Nova School of Business and Economics

Case 01

Author:
Benjamin C. Herbert (45775)
Carolina Domingues (31951)
David Neves (31927)
Rian Yan (28823)
Sebastian Rupp (46093)
Thomas Dornigg (41727)

Instructor:
Prof. Carlos Daniel SANTOS

An assignment submitted for the course
Big Data Analysis

Second Semester, T4
2021

April 30, 2021

A) Data Project Plan

1. In the past decades, the processing of data and the informed decision-making through its analysis has become indisputable. At the same time, the amount of collected and accessible data has continuously increased, and the opportunities for companies to leverage this data to drive smarter decisions is growing at a high pace [1].

In terms of value creation, this report will focus on two potential benefits regarding the analysis of the present dataset: 1) Understanding demand: the "big" amount of data allows the company to gain knowledge on customers' preferences, behaviour and features, and move from a guessing framework to targeted marketing and product strategies to customers in micro-segments [2] and thereby benefiting from growth in sales and increase in ROI [3]. 2) Adjusting supply: firstly, the company can better customize and tailor its services and product offerings to customers from different segments. Secondly, the company can improve the effectiveness of marketing and communication, being more cost efficient in terms of advertising, which will lead to a reduction in overall expenses.

Thus, the aim of our analysis is to promote data driven decision making and gain a competitive advantage in the telecommunications market by 1) constructing a predictive model that determines the customer profile that is more likely to respond positively to product offerings, 2) identify which customer features have greater influence in the response to offers.

2. To increase marketing efficiency and ROI, it is necessary to define the targeting segmentation of the three preliminary product offers and promote them to different corresponding targeting segmentations with highest sales potential as possible. We assume that the telephone company performed a massive product test in the market. As a result, a dataset that provides the information of 4000 observations randomly-selected from the overall 5000 company's customer base, having 132 variables in total, within which the sales result of 3 product offers are included.

Considering that the variable size is remarkably large, and not all of the variables are useful to define a targeting segmentation due to the variable attributes, 41 variables are conceptually selected through a marketing and business orientated approach to build a more accurate predictive model in the further process. The variable "Customer ID" (*custid*) is selected as an **identifier variable**. And variables *response_01*, *response_02* and *response_03* are defined as **outcome variables**. Among the rest of the 37 **explanatory variables**, the logarithmic configuration version of certain variables are selected in order to facilitate the data analysis process and dilute the effect from value dispersion. Nonetheless, when it comes to the case scenario where a certain category presents the dummy configuration, the scale dimension of the variable is selected to access more information and avoid the missing value brought by the logarithmic configuration. As a result, according to the marketing segmentation process, those 37 explanatory variables selected are divided into three groups, namely: **Demographic Features**, **Geographic Features** and **Behavioral Features**.

To start with, variables under **Geographical Feature** group potential illustrate the sales potential of specific geographical areas, which includes *region* and *own size*. Then, variables from

Demographic Feature group demonstrates the basic customer profiles about who exactly the telephone company should target and indicate the purchase power through level of disposable income as a result of income and debt. Those variables representing basic customer profiles are *gender*, *age cat*, *ed cat*, *job cat*, *marital*, *spouse*, *spousedcat*, *empcat* and *retire.* And variables that illustrate purchasing powers through level of disposable income are *lninc*, *debtinc*, *inccat*, *lncreddebt*, *lnothdebt* and *carcatvalue.*

As for the last group, **Behavioral Feature**, three sub-groups can be defined as **Purchase Occasions**, **Benefit Sought** and **User Status**. Variables in the **Purchase Occasion** sub-group indicate the case scenario when and where the customers will use the product offered by the telephone company, which includes *telecommute*, *cardmon*, *wiremon*, *multline*, *voice*, *internet*, *callwait*, *forward* and *confer.* As for the sub-group **Benefit Sought**, purchase motivations and purchase potentials are expected to be directly or indirectly indicated through variables including: *reason*, *cardspent*, *churn*, *longmon*, *tollmon*, *equipmon*, *ownpc*, and *ebill.* Lastly, sub-group **User Status** includes variables that demonstrate the conditions or limitations shown by potential targets, which are *reside*, *hometown*, *hometype* and *addresscat.*

After the conceptual selection process of explanatory variables, several preliminary data processing procedures including **data organization**, **data transformation** and **data exploration** should be conducted so that a quality dataset can be obtained ready for further investigation and modelling analysis, and a general preliminary understanding on the dataset can be acquired as well.

