## horizontal line



Portuguese Bank Marketing

12.10.2021

**─**

Lisa Patel

Springboard Mentor

Dipanjan Sarkar

# Introduction | Motivation

A Portuguese retail bank has recently seen a revenue decline and following a further investigation, found that the root cause is that their clients are not purchasing enough term deposits. We know that term deposits, which are the cash investments held at a financial institution, are a major source of income for banks since they allow banks to invest in other, more profitable financial products. The dataset in discussion here is consists of data from the Portuguese banking institution, which collected records of their direct marketing campaign phone calls, and the final outcomes indicating whether success campaigns in a binary format (yes/no). A success campaign indicates the customer has finally subscribed a term deposit at the end of the campaign.

**Our business goal here is to build a machine learning classification model that can identify the elements for a success campaign, from which we can improve marketing effectiveness by targeting the right customer. In other words, we want to find what kind of campaign strategies/history combined with what kinds of customers will bring about high propensities for subscribing to term deposits.** Through building predictive models, we can therefore increase revenues and lower labor costs by having more efficient marketing strategies without harming customer relationship.

# Dataset Background

The Bank of Portugal has collected a huge amount of data that includes customers profiles of those who have subscribed to term deposits and the ones who did not subscribe to a term deposit. The dataset[[1]](#footnote-1), sourced from the UCI Machine Learning Repository, contains 41,118 instances and 20 features, which are described below.

1 — age (numeric)

2 — job: type of job (categorical: ‘admin.’,’blue-collar’,’ entrepreneur’,’ housemaid’,’ management’,’ retired’,’ self-employed’,’ services’,’ student’,’technician’,’ unemployed’,’ unknown’)

3 — marital: marital status (categorical: ‘divorced’,’ married’,’ single’,’ unknown’; note: ‘divorced’ means divorced or widowed)

4 — education (categorical): ‘basic.4y’,’basic.6y’,’basic.9y’,’ high.school’,’ illiterate’,’professional.course’,’ university. degree’,’ unknown’)

5 — default: has credit in default? (categorical: ‘no’,’yes’,’ unknown’)

6 — housing: has a housing loan? (categorical: ‘no’,’yes’,’ unknown’)

7 — loan: has a personal loan? (categorical: ‘no’,’yes’,’ unknown’)

The following are related to the **last contact of the current campaign**

8 — contact: contact communication type (categorical: ‘cellular’,’telephone’)

9 — month: last contact month of the year (categorical: ‘Jan’, ‘Feb’, ‘mar’, …, ‘Nov’, ‘Dec’)

10 — day\_of\_week: last contact day of the week (categorical: ‘mon’,’tue’,’wed’,’thu’,’fri’)

11 — duration: last contact duration, in seconds (numeric). Important note: this attribute highly affects the output target (e.g., if duration=0 then y=’no’). Yet, the duration is not known before a call is performed. Also, after the end of the call y is known. Thus, this input should only be included for benchmark purposes and should be discarded if the intention is to have a realistic predictive model.

**other attributes:**

12 — campaign: number of contacts performed during this campaign and for this client (numeric, includes the last contact)

13 — days: number of days that passed by after the client was last contacted from a previous campaign (numeric; 999 means the client was not previously contacted)

14 — previous: number of contacts performed before this campaign and for this client (numeric)

15 — outcome: outcome of the previous marketing campaign (categorical: ‘failure’,’ nonexistent’,’ success’

**Social and economic context attributes**

16 — emp.var.rate: employment variation rate — quarterly indicator (numeric) 17 — cons.price.idx: consumer price index — monthly indicator (numeric)

17 — cons.conf.idx: consumer confidence index — monthly indicator (numeric)

18 — euribor3m: Euribor 3 month rate — daily indicator (numeric)

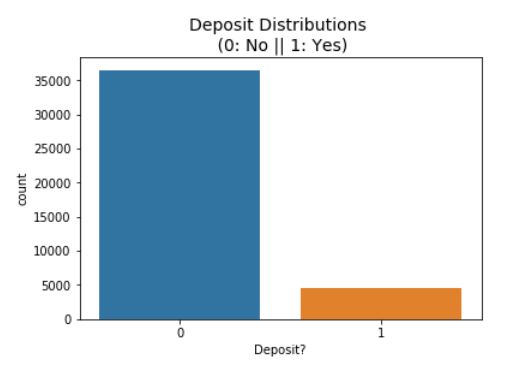
19 — nr.employed: number of employees — quarterly indicator (numeric) Output variable (desired target):

