Superstore Marketing Campaign Data Analysis

Evangelia Zaou
Department of Mathematics and Statistics
University of Cyprus
Nicosia, Cyprus
zaou.evangelia@ucy.ac.cy

Grigorii Turchenko
Department of Business and Public Administration
University of Cyprus
Nicosia, Cyprus
turchenko.grigorii@ucy.ac.cy

Panikos Christou
Department of Computer Science
University of Cyprus
Nicosia, Cyprus
christou.panikos@ucy.ac.cy

1. INTRODUCTION

1.1 Problem statement

Marketing has always been one of the central activities to any business, especially for B2C (business-to-consumer) companies. High competition and product homogeneity, which is quite often the case, forces marketing managers to constantly create unique offers in pursuit of attracting more and more customers. In such a specification, it would be useful for the decision-makers to identify which factors influence the marketing campaign response, so that an offer can be created in a more customized way resulting in effective and efficient targeting.

In order to carry out a profound analysis, we plan to answer the following questions based on the historical data of Superstore marketing campaigns:

- Which characteristics of the customers and their consumer behavior are important in the context of marketing campaign response?
- Is it possible to create a robust model, capable of accurately predicting the decision of a customer to respond?

1.2 Goals

The objective of the analysis is to determine the significant factors influencing the decision of a superstore customer to respond to the marketing campaign offer. Through an exploratory data analysis process the patterns in the data were studied, as well as the cleaning procedures were carried out. Consecutively, with the methods of statistical analysis, the modelling was carried out, and based on the results the research questions were addressed.

2. DATA DESCRIPTION

The table below provides a brief explanation of each variable present in the dataset.

Table 1 - 'Feature' Description

Variable name	*Variable meaning*
Id	Unique ID of each customer
Year_Birth	Age of the customer
Education	Customer's level of education
Marital_Status	Customer's marital status
Income	Customer's yearly household income
Kidhome	Customer's of small children in customer's household
Teenhome	Number of teenagers in customer's household
Dt_Customer	Date of customer's enrollment with the company
Recency	Number of days since the last purchase
MntWines	The amount spent on wine products in the last 2 years
MntFruits	The amount spent on fruits products in the last 2 years
MntMeatProducts	The amount spent on meat products in the last 2 years
MntFishProducts	The amount spent on fish products in the last 2 years
MntSweetProducts	Amount spent on sweet products in the last 2 years
MntGoldProds	The amount spent on gold products in the last 2 years
NumDealsPurchases	Number of purchases made with discount
NumWebPurchases	Number of purchases made through the company's website
NumCatalogPurchases	Number of purchases made using catalog (buying goods to be shipped through the mail)
NumStorePurchases	Number of purchases made directly in stores
NumWebVisitsMonth	Number of visits to company's website in the last month
Complain	1 if the customer complained in the last 2 years

In total, there are 2,240 observations of 22 different variables. In case we exclude the Id column, we observe 184 duplicated observations. We assume that these observations refer to the same person as it is not that common for a person to have the same characteristics and customer behaviour with another person. Therefore, these variables are removed from the dataset that will be used for the analysis. In addition, there are 24 values that are missing for the Income variable, accounting for approximately 1.2% of total observations which our team interpolated by filling in the mean of every class.

3. EXPLORATORY DATA ANALYSIS

In this section the above-mentioned variables will be presented in the greater detail. Since some of the predictors have similar semantic meaning, they will be presented together for the purposes of concision. Moreover, we will only display a subset we thought would be better for the report.

3.1 Variables Exploration

3.1.1 Response

The target variable is binary and takes value of “1” if the given customer responded to a previous marketing offer, and “0” otherwise. The bar chart below represents its distribution.

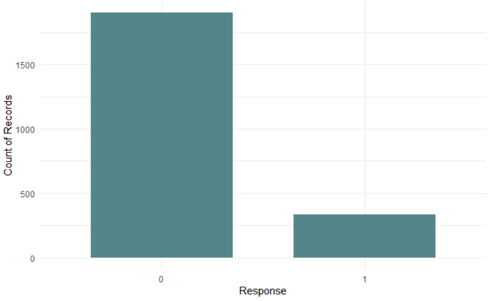


Figure 1 - 'Response' Bar plot

The chart gave us an indication of a notable imbalance of the classes, which had been taken into account at the modelling stage.

3.1.2 Year of Birth

The histogram below shows the distribution of the birth year with respect to responders and non-responders.

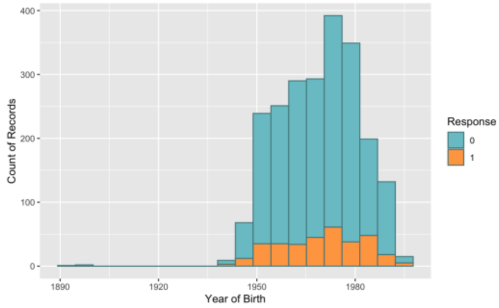


Figure 2 - 'Year of Birth' Histogram

The issue with this variable is there are people who were born in the 19th century, which is highly unlikely to be true. There were 4 of such people, and hence 4 rows were dropped. In general, both samples have approximately similar distribution with the majority of our customers born between 1959 and 1977.

3.1.3 Education

The distribution of the highest education achieved by a customer is represented in the bar chart.

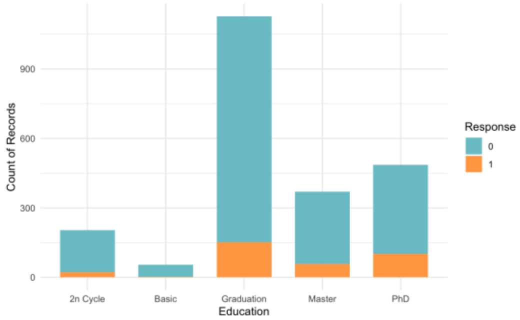


Figure 3 – 'Education' Bar plot

The "2n Cycle" education was combined with the "Master", since essentially, they both represent the same step in the educational system. In general, most of the customers have completed the first cycle of higher education, however the share of graduates and postgraduates is also quite substantial.

3.1.4 Marital Status

The dataset provides quite broad information on the client's marital status, which is summarized in the following bar chart.

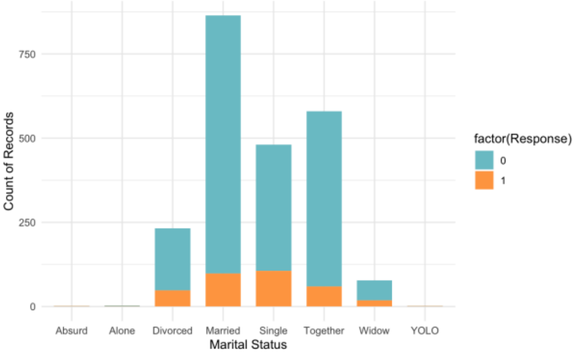


Figure 4 - 'Marital Status' Bar Plot

It is evident there are 3 very poorly represented categories, which need to be dealt with. People with "Absurd" and "YOLO" statuses were disregarded from the analysis, since it is unclear what these can possibly indicate. As for the "Alone" one, it was combined with the "Single" marital status, because they provide more or less the same information.

