Individual report – Marija Tosic

Haiku: *If you ever feel sad,*

Please feel free to take my hand

I'll go till the end

A response model can provide a significant boost to the efficiency of a marketing campaign by increasing responses or reducing expenses. The objective is to predict who will respond to an offer for a product or service. The dataset chosen for the analysis was the marketing campaign dataset. This dataset was retrieved from the Kaggle. The reason for choosing this particular dataset was mainly the structure of dataset, different data types among features, long list of features, as well as many samples that would secure the possibility for model to learn and be able to do well on unseen data. All of this represented a great challenge as well as an opportunity for a deep and interesting analysis. The main questions that lead this analysis were:

1. What features contribute the most to predicting if a customer will respond positively to the marketing campaign?
2. Is the full model better than the reduced model?

Introduction to the variables

The dataset consists of 2240 samples and 29 features (26 features were numerical type, while the rest were characters). The list of features and the details on them can be seen in the Appendix. One of the main challenges that came with this dataset was dealing with different type of categorical data, as well as handling "date" type of data. The detailed steps of cleaning the data will be explained in the next section.

Data cleaning and Exploratory data Analysis

The first step was to check for the missing values and decide which approach to take (delete, fill with mean/median, etc.). The result showed there were only 24 missing values, all belonging to the "Income" column. Since there were 2240 samples, it was decided to drop the samples containing missing income values. The next step was checking for the duplicate samples, which were not found in the dataset. After checking the histogram plot it was seen that two features had uniformed distribution, having the same value for each sample. Since these features ("Z_CostContact"," Z_Revenue") would have no contribution/impact they were dropped. The only normal distribution was found in "Year_Birth" column, having the peak around year 1970. Education and Income also had a slight normal distribution, both being skewed to the right. "Kidhome", "teenhome", "NumDealsPurchases", "NumWelPurchases", "NumCataloguePurchases", and all features related to the "amount spent" had Poisson distribution. "Recency" feature (numbers of days since the last purchases) had a very flat, almost uniformed distribution. Most of customers had at least Graduation degree, while even 38% have a Master or PhD degree, 64% are either married or in a relationship, while 21% are single, 10% are divorced, and 3% are widows. The rest of the features were related to whether the customers responded to the number of previous campaigns, which resulted in either "yes" or "no", so the distribution among these features was the Bernoulli.

The next step was to investigate on the outliers by plotting "box-plot" for each numerical data. The plots showed that there were reasonable outliers in the following features: "NumWebVisitsMonth", "NumCatalogPurchases", "NumWebPurchases", "NumDealsPurchases", "MntGoldProds",

“MntSweetProducts”, “MntFishProducts”, “MntMeatProducts”, “MntFruits”, “MntWines”, and “Income”. All the outliers were reasonable, showing the pattern of more than one sample being above the upper quantile (Q3). The decision was made not to drop the samples containing the outliers, since it was concluded that these could tell more about the pattern of behavior of customers responding positive to the marketing campaign, than just being the noise in the data.

Feature engineering

First step was to handle the “Marital_Status”, since the column had values such as “Married”, “Together”, “Single”, Divorced”, “Widow”, “Alone”, ”YOLO”, “Absurd”. One hot encoded was selected as the way to go, and this was done by using dummy variables. Since “Alone”, ”YOLO”, and “Absurd” together didn’t make 1% of the customers, these new columns were dropped.

Next step was to create separate columns for year, month and day the customer become a member of the company extracting each value from the original column “Dt_Customer”. After creating new three features, the original one was dropped. The further analysis on these showed that the data collected was mostly evenly distributed among all the months of the year, as well as days, while the 2013 out of 2012, 2013, and 2014 was the most popular, having more than double of the samples when comparing to the other years (1173 samples).

The final feature that needed to be manipulated was “Education”. Since the values regarding education are ordinal, the corresponding values with appropriate weights were assigned to each value. This meant that phD values were replaced with the highest value – 5, while the 2n Cycle value was replaced with the lowest value – 1. The rest of the values in between these were assigned in the same manner, depending on the level of the degree.

The last step in this part of the process was to scale the numeric data. Based on the research, it was decided not to scale new features related to Education, since the difference between different levels of degrees would be lost. The numeric features and others related to days, month, and years were scaled. One-hot encoded data remind untouched as well.

