An Exploratory Analysis of Customer Profiles for Telecommunications Marketing

25551995: Benedict E. Brunker

School of Transdisciplinary Innovation

36103: Statistical Thinking for Data Science

1. An Exploratory Analysis of Customer Profiles for Telecommunications Marketing

Telecommunications firms have an obvious interest in finding customers to sign up for their products. Since a mobile-phone plan is usually an ongoing subscription service, there is likely a relatively small market of potential customers. We hypothesize that the market is limited to three main groups:

1. Do not currently have a phone plan and are looking for one.
2. Are actively looking for a better deal than their current plan.
3. Are open to hearing about a better deal than their current plan.

Besides the price of the plan on offer, there is a certain administrative burden involved in switching from one plan to another. For those in groups 1 and 2, remaining on their current plan can be conceived as "the path of least resistance". Some approaches to this problem might be to improve techniques of persuasion, improve the quality-to-price ratio of the product relative to competitors, or reach a wider audience through broad (non-targeted) marketing campaigns. This report focuses on a different approach, namely, to analyze an existing dataset and find patterns relating information about customers (predictor variables) with information about whether or not they subscribed to the plan (our response variable). If it is possible to discover consistent patterns relating predictors to response, telecommunications firms will be able to target their advertising campaigns to those more likely to subscribe to a plan. Such an approach could make advertising campaigns far more cost-efficient, while avoiding unwanted outreach-efforts to segments of the population unlikely to subscribe to a plan (and therefore more likely to feel annoyed by unwanted advertising).

2. Data Preprocessing

2.1 Unknown Values

Our dataset contains unknown values for *job, marital, education, default, housing,* and *loan,* displayed in *Table N.* These are all categorical data types. There are broadly four main strategies for dealing with unknown values:

1. Remove the whole row with the unknown value.
2. Remove the whole column containing unknown values.
3. Impute a new value to the unknown value.
4. Treat *unknown* as a categorical variable in its own right.

The selection of an appropriate strategy depends on:

a. The number of unknowns in the row or column.
b. A judgment regarding the randomness or non-randomness of unknowns (Hou *et al.*: 2022).

Table 1: Unknown Values

Variable	*No. of Unknowns*	*~% of total*
Job	330	0.8
Marital	80	0.2
Education	1731	4.2
Default	8596	20.9
Housing	990	2.4
Loan	990	2.4

In cases where unknowns constitute only a very small percentage of the dataset, and no clear relationships are exhibited between the frequency of unknowns, we can safely judge these unknowns to be Missing Completely At Random (MCAR), meaning that "the propensity for a data point to be missing is completely random" (Hou *et al.,* 2002, p.4). We judge *job* and *marital* to fall into this category, and

simply drop these rows from our dataset, reducing the dataset from 41180 to 40779 rows (a loss of 401 rows).

Housing refers to whether or not the client has a home-loan, while *loan* refers to whether or not the client has a personal loan. The high degree of correspondence between unknowns for *housing* and *loan* (cf. Table N) suggests these values are Missing At Random (MAR), meaning that unknowns in these two columns are related to each other (Hou *et al.*: 2022). We speculate that those unwilling to provide information on their home-loan would also be unwilling to provide information on personal loans. Because of the high degree of correspondence between unknowns in the two columns, we were able to drop these rows from the dataset without losing too much of our overall data.

For unknowns in *education,* we experimented with transforming the values to ordinal ranks (see §2.2.2 below on *Ordering Nominal Variables*), and with imputing the mean of all ranks to unknowns. We decided against this, since we couldn't identify any patterns based on this imputation, and couldn't be sure that unknowns were Completely Random, and therefore likely to correspond with the mean distribution. This was confirmed by looking at the rows containing *unknown* for these two columns side-by-side. We therefore simply dropped unknowns for this variable.