3.1.5 Income

For the Income distribution, a long-thin tail, positive skewness and high variance and kurtosis is observed. In cases where high kurtosis occurs, there is an indication that the high variance is resulted from outliers in the data. As shown in the boxplot below, there is an extreme observation in the income data.

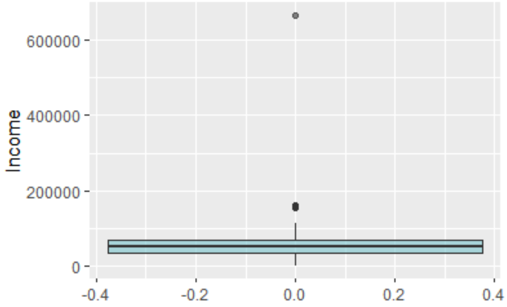


Figure 5 - 'Income' Box Plot

There is a significant decrease in both kurtosis and variance for the current variable if the extreme observation mentioned before is excluded.

Table 2 - 'Income' Summary Statistics

	Mean	Variance	Skewness	Kurtosis
Income	52,330	652,787,278	7.055	163.855
Income without extreme value	52,027	466,643,240	0.368	0.843

As shown above, this extreme observation has a significant impact in the summary statistics. Having this in mind, the influence of this observation in the model will also be tested in the statistical analysis stage.

3.1.6 Amount Spent

There are 6 variables in the dataset indicating the total spending on various products, such as wine, fruits, fish, meat, sweets, and gold. First the distributions were inspected.

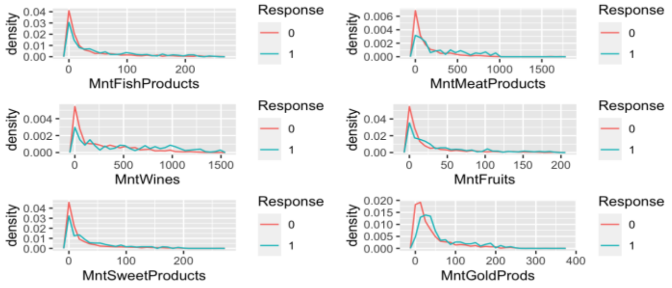


Figure 6 - Distribution of amount spent on various products

The behavior of these variables' distribution is very similar in general. It must be noted though that different behavior for each of these features is observed between the target variable's classes. For instance, higher probability is allocated for higher amount spent in wines by the positive instances rather than the negative instances. Due to these differences, it is expected that some of these variables will contribute to the model development.

Moreover, Pearson correlation was calculated in order to identify covariance between these features.

Table 3 - Correlation Matrix of Amount spent on products

	Wines	**Fruits**	**Meat**	**Fish**	**Sweets**	**Gold**
Wines	1	0.39	0.56	0.4	0.39	0.39
Fruits	0.39	1	0.54	0.59	0.57	0.39
Meat	0.56	0.54	1	0.57	0.52	0.35
Fish	0.4	0.59	0.57	1	0.58	0.42
Sweets	0.39	0.57	0.52	0.58	1	0.37
Gold	0.39	0.39	0.35	0.42	0.37	1

The resulted correlations are quite high, indicating the need to check the model for multicollinearity at the later stages.

3.1.7 Number of Purchases

The superstore sells its products across 4 channels, which are deals, store, catalog, and web.

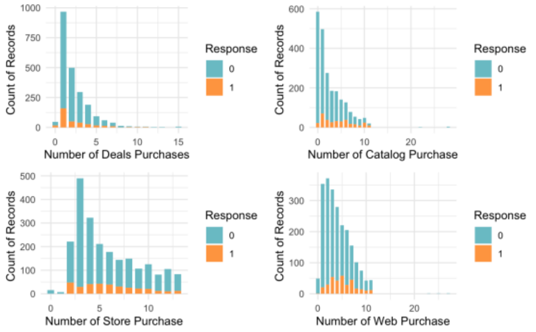


Figure 7-Frequency of Purchases on each channel

The distribution of the number of purchases from different channels varies according to the target variable classes. For instance, for web pages, there is a rational decrease in the number of purchases of negative class compared to the positive class. Moreover, for the specific variable any observation greater than 10, refers to a negative instance. Therefore, it is expected for these variables to be statistically significant in the statistical analysis that will be performed later.

Pearson correlation among these variables have also been calculated.

Table 4 - Correlation between number of purchases for each channel

	NumDealsPurchases	**NumWebPurchases**	**NumCatalogPurchases**	**NumStorePurchases**
NumDealsPurchases	1	0.23	-0.01	0.07
NumWebPurchases	0.23	1	0.38	0.5
NumCatalogPurchases	-0.01	0.38	1	0.52
NumStorePurchases	0.07	0.5	0.52	1

Besides quite low correlation of deals with catalog and store purchases, the rest of the correlations are notably high, which once again identifies potential multicollinearity early.

3.1.8 Recency

The distributions of recency for the 2 groups are quite different. For Negative Responders, higher probability is allocated to higher recency values, meaning that the majority of the negative responders have not purchased something recently. For positive responders, the opposite behavior is observed, as higher probability is allocated to lower values of recency.

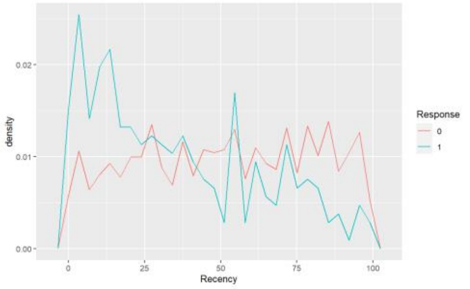


Figure 8 - Distribution of Recency for each class

3.1.9 Dt Customer

The date of enrolment with the Superstore does not provide substantial insights. However, bringing the observations down to the month and year level provided some useful information regarding monthly seasonality.

4. MODELING

In Section 4 of this report, the comprehensive modelling approach is presented. The techniques that were mainly used to build an interpretable and accurate model were logistic regression, naive Bayes, and k-nearest neighbours. The model performance will be evaluated based on various classification metrics. It must be noted though that emphasis will be given to Recall (a measure of classifier's completeness) and the F1-score (harmonic mean of precision and recall) that is the most suitable for imbalanced dataset and the fact that positive instance classification is more important in this problem.

