Final_Report

January 14, 2021

1 Project: Bank Marketing

By:

Esraa Gaber 201700429

Ahmad Sami Muhammad 201700259

Aiad Asaad 201700790

Table of Contents

Problem definition and motivation

Marketing is such a crucial thing for any business. It may determine if the business is going to succeed or not. The project we are trying to tackle is the effectiveness of a certain marketing technique with a certain target audience. More specifically, we want to predict if a bank's marketing campaign will drive people to deposit money in the bank with the knowledge of their personal information such as age, job, marital status and education, and other information as the last time he was contacted by us. The data was collected for a marketing campaign from bank clients to determine whether the client will subscribe to a term deposit or not. It consists of 45,211 rows and 17 columns The data attributes are divided into input variables - information about the client - and output which is a subscription to a term deposit.

General Properties

The data are collected via phone calls by a Portuguese retail bank, from May 2008 to June 2013.

Features: 1 - age (numeric) 2 - job: type of job (categorical: 'admin.', 'blue-collar', 'entrepreneur' ,'housemaid', 'management', 'retired', 'self-employed', 'services', 'student', 'technician', 'unem

ployed', 'unknown') 3 - marital: marital status (categorical: 'divorced', 'married','single', 'unknown'; note: 'divorced' means divorced or widowed) 4 - education (categorical: 'basic.4y', 'ba sic.6y', 'basic.9y', 'high.school', 'illiterate', 'professional.course', 'university.degree', 'unknown') 5 - default: has credit in default? (categorical: 'no','yes','unknown') 6 - balance: average yearly balance, in euros (numeric) 7 - housing: has a housing loan? (categorical: ‘no’,‘yes’,‘unknown’) 8 - loan: has personal loan? (categorical: ‘no’,‘yes’,‘unknown’) 9 - contact: contact communica tion type (categorical: ‘cellular’,‘telephone’) 10 - month: last contact month of year (categorical: ‘jan’, ‘feb’, ‘mar’, . . . , ‘nov’, ‘dec’) 11 - day_of_week: last contact day of the week (categorical: ‘mon’,‘tue’,‘wed’,‘thu’,‘fri’) 12 - 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.13 - campaign: number of contacts performed during this campaign and for this client (numeric, includes the last contact) 14 - 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) 15 - previous: number of contacts performed before this campaign and for this client (numeric) 16 - outcome: outcome of the previous marketing campaign (categorical: ‘failure’, ‘nonexistent’, ‘success’) **Label** 17- y - has the client subscribed a term deposit? (binary: “yes”, “no”)

The data contain categorical and numeric attributes. The numeric are age, balance, day, duration, campaign, pdays, and previous. The rest are categorical.

1.6 Approach and methodology

First, we divided the data set into the training set and test set then the random forest approach was applied and we calculated its accuracy and precision.

Then AdaBoostClassifier approach was applied and the accuracy and precision were calculated too.

Finally, we applied the XGBClassifier approach and calculated its accuracy and

precision. We compared the three approaches at the end.

1.6.1 Literature Review

In previous papers, they have used LOGISTIC REGRESSION, DECISION TREES, SVM, and NEURAL NETWORK. They have reached 91% as the best accuracy by using Nural networks. Here, We will try to use different models than the previously used models. We will use RandomForestClassifier, AdaboostClassifier and XgBoostClassifier. Moreover, We care about precision more than ROC as we concern about all the people who will make a deposit (True positive). We want the model to make a correct prediction for all the True positives with mean making false-positive near to zero precision= true positive/(true positive+false positive). We will target a precision score of 90%.

1.6.2 Oversampling

There are three ways to oversampling in imblearn.over_sampling Library. The first one is Random oversampling. It duplicates more samples from the abundant class with random sampling with replacement. However, this method may lead to overfitting. Because many samples of the majority class will not be included in the dataset as the minority class take place of them. The second one is Synthetic Minority Oversampling Technique (SMOTE).

It generates new samples from the abundant class based on the feature space similarities between existing abundant instances. That is happened by finding the K-nearest neighbors of each abundant instance. Then randomly selects one of them. Finally, calculate linear interpolations to produce a new minority instance in the neighborhood. The third one is Adaptive Synthetic Sampling (ADASYN). It also generates new samples based on density distributions. It also finds the K-nearest neighbors of each abundant instance. But it repeats this process and adaptively shifts the decision boundary to focus on those samples that are difficult to learn.

we will use the smote method to make our data balanced

XGBClassifier

Here is a simple explanation about how the XGBOOST WORKING.

There is a definition called similarity or quality score. It's calculated by the square (the summing of the residual values) and then divides the result by (their number + learning rate) Residuals is the difference between the observed and the predicted value gain= left branch similarity +right branch similarity -root similarity

- At first the XGBoost starts from a single leaf of the tree.
- Then start with an initial prediction with any value the difficulty is 0.5
- Calculate the residual and but the values in the single leaf
- Calculate the quality score or the similarity
- Then we try to do better by splitting the residuals into groups.
- Here we try to find the best split to reduce the residuals and split them into two groups • Then we will have two branches one on the left another on the right
- we calculate the similarity score for both branches
- then we need to quantify how much better the leaves cluster similar residuals than the root by calculating the gain
- then we go into the residuals again and split them and calculate the gain
- and then compare the gain again and see with is better
- we do this till find the best fit or reaching the asked splitting number

2 Discussion and analysis

At first, the accuracy of RandomForestClassifier, XGBClassifier, and AdaBoostClassifier was 91%, 88%, and 87 respectively. After hyperparameters tunning the accuracy becomes 91.1% (Random ForestClassifier) ,91% (AdaBoostClassifier) and 92.4% (XGBClassifier).

We finally found that XGboost Classifier is the best model. This model gives an accuracy of

92.4 %.

Ethical considerations

1. **Data**

Data was collected from an open-sourced, free source. It was collected from bank clients in Por Portuguese. These data are collected randomly which will ensure fairness while training and testing

the model.

Also, The correlation between the different features and the target was calculated and the features with low correlation were neglected during the training of the model (e.g. “default”, “month”, “age”, "day)

2. **The algorithm**

The algorithms we applied are free sourced and available in python libraries which can be accessed by anyone (No intellectual property theft)

3. **The results:**

It was clear that the data is biased as the no answers are much larger than the yes answers (unbalanced data). So in order to get accurate results, some processing was done first before training and testing the model.

3 Professional responsibility and accountability

We are supposed to recheck the concept drift in our models which refers to the unknown and hidden relationship between inputs and output variables.

There are some decisions required by the machine learning practitioner to deal with the concept drift such as (Future assumption about the data source, Possible change patterns, Mechanisms adaptivity, Model selection based on particular parametrization)

To address the concept drift, we can take many approaches

1. Do Nothing (Static Model)
2. Periodically Re-Fit
3. Periodically Update
4. Weight important data
5. Learn The Change
6. Data Preparation for the change

Conclusions

We can see that we have made a relatively efficient recommendation system for the bank marketing team. The recommendation system was based on a Random Forest ML model as it appeared to be more appropriate logically. However, we tried other models like AdaBoost Classifier and XGB Classifier, and the XGB showed the best results. It managed to achieve an accuracy of 92.4%.

3.1 Reference

S. Moro, P. Cortez and P. Rita. A Data-Driven Approach to Predict the Success of Bank Telemar keting. Decision Support Systems, Elsevier, 62:22-31, June 2014

Oversampling was acknowledge from:
https://towardsdatascience.com/sampling-techniques-for-extremely
imbalanced-data-part-ii-over-sampling-d61b43bc4879