École polytechnique fédérale de Lausanne

Bank Marketing Analysis

Vladimir Macko
Lausanne 2017

Introduction

The data set collected by a Portuguese bank institution is analyzed in order to determine and possibly predict whether a customer is interested in buying a bank product (bank term deposit) based on the attributes about the customer and details of telephonic contact between a bank and the customer [1]. Similarly to previous analysis described in [2], several data analysis tools are examined in order to select the one which provides the most reliable predictions. The selection of customers which are likely to accept the bank product may lead to a significant price reduction of the bank telemarketing campaign. Also, the most important indicators determining the customer decision are presented.

Data Material

The data set is associated with a direct marketing campaign of a bank. It contains personal (age, type of employment, marital status, reached education), and financial details (possession of credit, housing loan, personal loans) about 41188 bank customers as well as information about telephonic calls used for product offering (type of phone contact, month of contact, day of the week of the contact) and other attributes (number of contacts with costumer during the campaign and outcomes of previous campaign, the data set also contains the final call duration but this attribute is not used in the analysis as it is directly correlated with the outcomes). In addition, other social and economic markers are included in the dataset (quarterly indicator of of employment variation rate, monthly indicator of consumer price index, monthly indicator of consumer confidence index, daily indicator of Euribor 3 month rate and quarterly indicator of number of employees). The dataset also contains the final decision of the client about the offered term deposit which is a subject of interest.

Analysis Methods

Firstly, in order to make use of machine learning algorithms it is required to represent the information contained in the data set purely with a numerical set. Therefore, data features which have categorical character are sparsed into several categorical variables. For example, the marital status is represented by boolean variables *person is married*, *person is single*, *person is widowed* and *persons marital status is unknown*. Numerical variables are used with they natural number representation. Since sparcing significantly increased the number of analyzed variables and therefore the dimension of analysis task. Moreover, sparcing also increased the level of correlation in the data. As the there are many dimensions in the data which carry only very limited unique information, the principal component analysis is applied to reduce the dimentionality of the problem. At this point only first 8 significant components are selected for further processing. At this point only first 8 significant components are selected for further processing and data sample is split into

training sample (60%) and testing sample (40%) which are used to train and estimate performance of selected method which are tested.

Performance of a set of classifying tools is analyzed in order to select the one with highest performance. The following classifier methods are tested:

- Random Forest Classifier [3]
- Bagging Classifier [4]
- Extra Trees Classifier [5]
- Gradient Boosting Classifier [6]
- Keras Sequential Neuron Network [7]

In all cases, the classifier model is trained on training data set and testing data is used to determine accuracy of predictions provided by a given model. For each model several configurations are examined but default values of input parameters usually seem to perform the best. In case of neuron network, multiple options of number and types of layers and are tested until the one providing the best performance is selected.

Finally, once the classifier delivering the best performance among the set of classifiers tested is identified, this classifier is retrained on full data sample without principal component analysis transformation to examine whether a more accurate result is reachable.

Results

Principal component analysis is applied on the data sample. A visualization of data distribution as a function principal component values is provided in figures 1,2,3.

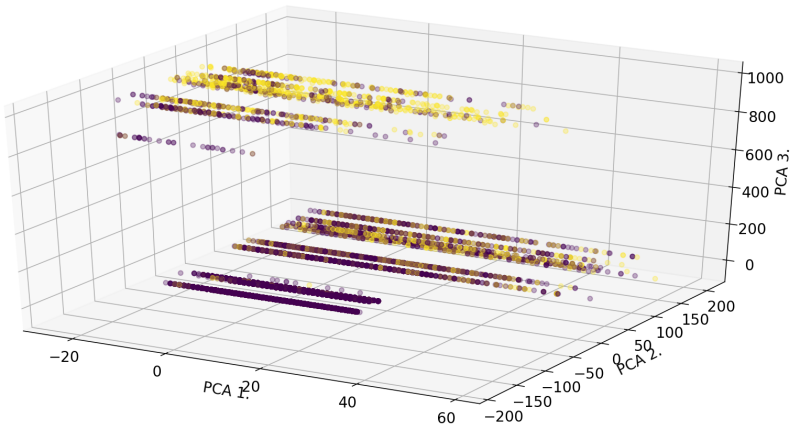

Figure 1: 3-dimensional visualization of data distribution in the first three major principal components.

For each classifier, only the best performance found after testing of various values of input parameters are reported. Table 1 summarizes the examined models and their corresponding accuracies. The data suggest that the best model to decide whether customer is willing to make a bank term deposit is Gradient Boosting Classifier as it provides the highest accuracy. To ensure the best performance, the Gradient Boosting Classifier is retrained on the full data sample (without applying PCA), the achieved accuracy is 90.2% which is better than 89.9% obtained using PCA. The importance of the