4.1 Logistic Regression

4.1.1 Initial modelling

a) 4.1.1.1 Extreme value in Income

From the exploratory analysis that was implemented before, there was an indication of one extreme value in the Income feature. It is important to examine whether this observation will be influential to the logistic regression model. Cook's distance for logistic regression is a measure of the influence of the $j-th$ observation to the model's coefficients.

The respective cook distances for each observation are presented below using an index plot.

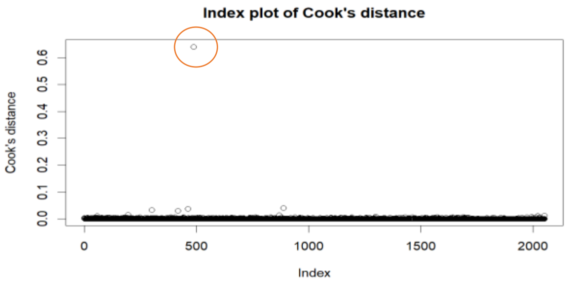


Figure 9 - Cook distance index plot

The observation with the higher cook distance was confirmed to be the one with the extreme value in the income feature.

We see the value having a cook distance of about 0.8 so we rerun the model without to see how it was changed. Another important thing that should be mentioned is that if the extreme observation was not excluded from the training dataset, the income coefficient was not considered statistically significant. However, if this observation was excluded from the training dataset, then the income coefficient was statistically significant with significance level of 0.05. Therefore, this observation was removed from the dataset.

4.1.2 Dealing with imbalance

Before proceeding with the imbalance elimination techniques, the datasets were separated into train (80% of the initial) and test (20% of the initial) datasets. The training dataset will be used for the modelling phase, as the classifiers will learn from the specific observations and the test dataset will be used for evaluation purposes. It was also confirmed that the resulted datasets have kept the imbalance ratio approximately the same with original dataset.

The training dataset (as the original) has an imbalance ratio of 85:15. Meaning that the 15% of the current dataset refers to the positive category which we are interested in. This will bias the model towards prediction of mostly negative instances of the target variable. One way of eliminating this is to choose a different threshold for the positive predictions. However, there is no optimal or formal procedure to choose this and thus it was decided not to be implemented. Another way for reducing the imbalance ratio is through down-sampling or up-sampling techniques. In the current project, the SMOTE technique was used to up-sample the positive instances in the training dataset. SMOTE generates a defined number of observations using the k-nearest neighbours (in this case $k = 3$). The desirable result was to transform the ratio to 2:1, meaning that 1/3 of the training dataset will refer to the positive class.

To evaluate the importance of the up-sampling technique, two logistic regression models were fitted, one trained in the original training dataset and another one trained in the up-sampled dataset. The quality assessment metrics for these models are represented in the table below.

Table 5 – Performance Evaluation of the Up-sampled dataset

Name	**Accuracy**	**Recall**	**Precision**	**Specificity**	**F1 score**
Original	0.859	0.338	0.595	0.957	0.431
SMOTE	0.795	0.585	0.400	0.835	0.475

If the classifier is trained in an up-sampled dataset, the Recall and F1-score are increased. It should be noted that Precision significantly decreased as the logistic regression classifier is more probable to classify a positive instance and thus the False Positive instances are increased too. However, it was decided to use through the analysis the Up-sampled dataset, as for the current problem, it is more valuable to identify a positive instance rather than misclassifying a negative instance.

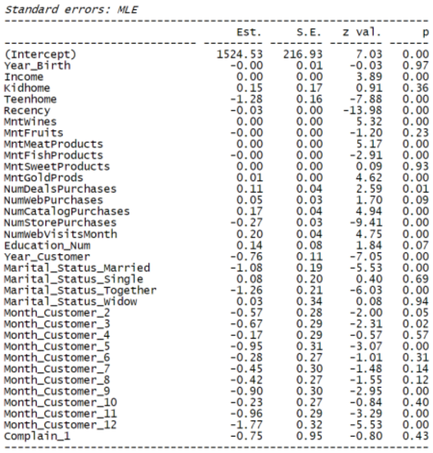
4.1.3 Logistic Regression – The Full Model

After training the Logistic Regression classifier with the up-sampled dataset with every available predictor, we have the following characteristics for the resulted model.

Table 6 - Resulted Deviances from Full Model

Deviance	**Value**	**Df**
Null	2750.4	2131
Residual	1790.2	2097

Null deviance correspond to how well the target variable is predicted by using only an intercept (random model). On the other hand, the residual deviance indicates how well the target variable is predicted using the corresponding predictors. In general, the lower the deviance the better. It is straightforward that the Residual deviance is lower than the Null Deviance, meaning that the current model is considered better than the random.



```
Standard errors: MLE
--------------------------------------------------------------
                                 Est.     S.E.   z val.      p
------------------------------ -------- -------- -------- ------
(Intercept)                    1524.53   216.93     7.03   0.00
Year_Birth                       -0.00     0.01    -0.03   0.97
Income                            0.00     0.00     3.89   0.00
Kidhome                           0.15     0.17     0.91   0.36
Teenhome                         -1.28     0.16    -7.88   0.00
Recency                          -0.03     0.00   -13.98   0.00
MntWines                          0.00     0.00     5.32   0.00
MntFruits                        -0.00     0.00    -1.20   0.23
MntMeatProducts                   0.00     0.00     5.17   0.00
MntFishProducts                  -0.00     0.00    -2.91   0.00
MntSweetProducts                  0.00     0.00     0.09   0.93
MntGoldProds                      0.01     0.00     4.62   0.00
NumDealsPurchases                 0.11     0.04     2.59   0.01
NumWebPurchases                   0.05     0.03     1.70   0.09
NumCatalogPurchases               0.17     0.04     4.94   0.00
NumStorePurchases                -0.27     0.03    -9.41   0.00
NumWebVisitsMonth                 0.20     0.04     4.75   0.00
Education_Num                     0.14     0.08     1.84   0.07
Year_Customer                    -0.76     0.11    -7.05   0.00
Marital_Status_Married           -1.08     0.19    -5.53   0.00
Marital_Status_Single             0.08     0.20     0.40   0.69
Marital_Status_Together          -1.26     0.21    -6.03   0.00
Marital_Status_Widow              0.03     0.34     0.08   0.94
Month_Customer_2                 -0.57     0.28    -2.00   0.05
Month_Customer_3                 -0.67     0.29    -2.31   0.02
Month_Customer_4                 -0.17     0.29    -0.57   0.57
Month_Customer_5                 -0.95     0.31    -3.07   0.00
Month_Customer_6                 -0.28     0.27    -1.01   0.31
Month_Customer_7                 -0.45     0.30    -1.48   0.14
Month_Customer_8                 -0.42     0.27    -1.55   0.12
Month_Customer_9                 -0.90     0.30    -2.95   0.00
Month_Customer_10                -0.23     0.27    -0.84   0.40
Month_Customer_11                -0.96     0.29    -3.29   0.00
Month_Customer_12                -1.77     0.32    -5.53   0.00
Complain_1                       -0.75     0.95    -0.80   0.43
--------------------------------------------------------------
```

