**Advanced Data Analytics Report**

Pedro Guarda Chinen

Jose Luis Palacios Gonzalez

Sonika Arun Govil

Ela Yenikoylu

Haiyan Hu

Team 6

Hult International Business School

**Part I: Statistical tests and Regression**

1. What factors are significantly related to the number of store purchases? What would you suggest to CMO to improve store purchases?

*i) Regression output*

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| *Regression Statistics* | |  |  |  |  |  |  |  |
| Multiple R | 0.521053164 |  |  |  |  |  |  |  |
| R Square | 0.271496399 |  |  |  |  |  |  |  |
| Adjusted R Square | 0.270835925 |  |  |  |  |  |  |  |
| Standard Error | 2.778274325 |  |  |  |  |  |  |  |
| Observations | 2209 |  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |  |  |
| ANOVA |  |  |  |  |  |  |  |  |
|  | *df* | *SS* | *MS* | *F* | *Significance F* |  |  |  |
| Regression | 2 | 6345.82558 | 3172.91279 | 411.0625236 | 1.829E-152 |  |  |  |
| Residual | 2206 | 17027.6909 | 7.71880823 |  |  |  |  |  |
| Total | 2208 | 23373.5165 |  |  |  |  |  |  |
|  |  |  |  |  |  |  |  |  |
|  | *Coefficients* | *Standard Error* | *t Stat* | *P-value* | *Lower 95%* | *Upper 95%* | *Lower 95.0%* | *Upper 95.0%* |
| Intercept | 4.267300397 | 0.08003242 | 53.3196479 | 0 | 4.11035363 | 4.42424717 | 4.11035363 | 4.42424717 |
| AcceptedCmp5 | 0.684706433 | 0.2406246 | 2.84553791 | 0.00447458 | 0.21283198 | 1.15658089 | 0.21283198 | 1.15658089 |
| NumCatalogPurchases | 0.557418218 | 0.02132979 | 26.1333158 | 2.2041E-131 | 0.51558964 | 0.59924679 | 0.51558964 | 0.59924679 |

*ii) Analysis of output in no more than 350 words,*

Given our data, we identified that there should be direct relationship between number of purchases in the store against the number of purchases made in other sales channels. After running a regression analysis, we identified that the number of catalog purchases had a statistically significant relationship with the number of purchases in the stores. Also, we wanted to understand the impact of campaigns in the number of store purchases, so we added the campaign acceptance variables to the regression analysis. As result of this analysis, we found that campaign 5 acceptance was statistically sufficient too, the result of this analysis is a predictive model that can help us understand better the factors that influence the number of store purchase. The mod found out that 27% percent of the store purchases can be explained with the catalog purchases and Commercial Campaign 5.

Checking the p values for the x variables, they are lower than 0.05, we conclude that the effect Commercial 5 and Catalog purchases on Number of Store Purchases is statistically significant. For every person accepted the campaign 5 influences 0.68 times purchases happening. For every person makes a catalog purchase increases the number of store purchases 0.56 times.

We also made further inspection with residuals to see if our model is reliable. Checking the summary statistics for residuals we have seen that they have a normal distribution, meaning that there is no pattern or association left. Checking the autocorrelation for residuals we have seen that there is no significant correlation left. Lastly, we checked the residuals for homoscedasticity, and we have not seen a clear pattern between residuals and the prediction for the number of store purchases.

*iii) Suggestions to CMO in no more than 500 words.*

Based on out regression analysis on store purchases:

If we run campaign 5 one time and use 10,000 catalogs, we are 95% confident that it will drive about 554 – 566 store purchases.

We noticed that your catalogues have significantly risen through positive reinforcement and would suggest promoting and advertising more via such mediums offline and online for the future campaigns.

27% of your store purchases influenced by your catalog purchases and your commercial campaign 5. Considering all other factors can affect the store purchases, 27% is a high number. We would suggest you focus on creating more commercial campaigns like campaign 5 and use consistent content in your catalogs design referring to the commercial. Also, increasing your channels to deliver your catalogs to your customer base would be impactful in increasing your store purchases.

In one of the regressions, we created a column with 5 different age categories, beginning with lower than 29 years old being as 1, adding 10 years for the others, until number 5 which is customers with ages higher than 59 years old. We found that for each unit increase in the age category we would increase the store sell in 0.27 on average, which tells us that people older tend to visit more the store.

The Recency of purchase in the last one month can help us understand how to target these set of buyer personas, through different mediums, Store or Web purchases.

1. Does US fare significantly better than RoW (Rest of the World) in terms of total purchases.

*i) Regression output,*

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| SUMMARY OUTPUT |  |  |  |  |  |  |
|  |  |  |  |  |  |  |
| *Regression Statistics* | |  |  |  |  |  |
| Multiple R | 0.009215378 |  |  |  |  |  |
| R Square | 8.49232E-05 |  |  |  |  |  |
| Adjusted R Square | -0.000368142 |  |  |  |  |  |
| Standard Error | 602.8618959 |  |  |  |  |  |
| Observations | 2209 |  |  |  |  |  |
|  |  |  |  |  |  |  |
| ANOVA |  |  |  |  |  |  |
|  | *df* | *SS* | *MS* | *F* | *Significance F* |  |
| Regression | 1 | 68124.16451 | 68124.16451 | 0.1874414 | 0.66509701 |  |
| Residual | 2207 | 802117521.4 | 363442.4655 |  |  |  |
| Total | 2208 | 802185645.6 |  |  |  |  |
|  |  |  |  |  |  |  |
|  | *Coefficients* | *Standard Error* | *t Stat* | *P-value* | *Lower 95%* | *Upper 95%* |
| Intercept | 605.4043768 | 13.14926504 | 46.04092891 | 0 | 579.6181493 | 631.1906043 |
| Country\_Cat | 25.86665125 | 59.74580929 | 0.432945031 | 0.66509701 | -91.29723769 | 143.0305402 |

*ii) Details of how you computed Rest of the World and Total purchases, in no more than 100 words,*

We created a column named Country\_cat, which attribute 1 if the country was US and 0 if was different then US, which represent the rest of the world.