The variable *default* refers to whether the client has credit in default. This data is likely taken from survey, not corroborated by any third-party financial information. This would explain the high degree of unknowns for this variable: far exceeding any other at ~21% of the column, in comparison to only 3 *yes* responses. It seems likely that survey respondents may be unwilling to disclose financial information of this sort if they feel it may impact their credit rating. We therefore judge *unknowns* as missing not at random (MNAR); there is some non-random reason for the propensity of unknowns (Hou *et al.*, 2002). We have decided to keep these unknowns in the dataset for the time being as values in their own right, in order to see if there is any relationship between a client's unwillingness to disclose this information and any other variables. The other reason for this decision is that dropping these rows would eliminate fully one-fifth of our whole dataset.

2.2 Data Transformations

Sometimes better insights can be gleaned from data by transforming data from one type to another. We have done so for the cases below, with a brief explanation:

2.2.1 Binarization

Binary data are those which can take only one of two possible values. Our response variable, *subscribed,* is such an example, taking only the values *yes* or *no*. In order to be able to better compute this data mathematically, we have created a new column called *subscribed_binary,* transforming *yes* to *1,* and *no* to *0.* When visualizing this data, however, we use the original *subscribed* column.

2.2.2 *Ordering Nominal Variables*

Fig. 1

```
'illiterate': 0
'basic.4y': 1
'basic.6y': 2
'basic.9y': 3
'high.school': 4
'professional.course': 5
'university.degree': 6
```

We call a variable *nominal* if it is a classifier –a word representing some category – and it is arbitrary with respect to order. An example from our dataset is the variable *contact,* which takes the values 'telephone' and 'cellular'. Sometimes a variable presents as nominal in a dataset, but is really non-arbitrary with respect to order. An example is *education*, which refers to the client's highest level of educational attainment. But educational attainment is ordinal: we naturally progress through stages of education (that is, someone who has completed a University Degree has almost certainly completed High School, *et cetera*). We have tried to capture this ordinal character by transforming the raw text-strings in *education* into ranks, from 0 to 6 (cf. fig N.N.). 4.2% of the values in this column are *unknown.* We initially thought to replace all *unknown* values with the mean of all ranks for education. However, we decided against this, since we of course don't know how well this data really is distributed around the mean, and since stipulating these values as mean didn't reveal any patterns in the education data.

3. Exploratory Data Analysis

Below, we analyze our response variable and all predictor variables. Some variables have deserved more attention than others. We note some relationships between predictors and response, and between predictors.

3.1 Response

This is our response variable, which has two values: *Y* to signify that the respondent has subscribed to a plan (the campaign has succeeded for this customer); *N* to signify the opposite. Fortunately there are no missing data in this column, and no clear errors. All values are either *Y* or *N.* The data show that 88.74% of respondents did not subscribe, as against 11.36% who did. A success rate of 11.36% for a marketing campaign of this kind appears facially plausible. Now the task is to discover patterns in the profiles of this ~11% of the sample.

3.2 Demographic Indicators

A number of our variables provide information about the demographic profile of the client. These indicators are nominal variables, and include: *age, job, marital status* and *education.* We have collated these as *demographics.* We briefly discuss each in turn.

3.2.1 *Age*

The ages of respondents surveyed range from 17 to 98, with a mean of ~40, standard deviation of ~10, and median of 38. By comparison, the mean and median age in the EU are both 44.5 (*Eurostat,* 2024). Since our sample (naturally) excludes those aged 0 to 16, a representative sample should see a mean and median higher than that of the EU population. This may suggest (provisionally) that the campaigns have failed to reach older age-groups. The precise distribution of ages is displayed in *figure 3.1.*

Figure 3 shows the age distribution coloured by the ratio of subscribers to non-subscribers, with darker blues representing higher ratios. The darker blues grouped around the tails of the distribution could suggest an untapped market for these older age groups in particular, but more data for these groups

would be needed to confirm this conjecture, since a higher ratio could easily be an effect of a smaller count. We do note in *fig.3* the very steep cliff that seems to separate over 60s from other age groups, and suggest that future campaigns should aim for a more normal distribution of age.

