | -0.126\*\*\* (0.008) |  |  |
| zipcodeUrban | 0.112\*\*\* (0.010) | 0.108\*\*\* (0.010) | 0.138 (0.088) |
| campaignMens E-Mail | -0.011 (0.010) | -0.014 (0.010) | -0.085 (0.374) |
| campaignWomens E-Mail | 0.054\*\*\* (0.010) | 0.062\*\*\* (0.011) | 0.501 (0.418) |
| mens |  |  | 0.498\*\* (0.238) |
| womens |  |  | 0.212 (0.232) |
| history |  | 0.00004\*\*\* (0.00001) | -0.00004 (0.0002) |
| newcustomer1 | 0.020\*\*\* (0.007) | 0.009 (0.008) | -0.247 (0.184) |
| visit | 0.022 (168.068) |  |  |
| conversion1 | 25.062 (155.945) |  |  |
| channelPhone |  | 0.070\*\*\* (0.012) |  |
| channelWeb |  | 0.068\*\*\* (0.012) |  |
| channelphone |  |  | -0.327 (0.234) |
| channelweb |  |  | -0.306 (0.230) |
| campaignMens E-Mail:mens |  |  | -0.302 (0.279) |
| campaignWomens E-Mail:mens |  |  | -0.768\*\* (0.309) |
| campaignMens E-Mail:womens |  |  | -0.171 (0.273) |
| campaignWomens E-Mail:womens |  |  | -0.853\*\*\* (0.304) |
| campaignMens E-Mail:newcustomer1 |  |  | 0.319 (0.214) |
| campaignWomens E-Mail:newcustomer1 |  |  | 0.286 (0.224) |
| campaignMens E-Mail:channelphone |  |  | 0.183 (0.278) |
| campaignWomens E-Mail:channelphone |  |  | 0.386 (0.298) |
| campaignMens E-Mail:channelweb |  |  | 0.229 (0.276) |
| campaignWomens E-Mail:channelweb |  |  | 0.317 (0.296) |
| history:campaignMens E-Mail |  |  | 0.0001 (0.0003) |
| history:campaignWomens E-Mail |  |  | 0.0001 (0.0003) |
| Constant | -20.379 (62.675) | 4.544\*\*\* (0.017) | 4.619\*\*\* (0.332) |
|  | | | |
| Observations | 64,000 | 64,000 | 64,000 |
| Log Likelihood |  | -27,366.310 | -5,457.253 |
|  | | | |
| *Note:* | \*p<0.1; \*\*p<0.05; \*\*\*p<0.01 | | |

We are taking zip model as the best model among the three. It passes all the assumptions which are tested below. We will be answering the questions based on this model.

Testing the Assumptions of the Zero-inflated model-

GLM models can withstand violations of multivariate normality, linearity, and homoscedasticity.

1. Test for Independence
2. Durbin-Watson test
3. data: zip
4. DW = 2.0062, p-value = 0.784
5. alternative hypothesis: true autocorrelation is greater than 0

In this case, the Durbin-Watson test has a value of 2.0062 and a p-value of 0.7262, indicating that there is no significant autocorrelation present in the residuals of the hurdle poisson model. Therefore, we can assume that the independence assumption of the model is not violated by autocorrelation.

2.Excess zeroes and Overdispersion– Zero inflated models are robust to excess zero problem and overdispersion.

3. Multicollinearity: Passed

vif(zip)

GVIF Df GVIF^(1/(2\*Df))

recency 3.442896 1 1.855504

history 13.773853 1 3.711314

zipcode 5.747842 1 2.397466

campaign 1795.120907 2 6.509137

mens 29.091334 1 5.393638

womens 30.055650 1 5.482303

newcustomer 13.648459 1 3.694382

channelphone 27.154264 1 5.210975

channelweb 30.067906 1 5.483421

campaign:mens 146.799809 2 3.480818

campaign:womens 196.854084 2 3.745727

campaign:newcustomer 32.887041 2 2.394728

campaign:channelphone 140.983871 2 3.445818

campaign:channelweb 195.138858 2 3.737541

history:campaign 39.369541 2 2.504898

Warning message:

In vif.default(zip) : No intercept: vifs may not be sensible.

**Question 4:**

* How did the promotion campaigns work relative to the control group? Did the men's promotions work better than the women's promotion (or vice versa) and by how much?

In order to answer this question, we need to take into account of all the factors that contribute to the mens and also womens campaign and then then the difference in the marginal effects of mens and womens campaign. If we consider these factors as constant, the marginal difference between mens Email and womens email is -0.59. The Men's Campaign performed 10% worse than the No Campaign and 59% worse than the Women's Campaign.

* Should we target these promotions to new customers (who joined over the last 12 months) rather than to established customers, or vice versa?

In comparison to not receiving a campaign, new consumers who received the men's campaign spent 6% more money overall, and those who received the women's campaign spent 3% more. New customers have a -25% e relative to established customers in the no campaign group.

* Should we target these promotions to customers who have a higher (or lower) history of spending over the last year?

History estimate was very small and we could say it doesn't have any effect.

* Did the promotions work better for phone or web channel?

If customers didn't receive a campaign, neither the phone nor the web channel performed well . When compared to women's campaigns, men's campaigns saw a rise in phone spending of -14% and a decrease in web spending of -7%. Hence, compared to no marketing, the women's campaign definitely increased customer spending. Men's campaigns decreased deficit spending when compared to none, yet they still had a detrimental impact.

* Will the promotions work better if the men's promotion is targeted at customers who bought men's merchandise over the last year (compared to those who purchased women's merchandise), and if the women's promotion would work better if targeted at customers who bought women's merchandise over the last year?

Men's advertising campaigns that targeted customers who had previously bought men's products had a 30% negative impact on spending compared to no advertising, whereas men's advertising campaigns that targeted customers who had previously bought women's products had a -17% negative impact. When compared to no advertising, women's campaigns that targeted customers who had previously purchased women's products had a negative impact of -85%, while those that targeted customers who had previously purchased men's products had a negative impact of -76%. So, it seems that these adverts are most effective when they are directed at new customers rather than those who have previously made purchases. The impact of the women's campaign was far worse than that of the men's.

Q5) Reflect on the quality of your analysis, and comment on things you can do to further improve this analysis.

It would be helpful to conduct more rigorous statistical tests (t-tests) to ensure that the observed differences in spending between different groups are significant and not due to chance.