

**Module 3 Midweek Project**

**ALY 6020**

Name: Kavitha Palthayvath

Instructor’s name: Justin Grosz

Due Date: 08/01/2021

**Abstract**

Anomaly detection is a good application of data analysis, in this report we are going to see one such application of this concept and analyze which features or variables were responsible for the decline in the number of subscriptions of the magazine, how important are these variables and how does it affect the subscription rate in general.

**Introduction**

The magazine company has been seeing a decline in subscriptions and wants to find the reason behind the same. Us as analysts have been given data collected over the previous years and find what has been affecting the subscription rate. We start of by looking at the dataset and realize we will be using logistic regression to examine the association between ‘Response’ the target variable and other variables that are responsible for predicting response.

**Analysis**

**Data Cleaning:**

While going over various columns in our dataset we came across ‘NumWebPurchases’ which had outliers (value more than 12) we replaced these outliers with the mean value of the datapoints less than or equal to 12. We found similar outliers (value more than 10) in ‘NumWebVisitsMonth’ and replaced them with mean of all points with value less than 10. We did not find any outliers with continuous variables, however we had to convert the ‘Dt\_Customer’ variable to separate out Year, month, and day as separate columns and remove the original ‘Dt\_Customer’ column. Furthermore, we had to perform one hot encoding for ‘Marital\_Status’ and ‘Education’ columns as these were continuous variables.

**Model Training:**

After preparing the data ready to be trained a logistic regression model with the dataset and printed out how is our model doing. As per Table 1, we saw thatthe variables 'ID', 'Year\_Birth', 'Income', 'Kidhome', 'MntFruits', 'MntFishProducts', 'MntSweetProducts', 'MntGoldProds', 'NumWebPurchases','Complain','Z\_Revenue','Marital\_Status\_Absurd','Marital\_Status\_Alone','Marital\_Status\_Divorced','Marital\_Status\_Married','Marital\_Status\_Single','Marital\_Status\_Together','Marital\_Status\_Widow','Marital\_Status\_YOLO','dt\_day' are not as important as other variables because their p-value is greater than 0.05, also the coefficient of these variables were lesser in numerical value as compared to other variables. Another observation we had is that a few variables like ‘Education\_2n Cycle’, ‘Education\_Basic’, ‘Education\_graduation’, ‘Eduation\_Master’, ‘Education\_PhD’ had p-value as Nan, we decided to keep these variables for another round of feature selection.

Since these variables did not seem important, we removed these variables from the dataset and retrained the model, on retraining the model, we saw the summary of the model for feature selection and as per Table 2, we saw that the variables ‘MntWines’, ‘Z\_CostContact’, ‘Education\_2n Cycle’, ‘Education\_Basic’, ‘Education\_graduation’, ‘Eduation\_Master’, ‘Education\_PhD’ were unimportant as compared to other variables because its p-value was greater than 0.05 and its coefficients were lesser in values as compared to other variables.

We retrained the model and performed another round of feature selection, and as per Table 3, we eliminated the variable ‘AcceptedCmp2’ because its p-value was greater than 0.05 and its coefficients were lesser in values as compared to other variables. After removing this variables, as per Table 4, we got our model with the best features, these features were, ‘Teenhome’, ‘Recency’, ‘MntMeatProducts’, ‘NumDealsPurchases’, ‘NumCatalogPurchases’, ‘NumWebVisitsMonth’, ‘AcceptedCmp3’, ‘AcceptedCmp4’, ‘AcceptedCmp5’, ‘AcceptedCmp1’, ‘dt\_year’, ‘dt\_month’.

**Best features and Business Implication**

On training the dataset with feature selection we realized that, the variables with ‘Accepted’ string in them have the most positive impact on the likelihood of subscription of magazines and the variables dt\_year and Teenhome have a negative impact on the likelihood of subscription of magazines. This means that weather a customer accepted a magazine subscription in 1st, 3rd, 4th, and 5th campaign directly benefits business and attracts more subscriptions. The other aspect that can be understood is that the newer the subscription of a customer with the company the lesser probability of the customer to subscribe again as per the negative impact of dt\_year variable. Furthermore, the business have negative impact if the subscription is from a household where there are teenagers and have lesser probability of getting magazine subscriptions from a household that has teenagers in it.

