SCHOOL OF BUSINESS

DEPARTMENT OF MANAGEMENT SCIENCE & TECHNOLOGY

ATHENS UNIVERSITY OF ECONOMICS AND BUSINESS

ACADEMIC YEAR OF 2021 – 2022

STATISTICS FOR BA II – ASSIGNMENT 1

NINAS KONSTANTINOS

f2822108

SUPERVISING PROFESSOR: KARLIS DIMITRIOS

Table of Contents

1 - Introduction

We are provided a dataset that contains the call history of a telemarketing company that is promoting a new product on behalf of a retail bank. The company's agents make daily phone calls to lists of the company's existing customers to promote and sell a new product. At the same time, clients may call the company's center for other reasons, during which the agents that serve them also try to promote the company's new product. It should be mentioned that many customers were contacted more than once, to persuade them to finally buy the product.

The dataset contains data regarding almost 40K phone calls that were conducted from May 2008 to June 2010. The dataset includes data that describe the bank's client (such as age, job, marital status, education and more), the date and the duration of the last call with each client, attributes that concern the telemarketing company's campaigns (for example number of calls per client and more), social and economic attributes, like consumer price monthly indexes and finally, the client's final decision on buying the product. The main purpose of this analysis is to pinpoint those variables that contribute to a significant degree to a successful contact (the client buys the product).

2 – Data Cleaning

The first step, prior to the analysis of the data, is the thorough clean process of the data. Specifically, the client's job as well as their marital status, education, credit default index, housing and personal loan indexes were updated to categorical variables. The same was applied for the communication type ('cellular', 'telephone'), the month and the weekday of the last contact with each client and outcome of the previous campaign with them. The values of 'subscribe' variable, which index whether the client finally purchased the product, initially were of the format no/yes. They were updated to 0/1 and to numeric type. Additionally, clients that were not contacted in the telemarketing company's previous campaigns received the value '999' as a placeholder in the variable 'pdays', which index the number of days number that passed by after the client was last contacted from a previous campaign. In fact, it was found that only 1.000 out of almost 40.000 customers in the dataset had been contacted at least once in previous campaigns. As a result, the rest 39.000 customers would have blank values in this specific variable. To resolve this issue, it made more sense to create an index of whether each customer was contacted at least once in any previous campaign. Although, it was found that the variable 'poutcome' indexed whether a customer was contacted in the previous campaign, so it was not needed to create a new variable. As a result, the variable 'pdays' was dropped completely from the dataset. Also, it is noticed that all the values in the consumer confidence index variable are negative, which does not make any sense, so they were all updated with their corresponding

positive numbers. Furthermore, an additional variable is created based on data from existing one. In detail, the season of the last contact was added as a variable, by identifying the values of the 'month' variable.

3 – Model Creation

To create a model that can recognize those variables that contribute to a significant degree in selling this specific product, first a constant and a full model will be produced. All models are produced based on a Bernoulli distribution since the response of the final model will be of type yes/no. A constant model is a model that only includes an intercept, while the full model includes every possible variable that is included in the dataset. This action is performed because these two models will constitute the bases on which the final model will be produced. In detail, a lasso and a stepwise method will be applied to identify the best inferential model.

The lasso method shrinks the coefficients of the unnecessary variables to 0 to remove them from the final model it screens based on a tuning parameter λ (lambda). Very big values of that parameter can set a great number of coefficients to zero, while a small value of that parameter can lead to over-fitted models (models with unnecessary explanatory variables). The model that is selected as the best has a Mean Classification Error (MCE) that is within one standard error of its minimum. It should be mentioned that the lasso process, unlike stepwise methods, does not produce models, and only screens them. The lasso process screened a model that drops the variables that explained the customer's age, the number of contacts performed before this campaign and for this client, the consumer confidence index and the Euribor 3-month rate from the full model. (Tables 1 & 2)