Feature reduction

After cleaning the data and doing feature engineering process dataset had 33 features and 2216 observations. Seeing the correlation between features would help at detecting if there are highly correlated features, meaning they would contribute giving the same information and even making the model weaker by on feature constantly affecting the other. The correlation plot (see figure 1.) showed there aren’t any strong correlations (>0.8), while the highest correlation was 0.73 between “MntMeatProducts” and “NumCataloguePurchases”. This wasn’t enough to drop one of these.

The Principal Component Analysis (PCA) was next, using only numeric features, as the factor data wouldn’t give us the best results using PCA. The results were not what it was expected, having first two Principal components explaining only 34.46% of data variance. This result was not nearly enough what it would need to be to transform raw data onto the first two principal components and reduce the dimensionality. Valuable information gotten from plotting PCA was that “Kidhome” feature and “NumCatalogue” were some of the feature contributing the most to the PC1, while still giving different information, while “NumDealsPurchases, and “year” were one of the most important when contributing to PC2. Many features on the left side overlapped, showing that these revel somehow similar information. The color of data points was assigned based on the target label. PCA doesn’t show any obvious clustering, except for many negative responses on the right side of the plot.(see Figure2)

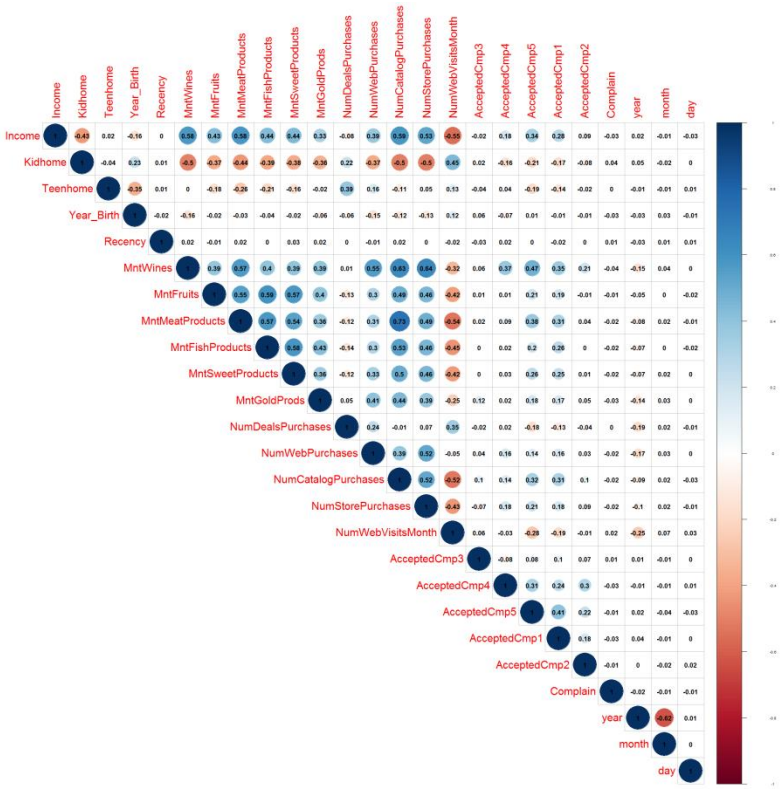

Figure 1 correlation plot

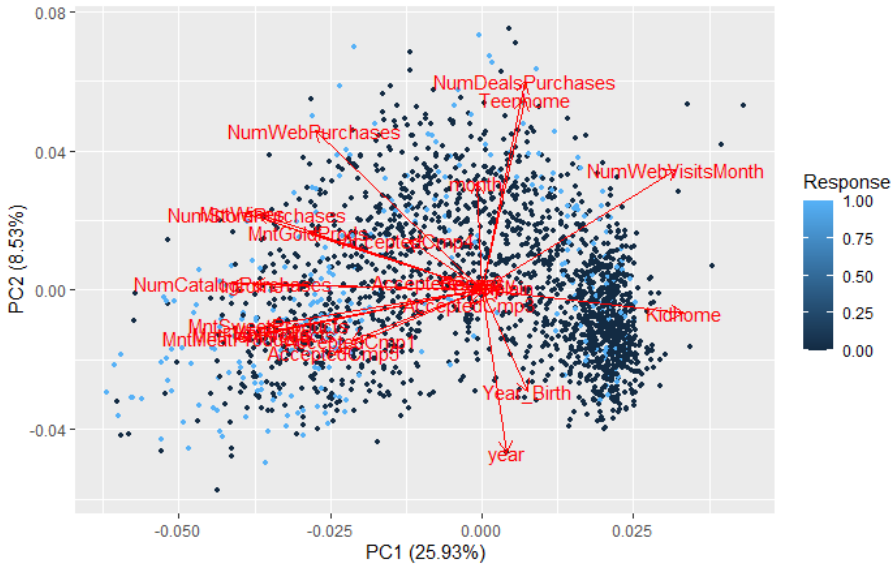

Figure 2 PCA plot

