Udacity Machine Learning Nanodegree

Bank Marketing Campaign Predictive Analysis

Sai Bhargav Kommuru

June 14th, 2019

Proposal

Domain Background

In banks, huge data records information about their customers. This data can be used to create and keep clear relationship and connection with the customers in order to target them individually for definite products or banking offers. Usually, the selected customers are contacted directly through: personal contact, telephone cellular, mail, and email or any other contacts to advertise the new product/service or give an offer, this kind of marketing is called direct marketing. In fact, direct marketing is in the main strategy of many of the banks and insurance companies for interacting with their customers [1].

Historically, the name and identification of the term direct marketing suggested first time in 1967 by Lester Wunderman, which he is considered to be the father of direct marketing [2].
In addition, some of the banks and financial-services companies may depend only on strategy of mass marketing for promoting a new service or product to their customers. In this strategy, a single communication message is broadcast to all customers through media such as television, radio or advertising firm, etc. [3]. In this approach, companies do not set up a direct relationship to their customers for new-product offers. In fact, many of the customers are not interesting or respond to this kind of sales promotion [4].

Accordingly, banks, financial-services companies and other companies are shifting away from mass marketing strategy because its ineffectiveness, and they are now targeting most of their customers by direct marketing for specific product and service offers [1, 4]. Due to the positive results clearly measured; many marketers attractive to the direct marketing. For example, if a marketer sends out 1,000 offers by mail and 100 respond to the promotion, the marketer can say with confidence that the campaign led immediately to 10% direct responses. This metric is known as

the 'Response Rate', and it is one of many clear quantifiable success metrics employed by direct marketers.

From the literature, the direct marketing is becoming a very important application in data mining these days. The data mining has been used widely in direct marketing to identify prospective customers for new products, by using purchasing data, a predictive model to measure that a customer is going to respond to the promotion or an offer [5]. Data mining has gained popularity for illustrative and predictive applications in banking processes.

Problem Statement

All bank marketing campaigns are dependent on customers' huge electronic data. The size of these data sources is impossible for a human analyst to come up with interesting information that will help in the decision-making process. Data mining models are completely helping in the performance of these campaigns.

The purpose is increasing the campaign effectiveness by identifying the main characteristics that affect a success (the deposit subscribed by the client) based on a handful of algorithms that we will test (e.g. Logistic Regression, Gaussian Naive Bayes, Decision Trees and others). We the experimental results we will demonstrate the performance of the models by statistical metrics like accuracy, sensitivity, precision, recall, etc. We the higher scoring of these metrics, we will be able to judge the success of these models in predicting the best campaign contact with the clients for subscribing deposit.

Datasets and Inputs

The dataset(s) and/or input(s) to be used in the project are thoroughly described. Information such as how the dataset or input is (was) obtained, and the characteristics of the dataset or input, should be included. It should be clear how the dataset(s) or input(s) will be used in the project and whether their use is appropriate given the context of the problem.
The data set is well known as bank marketing from the University of California at Irvine (UCI)[6].

Abstract:

The data is related with direct marketing campaigns (phone calls) of a Portuguese banking institution. The classification goal is to predict if the client will subscribe a term deposit (variable y).

Category	Value
Data set characteristics	Multivariate
Number of Instances	45211
Area	Business
Attribute Characteristics	Real
Number of Attributes	17
Associated Tasks	Classification
Missing Values	Labelled as “unknown”

Source:

[Moro et al., 2014] 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

Data Set Information:

The data is related with direct marketing campaigns of a Portuguese banking institution. The marketing campaigns were based on phone calls. Often, more than one contact to the same client was required, in order to access if the product (bank term deposit) would be ('yes') or not ('no') subscribed.

Data files :

‘bank-additional-full.csv’ with all examples (41188) and 20 inputs, ordered by date (from May 2008 to November 2010), very close to the data analyzed in [Moro et al., 2014]
The classification goal is to predict if the client will subscribe (yes/no) a term deposit (variable y).

Attribute Information:

Input variables:

Bank client data:

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 housing loan? (categorical: 'no','yes','unknown')
7. loan: has personal loan? (categorical: 'no','yes','unknown')

Related with the last contact of the current campaign:

1. contact: contact communication type (categorical: 'cellular','telephone')
2. month: last contact month of year (categorical: 'jan', 'feb', 'mar', ..., 'nov', 'dec')
3. day_of_week: last contact day of the week (categorical: 'mon','tue','wed','thu','fri')
4. 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 obviously 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:

1. campaign: number of contacts performed during this campaign and for this client (numeric, includes last contact)
2. pdays: number of days that passed by after the client was last contacted from a previous campaign (numeric; 999 means client was not previously contacted)