As for data organization, taken into account that the dataset is already well structured and labeled with correct measurement, logarithm configuration applied to certain variables and multiple categorized variables created, first and foremost, data validation will be performed in order to deal with potential duplicated or incomplete identifiers of the variable *custid*, and also evaluate the amount of missing values among explanatory variables. In case there exists a significant amount of missing values, a **MACR test** will be performed to understand the reasons of missing values, whether they are missing completely at random or if there exists any relationship among the very same. Consequently, missing values with a threshold larger than 10% should be eliminated either listwise or pairwise.

After performing data organization, data transformation will be implemented through data reduction. Data reduction strategy includes dimensional reduction and numerosity reduction. The predefined conceptual variable selection process can be considered as a **dimensional reduction** process. And a **numerosity reduction** process will be implemented based on the result of missing data analysis as is mentioned previously and the result of outlier analysis which will be determined through data visualization processes.

Next, in order to explore the dataset and report the main features and characteristics, data exploration will be implemented mainly through summative statistics and data visualization so that key patterns among customer information can be identified and potential insights for defining targeting segmentation purposes can be generated as well. To start with, univariate profiling will be performed through several tools, including: summative statistics, descriptive statistics, frequency table and histogram in order to check the general characteristics of the variable, define the sample distribution and test for normality. Meanwhile, outliers should be detected based on the result of data visualization through assessing the value extremely deviant from the sample distribution. As a result, the outliers attribute should be defined and bad outliers, mistakes during data collection, should be discarded as part of the numerosity data reduction process. Next, bivariate analysis shall be performed in order to assess the links between two possibly connected variables, in which several tools will be adopted, such as: Scatterplots and Pearson correlation (two metric variables), Cross-tabulation (two nominal variables), Spearman correlation (ordinal variables), and Boxplots (metric non-metric variables), such that the relationship between two variables and be visualized and quantified.

---

3. In order to predict if a customer will likely buy a subscription, the group thinks that the best option to model such outcome is to use supervised machine learning techniques. Generally, supervised machine learning, which is a subcategory of machine learning and artificial intelligence, can be categorized into two types of problems:

- **Regression:** This technique is used when the goal is to predict continuous outcomes. For instance, by using regression methods such as a simple linear regression, it is possible to predict the sales growth for the telephone company.
- **Classification:** For predicting binary classes, e.g. where (1) means buying a subscription and (0) means not buying a subscription, classification algorithms such as Logistic Regression, Naive Bayes, Decision Trees or Random Forests will be a better choice than compared with the regression approach.

As mentioned already, one can distinguish between multiple classification models. Most widely used in the industry, among othes, are the following:

a. A **Logistic Regression** models the probabilities for classification problems with a binary outcome. It can be regarded as an extension of the linear regression model for classification problems. Instead of fitting a straight line or hyperplane, the logistic regression model uses the logistic function to squeeze the output of a linear equation between 0 and 1 [4]. The logistic sigmoid function produces class membership probabilities for the target class (y=1) given features x.

b. The **Naïve Bayes Classifier** is a classification approach that adopts the principle of class conditional independence from the Bayes Theorem. This means that the presence of one feature does not impact the presence of another in the probability of a given outcome, and each predictor has an equal effect on that result [5].

c. The **Decision Tree** is a supervised learning algorithm, which can handle both classifications and regressions, thus it can be used to predict a client's probability of subscribing to a given service. One of the decision tree's advantages is that it can be easily used on a dataset consisting of numerical and categorical values.

d. A **Random Forest** is an ensemble learning method, which is basically a group of decision trees trained on different subsets of a given dataset, where a majority or soft majority vote decides the classification of the target. It not only has the same advantage of a decision tree,

> but also has a higher accuracy as a single tree. However, the main limitation of a random forest is that a large number of trees can make the algorithm too slow and ineffective for real-time predictions [6].