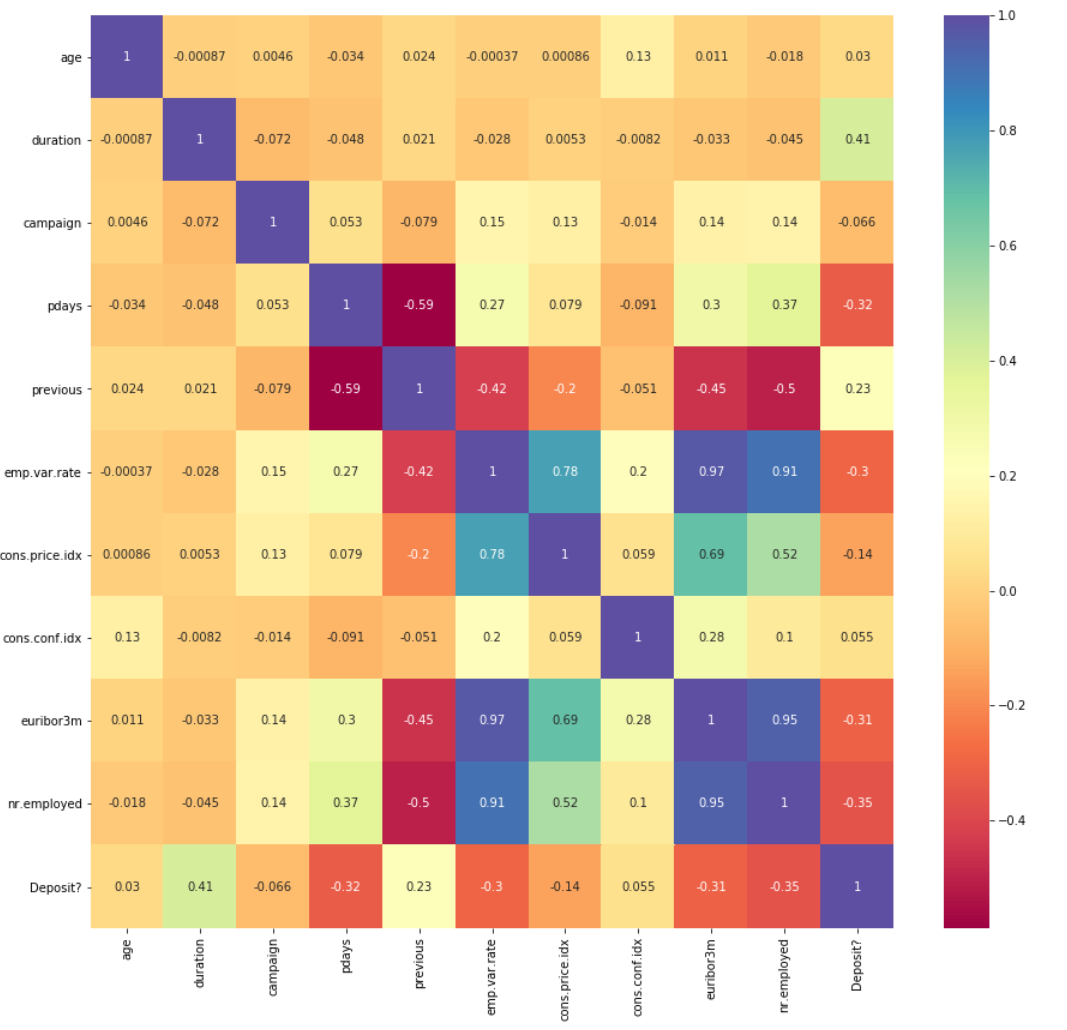
20 — y — has the client subscribed to a term deposit? (binary: ‘yes’,’ no’)