Figure 10 - Summary of Logistic Regression Full Model

The last column of the table indicates the resulting p-value of the z-statistic for each coefficient. The null hypothesis that is tested here is if the coefficient for the respective predictor is equal to zero. It is important to mention that for the dummy variables (Marital Status, Month Customer and Complain), the baseline category is included in the intercept. Specifically, in the intercept the information of Marital Status Divorced, Month Customer 1 and Complain 0 is included. Therefore, the interpretation of the coefficients for the other dummy variables are slightly different. For instance, the coefficient of the marital status 'Married' expresses the difference between the 'Divorced' category (that is included in the intercept) and the Married category. Having that in mind, the coefficients for the following features are not considered statistically significant in level of 0.05: Year of Birth, Kids at home, Amount of Fruits purchased, Amount of sweets purchased, Number of web purchases, Level of education, difference in marital status between single and divorced and widow and divorced, difference between complain occurrence with no complain and with Month Customer 1 with all of the months 4, 6, 7, 8 and 10. Some features, such as amount of fruits or sweets purchased, are not considered statistically significant as their information is also included in other features as we have already identify correlation between these features.

4.1.4 Investigation of Multicollinearity

By definition, logistic regression assumes to have little to no multicollinearity between the predictors. As it has already stated before, there are some features that appear to be correlated with each other. This should be investigated further in order to identify and eliminate (if any) multicollinearity from the training dataset.

In cases were genderized linear models are used and there are categorical predictors, Generalized Variance Inflation Factor (GVIF) is the indicator that is used for the identification of multicollinearity in the resulted model. The the GVIF represents inflation in the squared hypervolume of the multidimensional confidence ellipsoid for the coefficients. It has to be noted though, that as the number of predictors increases, the size of the GVIF tends to grow too and thus the (2p)-th root of the GVIF is mainly used. Using the vif function in R, we were able to estimate the resulting $(GVIF)^{1/2p}$ for each predictor.

Table 7 - Features with highest GVIF (subset)

Predictor	**GVIF**	**Df**	**GVIF^1/2p**
Income	5.063475	1	2.250217
NumWebVisitsMonth	3.427989	1	1.851483
NumCatalogPurchases	3.002357	1	1.732731
MntWines	2.849053	1	1.687914
MntMeatProducts	2.83355	1	1.683315

We are going to use the rule of thumb of having $(GVIF)^{1/2p} \geq 2.2$ to be an indication of multicollinearity. As we can see from the above extract it seems that this threshold is violated for Income coefficient, which means that the predictors exhibit collinearity.

Nevertheless, this approach just gives us an indication of the presence of multicollinearity in the current model and decisions for variables exclusions should not be taken up to this stage. To penalize or limit the multicollinearity problem, feature selection and penalized logistic regression will be implemented in a later stage.

4.1.5 Feature Importance and Feature Selection

a) 4.1.4.1 Feature Importance

The following plot indicates the percent of the total deviance explained by each variable alone and cumulatively with other variables in the model. The higher the percentage, the more important the variable is considered.

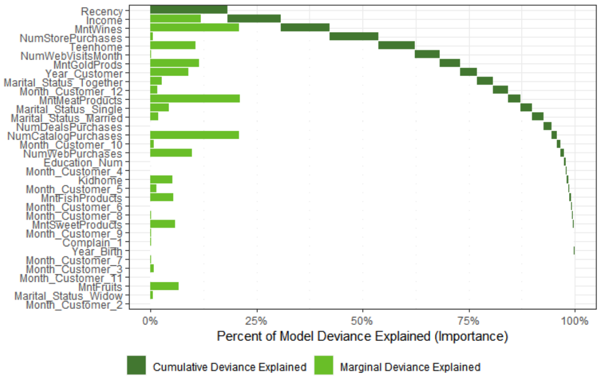


Figure 11 - Model deviance explained in accordance with predictors

We observe that the features with the higher marginal deviance are Recency, Income, Amount of Wines purchased, Number of Store purchases etc. The chart below represents another approach to computing features importance using absolute values of the t-statistics.

b) 4.1.4.2 Best Subset Selection

The current best subset selection that was used searches through a range of non-continuous model sizes. The criterion is going to use for this alternative version of Best Subset Selection is the BIC which penalizes the addition of new predictors in the model.

The result of the current algorithm is 21 features, meaning that 13 features have been removed. The model was re-fitted with this selection of features and evaluated using the test dataset.

The table below provides a comparison of the "full" logistic regression model to the logistic regression with reduced features according to best subset selection technique.

Table 8 - Performance Evaluation of Best Subset Features

Name	**Accuracy**	**Recall**	**Precision**	**Specificity**	**F1 score**
LR Smote	0.795	0.585	0.400	0.835	0.475
LR SM Subset	0.802	0.600	0.415	0.841	0.491

Classification metrics results have sufficiently increased. This means that the reduced model, was able to capture the valuable patterns for the separation of the 2 classes of the target variable by using only 21 out of the 34 available predictors.

c) 4.1.4.2 Forward Feature Selection

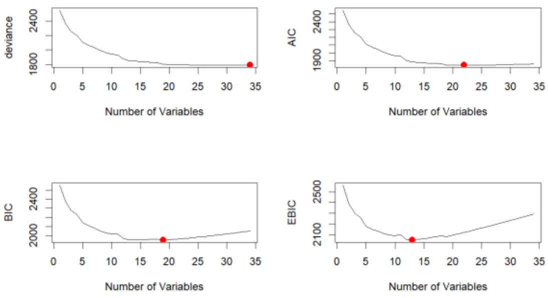


Figure 12 - Results of Forward Feature Selection

There are 2 other feature selection methods, which are computationally less expensive than best subset selection, namely forward and backward selection. After implementing the former, 19 features were selected based on the BIC criterion. However, we observe that 13 predictors might also be a sufficient number to make accurate prediction. To evaluate this, we fitted a new logistic regression by including only the 13 predictors.

Table 9 - Performance Evaluation of 13 predictors

Name	Accuracy	Recall	Precision	Specificity	F1 score
LR Best Sub	0.802	0.600	0.415	0.841	0.491
LR FW 13	0.805	0.615	0.421	0.841	0.500

The results are improved in terms of Recall, F1 score and Accuracy.