**Training a SVM Model**

We trained a SVM model on the cleaned dataset without feature selection to see which model did the best for our use case.

**Performance Comparison**

We used metrics class to gauge the performance of Logistic regression model without feature selection, with feature selection, and the SVM model and received the following values:

|  |  |  |  |
| --- | --- | --- | --- |
| **Model** | **Accuracy** | **Precision** | **Recall** |
| Logistic Regression without Feature Selection | 0.8392 | 0.4444 | 0.1142 |
| Logistic Regression with Feature Selection | 0.8459 | 0.5151 | 0.2428 |
| SVM | 0.8482 | 0.5384 | 0.2 |

We would recommend using the SVM model for our use case, as it has the best Precision and the best Accuracy among the three models. It also has the best ability to find the correct data points of interest.

**Conclusion**

After a thorough analysis to help the magazine company, it should introduce more discounts on the price for customers who are more loyal and have been having repetitive subscriptions to the magazine. This will ensure that the magazine will not lose customers and have older customers still renew their subscription. Moreover, the magazine company should launch campaigns more often as campaigns attract more subscriptions. Another important suggestion we can make to the magazine company is to target households with no teenagers since they see lesser subscriptions from households with teenagers.

**Appendix**

Table 1: Logistic Regression Results of Model without Feature Selection

Logit Regression Results

==============================================================================

Dep. Variable: Response No. Observations: 2240

Model: Logit Df Residuals: 2202

Method: MLE Df Model: 37

Date: Sun, 01 Aug 2021 Pseudo R-squ.: 0.4310

Time: 21:19:43 Log-Likelihood: -536.78

converged: True LL-Null: -943.39

Covariance Type: nonrobust LLR p-value: 8.193e-147

===========================================================================================

coef std err z P>|z| [0.025 0.975]

-------------------------------------------------------------------------------------------

