Data Science Essentials

Project Fall 2016

Bank Marketing Dataset

Guided by:
Pr. Mike Bernico

Group Members:

Aman Chandrakar

UNIVERSITY OF ILLINOIS SPRINGFIELD

Contents

1. Introduction 3
2. Dataset……...3
3. Exploratory Data Analysis (EDA)..4
4. Pre - processing... ..5
5. Models 5
6. Results…..5
7. References…...6

1. Introduction:

The Bank Marketing dataset was created by Paulo Cortez (Univ. Minho) and Sérgio Moro (ISCTE-IUL) @ 2012. 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 (or not) subscribed. The classification goal is to predict if the client will subscribe a term deposit or not.

2. Dataset:

The Bank marketing dataset is available to download by http://archive.ics.uci.edu/ml/datasets.html. The dataset consists of 45211 instances and 16 + the output attribute.

Attribute information are as follows:

Independent variables:

1) age: Age of the person contacted (numeric)

2) job: type of job (categorical: "admin.","unknown","unemployed","management","housemaid","entrepreneur","student", "blue-collar","self-employed","retired","technician","services")

3) marital: marital status (categorical: "married","divorced","single"; note: "divorced" means divorced or widowed)

4) education: (categorical: "unknown","secondary","primary","tertiary")

5) default: has credit in default? (binary: "yes","no")

6) balance: average yearly balance, in euros (numeric)

7) housing: has housing loan? (binary: "yes","no")

8) loan: has personal loan? (binary: "yes","no")

# related with the last contact of the current campaign:

9) contact: contact communication type (categorical: "unknown","telephone","cellular")

10) day: last contact day of the month (numeric)

11) month: last contact month of year (categorical: "jan", "feb", "mar", ..., "nov", "dec")

12) duration: last contact duration, in seconds (numeric)

# other attributes:

13) campaign: number of contacts performed during this campaign and for this client (numeric: includes last contact)

14) pdays: number of days that passed by after the client was last contacted from a previous campaign (numeric: -1 means client was not previously contacted)

15) previous: number of contacts performed before this campaign and for this client (numeric)

16) poutcome: outcome of the previous marketing campaign (categorical: "unknown","other","failure","success")

Output variable (desired target):

17) y: has the client subscribed a term deposit? (binary: "yes","no")

3. Exploratory Data Analysis (EDA):

I conducted the EDA for all the attributes to check:

1) Whether the attributes are categorical or continuous?

2) To check for missing values?

3) To see the values counts for all the attributes.

4) To visualize the attributes in form of histogram (for continuous attributes) and bars (for categorical attributes).

By EDA I found that none of the attributes have missing values and the data is clean.

age, balance, day, campaign, pdays, duration, previous are continuous variable since these variables have range of values.

job, marital, education, default, housing, loan, contact, month, poutcome and y are categorical variables since these variables are further divided into categories.

4. Pre-processing:

There were 4 attributes with binary datatype, so I converted these binary datatype (Yes/No) attribute to continuous attributes. I used get_dummies function to return an indicator variable for each category.

5. Models:

I implemented 3 models to carry on my experiment with this bank marketing dataset: Logistic Regression, Random Forest Regressor and Random Forest Classifier, and tuned the hyper parameters using Grid Search with K-Fold Cross Validation, with 10 folds. For Logistic Regression and Random Forest Classifier, I split the dataset into training and test data, 80% was used as training model and 20% remaining was used as testing the model.

Logistic Regression is a predictive analysis and is one of the best model to be used when the dependent variable is binary. Random forest is an estimator that generates multiple decision trees and produce the result by averaging. I used Grid Search with K-Fold Cross Validation, with 10 folds to tune the hyper parameters of Random Forest classifier.

6. Results:

The following results were obtained for various model:

1) Logistic Regression:

Accuracy: 0.90

AUC = 0.66

2) Random Forest Regressor:

1) Only numeric variable (Benchmark)

oob_score_: 0.26695958282941801
roc_auc_score: 0.875243118011

2) With all variable i.e., continuous + categorical

a) Before tunning the parameters:

roc_auc_score: 0.921094484595

b) After tunning the parameters:

roc_auc_score: 0.935280343276

3) Random Forest Classifier with Grid Search

Accuracy: 0.906115227248
Precision: .63
Recall: .44
AUC: 0.92967008052
With (10-fold cross validation):
Score is 0.713023 +/- 0.121932

7. References:

[1] http://archive.ics.uci.edu/ml/datasets.html
[2] [Moro et al., 2011] S. Moro, R. Laureano and P. Cortez. Using Data Mining for Bank Direct Marketing: An Application of the CRISP-DM Methodology. In P. Novais et al. (Eds.), Proceedings of the European Simulation and Modelling Conference - ESM'2011, pp. 117-121, Guimarães, Portugal, October, 2011. EUROSIS.
[3] http://scikit-learn.org/stable/#
[4] https://en.wikipedia.org/wiki/Logistic_regression