# Data Cleaning | EDA

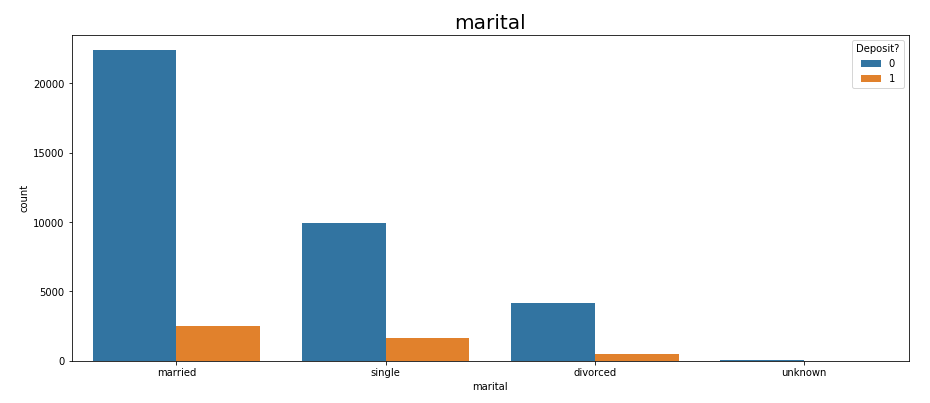
When checked for any missing or duplicate values in the dataset, there were none to be found. Due to the explanation about ‘Duration’ above, that feature was dropped from the set. The response variable (Yes/No) was converted to binary values (1/0) in this stage.

Another important aspect to consider initially is the class distribution of data. In our binary case, we can see from the figure below that this data is imbalanced as more than 85% of individuals who were called did not subscribe to a term deposit. This will certainly cause issues with the models as well as the evaluation process, so additional measures will have to be taken such as oversampling or under sampling to try to overcome those problems.