Nevertheless, the above results have been achieved with only one fit. In order to make more accurate conclusions we should evaluate the results of the forward selection through the Cross-Validation method. The function stepAIC in r enables to apply stepwise (both forward and backward) feature selection with having as a valuation metric the AIC criterion. After the implementation of the CV-stepAIC, 22 features were selected with all. It is important to mention here that the Best Subset Selection features are actually subset of these features too and the evaluation metrics are exactly the same as for the model fitted with 21 features.

4.1.6 Examination of Interactions

Another issue that we have is that the majority of our features refer to discrete or categorical data. Therefore, interactions were difficult to observed a-priori from plots.

However, we have observed that interaction term between TeenHome and Amount of sweets purchased, is actually considered statistically significant. We proceeded by evaluating the model performance using the training dataset and the results are improved according to all classification metrics. Moreover, we compare the resulted coefficients, from the baseline model and the model with interaction and we can see that by introducing an interaction term, the amount spent in sweets becomes significant too.

A linear model with this baseline performs better than without it.

Table 10 - Performance Evaluation by introducing interactions

Name	Accuracy	Recall	Precision	Specificity	F1-score
LR Smote	0.795	0.585	0.400	0.835	0.475
LR Interraction	0.802	0.600	0.415	0.841	0.491

Note that after introducing this interaction to the logistic regression model, we performed a step-wise feature selection using Cross-Validation in order to see its performance. Even if the selected features have changed compared, the performance of the model has not been improved.

4.1.7 Penalized logistic regression

We have seen that with feature selection, we manage to reduce our predictors to only 13 compared to the 34 that we initially had in our dataset. We now proceed to the application of the penalized logistic regression, to see how it performs. Ridge and Lasso are the regularizations methods that we are going to perform, which basically reduce/shrink the coefficients in the resulting logistic regression and achieving smaller mean squared errors for the coefficients.

4.1.6.1 Ridge Penalty

For the selection of the penalty term that is going to be used in the ridge logistic regression, we are implementing a Cross-Validation with 10 folds while trying to find the optimal lambda in accordance to the misclassification error. The resulted optimal lambda is $\lambda_{0,1} = 0.02$ which is very close to zero. We are also manually selecting another lambda from the resulted graph to evaluate the performance of the ridge logistic regression with higher penalty too $\lambda_{0,2} = e^{-3} \approx 0.05$

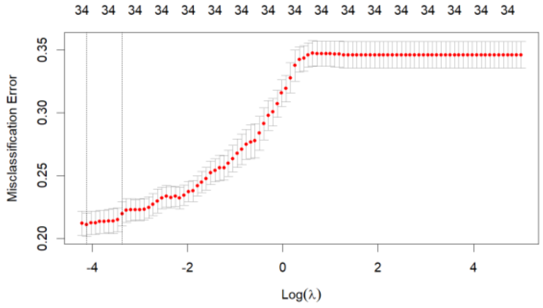


Figure 13 - Selection of Ridge Penalty

We proceed through the evaluation of these models using the test dataset.

Table 11 - Performance Evaluation of Ridge Logistic Regression

Name	Accuracy	Recall	Precision	Specificity	F1-score
$\lambda_{0,1}$	0.768	0.615	0.364	0.797	0.457
$\lambda_{0,2}$	0.780	0.374	0.569	0.910	0.451

From the above table we can conclude that Ridge regression failed to improve the classification metrics by shrinking the coefficients. We may have manage to achieve a higher recall by using the optimal penalty, but in terms of F1-score and Accuracy, the current model is not the best option. Precision is also significantly decrease which means that the current penalized classifier, does not provide “honest” results. It should also be noted that some coefficients have actually shrank compared to the full logistic regression model.

4.1.6.2 Lasso Penalty

Similarly, 10-fold CV was also performed here in order to select the lambda penalty. The optimal lambda is $\lambda_{1,1} = 0.001$, which is approximately equal to zero. It was initially expected that the results of Lasso Logistic Regression will be very similar to the ones introduced by the initial logistic regression model as the lambda is approximately zero. However, from the resulted plot, this is not the case, as we observe that for the optimal lambda we have 29 predictors. The variables that their coefficient shrunk to 0 are: Year of Birth, MntSweetProducts, Marital_Status_Widow, Month_Customer_7 and Complain_1 . By comparing these features with the feature selection implemented in a prior stage, we can see that indeed these features were rejected from the AIC and BIC feature selection methods too.

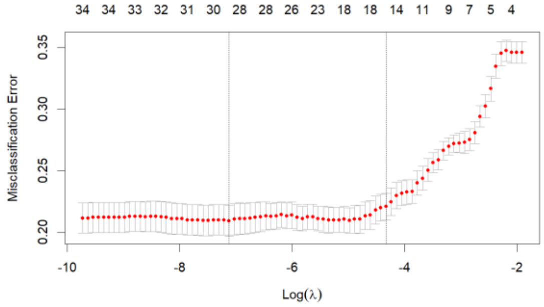


Figure 14 - Selection of Lasso Penalty

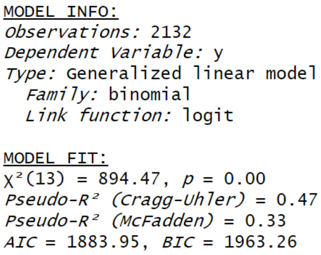
As for the evaluation metrics, we observe that the results according to Accuracy and F1-score are improved compared to the full model. We should also mention that the Recall and F1-score have significantly increased compared to every other model that was fitted.

Table 12 - Evaluation of Lasso Logistic Regression

Name	**Accuracy**	**Recall**	**Precision**	**Specificity**	**F1-score**
$\lambda_{1,1}$	0.788	0.692	0.402	0.806	0.509

4.1.8 Final Logistic Regression Model

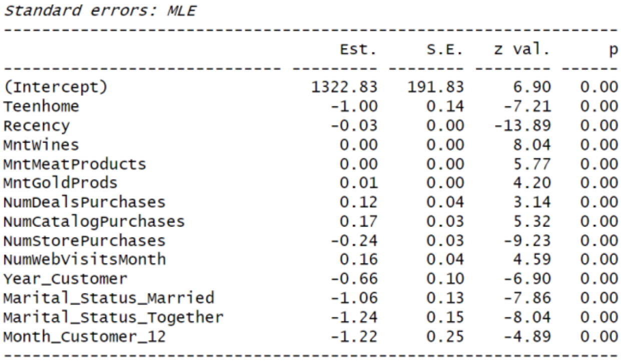
From all the analysis that was implemented before, we consider the optimal model to be the one selected through the forward selection with 13 predictors, which allows us to achieve also good results with minimum number of predictors.



```
MODEL INFO:
Observations: 2132
Dependent Variable: y
Type: Generalized linear model
  Family: binomial
  Link function: logit

MODEL FIT:
χ²(13) = 894.47, p = 0.00
Pseudo-R² (Cragg-Uhler) = 0.47
Pseudo-R² (McFadden) = 0.33
AIC = 1883.95, BIC = 1963.26
```