For the Total purchases we created a column which is the sum of the columns with the amount spent in each product like meat, fish, gold, etc.

*iii) Your analysis of the results in no more than 100 words.*

Based on that regression, we can’t reject the null hypothesis since the P-value is bigger than 0.05 and zero is in the interval. In 40% of the interval, the purchases made in US actually reduce the interval, so we cannot affirm that US is better in total purchases. With the model we have seen that significance F is very high, so we decided that our model isn’t statistically significant.

c. Your supervisor insists that people who buy gold are more conservative and as such people who spent **an above average amount** on gold in last 2 years would have more in store purchases. Justify/refute the statement using appropriate statistical tests.

*i) Output of test(s),*

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| *Regression Statistics* | |  |  |  |  |  |
| Multiple R | 0.410359747 |  |  |  |  |  |
| R Square | 0.168395122 |  |  |  |  |  |
| Adjusted R Square | 0.168018319 |  |  |  |  |  |
| Standard Error | 2.967695757 |  |  |  |  |  |
| Observations | 2209 |  |  |  |  |  |
|  |  |  |  |  |  |  |
| ANOVA |  |  |  |  |  |  |
|  | *df* | *SS* | *MS* | *F* | *Significance F* |  |
| Regression | 1 | 3935.986164 | 3935.986164 | 446.9045863 | 1.75509E-90 |  |
| Residual | 2207 | 19437.53036 | 8.807218106 |  |  |  |
| Total | 2208 | 23373.51652 |  |  |  |  |
|  |  |  |  |  |  |  |
|  | *Coefficients* | *Standard Error* | *t Stat* | *P-value* | *Lower 95%* | *Upper 95%* |
| Intercept | 4.90 | 0.076245297 | 64.27924481 | 0 | 4.751470064 | 5.050510134 |
| Gold > avg | 2.88 | 0.136028899 | 21.14011794 | 1.75509E-90 | 2.608908925 | 3.142424998 |

*ii) Your justification/rebuttal in no more than 100 words.*

He looks to be right, looking at the regression if the customer buys more than average for Gold products, he will spend more 2.88 dollars on average for each purchase that he does.

The P-value is lower than 0.05 and zero is not in the interval so we can reject the null hypothesis which mean is statistically significant.

With the model we have seen that significance F is very low, so we decided that our model is statistically significant.

1. *An* ***interaction effect*** *occurs when the effect of one variable on dependent variable depends on the value of some other explanatory variable. This effect is captured by* ***creating*** *an interaction variable, by multiplying the two relevant explanatory variables. Use your knowledge of interaction variables/effects to answer this part.* Fish has Omega 3 fatty acids, good for brain, accordingly, do "Married PhD candidates" have a significant relation with amount spent on fish? What other factors are significantly related to amount spent on fish?

*i)Regression output,*

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
|  |  |  |  |  |  |  |
| *Regression Statistics* | |  |  |  |  |  |
| Multiple R | 0.567968339 |  |  |  |  |  |
| R Square | 0.322588034 |  |  |  |  |  |
| Adjusted R Square | 0.32197388 |  |  |  |  |  |
| Standard Error | 184.6838654 |  |  |  |  |  |
| Observations | 2209 |  |  |  |  |  |
|  |  |  |  |  |  |  |
| ANOVA |  |  |  |  |  |  |
|  | *df* | *SS* | *MS* | *F* | *Significance F* |  |
| Regression | 2 | 35830990.08 | 17915495.04 | 525.2558548 | 2.6959E-187 |  |
| Residual | 2206 | 75242535.05 | 34108.13012 |  |  |  |
| Total | 2208 | 111073525.1 |  |  |  |  |
|  |  |  |  |  |  |  |
|  | *Coefficients* | *Standard Error* | *t Stat* | *P-value* | *Lower 95%* | *Upper 95%* |
| Intercept | -2544.342317 | 83.79086136 | -30.36539159 | 1.5988E-169 | -2708.659542 | -2380.025091 |
| Ln Income | 252.2816192 | 7.792650103 | 32.37430346 | 1.8279E-188 | 236.9999212 | 267.5633173 |
| Married PHD | -17.83259528 | 14.06796145 | -1.267603366 | 0.205073412 | -45.42042953 | 9.755238957 |

*ii) Your answer to the questions stated above.*

With that regression what we see is actually the opposite, if the customer has an PHD and is married he actually spend less $ 17.83 than costumers that are not PHD married.

But we also did the regression using the Log of the income and we found that is really relevant, meaning that people with higher income would spend more money on fish.

With the model we have seen that significance F is very low, so we decided that our model is statistically significant.

e, Do any other analysis you deem relevant to show to your CMO. For the purpose, **propose a hypothesis and perform the appropriate tests**. **Be creative and insightful.** We can ask any questions about last one. Just try to impress us!

We decided comparing the web shoppers and store shoppers to check if they have different profile. To do that, we checked how the campaigns worked for each and we discovered that for the web shoppers, the campaign #4 was more impactful and for the store shoppers it was the campaign #5. For the web shoppers, the campaign is more impactful because for each person accepting the camping sales go up by 1.7 times. For the store shoppers, the campaign is less impactful because for each person accepting it, sales go up by 0.7 times which is lower. We conclude that, campaigns for the web customers are more impactful, because they can target better online.

