Bank Marketing Data Analysis



Predicting New Product Subscription

Maja Pedigo – Data Scientist at XYZ

# Abstract

A bank is conducting a marketing campaign over the phone to place this year’s product, a term deposit subscription. The goal is to investigate previous experience in relation to clients’ features to predict what approach will lead to an increase in subscription compared to last year.

## Data

The data consists of 23 features including the target variable, and 45211 instances. It is entirely contained within one dataset in .csv format.

## Method

The goal is to predict a ‘yes’ or ‘no’ answer from a client. The method will be to find the best classification model to accurately predict this binary variable. This will be done through data wrangling, exploration, preprocessing and ultimately building of prediction models.

## Data Wrangling

Initial data inspection revealed no missing values in the dataset. This significantly accelerated the phase of data cleaning.

Several redundant and misleading features were identified and removed from the dataset. Those features include: ‘response’, ‘marital-education’, ‘age’, and ‘pdays’ columns.

## Exploratory Data Analysis

During this phase, using visualization tools it was discovered that:

* Bank’s current clients are mostly:
  + Married
  + In their 30s-40s
  + Managers or blue-collar profession
  + High-school graduates
  + Use cell phones
  + Do not have loans
  + Were mostly contacted in month of May
  + 89% said no to subscribing to term deposit in the previous campaign
  + Greater salary indicated more ‘yes’ answers
  + Longer call durations led to more positive answers

## Data Preprocessing

This phase required dealing with the two types of data in our dataset preparing them for machine learning process. When the dataset was ready, we built and tested three machine learning models, eventually choosing the best one for prediction purposes.

Firstly, we transformed categorical features into numerical in order to make them usable for the machine learning process.

Then we split the data into previously seen (used for machine learning) and unseen (used for testing the models).

In the next step, we scaled the numerical data with a high range of values.

With a processed dataset, we started building the first classification model:

**Logistic Regression:**

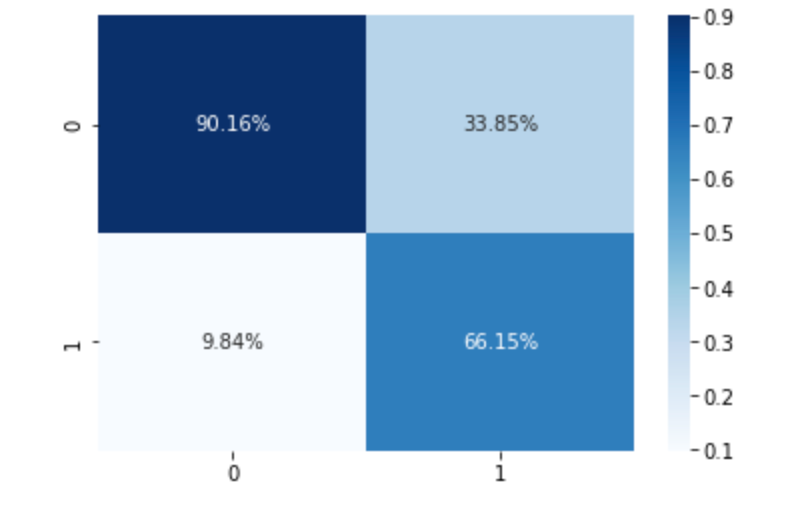
This model performed fast, and gave relatively good accuracy on the trained and test data:

Text

Description automatically generated

Table

Description automatically generated



Confusion Matrix

**Decision Tree:**

Accuracy went down with this model, as well as all other metrics.

**Text

Description automatically generated**

Table, calendar

Description automatically generated

Chart, treemap chart

Description automatically generated

Confusion Matrix

**Random Forest**

This model was our winner with all the metrics being the highest. We decided to choose this model and do some hyperparameter tunning with it to achieve best results.

Text

Description automatically generated

**Table

Description automatically generated**

**Chart, treemap chart

Description automatically generated**

Confusion Matrix

## **Modelling and Prediction**

In the final phase, we saved the Random Forest model, and tested it on the ‘unseen’ data we previously separated from the dataset.

The score we received was 88%.

We used this model to extract the most important features in the dataset in relation to our target:

1. Balance
2. Day (of the call)
3. Age group
4. Campaign (number of campaigns this person participated in)
5. Salary
6. Previous (number of contacts)
7. Housing (y/n mortgage)

## **Conclusions and Recommendations**

* older clients will subscribe more likely
* income of 40-50k is more likely to lead to subscription
* people who don't already have a mortgage would rather subscribe vs. those who do
* 1st and 10th day of contact in a month has the greatest positive impact
* contacting more than just a few times leads to negative impact to subscription
* higher balance means more chance to subscribe

Having in mind all the above we can recommend the following:

1. Collect more cell phone numbers from customers, as a large number of them is not filled
2. Duration of the call is highly correlated to the response, so any call over 300 seconds almost guarantees a subscription. Try to engage with a client.
3. Make calls on the 1st and 10th day of the month.
4. Do not make calls to one person more than 3 times.
5. Target clients who are in their 30s and 40s, married, with a balance over $1300, and who do not have a mortgage