Before going ahead with creating clusters, both elbow and silhouette methods were used to check for the best number of clusters based on the data. From the target label it was clear that 2 clusters would be the most appropriate (yes/no), so it was very good when both methods showed the same result. Clustering showed the two groups k-means was able to group together, but comparing these grouping with the real labels data had-"Response" was not matching. There was much more overlapping among the real labels, than among the ones it was grouped with K-means.

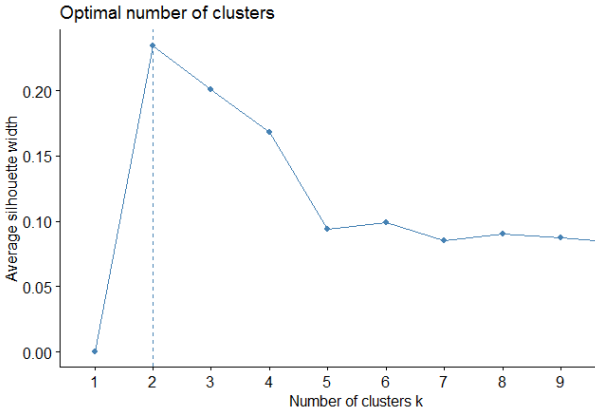

Figure 3 Silhouette calculation

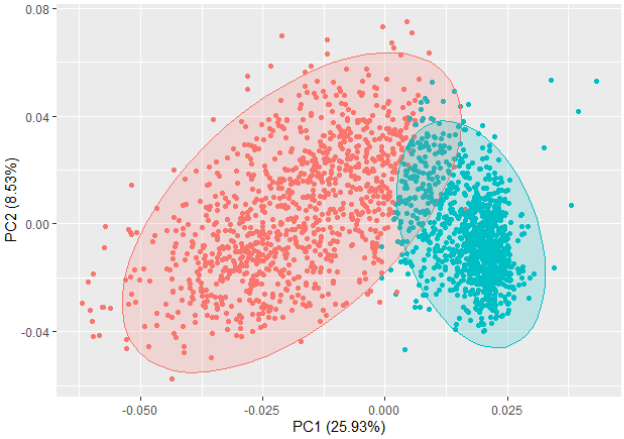

Figure 4 K-means cluster

The cluster on the right would be a good solution if there weren't already assigned labels for classification. Nevertheless, the clustering confirms the strong grouping of samples with a negative response mentioned earlier during the PCA analysis.

Model selection

For creating generalized linear model "glm" was used with binomial family and "logit" activation function. For choosing the best combination of features step function was used with backward direction. After running the step function **AIC** value went from **875.78** for the full model, to **852.41** AIC, leaving the best combination of features to be Education, Teenhome, Recency, MntMeatProducts , NumDealsPurchases, NumWebPurchases, NumStorePurchases, AcceptedCmp3, AcceptedCmp4 ,AcceptedCmp5, AcceptedCmp1, AcceptedCmp2, Marital_Status_Married , Marital_Status_Together, year, and month. This answered to both of the leading questions defined in the Introduction part.

This model was used to fit the training dataset and later on used to predict the "Response" on the test dataset, which was 30% of complete data. It was shown that the best threshold was 0.6, having values greater being classified as 1, else 0. The accuracy of the prediction was 89%.