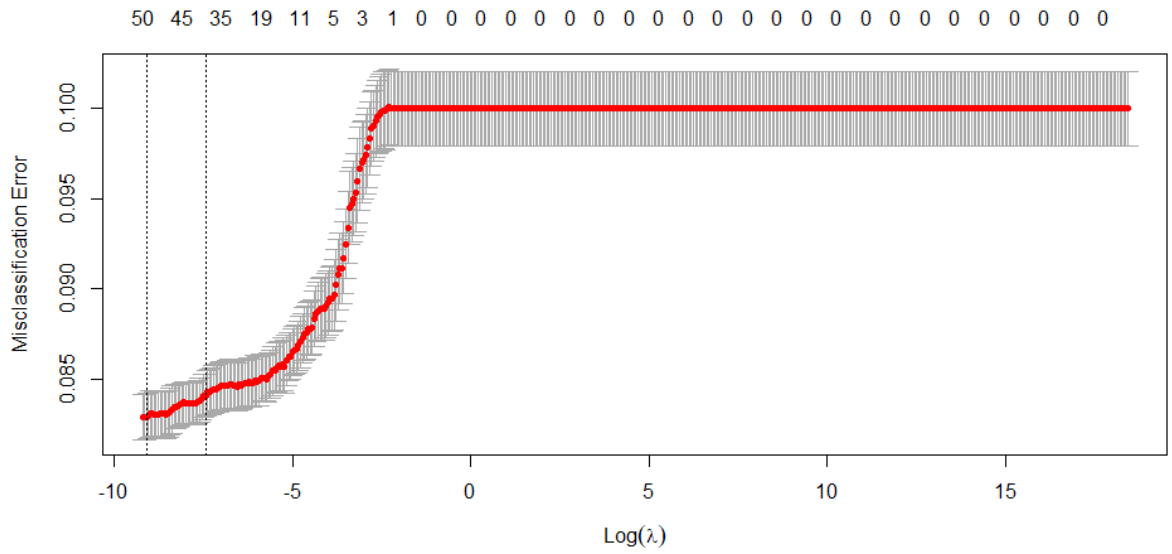

Table 1. Lasso method using cross validation

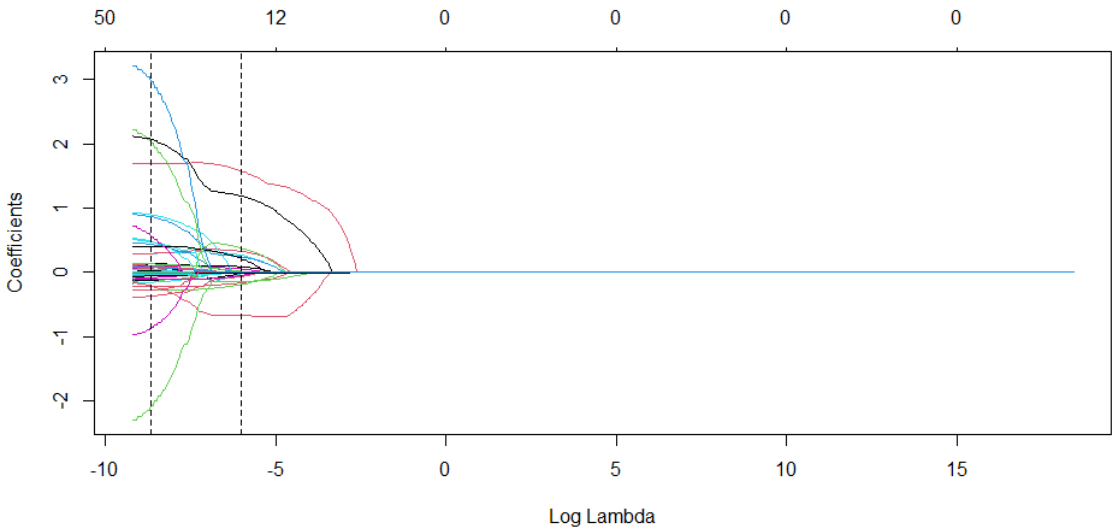

Table 2. Lasso method – Variable Selection

Following, a stepwise method will be applied to the model that was produced based on the screening of the lasso process. Since the aim of the analysis is to produce an inferential model, the Bayesian Information Criterion (BIC) will be used to explore the possible models for the analysis. Stepwise methods can be applied with a plethora of different techniques, for example with a forward direction, where the process starts with a constant model, and iteratively inserts variables from the dataset until it reaches its minimum possible AIC value. The model that was finally produced dropped the client's personal and housing loan indexes, their marital status, the weekday of their last contact, their educational status, and their work occupation.