3. previous: number of contacts performed before this campaign and for this client (numeric)
4. poutcome: outcome of the previous marketing campaign (categorical: 'failure','nonexistent','success')

Social and economic context attributes

1. emp.var.rate: employment variation rate - quarterly indicator (numeric)
2. cons.price.idx: consumer price index - monthly indicator (numeric)
3. cons.conf.idx: consumer confidence index - monthly indicator (numeric)
4. euribor3m: euribor 3 month rate - daily indicator (numeric)
5. nr.employed: number of employees - quarterly indicator (numeric)

Output variable (desired target):

1. y - has the client subscribed to a term deposit? (binary: 'yes','no')

Distribution of Data:

The dataset has 21 columns and 41188 rows, with 20 features, and one response variable. The number of customers who subscribed 4640 and the number of customers who didn't subscribe is 36548. Based on this the response rate of customers is about 11.27%, this makes the dataset very imbalanced. Fig. 1 shows the age distribution in the data is a normal distribution with slight left skewness. This hints that the majority of responses are in 25-40 age group.

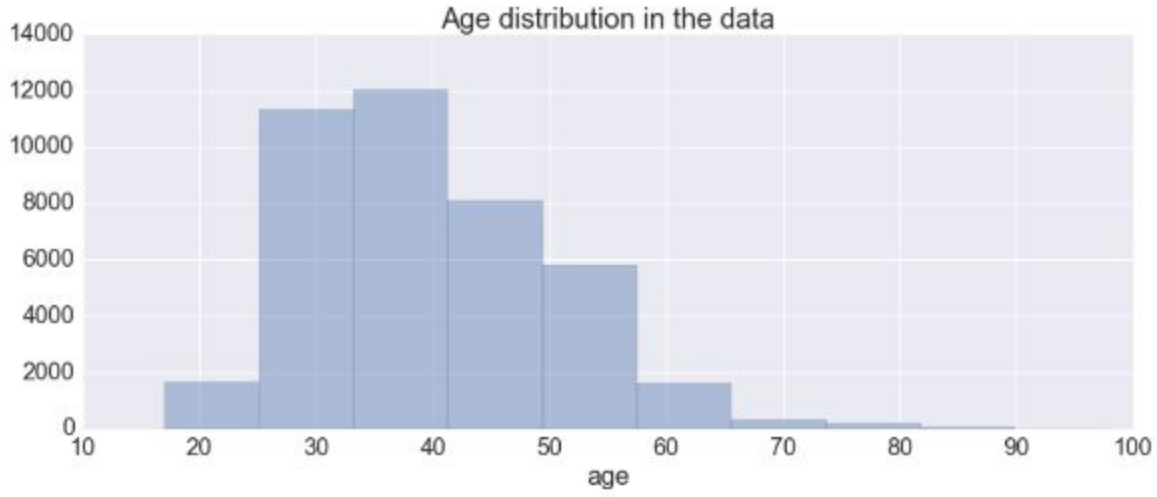

Fig. 1 Age distribution

Missing Attribute Values:

There are several missing values in some categorical attributes, all coded with the "unknown" label. These missing values can be treated as a possible class label or using deletion or imputation techniques.

Solution Statement

We will prepare the data by splitting feature and target/label columns and also check for quality of given data and perform data cleaning. To check if the model I created is any good, I will split the data into training and validation sets to check the accuracy of the best model. We will split the given training data in two ,70% of which will be used to train our models and 30% we will hold back as a validation set.

As described in the above section, there are several non-numeric columns that need to be converted. Many of them are simply yes/no, e.g. housing. These can be reasonably converted into 1/0 (binary) values. Other columns, like profession and marital, have more than two values, and are known as categorical variables. The recommended way to handle such a column is to create as many columns as possible values (e.g. profession_admin, profession_blue-collar, etc.), and assign a 1 to one of them and 0 to all others. These generated columns are sometimes called dummy variables, and we will use the pandas.get_dummies() function to perform this transformation.

We can also make subsets of original data using feature scaling techniques to normalize and scale data to try various iterations on the chosen models just to see if we see any differences in performance. Since the range of values of raw data varies widely, in some machine learning algorithms, objective functions will not work properly without normalization. For example, the majority of classifiers calculate the distance between two points by the Euclidean distance. If one of the features has a broad range of values, the distance will be governed by this particular feature. Therefore, the range of all features should be normalized so that each feature contributes approximately proportionately to the final distance.[7]

We don't know which algorithms would be fit for this problem or what configurations to use. So let's pick a few algorithms to evaluate.