ID -1.426e-05 2.48e-05 -0.575 0.565 -6.29e-05 3.43e-05

Year\_Birth -0.0071 0.007 -0.969 0.333 -0.021 0.007

Income 1.11e-06 3.82e-06 0.291 0.771 -6.37e-06 8.59e-06

Kidhome 0.1222 0.231 0.529 0.597 -0.331 0.575

Teenhome -1.1046 0.210 -5.272 0.000 -1.515 -0.694

Recency -0.0330 0.003 -10.679 0.000 -0.039 -0.027

MntWines -0.0010 0.000 -2.415 0.016 -0.002 -0.000

MntFruits 0.0025 0.002 1.046 0.295 -0.002 0.007

MntMeatProducts 0.0022 0.000 4.676 0.000 0.001 0.003

MntFishProducts -0.0028 0.002 -1.472 0.141 -0.007 0.001

MntSweetProducts 0.0017 0.002 0.700 0.484 -0.003 0.006

MntGoldProds 0.0025 0.002 1.434 0.152 -0.001 0.006

NumDealsPurchases 0.1223 0.049 2.502 0.012 0.026 0.218

NumWebPurchases 0.0778 0.041 1.879 0.060 -0.003 0.159

NumCatalogPurchases 0.2327 0.047 4.990 0.000 0.141 0.324

NumStorePurchases -0.2118 0.038 -5.510 0.000 -0.287 -0.136

NumWebVisitsMonth 0.2262 0.064 3.522 0.000 0.100 0.352

AcceptedCmp3 1.9191 0.241 7.950 0.000 1.446 2.392

AcceptedCmp4 1.3303 0.304 4.382 0.000 0.735 1.925

AcceptedCmp5 2.0840 0.306 6.804 0.000 1.484 2.684

AcceptedCmp1 1.5684 0.299 5.253 0.000 0.983 2.154

AcceptedCmp2 1.7541 0.593 2.958 0.003 0.592 2.916

Complain -0.2628 1.162 -0.226 0.821 -2.540 2.015

Z\_CostContact 70.2280 nan nan nan nan nan

Z\_Revenue 257.4817 1.81e+05 0.001 0.999 -3.55e+05 3.56e+05

Marital\_Status\_Absurd 3.1411 6.68e+06 4.7e-07 1.000 -1.31e+07 1.31e+07

Marital\_Status\_Alone 4.0980 6.68e+06 6.13e-07 1.000 -1.31e+07 1.31e+07

Marital\_Status\_Divorced 3.1622 6.63e+06 4.77e-07 1.000 -1.3e+07 1.3e+07

Marital\_Status\_Married 1.7446 6.64e+06 2.63e-07 1.000 -1.3e+07 1.3e+07

Marital\_Status\_Single 3.1767 6.68e+06 4.76e-07 1.000 -1.31e+07 1.31e+07

Marital\_Status\_Together 1.8024 6.66e+06 2.71e-07 1.000 -1.3e+07 1.3e+07

Marital\_Status\_Widow 3.2011 6.69e+06 4.79e-07 1.000 -1.31e+07 1.31e+07

Marital\_Status\_YOLO 2.4491 6.72e+06 3.65e-07 1.000 -1.32e+07 1.32e+07

Education\_2n Cycle 4.3348 nan nan nan nan nan

Education\_Basic 3.1174 nan nan nan nan nan

Education\_Graduation 4.7595 nan nan nan nan nan

Education\_Master 5.3213 nan nan nan nan nan

Education\_PhD 5.8757 nan nan nan nan nan

dt\_year -1.5093 0.190 -7.942 0.000 -1.882 -1.137

dt\_month -0.1310 0.031 -4.228 0.000 -0.192 -0.070

dt\_day 0.0054 0.009 0.578 0.563 -0.013 0.024

===========================================================================================

Table 2: Logistic Regression Results of Model with round 1 of Feature Selection

Logit Regression Results

==============================================================================

Dep. Variable: Response No. Observations: 2240

Model: Logit Df Residuals: 2220

Method: MLE Df Model: 19

Date: Sun, 01 Aug 2021 Pseudo R-squ.: 0.3839

Time: 21:19:43 Log-Likelihood: -581.26

converged: True LL-Null: -943.39

Covariance Type: nonrobust LLR p-value: 2.573e-141

========================================================================================

coef std err z P>|z| [0.025 0.975]

----------------------------------------------------------------------------------------

Teenhome -0.9172 0.183 -5.004 0.000 -1.276 -0.558

Recency -0.0307 0.003 -10.481 0.000 -0.036 -0.025

MntWines -0.0006 0.000 -1.587 0.113 -0.001 0.000

MntMeatProducts 0.0020 0.000 4.604 0.000 0.001 0.003

NumDealsPurchases 0.1343 0.043 3.104 0.002 0.049 0.219

NumCatalogPurchases 0.2405 0.042 5.703 0.000 0.158 0.323

NumStorePurchases -0.1802 0.036 -5.072 0.000 -0.250 -0.111

NumWebVisitsMonth 0.2279 0.055 4.140 0.000 0.120 0.336

AcceptedCmp3 1.9042 0.229 8.332 0.000 1.456 2.352

AcceptedCmp4 1.1750 0.284 4.138 0.000 0.618 1.732

AcceptedCmp5 1.8848 0.289 6.526 0.000 1.319 2.451

AcceptedCmp1 1.4273 0.278 5.129 0.000 0.882 1.973

AcceptedCmp2 1.3921 0.552 2.522 0.012 0.310 2.474

Z\_CostContact 882.9558 4.53e+06 0.000 1.000 -8.88e+06 8.88e+06

Education\_2n Cycle 58.5941 1.36e+07 4.31e-06 1.000 -2.66e+07 2.66e+07

Education\_Basic 57.1617 1.36e+07 4.2e-06 1.000 -2.66e+07 2.66e+07

Education\_Graduation 59.0455 1.36e+07 4.34e-06 1.000 -2.66e+07 2.66e+07

Education\_Master 59.5042 1.36e+07 4.38e-06 1.000 -2.66e+07 2.66e+07

Education\_PhD 60.0131 1.36e+07 4.41e-06 1.000 -2.66e+07 2.66e+07

dt\_year -1.3462 0.176 -7.665 0.000 -1.690 -1.002

dt\_month -0.1207 0.030 -4.060 0.000 -0.179 -0.062

========================================================================================