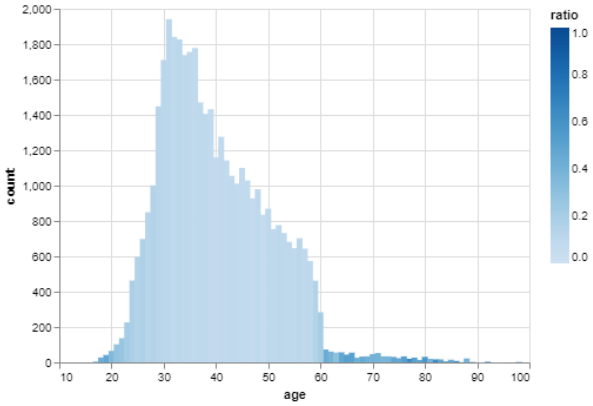


Fig 3: Success Rate by Age

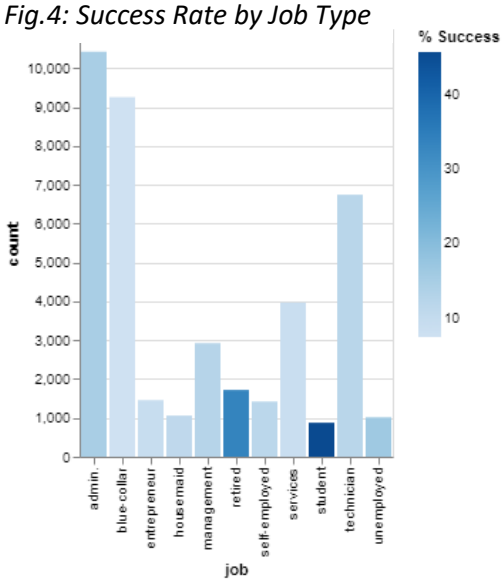
It is possible that the young (the 5th percentile) and the old (95th percentile) are less likely to have a satisfactory phone plan, are more open to advertising persuasion, or both. Conversely, it is possible that those aged 32 (Quartile 1) to 47 (Quartile 3) are more likely to have already found a satisfactory plan, are less open to persuasion, or both. Furthermore, average age in the European market is increasing (*Eurostat,* 2024). **However,** we stress again that confidence intervals are much smaller at the tails of the distribution, meaning than very small changes in response lead to large changes in correlation.

3.2.2 Job

The job variable is nominal and has twelve possible values: *admin*., *services, blue-collar, technician, housemaid, management, retired, self-employed, services, student, technician* and *unemployed.* Our dataset does not contain a variable for income or net wealth, which might plausibly be useful in predicting response. However, job type might be used as a proxy for income, since we might estimate an income for each job type based on the median income for that job in the European Union.

If job-type is a proxy for income, there is no reason as of now to believe that clients in more lucrative professions are more likely to subscribe. Figure 4 below displays how many clients of each job-type appear in our data, with the bars coloured to represent the proportion of subscribers.

Fig.4: Success Rate by Job Type



What these ratios seem to suggest is that those more likely to be currently unengaged in full-time work are more likely to subscribe to a plan. We would speculate that this is because subscribers are more likely to be those who do not currently have a phone plan (perhaps using a Pay-As-You-Go model), or are looking for a better deal on their current plan. Note that the lowest ratios occur for those in jobs which most likely require daily use of a mobile phone, like blue-collar workers, service workers and entrepreneurs. This is one (small and provisional) indication that an advertising strategy for increasing plan subscription should **not** focus on high-income earners, but on those less likely to have a good-value plan already. To confirm this however would require the construction of a good predictive model.

3.2.3 Education[1]

Like *job*, *education* is also a plausible proxy for income in our dataset (*Eurostat,* 2021). And as with *job*, there is no clear pattern signifying that those with a higher level of educational attainment (and therefore a higher probable income) have a higher propensity to subscribe. Note that Table 3 below might seem to suggest that illiterates are more likely to subscribe, but this is misleading given the small sample size of illiterates. That is, if only one of the four clients in the illiterate sample had failed to subscribe, the ratio would drop to $1.\bar{6}$.