In the next step, the model's explanatory variables (the ones trying to interpret the subscription index) multi-collinearity will be examined. The multi-collinearity of the variables is examined using the General Variance inflation factors method (GVIF). This method identifies linear relationships between explanatory variables. Specifically, the variables that have a GVIF^(1/(2·df)) (because the model includes at least one categorical variable with more than 2 levels) value greater than 3.16 have a multi-collinearity issue. It can be observed that some variables indeed have a multi-collinearity issue. To deal with this issue, the variable with the greatest multi-collinearity value will be removed, and the same test will be applied on the new model iteratively until the issue is resolved. It took one iteration of the above process to resolve this issue. As a result, the employment variation rate was dropped from the stepwise model, to produce the final model.

4 – Hypothesis Testing & Interpretation

Prior to interpreting the model's coefficients, the variable's significancy for the model will be examined, so that variables that do not contribute to the model's inference be dropped. To determine the significance of the variables in terms of their effect to the probability of a client buying the product, a threshold for a statistical significance equal to a = 5% will be held and compared with the p-values of each corresponding variable. Starting, the intercept, the duration and the quarterly indicator of the number of employees all seem to have a significant effect to the model since they all have a p-value < a.

The month during which the last contact was made will be first examined as a whole to determine whether it has a significant effect using a Wald test. A Wald test can be performed to examine the significance of any possible combination of a model's variables. It is found that the month variable is indeed significant for the model as a whole and should remain in it, since it has a p-value almost equal to 0 < a = 5%. The month August seems to have a significant effect as an individual month in the model, as well as July, June, March, May and October in comparison to April (April is included in the model's intercept). Exceptions to these months pose the months December and November and September (for any reasonable level of statistical significance a = 1%, 2%, 5%, 10%). For these three individual months, their effect to the probability of a successful sale can be considered to be equal to that of April, since they do not differ significantly.

Next, the outcome of previous campaigns will be examined. The rejection on previous campaigns is used as a threshold and is included in the intercept. First, the variable's significance will be examined as a whole using a Wald test. It is found that the variable as a whole is significant for the model's inference since it has a p-value almost equal to 0 < a = 5% and should not be removed from it. The variable that indexes whether a customer was not contacted in previous campaigns seems to have a significant effect as an individual variable in the model, since it has a p-value < a = 5%. It should be mentioned the index of whether the customer did buy the product in a previous campaign also has a significant effect in the model's inference as an individual variable.

Following, the variable that indexes whether a customer has credit in default will be examined. The customers that do not have credit in default are used as a threshold and are included in the intercept. But first, the variable's significance will be examined as a whole using a Wald test. It is found that the variable as a whole is also significant for the model's inference since it has a p-value < a = 5%. The variable that indexes whether it is known that a customer's credit defaulted seems to have a significant effect on the model as an individual variable. The same cannot be inferred for the index that shows whether the customer's credit defaulted since it has a p-value > a (for all reasonable values of a = 1%, 2%, 5%, 10%).

In other words, the effect of the index on whether the customer's credit indeed defaulted can be regarded equal to zero (in comparison to the effect of the index of whether the customer does not have credit in default) on the customer's decision to buy the product. As a result, the clients that have credit in default are equally likely to buy the product as those clients that do not.

The consumer price index does not have a significant effect on the model since it has a p-value > a, for all reasonable values of 'a'. Consequently, the effect of the consumer price index can be regarded equal to 0 on the customer's decision to finally buy the product. In fact, the variable can be dropped from the model since it is not a significant factor of the customers' final decision. That statement is confirmed since the coefficients of the rest of the model's explanatory variables essentially remained unchanged (some of them had small changes).