Table 3: Logistic Regression Results of Model with round 2 of Feature Selection

Logit Regression Results

==============================================================================

Dep. Variable: Response No. Observations: 2240

Model: Logit Df Residuals: 2225

Method: MLE Df Model: 14

Date: Sun, 01 Aug 2021 Pseudo R-squ.: 0.3606

Time: 21:19:43 Log-Likelihood: -603.23

converged: True LL-Null: -943.39

Covariance Type: nonrobust LLR p-value: 4.073e-136

=======================================================================================

coef std err z P>|z| [0.025 0.975]

---------------------------------------------------------------------------------------

const 2442.9524 341.003 7.164 0.000 1774.599 3111.306

Teenhome -0.7806 0.175 -4.452 0.000 -1.124 -0.437

Recency -0.0298 0.003 -10.392 0.000 -0.035 -0.024

MntMeatProducts 0.0020 0.000 4.634 0.000 0.001 0.003

NumDealsPurchases 0.1337 0.042 3.157 0.002 0.051 0.217

NumCatalogPurchases 0.2149 0.037 5.739 0.000 0.142 0.288

NumStorePurchases -0.1841 0.033 -5.662 0.000 -0.248 -0.120

NumWebVisitsMonth 0.2157 0.053 4.052 0.000 0.111 0.320

AcceptedCmp3 1.7978 0.222 8.099 0.000 1.363 2.233

AcceptedCmp4 1.0989 0.268 4.106 0.000 0.574 1.623

AcceptedCmp5 1.7217 0.266 6.466 0.000 1.200 2.244

AcceptedCmp1 1.2696 0.268 4.735 0.000 0.744 1.795

AcceptedCmp2 1.2885 0.538 2.397 0.017 0.235 2.342

dt\_year -1.2145 0.169 -7.175 0.000 -1.546 -0.883

dt\_month -0.1090 0.029 -3.764 0.000 -0.166 -0.052

Table 4: Logistic Regression Results of Model with perfect features

Logit Regression Results

==============================================================================

Dep. Variable: Response No. Observations: 2240

Model: Logit Df Residuals: 2226

Method: MLE Df Model: 13

Date: Sun, 01 Aug 2021 Pseudo R-squ.: 0.3575

Time: 21:19:43 Log-Likelihood: -606.14

converged: True LL-Null: -943.39

Covariance Type: nonrobust LLR p-value: 9.640e-136

=======================================================================================

coef std err z P>|z| [0.025 0.975]

---------------------------------------------------------------------------------------

const 2421.3812 339.839 7.125 0.000 1755.309 3087.453

Teenhome -0.7730 0.175 -4.418 0.000 -1.116 -0.430

Recency -0.0296 0.003 -10.372 0.000 -0.035 -0.024

MntMeatProducts 0.0019 0.000 4.575 0.000 0.001 0.003

NumDealsPurchases 0.1285 0.042 3.042 0.002 0.046 0.211

NumCatalogPurchases 0.2181 0.037 5.861 0.000 0.145 0.291

NumStorePurchases -0.1787 0.032 -5.555 0.000 -0.242 -0.116

NumWebVisitsMonth 0.2242 0.053 4.231 0.000 0.120 0.328

AcceptedCmp3 1.8343 0.221 8.307 0.000 1.401 2.267

AcceptedCmp4 1.2586 0.255 4.936 0.000 0.759 1.758

AcceptedCmp5 1.7546 0.264 6.640 0.000 1.237 2.273

AcceptedCmp1 1.2720 0.265 4.792 0.000 0.752 1.792

dt\_year -1.2038 0.169 -7.136 0.000 -1.534 -0.873

dt\_month -0.1083 0.029 -3.756 0.000 -0.165 -0.052

=======================================================================================