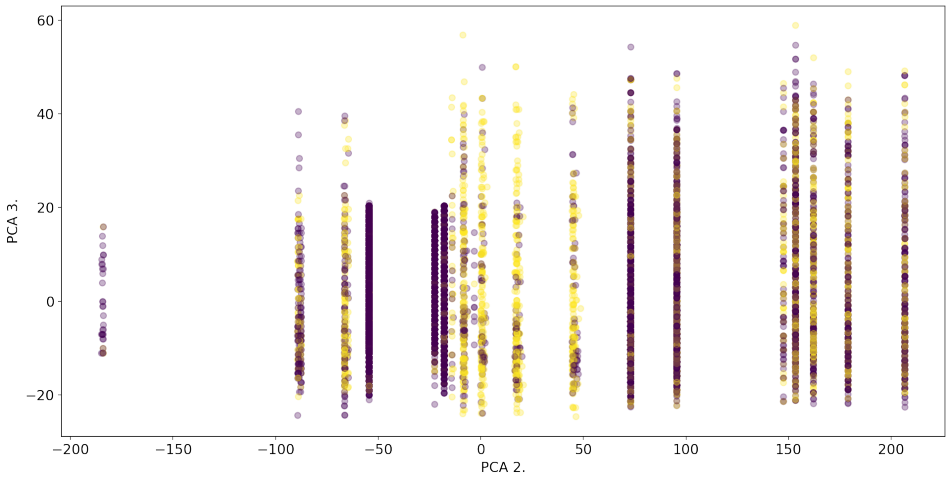

Figure 2: 2-dimensional visualization of data distribution in the 2nd. and 3rd. major principal components.

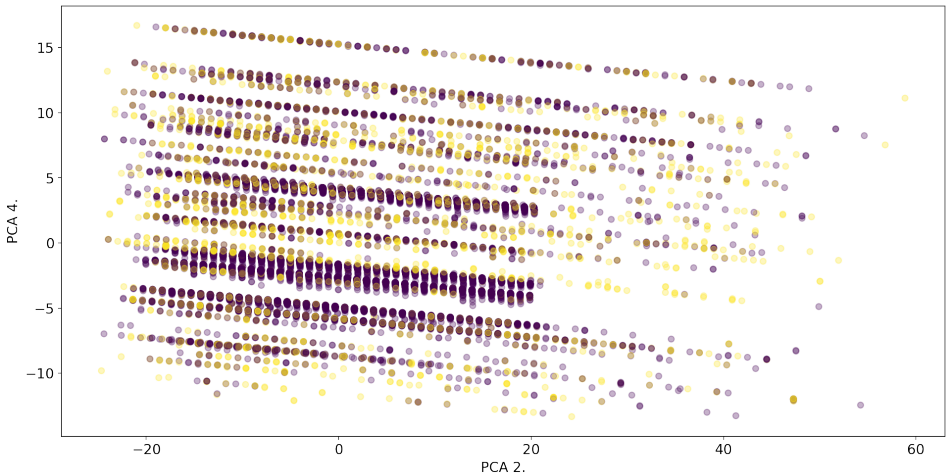

Figure 3: 2-dimensional visualization of data distribution in the 2nd. and 4th. major principal components.

Classifier	Accuracy
Random Forest Classifier	89.0%
Bagging Classifier	88.7%
Extra Trees Classifier	88.4%
Gradient Boosting Classifier	89.9%
Keras Sequential Neuron Network	89.6%

Table 1: Classifiers and corresponding accuracy.

features is estimated as feature imporatance of this classifierand these importances are summarized in table 2.

Feature	Relative imporance
Age	12.1%
Type of employment	3.4 %
Marital status	2.0%
Reached education	1.8%
Possession of credit	0.4%
Housing loan	0.5%
Personal loans	0.7%
Type of phone contact	2.2%
Month of contact	7.3%
Day of the week of the contact	2.7%
Number of days from the last contact	5.7%
Number of contacts with costumer during the campaign	0.6%
Outcomes of previous campaign	5.1%
Employment variation rate	3.1%
Consumer price index	3.6%
Consumer confidence index	7.0%
Euribor 3 month rate	20.1%
Number of employees	14.5%

Table 2: Estimated features importances for Gradient Boosting Classifier.

Based on the features importance in is possible to find the factors which may contribute to profile customers and predict their reaction to product offering. Moreover, as data suggest that some months and days of the week are more preferential as well as Euribor rate and number of employees are playing role in decision, it is favourable to conduct selling campaign only during the time during which those factors are favourable.

Conclusions

The banking data set is analyzed in order to predict costomer decisions regarding produc offer of term deposit. The Gradient Boosting Classifier is identifier is identified as the best performing classifier since it delivers the highest accuracy. However, other classifiers also deliver competitive results. The feature importance is estimated in order to possibly optimize phone call campaign in terms of selecting prefered days and year time indicated with month and economical indicators.

Bibliography

[1] S. Moro, P. Cortez and P. Rita
http://archive.ics.uci.edu/ml/datasets/Bank+Marketing(1999)

[2] S. Moro, P. Cortez and P. Rita
A Data-Driven Approach to Predict the Success of Bank Telemarketing.;

[3] Sklear: RandomForestClassifier
http://scikit-learn.org/stable/modules/generated/sklearn.ensemble.RandomForestClassifi

[4] Sklear: BaggingClassifier
http://scikit-learn.org/stable/modules/generated/sklearn.ensemble.BaggingClassifier.htm

[5] Sklear: ExtraTreesClassifier
http://scikit-learn.org/stable/modules/generated/sklearn.ensemble.ExtraTreesClassifier.h

[6] Sklear: Gradient Boosting Classifier
http://scikit-learn.org/stable/modules/generated/sklearn.ensemble.GradientBoostingClas

[7] Keras: Sequential Neuron Network
https://keras.io/;