The index of the communication type has a significant effect on the model since it has a p-value < a = 5%. Finally, the number of number of contacts performed during this campaign for each client has a significant effect to their decision in buying the product since it has a p-value < 5%.

The final model can be described from the following equation: $logit\ SUBSCRIBED\ = 79.45 + 0.05duration - 0.016nr.employed + 0.637monthaug + 0.19monthdec + 0.494monthjul + 0.6monthjun + 1.26monthmar - 0.73monthmay + 0.04monthnov + 0.384monthoct - 0.18monthsep + 0.45poutcomenonexistent + 1.78poutcomesuccess - 0.37defaultunknown - 7.45defaultyes - 0.23contacttelephone - 0.04campaign,\ Yi|xi \sim Bernoulli(pi = P(SUB\ = 1))$.

In the following step, each variable's effect to the probability of a successful sale will be interpreted. According to the intercept, if the last call's duration was equal to 0 seconds and took place in April, the selling attempt failed in the previous campaign, the client does not have credit in default, the consumer price index is equal to 0, the last contact was conducted through the client's cellular and the number of contacts performed during this campaign and for this client is equal to 0 then the probability that the customer buys the product is equal to $pi = \frac{e^{\beta o+\beta 1\chi 1i+\cdots+\beta v\chi vi}}{1+\left(e^{\beta o+\beta 1\chi 1i+\cdots+\beta v\chi vi}\right)} = \frac{1}{1+e^{79.91}} \approx 0$. In case the agency called two different customers that have identical and unchanged variables but have a 1 second difference in the duration of their last call with the agency, the one with the greater duration has a $e^{0.005} \approx 1$-times greater relative odds to finally buy the product. As a result, calls of greater duration seem to have a positive effect on the customers decision to buy the product. Specifically, if the quarterly indicator increases by one unit, the relative odds of a client buying the product will drop to 0.98, which means that the index has a negative effect in the customer's final decision.

If the agency has contacted two different clients that have identical and unchanged variables, but one was contacted last in April and the other in August, the latter has almost $e^{0.64} \approx 1.9$-times greater relative odds to buy the product in comparison to the first one. If the latter client was contacted last in July, his relative odds would be greater by 1.6 times instead. Identically, if he was last contacted in June, his relative odds would be greater by 1.8 times instead. Following, if he was last contacted in March, his relative odds to buy the product would be greater by 3.5 times instead. If he was last contacted in May, his relative odds would drop to 0.5. Finally, if he was last contacted in October, his relative odds would be greater by 1.5 times instead. In other words, clients that were last contacted during December, November or September are equally likely to buy the product as those that were last contacted in April. Additionally, clients that were last contacted in May are less likely to buy the product, while those that were contacted in July, June, March, May or October are more likely to buy the product in comparison to April.

For two customers that were contacted with the agency and had identical and unchanged variables, but one refused to buy the product in the previous campaign and the other was not contacted at all, the latter has greater relative odds to buy the product by almost 1.6 times. On the other hand, if the latter bought the product in a previous campaign, his relative odds to buy the product would be greater by almost 5.9 times. As a result, it can be deduced that clients that were not contacted in previous campaigns are more likely to buy the product than those that were contacted and refused. Additionally, clients that were contacted in a previous campaign and bought the product, are the most likely to buy the product again in comparison to the rest.

If the agency contacted two customers with identical and unchanged variables, but one has no defaulted credit, and the other one is unknown whether he has credit defaulted, the latter one has decreased relative odds to buy the product equal to 0.69 times. Hence, clients for which it is unknown whether they have credit in default the probability that they buy the product is reduced in comparison to the rest cases.