Table 3: Success Rate by Education

Educational Attainment	No	Yes	Success (%)
Illiterate	14	4	28.57
University Degree	10497	1669	15.9
Professional Course	4647	594	12.78
High School	8482	1031	12.16
4 Years	3747	428	11.42
6 Years	2104	188	8.94
9 Years	5571	473	8.49

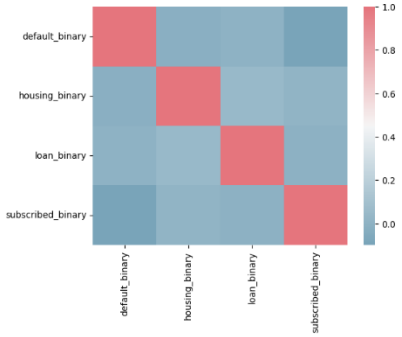
3.2.4 Marital

The dataset showed a success rate of ~14% for singles, as against ~10% for both divorced and married clients. This difference may well be insignificant, but it does seem to correspond to other demographic data showing a better response among students and young people.

3.3 Financial Profiles

Three of our predictors provide direct information about the financial profile of the client: *default* (whether or not the client has credit in default), *housing* (whether client has a home loan), and *loan* (whether client has a personal loan. We discuss issues with *default* data in §2.1 above. After transforming these data into binary numbers, our exploratory analysis did not reveal any noteworthy patterns within this set of variables or with the response variable, except the seeming correspondence between unknowns discussed in §2.1 above.

Figure 5: Correlations between financial indicators and subscription

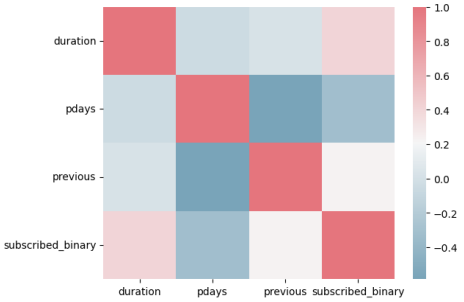


3.4 Nature of Contact

Eight of our predictors provide information about how and when the client was contacted by marketers. *Contact* tells us how the client was contacted: either by 'telephone' or 'cellular'; *duration* refers to the duration of the call; *campaign* records the number of contacts performed during this campaign for this client; *pdays* is the number of days since client was last contacted from a previous campaign, with -1 indicating client was not previously contacted; *previous* is the number of contacts performed before this campaign and for this client; and *poutcome* is the outcome of the previous marketing campaign, with values 'nonexistent', 'failure' and 'success'. We transformed *poutcome* to ternary data as *poutcome_ternary* to perform our analysis, with -1 for 'failure', 0 for 'nonexistent', and 1 for 'success'.

[1] Cf. section N.N above for a summary of data transformation performed for this variable

Figure 6: Contact Correlations



We took *duration, pdays, previous* and *poutcome_ternary* together as numerical data, grouping them as *contact_numeric.* By doing this, we could observe some correlations between these variables and with *subscribed,* visualized in *fig.6.* We can see here a fairly significant correlation between the duration of the call and probability to subscribe. This is hardly surprising: those who remain on the call for longer are more likely to subscribe to a plan. This likely doesn't provide us with actionable information in itself. We could infer that marketers should try to keep clients on the call for longer, but this is a banality, and probably confuses cause for effect: it is at least as likely that those looking for a better phone plan are more likely to stay on the call for longer, than that longer calls *cause* the increase in subscription uptake. *Duration* could be used as a proxy for likelihood-of-subscription, if we observed correlations between *duration* and other numerical data, but unfortunately we don't.