- Logistic Regression (LR)
- K-Nearest Neighbors (KNN)
- Classification and Regression Trees (CART)
- Gaussian Naive Bayes (NB)
- Support Vector Machines (SVM)
- Random Forests (RF)
- XGBoost (XGB)

We are using 5-fold cross validation to estimate accuracy. This will split our dataset 5 parts, train on 4 and test on 1 and repeat for all combinations of train-test splits.

Also, we are using the metric of accuracy to evaluate models. This is a ratio of the number of correctly predicted instances in divided by the total number of instances in the dataset multiplied by 100 to give a percentage (e.g. 95% accurate). We will be using the scoring variable when we run build and evaluate each model next.

Benchmark Model

The given dataset is a typical supervised learning problem for which tree type models perform a lot better than the rest. So we will pick Extreme Gradient Boosting (XGB) as benchmark and try to beat the benchmark with hyperparameter tuning. We will also try Ensemble methods if the hyperparameter tuning does not improve the score.

Evaluation Metrics

Student proposes at least one evaluation metric that can be used to quantify the performance of both the benchmark model and the solution model presented. The evaluation metric(s) proposed are appropriate given the context of the data, the problem statement, and the intended solution.

The performance of each classification model is evaluated using three statistical measures; classification accuracy, sensitivity and specificity. It is using true positive (TP), true negative (TN), false positive (FP) and false negative (FN). The percentage of Correct/Incorrect classification is the difference between the actual and predicted values of variables. True Positive (TP) is the number of correct predictions that an instance is true, or in other words; it is occurring when the positive

prediction of the classifier coincided with a positive prediction of target attribute. True Negative (TN) is presenting a number of correct predictions that an instance is false, (i.e.) it occurs when both the classifier, and the target attribute suggests the absence of a positive prediction. The False Positive (FP) is the number of incorrect predictions that an instance is true. Finally, False Negative (FN) is the number of incorrect predictions that an instance is false. Table below shows the confusion matrix for a two-class classifier.

	Predicted No	**Predicted Yes**
Actual No	TN	FN
Actual Yes	FP	TP

Classification accuracy is defined as the ratio of the number of correctly classified cases and is equal to the sum of TP and TN divided by the total number of cases (TN + FN + TP + FP).

$$\text{Accuracy} = \frac{TP+TN}{TN+FN+TP+FP}$$

Precision is defined as the number of true positives (TP) over the number of true positives plus the number of false positives (FP).

$$\text{Precision} = \frac{TP}{TP+FP}$$

Recall is defined as the number of true positives (TP) over the number of true positives plus the number of false negatives (FN).

$$\text{Recall} = \frac{TP}{TP+FN}$$

Sensitivity refers to the rate of correctly classified positive and is equal to TP divided by the sum of TP and FN. Sensitivity may be referred as a True Positive Rate.

$$\text{Sensitivity} = \frac{TP}{TP+FN}$$

Specificity refers to the rate of correctly classified negative and is equal to the ratio of TN to the sum of TN and FP

$$\text{Specificity} = \frac{TN}{TN+FP}$$

Project Design

The workflow of solving this problem will be in the following order:

- Exploring the Data
    - Loading Libraries and data
    - Peek at the training data
    - Dimensions of data
    - Overview of responses and overall response rate
    - Statistical summary
- Data preprocessing/cleaning
    - Preprocess feature columns
    - Identify Feature and Target columns
    - Data cleaning
    - Training and Validation data split
    - Feature Scaling - Standardization/Normalizing data
- Evaluate Algorithms
    - Build models
    - Select best model
    - Make predictions on the validation set
    - Feature importance and feature selection
- Model Tuning to Improve Result
- Final conclusion

Visualizations will be provided in some sections as needed.

References

[1] Ou, C., Liu, C., Huang, J. and Zhong, N. 'One Data mining for direct marketing', Springer-Verlag Berlin Heidelberg, pp. 491–498., 2003.

[2] http://en.wikipedia.org/wiki/Direct_marketing . Wikipedia has a tool to generate citations for particular articles related to direct marketing.

[3] O'guinn, Thomas.” Advertising and Integrated Brand Promotion”. Oxford Oxfordshire: Oxford University Press. p. 625. ISBN 978-0-324-56862-2. , 2008.

[4] Petrison, L. A., Blattberg, R. C. and Wang, P. ‘Database marketing: Past present, and future’, Journal of Direct Marketing, 11, 4, 109–125, 1997.

[5] Eniafe Festus Ayetiran, “A Data Mining-Based Response Model for Target Selection in Direct Marketing”, I.J.Information Technology and Computer Science, 2012, 1, 9-18.

[6] https://archive.ics.uci.edu/ml/datasets/bank+marketing

[7] https://en.wikipedia.org/wiki/Feature_scaling