If the agency contacted one customer to their cellular, and the other to their telephone, and both have all the rest of the variables identical and unchanged, then the latter has decreased relative odds to buy the product equal to 0.8 times. So, it can be deduced that it is preferable for the agency to call the clients to their cellular instead of their telephone. If the agency has contacted two different customers, and one of them was contacted one time more than the other, his relative odds to buy the product dropped to 0.96. In other words, the additional calls to the clients have a negative effect for their decision to buy the product.

5 - Goodness of Fit Tests

Finally, the model's goodness of fit will be determined. Starting, it will be determined whether the final model fits significantly better the data in comparison to the constant model. Since p-value = 0 < 5%, the null hypothesis that the constant and the final model fit the data equivalently is rejected, and thus the final model fits the data significantly better.

Following, it will be determined whether the final model differs significantly from the 'saturated' model. The saturated model is that which maximizes the loglikelihood assuming one parameter per datum. In other words, it is the best possible model. Since p-value = 1 > 5%, the null hypothesis that the final model fits well cannot be rejected, thus, it is inferred that the model is a good fit for the data.

Next, the model's deviance will be evaluated. Models that have deviance values that are close to one (1) are considered as models that fit the data well. It can be observed that the final model's deviance is equal to 15610/39831 = 0.4. Thus, the model's deviance value is not acceptable, and the model cannot be considered a good fit for the data according to it.

The model's residuals will also be evaluated. Starting, the sum of the model's squared residuals should be approximately close to n – p (39.883 – 7 = 39.876). It is observed that the value is not close (it is equal to 15.978) to the desired one, and so, it can be inferred according to this evaluation that the model is not a good fit for the data. The previous statement about the model's fit can be backed up by observing the residuals' plots (Tables 3 & 4) since they do not have the desired form. In detail, they residuals do not seem to have a pattern which is desirable, but they should not touch the horizontal line where y = 0, and in most plots they do.

