Telemarketing of a Portuguese banking institution

Jirayu Burapacheep

Often time, banks are using telemarketing to promote their campaign to the clients. Although telemarketing provides interaction between the firm and clients, this method can be time consuming. My goal in this project is to understand the relation between client's background and their decision on subscribing the campaign and predict how likely the client will subscribe. This prediction can help the firms strategize their telemarketing plan.

Moro et al. (2011) published the data of telemarketing of a Portuguese banking institution in his paper. We can find this dataset at the University of California-Irvine machine learning repository: http://archive.ics.uci.edu/ml/datasets/Bank+Marketing. The dataset consists of 21 columns of 3 groups: (1) client demographic such as age, job, whether they have a credit in default, etc., (2) the information of the previous contact with each client and (3) social and economic in Europe at that time.

First, let's study the impact of call duration on the client decision whether to subscribe to the campaign or not. The blue line in Figure 1 shows us the number of subscribed clients over the duration of the call. Similarly, the yellow line shows the number of not-subscribed clients. From the graph, we see that even though there are only few clients (8%) with call duration longer than 10 minutes, for those clients, the subscription rate is high (48%) compared to the overall rate (11%) as the blue and yellow lines are really close to each other. We observe the effect of call duration on the decision; however, we will not use this feature to predict the outcome in this report. This is because, in real world, we can't estimate the call duration before actually calling the customers.

We select 7 features for our analysis: (1) client has credit default, (2) number of prior contacts during this campaign, (3) number of days since the last contact, (4) number of contacts before this campaign, (5) outcome of the previous campaign, (6) contact type (cellular or telephone), and (7) Euribor 3 months rate. The Euribor rate is short for the Euro Interbank Offered Rate, the interest rate at which banks lend money to each other. Since there are two option of contact type, we translate contact type into two one-hot columns (6.1) cellular and (6.2) telephone. We also add a new feature (8) client income based on the normal distributed income of the client's job position that we scrape from http://www.salaryexplorer.com. Note that we drop the rows with a missing job position. Our resulted dataset has dimension 40,858 rows × 9 columns.

Next, we use a principal component analysis to find the dimensionality the data. The yellow line of Figure 2 shows that without scaling the data, we have very high explained variance (97%) even with one dimension. So, we apply StandardScaler on every column. The blue line shows the explained variance after scaling, we observe that, using PCA, we can reduce the dimension of our dataset from 9 to 5 columns which explain 86% of the variance.

To make the subscription prediction model, we create an sklearn pipeline with BankTransformer and a sklearn LogisticRegression. Inside BankTransformer, we use SimpleImputer for (1) client has credit default and (5) outcome of the previous campaign features to replace the unknown and non-existed value with the most frequent value appears in their column. We perform a 70/30% train/test split on our data. Our model achieves 90.1% accuracy. The false positive rate and false negative rate are 0.127% and 0.873%. With small false positive rate, this model is perfect for our goal which is detecting the client who tends to not subscribe.

The coefficient weights for each of the features in our model are presented in Figure 3. We can see that the feature that has the most impact on the prediction is the Euribor 3 months rate with weight -0.731. Since the weight is negative, the higher the interest rate between banks leads to the lower rate of subscription of the client. We also observe that, with weight -0.331, number of days since the last contact (pdays) significantly influences the decision as well: the longer the follow-up call, clients tend to not reject the plan.

To conclude, our analysis shows that the call duration highly correlate to the number of subscriptions with nearly half of the clients subscribe to the plan if the call duration is at least 10 minutes. However, the call duration can't be used as a factor for our prediction model because we don't know the call duration beforehand. With our selected features, we observe that the Euribor 3 months rate and the number of days since the last contact to the customer are an important factor. Both factors have negative correlation with the client decision.

Figure

Figure 1: call duration effect on subscription

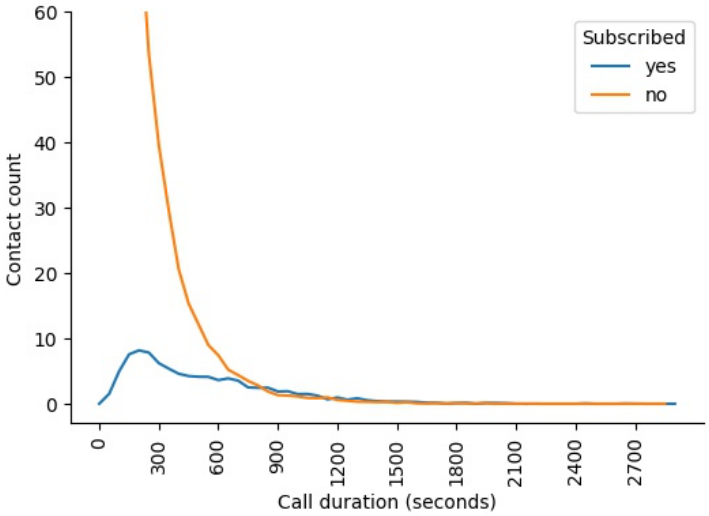

Figure 2: Principal component analysis

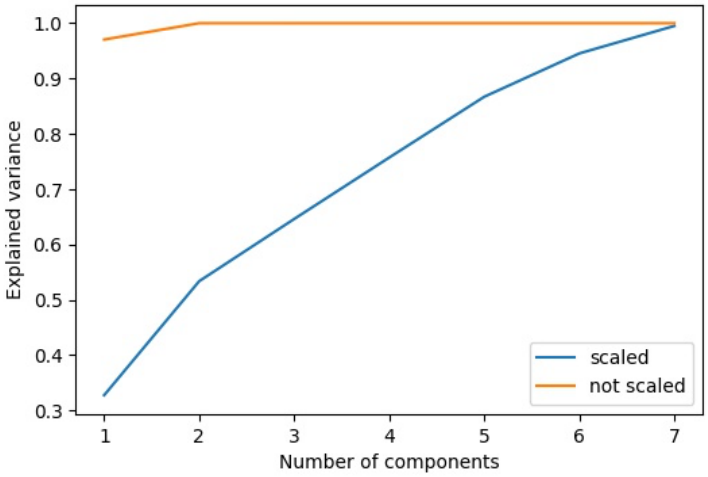

Figure 3: Logistic Regression Coefficients