To properly evaluate the performance of these algorithms, it is important to split the dataset at the beginning of the modelling into a training and test set, perform cross-validation and then compute evaluation-metrics such as confusion matrix, AUC-ROC and Precision-Recall curves. Thus, it is not possible to already make an informed decision on the best classifier to use, right from the beginning. All mentioned models must be trained and evaluated individually, and in the end compared by standard-metrics, as explained above.

In addition to that, the group also looked into the other models which were discussed in class and has strong arguments not to choose any of them. Firstly, the Look Up Model would be too complex given the array of variables and observations. Secondly, the RFM model cannot be built around this data since Recency, Frequency and Monetary value of purchases are out of the scope of this analysis. Thirdly, the Multiple Linear Regression Model is not used for classification problems by definition, thus it cannot be applied here.

4. Given the previously described tasks, the fundamental technical requirements to run the data project are divided into two categories: hardware and software. For the hardware, the key elements are storage, processing, memory, and relevant networks for parallel processing. Usually, the traditional hard drive for data storage is based on a hard disk, which is very precise but relatively slow. As an alternative, disk arrays with speeds of 1GB/sec and SSD (solid-state drive) will be faster than hard disk and more costly. Besides that, good performance of CPU and Memory (RAM) is essential for our analysis process. In terms of software tools, the main focus will be on SPSS tools.

Regarding Human resources, one would need to recruit data scientists and data analysts with a background in business, capable of transforming data insights into knowledge and value-added to the business that can be integrated into daily decision making. All in all, having a good understanding of statistics and business is crucial for a good analysis and mastering SPSS tools, which could require further training.

B) Reporting of results

At this stage, the tasks defined in A.2 were performed using SPSS tools and Python. In terms of data validation, all variables have passed the requested basic checks except for the variable "Primary reason for being a customer here" (*reason*) that reported more than 70% of the observations as missing. Handling missing data is an important part of any data analysis process. If not handled properly, part of the analysis can be biased and may provide misleading results – thus, we proceed to checking for missing values and testing the hypothesis that they are missing at random. As a result, we observed that "Primary reason for being a customer here" (*reason*) had actually 81,5% of missing values (3258 in total), and two other variables also raised a concern:

"Spouse level of education" (*spousedcat*) with 51% of the observations missing (2040 in total), and "Primary vehicle price category" (*carcatvalue*) with 9,9% missing values (395 in total). The remaining variables reported had only minor problems of missing values (<1%). Analysing the cross-tabulations results, we observe that - for the three problematic variables - their missing data across all categories varied little around their respective percentage of total missing data, hence we fail to reject MCAR. To deal with the random missing data, we opted for a pairwise deletion of the observations that contain missing data in *carcatvalue* and decided to drop the columns *spousedcat* and *reason* from our dataset since no statistically significant information could be inferred given their nature and great percentage of missing values. Additionally, we also checked for duplicates in the dataset but no duplicates were found.