Table 4: Subscription Relative to Previous Outcome

Previous Outcome	*Subscribed?*	*Count*
Failure	No	3624
Nonexistent	No	32099
Success	No	464
Failure	Yes	595
Nonexistent	Yes	3114
Success	Yes	883

The data for *poutcome* is a little ambiguous. *Poutcome* was defined in our data as "Outcome of the previous marketing campaign", taking the variables 'success', 'failure' and 'non-existent'. Exploratory analysis reveals 464 cases where *poutcome* was a 'success', but the client is recorded as a 'no' for *subscribed* (cf. *Table 4*). Probably the reason for this is that the *subscribed* variable only refers to the success of the current and not the previous campaign, so a previous campaign may have succeeded for a client where the current one didn't.

3.5 Macro-Economic Indicators

Figure 7: Correlations among economic indicators

The final group of variables to consider are figures about macro-economic conditions when contact was made with the client. We grouped here: *emp.var.rate, cons.price.idx, euribor3m* and *nr.employed.* These variables are distinct in that they provide information not about the clients themselves, but about the broader economic climate. All these variables take decimal numbers. Since they are all quarterly indicators, they overlap with the variable *month.* The heatmap in *fig.7* visualizes strong correlations between this variables, which are unsurprising.

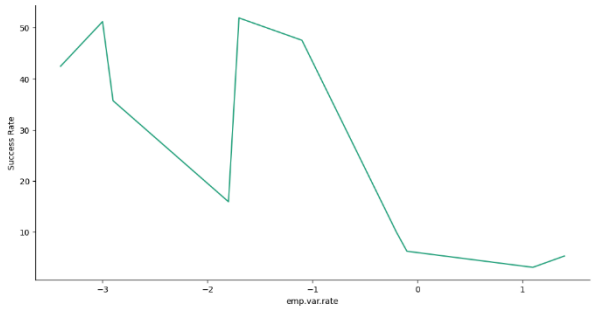
Table 5: Success Rate by EVR

EVR	*Total Subscribed*	*Overall Total*	*Success Rate (%)*
-3.4	447	1053	42.45
-3.0	87	170	51.18
-2.9	585	1638	35.71
-1.8	1455	9128	15.94
-1.7	396	763	51.9
-1.1	298	627	47.53
-0.2	1	10	10.0
-0.1	230	3672	6.26
1.1	238	7648	3.11
1.4	855	16070	5.32

Emp.var.rate is an abbreviation of Employment Variation Rate (EVR), which indexes the amount of hiring and firing in the economy (how many people are changing jobs), or the degree of churn in the labour market. According to Hou *et alia* (2004), "It has a negative impact on clients' purchasing decision [sic], which means large employment rate change makes clients less likely to subscribe a term deposit" (p.12). Although Hou *et al.* deal with subscriptions to term deposits for banks, not subscriptions to phone plans, Table 5 might suggest some inverse correlation between EVR and the success rate (SR) for securing subscriptions: a higher success rate seems to correspond to lower EVR, though we note too significant deviations, for example, the sudden drop to a ~16% SR at EVR of -1.8 followed by a rebound back to SR of ~52% at EVR of -1.7. Certainly there is no univariate linear relationship between EVR and response. *Fig.8* (right) charts the relationship, showing an erratic zig-zag pattern but with a possible downward trend as EVR rises. The range of EVR in our data is somewhat limited, so this trend would have to be confirmed by more data.

Cons.price.idx is an abbreviation for Consumer Price Index (CPI) which measures overall household inflation in the economy (ABS, 2023). Specifically, CPI measures the price change for a set basket of household goods, where a CPI of 100 means no change, and CPI below 100 means deflation for those goods. The CPI range in our data is small, at 2.566. The maximum is 94.767, which means our data can only track this variable over a period of general deflation for CPI. CPI correlates less with other economic variables than do the others, as visualized in *fig.7* above, but shows some correlation with *nr.employed* (discussed below).

Figure 8: Time-Series plotting Success Rate over Employment Variation Rate



Cons.conf.idx is an abbreviation for Consumer Confidence Index (CCI) which measures consumers' overall willingness to spend relative to their willingness to save. According to the OECD (2024):

> An indicator above 100 signals a boost in the consumers' confidence towards the future economic situation, as a consequence of which they are less prone to save, and more inclined to spend money on major purchases in the next 12 months. Values below 100 indicate a pessimistic attitude towards future developments in the economy, possibly resulting in a tendency to save more and consume less.