Appendix

List of all variables:

1. ID - Customer's id - (integer)
2. Year_Birth - Customer's year of birth - (integer)
3. Education - Customer's level of education (character)
4. Marital_Status - Customer's marital status (character)
5. Income - Customer's yearly household income - (integer)
6. Kidhome - Number of small children in customer's household - (integer)
7. Teenhome - Number of teenagers in customer's household - (integer)
8. Dt_Customer - Date of customer's enrolment with the company - (character)
9. Recency - Number of days since the last purchase - (integer)
10. MntWines - Amount spent on wine products in the last 2 years - (integer)
11. MntFruits - Amount spent on fruits products in the last 2 years - (integer)
12. MntMeatProducts - Amount spent on meat products in the last 2 years - (integer)
13. MntFishProducts - Amount spent on fish products in the last 2 years - (integer)
14. MntSweetProducts - Amount spent on sweet products in the last 2 years - (integer)
15. MntGoldProds - Amount spent on gold products in the last 2 years - (integer)
16. NumDealsPurchases - Number of purchases made with discount - (integer)

17. NumWebPurchases - Number of purchases made through company's web site - (integer)
18. NumCatalogPurchases - Number of purchases made using catalogue - (integer)
19. NumStorePurchases - Number of purchases made directly in stores - (integer)
20. NumWebVisitsMonth - Number of purchases made through company's web site - (integer)
21. AcceptedCmp3 - 1 if customer accepted the offer in the 3rd campaign, 0 otherwise - (integer)
22. AcceptedCmp4 - 1 if customer accepted the offer in the 4th campaign, 0 otherwise - (integer)
23. AcceptedCmp5 - 1 if customer accepted the offer in the 5th campaign, 0 otherwise - (integer)
24. AcceptedCmp1 - 1 if customer accepted the offer in the 1st campaign, 0 otherwise - (integer)
25. AcceptedCmp2 - 1 if customer accepted the offer in the 2nd campaign, 0 otherwise - (integer)
26. Complain - 1 if customer complained in the last 2 years - (integer)
27. Z_CostContact - Cost to contact a customer - (integer)
28. Z_Revenue - Revenue after client accepting campaign - (integer)
29. Response (target) - 1 if customer accepted the offer in the last campaign, 0 otherwise - (integer)

The summary of the final model:

```
my_model <- readRDS("model.rds")
summary(my_model)

##
## Call:
## glm(formula = Response ~ Education + Teenhome + Recency + MntMeatProducts +
##     NumDealsPurchases + NumWebPurchases + NumStorePurchases +
##     AcceptedCmp3 + AcceptedCmp4 + AcceptedCmp5 + AcceptedCmp1 +
##     AcceptedCmp2 + Marital_Status_Married + Marital_Status_Together +
##     year + month, data = df[2:33])
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -0.81136  -0.16955  -0.05691   0.06775   1.04947
##
## Coefficients:
##                          Estimate Std. Error t value Pr(>|t|)
## (Intercept)              0.118477   0.021828   5.428 6.34e-08 ***
## Education                0.029604   0.005606   5.281 1.41e-07 ***
## Teenhome                -0.034216   0.007174  -4.769 1.97e-06 ***
## Recency                 -0.070381   0.006220 -11.315  < 2e-16 ***
## MntMeatProducts          0.046640   0.008029   5.809 7.19e-09 ***
## NumDealsPurchases        0.023330   0.007125   3.274 0.001075 **
## NumWebPurchases          0.027789   0.007731   3.595 0.000332 ***
## NumStorePurchases       -0.059163   0.008063  -7.338 3.04e-13 ***
## AcceptedCmp3             0.071450   0.006366  11.224  < 2e-16 ***
## AcceptedCmp4             0.031163   0.006957   4.480 7.86e-06 ***
## AcceptedCmp5             0.061978   0.007497   8.267 2.35e-16 ***
## AcceptedCmp1             0.053904   0.007097   7.595 4.51e-14 ***
## AcceptedCmp2             0.023907   0.006629   3.606 0.000318 ***
## Marital_Status_Married  -0.104761   0.014461  -7.244 5.98e-13 ***
## Marital_Status_Together -0.108932   0.016075  -6.776 1.58e-11 ***
## year                    -0.084663   0.008279 -10.226  < 2e-16 ***
## month                   -0.037012   0.008017  -4.617 4.12e-06 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for gaussian family taken to be 0.08528501)
##
##     Null deviance: 282.96  on 2215  degrees of freedom
## Residual deviance: 187.54  on 2199  degrees of freedom
## AIC: 852.42
##
## Number of Fisher Scoring iterations: 2
```