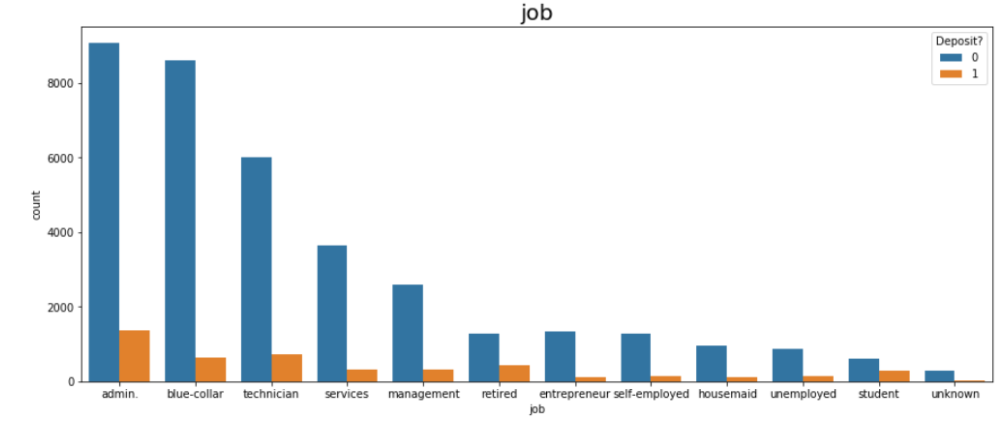


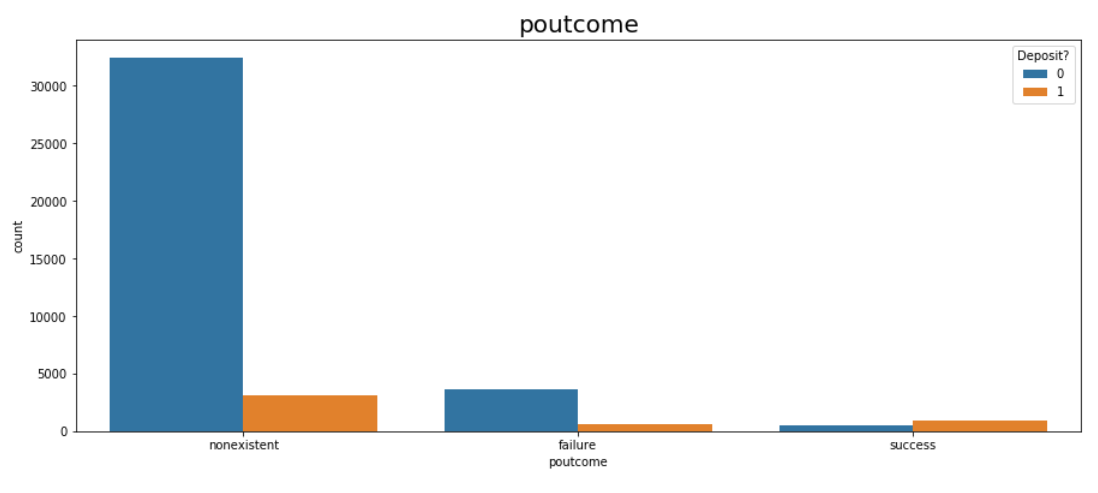
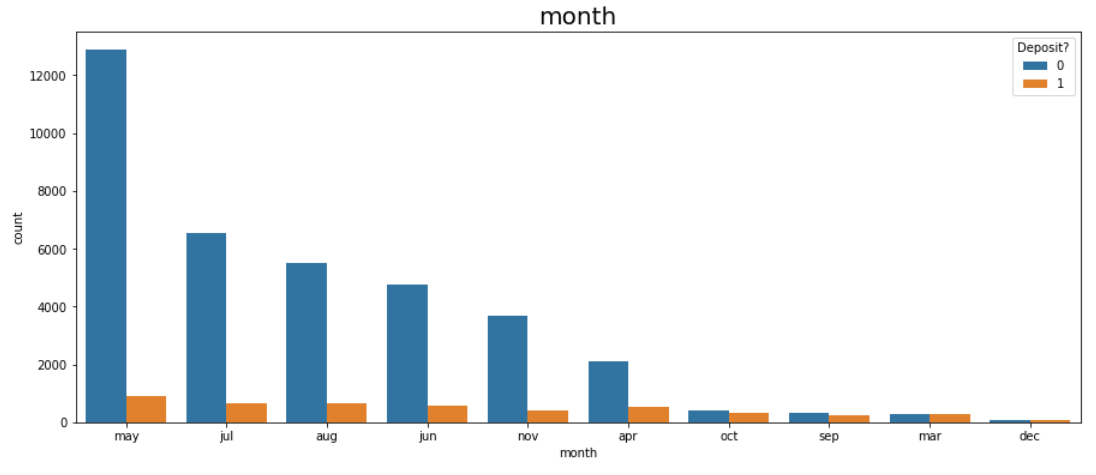
*An investigation of the covariance structure of the predictive measures in the dataset shows that the majority of the measures are not highly correlated. The few that have a high positive correlation are features related to socio-economic context attributes, which is understandable and expected. ‘Eurobor3m’ is highly and positively correlated with the number of employees, the employment variation rate and the consumer price index. There also seems to be moderate negative correlation between ‘pdays’ and ‘previous’ . With respect to the response variable, except for ‘Duration’ which will be dropped, there isn’t a strong relation with any other features.*



The biggest segment that was contacted were married individuals, followed singles and divorcees. However, when comparing the rate of ‘yes’, single individuals are slightly more likely to subscribe (14% compared to 10% for married and divorced.)

The target customers are admins, blue-collars and technicians but the frequency of students and retired people subscribed to the term deposit are pretty high (28.68% for students and 22.79% for retired people). This is understandable since those individuals have a greater need for a stable and reliable investment.

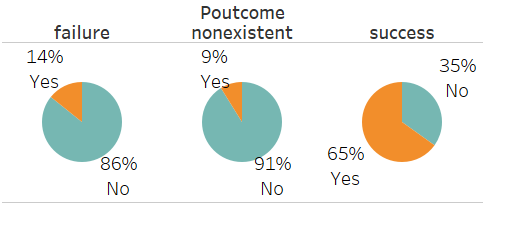




The outcome of the previous marketing campaign does appear to have an effect. Even though it is a very small number, prior success seems to suggest later success as 65% of those individuals did subscribe when contacted again. And, those that were contacted and failed also are more successful, indicating that returning to previous targets may be valuable.

The majority of contact calls were made between May and August, but month does not appear to have a very significant effect. There are months with very high success rates, but those are often months that also have very low relative counts.

[Grab your reader’s attention with a great quote from the document or use this space to emphasize a key point. To place this text box anywhere on the page, just drag it.]



## 

Customers contacted via cellular phone appear to be much more likely to be successful prospects. Cellular customers were also contacted more, but not with too much imbalance. Distribution of the days of the week feature was considered as well but there was little difference between any of the days.

## Preprocessing

Now that the main points of exploratory data analysis are covered, we can transition to what the processing steps consisted of and then the modeling. Processing involves encoding the raw data to well-formed datasets so that machine learning algorithms can be suitably applied. ML algorithms can only read numerical values, so categorical features, such as job, marital, education, housing, loan, contact, month, day of week and poutcome, were transformed through one-hot encoding.

Next steps were to scale the data so that the variables have similar scale through the Standard Scaler function and then to split the data into our training and test sets (70/30 split).

# Modeling

To fully evaluate and interpret the results of the models, we have to first consider which evaluation metric is best suited to our specific problem. The 4 metrics below are the most common one used for classifications models:

1. Accuracy: the proportion of true results among the total number of cases examined. it should be noted that while the prediction accuracy is the most common metric used for classification tasks, but it becomes inappropriate and misleading when it is used on an imbalanced data set. The model overfits to the class that’s represented more in your dataset and become oblivious to the existence of the minority class since the algorithm decides to classify everything in the majority class to get a good accuracy score. So, an alternate performance metrics must be used. For imbalanced datasets, the recall score and precision are much better metrics.