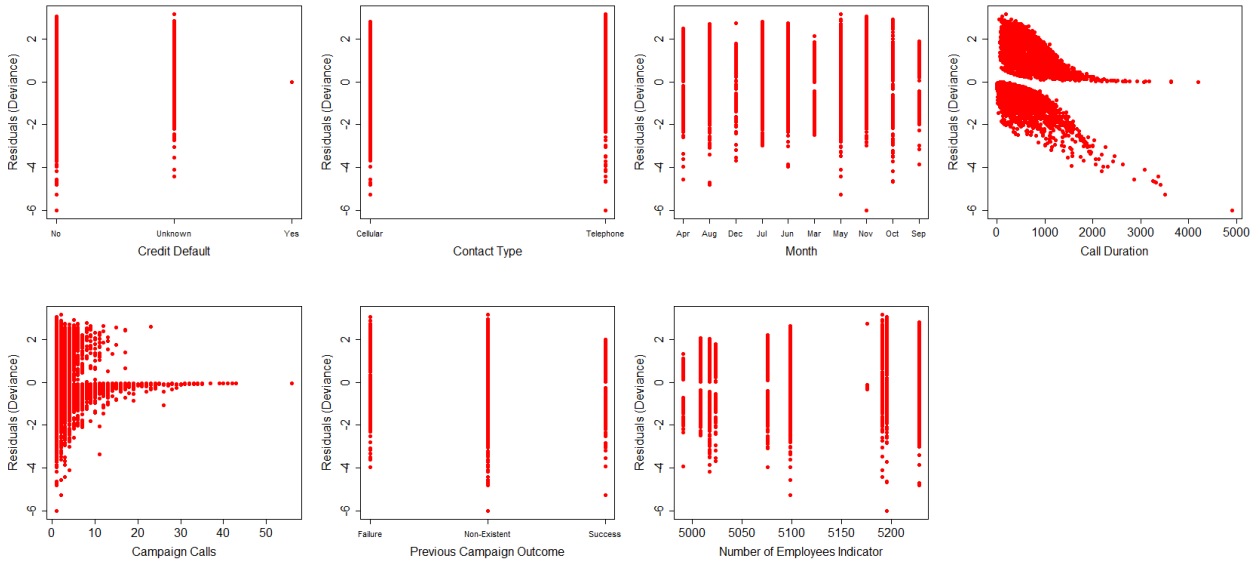

Table 3. Deviance Residual Plots

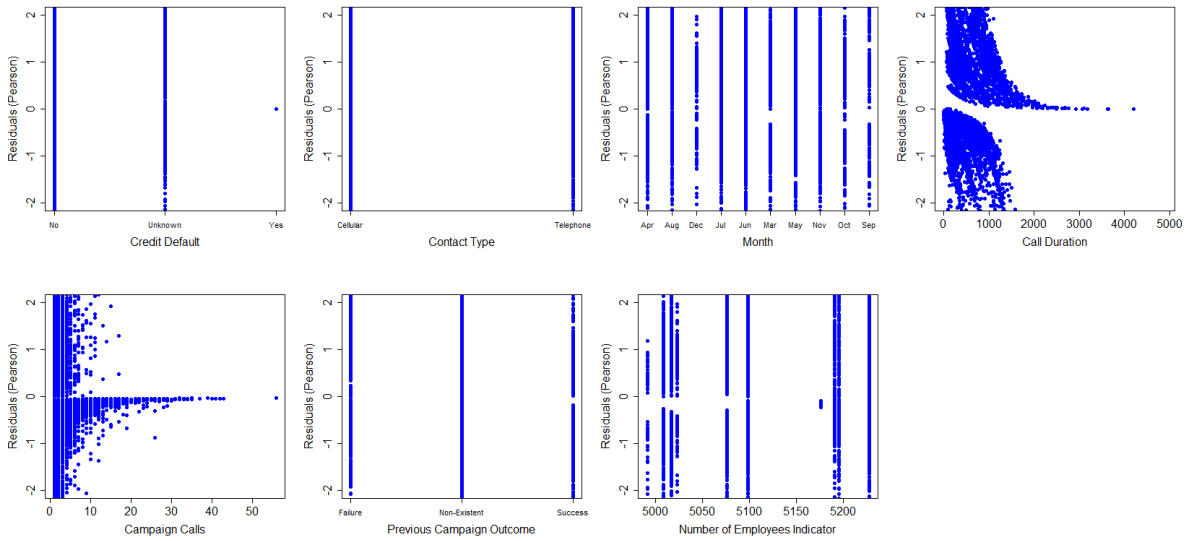

Table 4. Pearson Residual Plots

6 - Conclusions

Summarizing, it can be inferred that the bank's customers seem to be more receptive to buy the bank's product when their calls are of long duration and when they are contacted in July, June, March, May or October. At the same time, clients that are contacted during December, November or September seem to be less receptive to buy the product. Additionally, clients that have either bought the product in a previous campaign, or those for whom it is unknown whether they did buy the product in a previous campaign, have greater odds to buy the product than those who rejected past proposals. Clients for whom it is unknown whether they have credit in default are less likely to buy the bank's product in comparison to those that have credit in default and to those that don't. Also, clients that are contacted to their cellular instead of their telephone seem to be more receptive to buy the product. Furthermore, the increases of the quarterly indicator of the number of employees seem to have a negative effect on the clients' decision to buy the bank's product. Concluding, clients that are contacted more frequently seem to be less likely to buy the product in comparison to those that are contacted less frequently.

To produce the final model, a cross-validation lasso and a stepwise process were performed to an initially full model. The model produced from the stepwise process had a multi-collinearity issue which was resolved by removing an explanatory variable (employment variation rate) from the model. Afterwards, it was found that the cons.price.idx variable (consumer price index), was not affecting the response of the model significantly, so it was also removed from the model. The model seemed to be significantly better than the constant model, while at the same time it did not differ significantly from that of the saturated model. As a result, according to those tests, the model can be considered a good fit for the data. Although,

according to the model's deviance and residuals (both the Pearson and the Deviance residuals), the model cannot be considered a good fit for the data. Concluding, it should be noted that with further research and analysis on the final model and the sample, that it is possible that a model can be constructed that satisfies all the goodness of fit tests that were conducted in this analysis.