As part of preprocessing and data exploration, the group started by running a descriptive statistics analysis. This initial examination of the data is a good practice that allows one to get a sense of the data's overall distribution. In this case, the descriptives chosen were the mean, the standard deviation, the minimum and maximum, and the 25%, 50%, and 75% quartiles. These descriptive statistics were calculated for the numeric scale variables (among the set of explanatory variables chosen in part 2A), and for the three outcome variables *response_1*, *response_2*, and *response_3*. For the latter, it makes the most sense to look at their mean, which represents the proportion of people accepting each kind of product. For product 1, 2, and 3, respectively, 8%, 13%, and 10% of respondents accepted the offer that was being sold to them. Looking at the first 7 variables featured in the analysis, it's possible to assess their skewness (to a certain extent) by comparing their mean to their 50% quartile (median). Most reflect little to no skewness as their means and medians are practically the same, except for *deptinc* and *cardspent*, that may be skewed to the left (mean > median). Additionally, specifically about the variable *reside*, even though the comparison between the median and the mean would not entail a skewed set of observations, the maximum value for this variable seems unusually high when compared to the mean and standard deviation. This might be a sign that the variable has one or more outliers. Then for *longmon*, *tollmon*, *cardmon*, and *wiremon*, it may be relevant to look at the quartiles to roughly understand the amount of participants using certain telephone services during the last month. For instance, from *tollmon* and *wiremon* one can say that at least 50% of indivuduals did not enjoy a toll free service or wireless service last month, respectively. "cardmon" shows that at least 25% of participants did not use a calling card last month. Contrarily, all observations made a long distance call/connection last month.

	lninc	debtinc	lncreddebt	lnothdebt	reside	cardspent	tenure	longmon	tollmon	cardmon	wiremon	response_01	response_02	response_03
count	4,000.00	4,000.00	3,999.00	3,999.00	4,000.00	4,000.00	4,000.00	4,000.00	4,000.00	4,000.00	4,000.00	4,000.00	4,000.00	4,000.00
mean	3.70	9.95	-0.12	0.70	2.23	340.64	38.03	13.51	13.41	15.20	10.86	0.08	0.13	0.10
std	0.76	6.39	1.27	1.13	1.41	252.42	22.70	13.13	16.43	14.89	19.84	0.27	0.34	0.30
min	2.20	0.00	-5.68	-4.09	1.00	0.00	0.00	0.90	0.00	0.00	0.00	0.00	0.00	0.00
25%	3.18	5.20	-0.95	-0.02	1.00	184.24	18.00	5.70	0.00	0.00	0.00	0.00	0.00	0.00
50%	3.64	8.80	-0.08	0.74	2.00	278.77	38.00	9.45	0.00	13.50	0.00	0.00	0.00	0.00
75%	4.20	13.50	0.72	1.48	3.00	423.54	59.00	16.30	24.75	22.50	21.96	0.00	0.00	0.00
max	6.98	43.10	4.69	4.95	9.00	3,926.41	72.00	179.85	173.00	188.50	186.25	1.00	1.00	1.00

Table 1: Descriptive statistic's table of numeric features

Identification and removing of potential outliers, is another necessary task to perform. Outliers can indicate bad data due to incorrect coding or misunderstanding of a survey. If an observation is determined to be erroneous, this outlier need to be removed. To quantify outliers in our dataset, we followed the approach suggested by Iglewicz and Hoaglin [7].

$$Q_1 = q_{0.25}(X_i); \quad Q_2 = q_{0.75}(X_i); \quad IQR = Q_3 - Q_1$$
$$Outlier_{lower}(X_i) \; = x_i < (Q_1 - 3 * IQR)$$
$$Outlier_{upper}(X_i) \; = x_i > (Q_3 - 3 * IQR)$$

Table 2 provides an overview of the amount of outliers statistically detected for each variable. After manually checking whether the detected values are outliers, we deleted the 190 rows in which one or more occurred.

	lninc	debtinc	lncreddebt	lnothdebt	reside	cardspent	tenure	longmon	tollmon	cardmon	wiremon
count	0	4	0	0	0	53	0	105	2	10	23