*Accuracy*=*TP*+*TN / TP*+*TN*+*FP*+*FN*

1. Precision: Refers to how accurate the model is at predicting positive cases correctly. Precision is a good metric to use when the **costs of false positives are high**. Example: Evidence that suggests a person is guilty of a crime (like DNA matching).

*Precision*=*TP / TP*+*FP*

1. Recall: Refers to how many actual positives the model captures by labeling it correctly. Recall is a good metric to use when the **costs of false negatives are high**. Example: Detecting fraud or serious illnesses.

*Recall*=*TP / TP*+*FN*

1. F1 score: a number between 0 and 1 and is the harmonic mean of precision and recall.

*F*1=2*TP /* 2*TP*+*FP*+*FN*

To figure out which metric is suitable for this case, the costs of false positives and false negatives have to be considered. The cost of false positives, or predicting a customer will subscribe but they won’t in actuality, is that the bank unnecessarily targets those individuals in future marketing campaigns, resulting in an increase of costs, time and resources. Additionally, they could potentially get annoyed due to frequent contacts, which may discourage them from subscribing in the future.

On the other hand, the cost of a false negative, or predicting a customer would not subscribe when they actually would, is the bank does not target them with additional marketing when it may be productive. Then bank would then miss out on gaining this person as a customer. In this case, we would want to focus on the **recall metric** of our model because we should try to predict as many actual positives as we can. A misclassification of a customer who actually wanted to make a deposit can mean a lost opportunity/revenue.

The three models that were run on this dataset were Logistic Regression, Decision Tree Classifier and Random Forest Classifier since they are some of the most common algorithms for binary classification problems. The logistic regression model is well suited for binary classification as it attempts to fit a curved line to the binary data by estimating coefficients for each feature in our dataset.

Decision Tree and Random Forest Classifiers are both built top-down from a single node that uses information gain, or entropy, to iteratively branch into smaller and smaller subsets of the data, ultimately ending in "leaves" that have gained the maximum amount of information based on the values of our variables. These models are "strong learners," meaning that at each decision point the model will seek to gain the split that gains the greatest amount of information possible, and then pursues that "branch" until it is exhausted, or meets constraints set by model parameters.

Random Forest is slightly different from a basic decision tree in that it is an ensemble of trees. A Random Forest model will create a series of smaller trees using only a few variables or features. Upon completion of constructing its trees, the trees with more predictive power get a stronger "vote" for predicting label data.

All three were run and the results of the models are summarized in the table and graph below. Since the data had the issue to being imbalanced, a more advanced oversampling method was used on the data to balance the distribution and potentially help remedy the problems it causes. The particular one used here is called Synthetic Minority Oversampling Technique (SMOTE), which generates new sample data for the minority class by creating 'synthetic' samples that are combinations of the closest minority class observations. The results of those models are added to the summary figures as well to make the comparison easier.

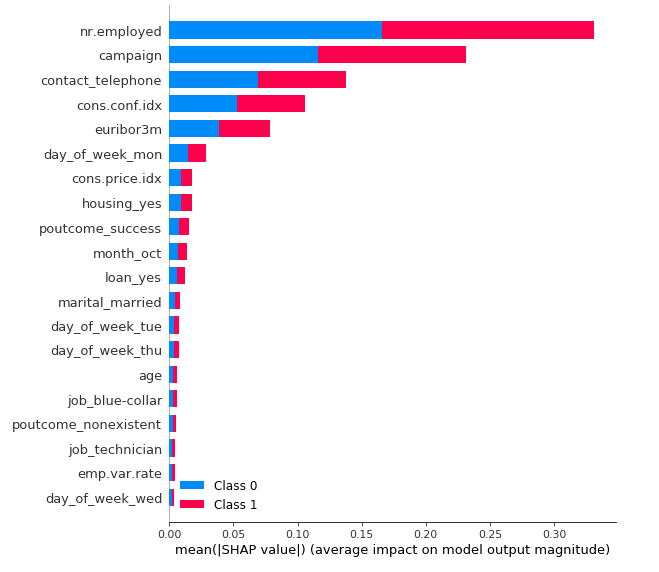
In terms of performance, it seems like most perform well for classifying the negative class given the high precision score, but not for classifying the positive class as the recall scores for all except Decision Tree with SMOTE are below 50%. They all result in more false positives than true positives. That should be expected since there is a significantly imbalanced distribution of the classes, but even with SMOTE, the Logistic Regression performed worse. The Random Forest model had a moderate improvement with SMOTE in recall given it increased from 22% to 49%. However, it is clearly seen in the summary graph that the greatest improvement in terms of all of the metrics was with the Decision Tree with SMOTE model. It is able to 83% percent of potential customers and miss the other 17%, which is much better than any of the others. It does well in accuracy, precision and F1 score as well, so it is the optimal model for this dataset.