For the model fitting, we observe that our model is better than the random model with these predictors in every known significance level.



```
Standard errors: MLE
-----------------------------------------------------------------
                                   Est.      S.E.   z val.      p
---------------------------- --------- -------- -------- ------
(Intercept)                     1322.83    191.83     6.90   0.00
Teenhome                          -1.00      0.14    -7.21   0.00
Recency                           -0.03      0.00   -13.89   0.00
MntWines                           0.00      0.00     8.04   0.00
MntMeatProducts                    0.00      0.00     5.77   0.00
MntGoldProds                       0.01      0.00     4.20   0.00
NumDealsPurchases                  0.12      0.04     3.14   0.00
NumCatalogPurchases                0.17      0.03     5.32   0.00
NumStorePurchases                 -0.24      0.03    -9.23   0.00
NumWebVisitsMonth                  0.16      0.04     4.59   0.00
Year_Customer                     -0.66      0.10    -6.90   0.00
Marital_Status_Married            -1.06      0.13    -7.86   0.00
Marital_Status_Together           -1.24      0.15    -8.04   0.00
Month_Customer_12                 -1.22      0.25    -4.89   0.00
-----------------------------------------------------------------
```

Figure 15 - Summary of Logistic Regression Model with 13 predictors

Moreover, we observe that all the included predictors are considered statistically significant in significance level of 0.05. We now proceed by describing in detail what these coefficients mean and how they affect our prediction model on a subset of the features we have on the model.

- Teenhome: The negative coefficient (-0.9966) suggests that for each additional teenager at home, the log-odds of the Response (class 1) decreases. In other words, the presence of more teenagers at home is associated with a lower probability of the event.
- Recency: The negative coefficient (-0.0325) indicates that as the recency increases by one unit, the log-odds of the Response (class 1) occurring decreases. This means that higher recency values are associated with a lower probability of a client to respond positively in the marketing campaign. This was also observed in the exploratory analysis.
- NumCatalogPurchases: The positive coefficient (0.1728) suggests that more catalogue purchases are associated with a higher probability of a customer positively responding to our offer. Needlessly to say, this is also applied to the amount of Gold and Discounted products a Customer purchases.

Similarly, we can interpret the other numerical variables in the current logistic regression model. On the other hand, we should be careful with the interpretation of dummy variables in the current model, as we have many categories absent.

For the marital status, we have only 2 dummy variables left, compared to the full model which contained 4 dummy variables and the ‘Divorced’ category in the reference group. Now that the model is re-fitted, then we end-up having 3 new categories for the Marital Status: Married, Together and Other that contains the rest of the categories (Divorced, Widow and Single) which is the new reference group. Therefore, the coefficients now refer to the differences between the included categories with the synthetic ‘Other’ category.

- Marital_Status_Married: The negative coefficient of (-1.0603) is related with the change of the Marital Status Other to Marital Status Married. Therefore, the log-odds associated with having a positive instance are lower for people that are Married compared to the people that have one of the marital statuses included in the ‘Other’ category as defined above.

- Marital_Status_Together: The negative coefficient of (-1.2439) indicates a decrease in the log-odds of the positive class (1) compared to the people with marital status ‘Other’.

Similarly, from the month customer dummies, we observe that only Month Customer 12 (December) is included in the optimal model. The reference category that is now included in the intercept, is any other month different from December. Therefore, the interpretation of this coefficient also changes.

- Month Customer 12: The negative coefficient of (-1.215) implies that customers who joined in December have a lower log-odds of the event (class 1) occurring compared to the reference group, that in this case is every other month. In other words, customers who became clients in December have a lower probability of the event compared to clients who joined in other months.

In general, the results of the logistic regression model align with the observations made during the Exploratory Data Analysis (EDA) phase. Furthermore, the model offers valuable insights into the characteristics of customers targeted by the store. To effectively guide the store in targeting specific customer groups, we would advise them to focus on customers who exhibit attributes associated with the positive coefficient features. Simultaneously, it would be beneficial to avoid customers who exhibit attributes related to the negative coefficient features. This strategy will help the store to better engage and serve their target audience.

4.1.9 PCA

As we have already examined in a previous section, we had an indication of multicollinearity in our data. Another way of dealing with this is the application of PCA, as the resulted components are orthogonal and uncorrelated. However, the majority of our features are distinct/categorical and as PCA is a rotation of data from one coordinate system to another, continues variables are required for its implementation.

Therefore, even if PCA is possible to be implemented when we apply one-hot encoding in our dataset (without getting an error in the code), this should not be done.

Nevertheless, for educational purposes, we implemented PCA. We used the scaled and up-sampled training dataset in order to ensure that all variables have equal importance in the PCA, regardless of their scale or variance. This helped prevent variables with high variances from dominating the principal components and can improve the accuracy of the resulting model.

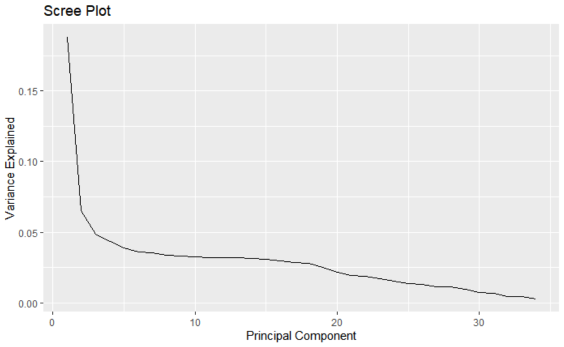


Figure 16 - Variance explained by the addition according to the Number of Principal Components

Based on the plot above, from the 34 components we have created there seems to be an elbow effect at 4. However, the corresponding scatter-plots between these components do not give us a clear separation of the two classes.

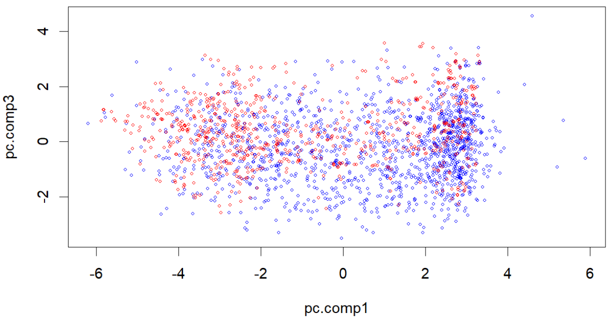


Figure 17 - Scatter plot of 1st and 3rd Principal Components

4.2 LDA and QDA

For LDA and QDA we have an assumption that our independent variables follow the normal distribution. As most of our features refer to categorical or discrete variables and only Income and Amount of products purchased can be considered continuous, the normality assumption does not hold. Thus, we cannot implement these 2 methods.