Table 2: Count of outlier in numeric variables excluding responses

Apart from missing value analysis, a correlation heatmap has been plotted to establish an understanding of how selected features interact with each other. Looking at the heatmap, the darker the blue of the spots, the higher the correlation between the two variables. As one can see, the highest correlation exists between the features “Log-credit card debt” and “Log-Other debt”. Considering it from a business point of view, this makes sense because if someone is taking on credit debt, one is also more likely to take on other debt. Depending on which model is applied, one might consider removing one of both variables because the high correlation introduces collinearity in our model. Moreover, there seems to be also a relationship between income and debt in general, which means the more people earn, the higher their debt levels. This also seems reasonable because the higher the income, the higher the credit you might obtain from your bank. Other variables which are highly correlated are *long-distance* and *tenure* as well as *toll-free last month* and *wireless last month.*

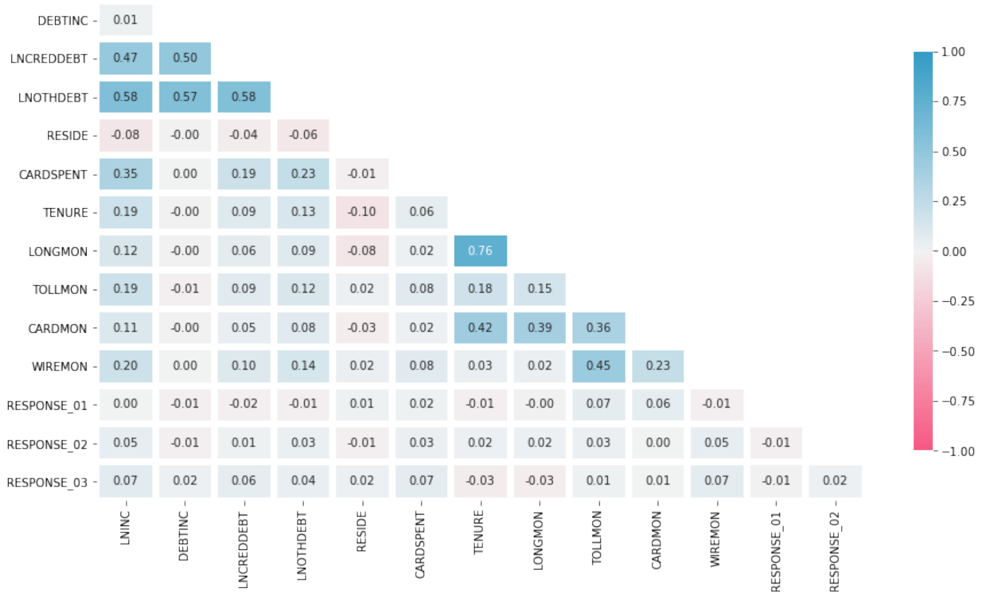

Figure 1: Correlogram

References

[1] Anuj Tripathi et al. “Big Data-Driven Marketing enabled Business Performance: A Conceptual Framework of Information, Strategy and Customer Lifetime Value”. In: *2021 11th International Conference on Cloud Computing, Data Science & Engineering (Confluence)*. IEEE. 2021, pp. 315–320.

[2] Alexandre Borba Salvador and Ana Akemi Ikeda. “Big data usage in the marketing information system”. In: *Journal of Data Analysis and Information Processing* 2014 (2014).

[3] Ziqi ZHONG, ZHOU Sanping, and LI Yuzhen. “Research on the Precise Marketing Method of Goods Based on Big Data Technology”. In: *2021 2nd International Conference on E-Commerce and Internet Technology (ECIT)*. IEEE. 2021, pp. 364–367.

[4] Christoph Molnar. “Interpretable Machine Learning - A Guide for Making Black Box Models Explainable”. In: *Leanpub* (Apr. 2021).

[5] IBM Cloud Education. *Supervised Learning*. URL: https://www.ibm.com/cloud/learn/supervised-learning. (accessed: 21.04.2021).

[6] Niklas Donges. *A complete guide to the Random Forest algorithm*. June 2019. URL: https://builtin.com/data-science/random-forest-algorithm. (accessed: 23.04.2021).

[7] Boris Iglewicz; David Hoaglin. “How to Detect and Handle Outliers”. In: *The ASQC Basic References in Quality Control: Statistical Techniques* (1993).