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
|  | Logistic Regression | Logistic Regression (SMOTE) | Decision Tree | Decision Tree (SMOTE) | Random Forest | Random Forest (SMOTE) |
| Accuracy | 0.9 | 0.58 | 0.66 | 0.83 | 0.9 | 0.97 |
| Precision | 0.67 | 0.93 | 0.66 | 0.83 | 0.67 | 0.33 |
| Recall | 0.22 | 0.18 | 0.22 | 0.74 | 0.22 | 0.49 |
| F1 | 0.33 | 0.3 | 0.49 | 0.82 | 0.33 | 0.66 |
| TP/TN | 308/10817 | 1988/10746 | 350/10780 | 8170/10082 | 352/10743 | 3606/10775 |
| FP/FN | 1081/151 | 9048/147 | 1039/188 | 2866/811 | 1037/225 | 7430/118 |

Even though they are useful, the classic ML metrics like accuracy, recall, precision, r2 score, etc does not give a detailed or meaningful insight into the performance of the model. Given that ML models are commonly getting used to solve many problems with important real-world consequences, it is becoming crucial peer inside the black box and understand the performance of the models to better interpret them.

One of the many python libraries available to debug model to better understand the model and its performance on any sample of the data is SHAP, which stands for SHapley Additive exPlanations and uses the approach of game theory to interpret model predictions with Shapely values. These are measures of contributions each predictor (feature) has in a machine learning model.

The SHAP for the Decision Tree model with SMOTE is shown below. In it, the 5 most important features contributing to the model prediction are nr.employed, campaign, contact\_telephone, con.conf.idx and eurobor3m. This indicates the number of employees of the retail bank and the number of times they are contacted matters a great deal in the perception of the consumer. This number of employees feature can also be linked to the broader economy consumer confidence index, which measures how optimistic or pessimistic consumers are regarding their expected financial situation, and eurobor3m, the interest rate at which a selection of European banks lends to one another fund denominated in Euros. It makes sense that with optimistic consumer perception of the future economy and a high interest rate, individuals would be more inclined to purchase a term deposit that will ensure them a guaranteed return on investment. Given this, we can clearly see how valuable a library like SHAP is in enabling us to create a deeper and more meaningful interpretation of our ML models.



# Conclusions / Recommendations

The overall goal of this project was to build a predictive classification model to increase the effectiveness of the bank’s telemarketing campaign to increase revenue and gaining an insight into the profile of the individuals who did end up subscribing to a term deposit. From analyzing the data, we learned it may be productive to focus on:

* Narrowing the job criteria since students and retirees have a slightly higher likelihood of subscribing, which is probably influenced by their need for a safe, stable financial investment.
* Reaching out to individuals from a previous campaign as people were significantly more willing to buy if multiple contacts were made. Combining this finding with the month data could also lead to a reimaging and experimenting of how campaign calls are conducted through out the year. A campaign where customers towards the beginning of the year and then towards the end might lead to interesting results and greater effectiveness of the campaigns.
* Conducting more calls through cellphone since the calls have a greater likelihood of being answered.

Overall, the retail bank can increase the effectives of their campaigns and attracting more customers by cultivating a more granular understanding of their customer base. From a purely speculative perspective, this could be achieved through as trying to get a sense of how much money the customers invested into the term deposits. Having data that could answer questions like the following could help them create a better, nuanced profitability analysis:

* What is the term of the deposit?
* What interest rate was offered? Simple or compound?
* How much money was invested per customer in the deposit?
* What is the total cost of each campaign for the bank?
* What is the cost per customer?
* What is the minimum amount of money required for initial investment?

1. S. Moro, P. Cortez and P. Rita. A Data-Driven Approach to Predict the Success of Bank Telemarketing. Decision Support Systems, Elsevier, 62:22-31, June 2014. https://archive.ics.uci.edu/ml/datasets/bank+marketing [↑](#footnote-ref-1)