Machine Learning Engineer Nanodegree
Capstone Project Proposal
Alfred Raju
Feburary22nd,2019

Proposal:

Predictive Analysis On Bank Marketing Campaign

Domain Background:

History:

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.

Bank marketing is known for its nature of developing a unique brand image, which is treated as the capital reputation of the financial academy. It is very important for a bank to develop good relationship with valued customers accompanied by innovative ideas which can be used as measures to meet their requirements[1].

The objective of **direct marketing** in retail **banking** is to attract new customers, to create a **direct** customer-**bank** communication to promote an offer or obtain customer information, and to strengthen a long-term relationship with the customer[2].

During 2008, banks increased direct mail offers cross-selling additional products and services to current clients by 57% over 2007. During the same period, acquisition direct mail rose 7%.

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 [3]. Data mining has gained popularity for illustrative and predictive applications in banking processes.

Problem Statement:

The customer’s huge electronic data is the source for all bank marketing campaigns. 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 , Adaboost classifier and Decision tree classifier). With the experimental results we will demonstrate the performance of the models by statistical metrics like accuracy, sensitivity, precision, recall, etc.With the highest scoring of these metrics mainly accuracy we can judge the success of these models in predicting the best campaign contact with the clients for subscribing deposit.

Datasets and Inputs:

The data set is well known as bank marketing from the University of California at Irvine (UCI)[4].

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
Date Donated	:	2012-02-14
Associated Tasks	:	Classification
Missing Values?	:	Yes, labelled as “unknown”
Number of Web Hits	:	386732

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-employe d' , '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)

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 a term deposit? (binary: 'yes' ,'no')

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:

Here we are trying to predict the best campaign to contact with the clients for subscribing deposit.

We shall first 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 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 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 don't know which algorithm would be best fit for this problem.We will try all the above mentioned algorithms and find their accuracy scores.

Benchmark Model:

Among all the models I have chosen,I consider Logistic Regression as my benchmark model.Here accuracy scores of the models are compared and the best model is the one with highest accuracy.

Evaluation Metrics:

The performance of each classification model is evaluated using three statistical measures; classification accuracy, recall and precision. 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).

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

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

Project Design:

Project is composed of different steps:

- 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
- Evaluate Algorithms
- Build models
- Select best model
- Make predictions on the validation set
- Final Conclusion.

References:

[1]https://www.tutorialspoint.com/bank_management/bank_management_marketing.htm

[2]https://www.retailbanking-academy.org/media/uploads/NEW-BRANDING-DOCS/modules/RBII/RB_II_New_201_Small.pdf

[3]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.

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