4.3 Naive Bayes

Naive Bayes can be actually used in cases were we have categorical or discrete variables, as the assumption behind the distribution of each variable is more "flexible".

We present below the evaluation metrics for the test dataset resulting from the Naïve-Bayes model.

Table 13 - Performance Evaluation of the Naive Bayes classifier (Full Model)

Name	**Accuracy**	**Recall**	**Precision**	**Specificity**	**F1-score**
NB- Test	0.693	0.492	0.256	0.730	0.337
LR FW 13 Feat	0.805	0.615	0.421	0.841	0.500

We see that it performs worse than the optimal Logistic regression model on all metrics.

From theory, Naive Bayes assumes in-dependency between the predictors. However, as we have seen before, we have some predictors that are highly correlated. Therefore, for now we will remove every amount of product purchase, except meat and Gold in order to eliminate the correlation between the amount of products purchased and re-fit the model in order to see how it performs.

Table 14 - Performance Evaluation of the Naive Bayes classifier on a redued dataset

Name	**Accuracy**	**Recall**	**Precision**	**Specificity**	**F1-score**
NB-Reduced Test	0.748	0.602	0.319	0.773	0.417

By removing correlated columns, we managed to achieve better performance than before. We see that we were right on removing the highly corelated columns and our performance increased significantly. Still, we are not as good as the Logistic Regression model, but we can see we have almost the same Recall.

The last thing we did, was scaling the numeric variables using a standardized scaler. After retraining the model, we see that Recall increased, but Precicion, Accuracy and F1-score significantly decrease.

Table 15 - Performance Evaluation of the Naive Bayes in a scaled dataset

Name	**Accuracy**	**Recall**	**Precision**	**Specificity**	**F1**
NB reduced scaled	0.727	0.508	0.292	0.768	0.371

Therefore, Naïve Bayes, is not the classifier that we would like to have for the specific problem.

4.4 K-Nearest Neighbors (KNN)

As a starting point, KNN was implemented for the original data, scaled and without any synthetic data points. Different values of k from 1 to 8 were used, and the results of the fits are depicted in the figure below.

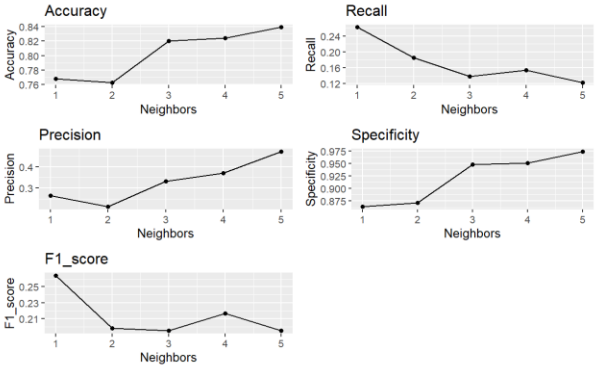


Figure 18 - K-NN with the original dataset

In general, KNN did not fit on the non-upsampled data well. It was able to identify negative instances better than the positive ones, which can be concluded by observing the increase in specificity. Recall as well as F1-score decreased as the number of neighbors increased. In addition, cross-validation was implemented to determine an optimal k, which resulted in $k = 20$. However, in this setting the model eventually switched to classifying all the instances as 0, which is why it could not be considered highly useful.

Similar procedure was conducted for the SMOTE-upsampled and downsampled data.

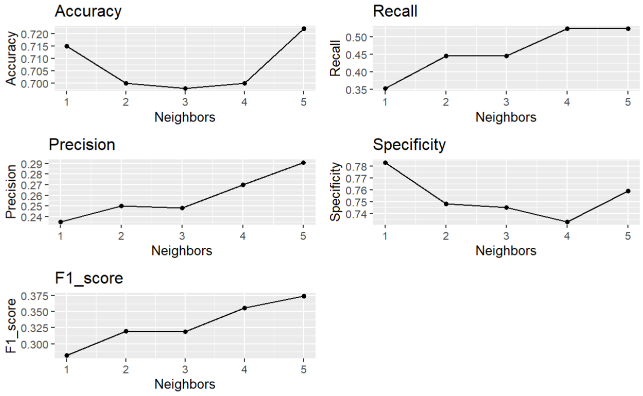


Figure 19 - K-NN with the up-sampled dataset (SMOTE)

While for the latter optimal number of neighbors was 25, which essentially indicated a highly similar result to the model fit on the original dataset, the former yielded a better optimal k equal to 2. The comparison between all the models is shown in the table below.

Table 16 - Performance Evaluation of K-NN Classifier

Name	**Accuracy**	**Recall**	**Precision**	**Specificity**	**F1**
KNN k=1	0.783	0.254	0.276	0.879	0.265
KNN DW k=5	0.617	0.794	0.258	0.585	0.389
KNN SM k=4	0.688	0.540	0.256	0.715	0.347

Neither of the models produced valuable results, which is why it can be concluded that KNN does not provide a good fit for the specific data.

4.5 Tree Structures

4.5.1 Model performance

Tree models are powerful tools that allow us to understand the importance of different features in a dataset. By using algorithms like Random Forest and Gradient Boosting, we can build models that not only predict outcomes accurately but also provide valuable insights into which variables are most influential.

However, sometimes simpler models can perform just as well as these complex ones. In the case of Logistic Regression, it's been proven to be an effective classification algorithm that can achieve high accuracy rates. When comparing the feature importance's obtained from a tree model with those of a Logistic Regression model, we can see that they are almost on par with each other. This highlights the robustness and versatility of Logistic Regression as it can be applied and improved in various scenarios with relatively reliable results.

Below we summarize the evaluation metrics for tree-based models compared to the current optimal logistic regression model.

Table 17 - Performance Evaluation of tree based models

Name	Accuracy	Recall	Precision	Specificity	F1
Logistic Regression	0.810	0.619	0.419	0.844	0.500
Decision Tree	0.817	0.446	0.426	0.887	0.436
Random Forest	0.844	0.429	0.491	0.919	0.458
Gradient Boosting Machines	0.837	0.603	0.475	0.879	0.531

An important observation here is that Gradient Boost Machine Model is able to make more accurate predictions in both classes as Precision and F1-score have both increased. In case that Gradient Boost machine is used as the classifier for the current model, then according to the Precision score, if a positive instance is predicted then an estimation of the probability of correctly identifying this would be about 0.48.

4.5.2 Feature importance of the Gradient Boosting Machines

As we have mentioned above, Tree-based classifiers enable us to get insights regarding the importance of the features. In this project,we will focus specifically on the Feature importance's seen on GBM since it had the best performance of the 3.

