Project being undertaken:

Bank Marketing (Campaign) --- Data Science

Group Name

The Data Team.

Members

Fawzi Khatib

Malipalema

Aurthur

Nigel

Team member details:

1. Malipalema Khang, malipalema@gmail.com, South Africa, University of the Witwatersrand, Data Science.
2. Nigel Chitere, tawanda.nigel@gmail.com, Zimbabwe, Rocapply, Data Analysis.
3. Fawzi El Khatib, fawzi_khatib@hotmail.com, Lebanon, Data Glacier, Data Science.
4. Arthur Mupfumira, artmupf@gmail.com, Zimbabwe, Data Glacier, Data Science.

Problem description

The goal of the project is to develop a model that will aid our client ABC Bank identify which clients are likely to be interested in their term deposit product before actually introducing this product into their customer base without prior research and client targeting.

Data understanding.

In order to actually get a better picture of how this data would be relevant and to who it would be beneficial towards the successful completion of the projects objectives, the following questions need to be answered:

To answer these questions we explored our data using **excel and jupyter notebook.**

The type of data we have:

We received raw and unprocessed data, which meant it had to be cleansed for machine learning suitability purposes.To fix this we used a variety of different techniques such as encoding, labelling, value dropping etc. The data was in Csv format.

Where we got the data:

The given data is a publicly available dataset available for research, which was given to us as per our project requirements. The dataset was retrieved from

1. [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, In press, http://dx.doi.org/10.1016/j.dss.2014.03.001

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.

3. P. Novais et al. (Eds.), Proceedings of the European Simulation and Modelling Conference - ESM'2011, pp. 117-121, Guimarães, Portugal, October, 2011. EUROSIS.

And is available at:

[pdf] http://hdl.handle.net/1822/14838

[bib] http://www3.dsi.uminho.pt/pcortez/bib/2011-esm-1.txt

[pdf] http://dx.doi.org/10.1016/j.dss.2014.03.001

[bib] http://www3.dsi.uminho.pt/pcortez/bib/2014-dss.txt

Data Background.

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 assess if the product (bank term deposit) would be (or not) subscribed.

How many data sets.

The file includes **two datasets**:

1. bank-additional-full.csv with all examples, ordered by date (from May 2008 to November 2010).
    - Number of Instances: 45211 for bank-full.csv (4521 for bank.csv)
    - Number of Attributes: 16 + output attribute.

2. bank-additional.csv with 10% of the examples (4119), randomly selected from bank-additional-full.csv.
    - Number of Instances: 41188 for bank-additional-full.csv
    - Number of Attributes: 20 + output attribute.

Examining the data features.

Combined the dataset has approximately close to 40 attributes.

These include **numerical** and **categorical** data.

Some of the key features vital to the building of a machine learning model to predict in the decision making include the following:

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

What are the problems in the data (number of NA values, outliers, skewed etc)

We discovered several missing values in the given data set features attributes,which were highlighted as unknown.

Fig.1 Image showing sample data missing ‘unknown’ values anomalies in the data set.

Finding the missing/null values highlight pandas calls/functions used.

Fig.2 Image showing the ‘poutcome’ feature having the most unknown values, hence we dropped this feature given it didn't have much of a bearing on the prediction model.

What approaches are you trying to apply on your data set to overcome problems like NA value, outlier etc and why?

To address the problem above we used the following techniques class labelling(dummies), deletion and imputation techniques.

Some of the data features were discrete data such as ‘pdays, previous’, hence despite dealing with the outliers it was difficult to still have a bull curve in the output graphs.

We also managed to smooth out the curves for features such as, ‘duration’, ‘balance’ ,etc.

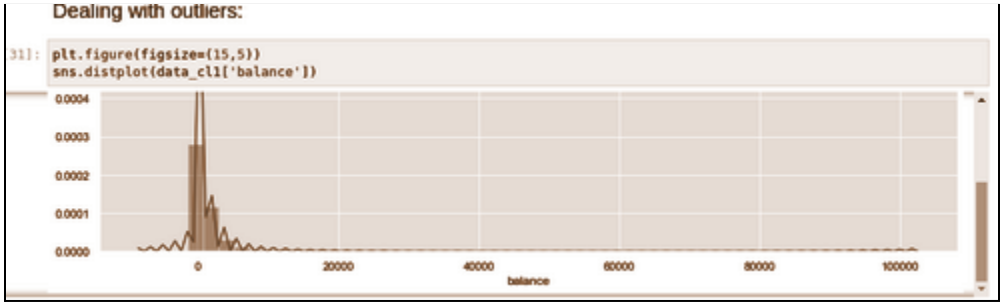

Fig.3 showing graphical representation of balance feature values before being worked on.

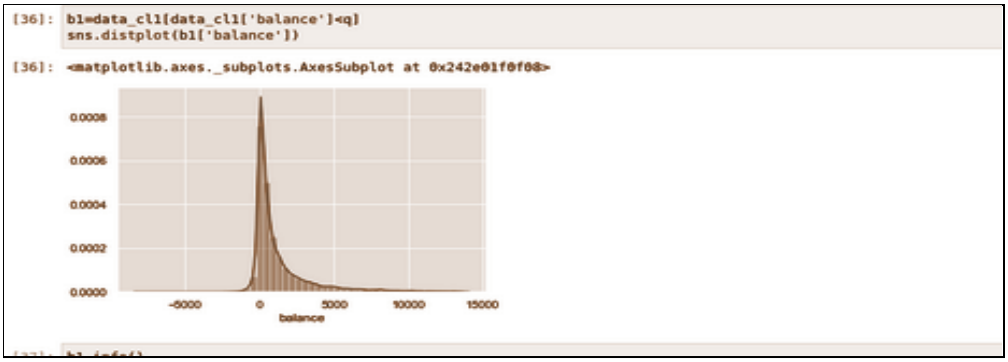

Fig.3 showing graphical representation of balance feature values before after being worked on.