Predictive Analytics for a Bank Marketing Campaign

Project Proposal

Robert Maheux

February 4 2019

**Research Question and Problem Framing**

Banks use telemarketing to increase the number of subscriptions to longer term investments such as term deposit bank accounts. This strategy works but it’s very time consuming and almost 90% of the phone calls don’t yield results. The bank could save time and money if it was able to focus its marketing efforts on the clients that are most likely to open a term deposit bank account.

**Background and Literature Review**

Banks have a wealth of digital data that can be used to reduce costs, increase revenues and improve customer experience. In order to remain competitive in today’s marketplace, banks that aren’t already leveraging predictive analytics need to begin doing so. The *Mobey Forum* expects that predictive analytics will soon be essential for banks to keep their position in the market and that other banks will be using predictive analytics as a competitive weapon1. A recent survey by *McKinsey & Company* revealed that almost every bank lists advanced analytics among its top five priorities2. Market Optimization and Cross-Selling are two uses of predictive analytics that have proven to be a benefit to financial institutions3.

**Dataset**

The dataset that will be used is obtained from the web site for the UC Irvine Machine Learning Repository**:** <http://archive.ics.uci.edu/ml/datasets/Bank+Marketing>.

The observations in the dataset are related to the direct marketing campaigns of a Portuguese national bank. The marketing campaigns were based on phone calls to offer a term deposit bank account. Each observation contains features related to a call made between May 2008 and November 2010. The dataset contains 41,188 observations with 21 features (including the output variable that indicates the client’s response).

It is based on a dataset that was originally published within the following publication: *S. Moro, P. Cortez and P. Rita. A Data-Driven Approach to Predict the Success of Bank Telemarketing. Decision Support Systems (2014), doi:10.1016/j.dss.2014.03.001.*

The original dataset was enriched by adding five social and economic features that were published by the Banco de Portugal and are publicly available at: <https://www.bportugal.pt/estatisticasweb>.

It contains the following features:

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | **Feature** | **Description** | **Type** | **Values** |
| 1 | age | Age of client | Numeric | 17 to 98 |
| 2 | job | Type of job | Categorical | admin, blue-collar, technician… |
| 3 | marital | Marital status | Categorical | divorced, married, single … |
| 4 | education | Education level | Categorical | high.school, university.degree… |
| 5 | default | Does the client have credit in default? | Categorical | yes, no, unknown |
| 6 | housing | Does the client have a housing loan | Categorical | yes, no, unknown |
| 7 | loan | Does the client have a personal loan? | Categorical | yes, no, unknown |
| 8 | contact | Contact communication type | Categorical | cellular, telephone |
| 9 | month | Last contact month of the year | Categorical | jan, feb, … dec |
| 10 | day\_of\_week | Last contact day of the week | Categorical | mon, tue, … fri |
| 11 | duration | Last contact duration, measured in seconds | Numeric | 0 to 4918 |
| 12 | campaign | Number of contacts during this campaign for this client | Numeric | 1 to 56 |
| 13 | pdays | Number of days since client was last contacted for a previous campaign | Numeric | 0 to 27, 999 |
| 14 | previous | Number of contacts before this campaign for this client | Numeric | 0 to 7 |
| 15 | poutcome | Outcome of the previous marketing campaign | Categorical | failure, nonexistent, success |
| 16 | emp.var.rate | Employment variation rate - quarterly indicator | Numeric | -3.4 to 1.4 |
| 17 | cons.price.idx | Consumer price index - monthly indicator | Numeric | 92.201 to 94.767 |
| 18 | cons.conf.idx | Consumer confidence index - monthly indicator | Numeric | -50.8 to -26.9 |
| 19 | euribor3m | Euribor 3 month rate - daily indicator | Numeric | 0.634 to 5.045 |
| 20 | nr.employed | Number of employees - quarterly indicator | Numeric | 4963.6 to 5228.1 |
| 21 | outcome | Does client want a term deposit account? | Categorical | yes, no |

*Feature Selection* techniques will be applied toidentify the features that explain most of the variance. The results will determine which features to include in the model.

There are several categorical features that will need to be converted into numerical values.

The values for the output variable are very imbalanced. Only 11% of the observations have an outcome of “yes”.

**Research Methodology**

The following steps will be taken to solve the problem:

Data Cleaning and Wrangling

The banking dataset contains missing values that will need to be imputed or discarded. There are several categorical features that will need to be converted to numerical features in order to be processed by machine learning algorithms. The observations that contain outliers will be discarded.

Identifying Useful Features

There are many features in the dataset but not all of them will be useful. *Feature Selection* techniques will be applied to identify the ones that should be included in the model.

Building Predictive Models

The Python programming language will be used to build four types of machine learning models using algorithms that are best suited to classification problems with a binomial output variable (“yes” or no”). The mathematical workings behind each model are beyond the scope of this document but here’s a rough idea of how each algorithm will work:

Logistic Regression

This algorithm accepts the features of an observation and predicts the probability that the output belongs to a certain class. It plots the training observations using a *sigmoid* which is a mathematical function with a characteristic "S"-shaped curve. This allows it to determine a boundary between two different classes of outcomes. A new observation is plotted and its outcome is predicted depending on which side of the boundary it was mapped.

Decision Tree

A decision tree breaks down the training dataset into smaller and smaller subsets and builds a set of *if-then-else* decision rules in the form of a tree structure. The final result is a tree with several decision nodes leading down to leaf nodes. Leaf nodes represent a classification or outcome.

K-Nearest Neighbours

This algorithm compares a new observation with all the observations in the training set and finds the K observations that most closely resemble it. It then takes the majority of the outcomes of these “nearest neighbours” to be the predicted outcome for the new observation.

Support Vector

This algorithm creates a representation of the training observations as points in space. They are mapped and are divided into two categories by a clear gap that is as wide as possible. New observations are then mapped into that same space and predicted to belong to one of the categories based on which side of the gap they fall.

Evaluating the Performance of the Models

Each model will be evaluated using the same set of metrics so that their performance can be compared. They will be evaluated using the Accuracy Score, Precision Score, Recall Score and Area under the ROC (Receiver Operating Characteristic) Curve.

**Conclusions**

This analysis will identify the data features that should be used to build a predictive model for the bank’s marketing team. It will build and compare four different types of models and find the most accurate one.

The predictive model will have some limitations. It can only be used for telemarketing campaigns to open term deposit bank accounts. In addition, it won’t determine the causes for a client’s willingness to open a new account. The factors that influence the client’s decisions will need to be investigated in order to improve the rate of subscriptions.

This analysis will result in the development of a valuable tool for the bank by using data that is already at its’ disposal. It will increase the efficiency of its marketing resources and thereby reduce costs.

1 - Mobey Forum - Predictive Analytics Workgroup,***Predictive Analytics in the Financial Industry***, (2016), [Mobey Forum](https://www.mobeyforum.org/predictive-analytics-financial-industry-art/)

2 - *McKinsey & Company,* ***Analytics in Banking: Time to Realize the Value*** (2017), [McKinsey & Company](http://www.mckinsey.com/industries/financial-services/our-insights/analytics-in-banking-time-to-realize-the-value?cid=eml-web)

3 - Banking CIO Outlook, ***Ten Ways Predictive Analytics Can Improve the Banking Sector*** (2018), [Banking CIO Outlook](https://www.bankingciooutlook.com/news/ten-ways-predictive-analytics-can-improve-banking-sector-nwid-383.html)