Cezar Anicai

Arnaud Nguyen

I download the data and pass it through a variable.

1. We understand and visualize the data: the first rows and the statistical description for numerical data.

1. As everything is labialize, the aim is to classify the data. We want to develop models that are accurate classifying the data.

First we need to do more descriptive analysis and more visualization. For example the age and the profession determines the balance. The difficulty is to deal with categorical data. For example it is difficult to do feature selection or algo like k nearest.

Feature Selection

We want to reduce the data set.

Technics: PCA, LDA = (linear) Kernel PCA (non linear)

Statistics test with numerical data to be sure that each feature is different

Classification algo:

K nearest neighbor

Decision Tree

Naives Bayesian Classification

After we should try some NN and SVM

Then we have to combine classifier in order to find better model.

Table des matières

[Introduction 3](#_Toc36216722)

[Dataset 3](#_Toc36216723)

[Methodology 3](#_Toc36216724)

[Results 3](#_Toc36216725)

[Discussion 3](#_Toc36216726)

# Introduction

The study is focused on a marketing campaign made by a Portuguese bank. The institution want to advertise the clients by Telemarketing call and make them subscribe to a term deposit. The employees from the bank call people and ask them if they want this contract. The bank has store all the data during the study keeping the clients anonymous in order to stick to the law and respect ethical issues.

We can presume that many people do not answer the phone because they do not have the time or they are not interested or think that they are not interested. Therefore, we can wonder how we can optimize the Telemarketing.

We will try to use big data and data science methods to predict the subscription of a client doing data mining and classification. If a model is accurate, that means the employee should focus on a specific part of population in order to save time and money. This type of problem is applied for a bank in this situation. However, the study could also help other marketing problem for other companies and not only the financial sector.

# Dataset

The dataset is available online at the following address:

<http://archive.ics.uci.edu/ml/datasets/Bank+Marketing?fbclid=IwAR2WJ-X-GYE9yYJBC3B3237ccQNHUslGXyY1vay7bdTVJYxYCxReOWDLcjI>

This is a multivariate dataset with 21 variables. It has 45211 samples with numerical and categorical data. There is on output variable y to know if the client has subscribed or not. So this is a binary classification problem.

###more description

# Methodology

# Results

# Discussion