The minimum CCI in our data is -50.8, maximum is -26.9, yielding a range of 23.9. This means our data is limited to a period of negative consumer confidence, when consumers in the aggregate prefer saving over spending. As with CCI above, this may limit the kinds of inferences we can draw from this data, since we can't examine subscriptions during periods of high consumer confidence. However, the range here *may* be sufficient to identify a potential influence on response. Low CCI and low CPI together may indicate a period of economic contraction or stagnation. We should not necessarily assume that these conditions would automatically result in a decrease in subscriptions, since demand for phone plans is likely inelastic: a phone plan is almost universally considered an indispensable need, and tough economic conditions may in fact prompt consumers to look for better value on their ongoing expenditures like subscriptions. Negative CCI means an overall propensity to save rather than spend, but a better deal on a phone plan may well constitute a saving if the consumer is able to spend less for more.

Euribor3m is an abbreviation for 3 months Euribor rate, which is the interest rate at which a selection of European banks lend one another funds denominated in euros whereby the loans have a maturity of 3 months" (Global-Rates, 2024). *Euribor3m* is a continuous numerical variable which denotes a

percentage. It functions as a general proxy for interest rates in the European market. The minimum in our data is 0.634, and maximum is 5.045, yielding a range of 4.411. Our data is likely taken from a period of overall low interest-rates (relative to economic history in the *longue duree*), but the range would likely be sufficient to observe any relationship with response that might obtain. *Euribor3m* correlates strongly with *nr.employed* and *emp.var.rate* (cf. *fig.7*).

Nr.employed is an abbreviation for Number Employed which measures employment. It correlates strongly with *euribor3m* and *emp.var.rate.* Whereas EVR measures churn, *nr.employed* measures total employment.

References

Australian Bureau of Statistics. (Dec-quarter-2023). Consumer Price Index, Australia. ABS. https://www.abs.gov.au/statistics/economy/price-indexes-and-inflation/consumer-price-index-australia/latest-release. Retrieved: 10/03/2024.

ChatGPT. (February-March 2024). https://chat.openai.com/.
(NB: ChatGPT was consulted for help in writing the Python code used in our analysis).

Colab AI. (March 2024). https://colab.research.google.com/.
(NB: Colab AI was consulted for help in writing the Python code used in our analysis).

Eurostat.

4. (March 2023). Marriage and Divorce Statistics. https://ec.europa.eu/eurostat/statistics-explained/index.php?title=Marriage_and_divorce_statistics. Retrieved: 05/03/2024
5. (April 2023). Education and Training Statistics at Regional Level. https://ec.europa.eu/eurostat/statistics-explained/index.php?title=Education_and_training_statistics_at_regional_level#:~:text=In%202022%2C%20more%20than%20two,non%2Dtertiary%20level%20of%20education. Retrieved: 05/03/2024.
6. (February 2024). Population Structure and Ageing. https://ec.europa.eu/eurostat/statistics-explained/index.php?title=Population_structure_and_ageing#:~:text=On%201%20January%202023%2C%20the%20median%20age%20of%20the%20EU's,the%20other%20half%20was%20younger. Retrieved: 06/03/2024.

Hou, S., Cai, Z., Wu, J., Du, H., & Xie, P. (2022). Applying Machine Learning to the Development of Prediction Models for Bank Deposit Subscription. International Journal of Business Analytics (IJBAN), 9(1), 1-14. http://doi.org/10.4018/IJBAN.288514.

Global-Rates. (2024). Euribor 3 Months. https://www.global-rates.com/en/interest-rates/euribor/2/euribor-interest-3-months/. Retrieved: 10/03/2024.

OECD. (2024). Consumer confidence index (CCI) (indicator). https://doi.org/10.1787/46434d78-en. Retrieved: 10/03/2024