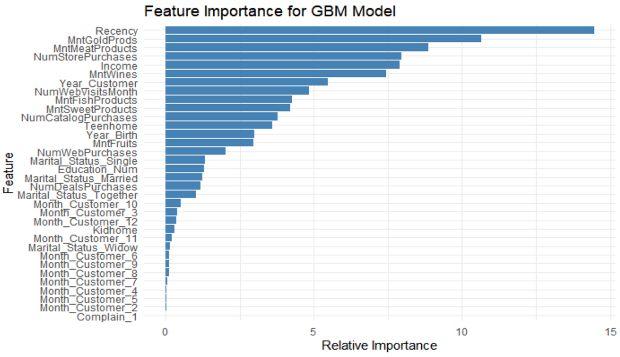


Figure 20 - Feature Importance in accordance with GBM

Here we see that again, Recency, amount of gold, meat and wines, Income and Year of becoming a customer were again the most influential features. Unlike logistic regression, we cannot interpret how much they affect the models' outcomes when it comes to probabilities or how they alter it's decision making. What we want to highlight though is that the extensive research and valuable insights we had on the Logistic Regression model have been proven really accurate in explaining how we are going to find positive instances of our target variable.

5. FINAL CONCLUSIONS AND CLOSING REMARKS

In this project, we embarked on a comprehensive data exploration and analysis journey, leading in the development of several classification models aimed at identifying customers who are likely to accept the supermarket offer. The dataset presented several challenges, including class imbalance, a wide variety of variable types, and multicollinearity issues. Despite these challenges, most models performed well, with the Logistic Regression model be the standout performer. The Logistic Regression model achieved high accuracy rates, and its feature importance analysis revealed that certain variables play a critical role in predicting customer response.

To further enhance the model's performance, we utilized several techniques such as up-sampling, down sampling, and penalized logistic regression. The up-sampling techniques involved using Synthetic Minority Over-sampling Technique (SMOTE) to generate more positive instances of customers accepting the supermarket offer. On the other hand, the down sampling technique involved randomly removing negative instances to balance the dataset. The technique most valuable was SMOTE since dropping data is not suggested, as important information might not be taken into account.

We also employed penalized logistic regression methods such as Ridge and Lasso penalties to deal with potential multicollinearity issues, which have shown to give the optimal results according to the Recall and F1-score.

Other classification models, such as Naive Bayes, K-NN Classifier, and Tree Structures, also showed promising results. The Naive Bayes model, which assumes that the presence of a particular feature is independent of other features, performed surprisingly badly. The K-NN Classifier model, which determines the class of a data point based on its proximity to other data points, also delivered unsatisfactory results. Finally, the Tree Structures, which included Decision Trees, Random Forest, and Gradient Boosting Machines, provided a little more confirmation in the feature importance we found from the Logistic Regression.

Lastly, regarding the most important factors influencing the response to a marketing campaign offer, some factors indicate that customers who attend the shops virtually by the catalog and make purchases on gold and discounted products are more likely to respond to the advertisement. This was further shown in section 4.1.8. as the coefficients of these specific variables were positive, meaning that the probability of responding to the marketing campaign is increased in accordance with these variables.

References

1. "12.1 - Logistic Regression | STAT 462," online.stat.psu.edu. https://online.stat.psu.edu/stat462/node/207/#:~:text=Cook%27s%20Distances&text=The%20residuals%20in%20this%20output (accessed Apr. 18, 2023).

   2. Nurunnabi, Abdul & Imon, A. & Nasser, M.. (2010). Identification of multiple influential observations in logistic regression. Journal of Applied Statistics. 37. 1605-1624. 10.1080/02664760903104307. Original Figure source found on https://www.researchgate.net/figure/Index-plots-of-a-Cooks-distance-b-difference-of-fits-DFFITS-and-c-generalized_fig5_258141262

3. "R: Calculation of filter-based variable importance," search.r-project.org. https://search.r-project.org/CRAN/refmans/caret/html/filterVarImp.html (accessed Apr. 18, 2023).

4. "Overview," R-Packages. https://cran.r-project.org/web/packages/tornado/vignettes/tornadoVignette.html (accessed Apr. 18, 2023).

5. C. Wen, A. Zhang, S. Quan, and X. Wang, "BeSS: An R Package for Best Subset Selection in Linear, Logistic and Cox Proportional Hazards Models," Journal of Statistical Software, vol. 94, pp. 1–24, Jun. 2020, doi: https://doi.org/10.18637/jss.v094.i04.

6. "Package 'BeSS' Type Package Title Best Subset Selection in Linear, Logistic and CoxPH Models," 2022. Accessed: Apr. 18, 2023. [Online]. Available: https://cran.r-project.org/web/packages/BeSS/BeSS.pdf

7. "Penalized Logistic Regression Essentials in R: Ridge, Lasso and Elastic Net - Articles - STHDA," www.sthda.com. http://www.sthda.com/english/articles/36-classification-methods-essentials/149-penalized-logistic-regression-essentials-in-r-ridge-lasso-and-elastic-net/

8. "Methods Implemented," R-Packages. https://cran.r-project.org/web/packages/logisticPCA/vignettes/logisticPCA.html (accessed Apr. 18, 2023).

9. "Lesson 6: Principal Components Analysis | STAT 897D," online.stat.psu.edu. https://online.stat.psu.edu/stat857/node/11/ (accessed Apr. 18, 2023).

10. "princomp function - RDocumentation," www.rdocumentation.org. https://www.rdocumentation.org/packages/stats/versions/3.6.2/topics/princomp (accessed Apr. 18, 2023).

11. finnstats, "PCA for Categorical Variables in R | R-bloggers," Nov. 20, 2022. https://www.r-bloggers.com/2022/11/pca-for-categorical-variables-in-r/ (accessed Apr. 18, 2023).

12. S. D. University Calvin, Chapter 11 Collinearity and Multicollinearity | STAT 245 Course Notes. Accessed: Apr. 18, 2023. [Online]. Available: https://stacyderuiter.github.io/s245-notes-bookdown/collinearity-and-multicollinearity.html

13. Y. Zablotski, "Logistic regression 5: multiple logistic regression with interactions," Dr. Yury Zablotski, Jan. 29, 2020. https://yury-zablotski.netlify.app/post/multiple-logistic-regression-with-interactions/#multiple-logistic-regression-with-higher-order-interactions (accessed Apr. 18, 2023).

14. "12.1 - Logistic Regression | STAT 462," online.stat.psu.edu. https://online.stat.psu.edu/stat462/node/207/

15. R. W. Nahhas, 5.19 Collinearity | Introduction to Regression Methods for Public Health Using R. Accessed: Apr. 18, 2023. [Online]. Available: https://bookdown.org/rwnahhas/RMPH/mlr-collinearity.html

16. S. Disci, "Feature Importance in Random Forest | R-bloggers," Jul. 01, 2021. https://www.r-bloggers.com/2021/07/feature-importance-in-random-forest/