MA429

ALGORITHMIC TECHNIQUES FOR DATA MINING

SUMMATIVE PROJECT

WE need a Title

Candidates
NEED CANDIDATE NUMS
NEED CANDIDATE NUMS
NEED CANDIDATE NUMS

May 3, 2020

Contents

Abstract

1 Introduction

This report will focus on analysing data of Portuguese direct marketing campaigns of banking institutions. The data related to the success of selling long term bank deposites through phone campaigns [3]. The data provided by the UC Irvine Machine Learning repository includes input variables relating to personal information, banking information, and contact information of potential clients. It includes 41,188 observations and 21 variables.

The analysis of banking data is extremely important for many reasons. This data can be used to promote equity, reduce fraud, and build a more resiliant economy. The data that we are analysing is Portuguese bank marketing including communication, banking, and demographic data. The analysis of such data though important comes with many pitfalls, demographic data has been purposefully used in conjunction with banking data to deny access to capital to certain disadvantaged groups such as in North American redlining of the mid 20th century [1]. Even if there is no malicious intent these models can have a significant impact on who gets access and who does not and therefor it is important to understand biases and shortcomings of any model that is built as well as imperfections in the data itself.

This report sets out to classify observations as whether an observation subscribed to a term deposit or not. We will focus on four classification methods, logistic regression, random forest, support vector machines, and neural networks. First we will discuss our data and preprocessing choices. Next we will focus on the data mining methods which we will then compare the effectiveness and interoperability of each method. Finally we will summarise our findings and look at limitations of our work.

2 Data Set

2.1 Data Selection

The dataset was chosen with the express goal of being able to test data mining techniques with a goal to compare their effectiveness in a classification task. The data was selected from the UCI Machine Learning Repository as it would be well studied and documented.

We chose the bank marketing dataset ... ADD SOME DETAIL HERE

2.2 Data Description

The dimensions of the dataset are 41188 rows 21 columns.It means we have 41188 observations and 21 unprocessed variables.The data is broken into four categories, client information, current campaign contact information, previous campaign information, social-economic context indicators. It contains a mix of categorical and numerical data with full descriptions available in appendix A. The data was collected between May 2008 and November 2010 from a Portuguese banking institution.

We could see the y feature are not well distributed since we have so many no outcomes and relatively small amount of yes outcomes. It is a reflection of the reality for people to subscribe to a deposite. The ratio is approximately 8:1 when comparing no to yes.

3 Pre-processing

3.1 Missing Values

We could see there are many unknowns in the variables: job, maritial, education, default, housing loan, personal loan, for the education and default variables,the missing values are relatively large. Therefore we are quite intereseted in how are the missing values distributed. After learning the missing pattern, we could possibly find how to clean this dataset.

3.2 Selecting Features

4 Data Mining Methods

4.1 Random Forest

4.2 Logistic Regression

Logistic regression is a better choice than a linear regression model in this case because our outcomes are strictly binary [2]. With a linear regression we could end up with a model that does not differentiate the categories as distinct while logistic takes this into account not allowing predictions half way between "yes" and "no" in our case. With logistic regression we code our categories as 0 or 1 and our prediction will be limited to the range $[0, 1]$. Using a non-linear logistic function our prediction will fit an S-curve which is how we keep the prediction between $[0, 1]$ [2].

4.3 Support Vector Machines

4.4 Neural Networks

5 Summary

6 Limitations

7 Conclusions

References

[1] Richard Harris and Doris Forrester. The Suburban Origins of Redlining: A Canadian Case Study, 1935-54. *Urban Studies*, 40(13):2661–2686, December 2003.

[2] Gareth James, Daniela Witten, Trevor Hastie, and Robert Tibshirani, editors. *An introduction to statistical learning: with applications in R.* Number 103 in Springer texts in statistics. Springer, New York, 2013.

[3] S. Moro, R. Laureano, and P. Cortez. *A Data-Driven Approach to Predict the Success of Bank Telemarketing. Decision Support Systems.* Elsevier, 2014.

A Table of Features

Client Data		
Feature	Description	Data Type
age	Age of a person	numeric
job	occupation of individual	categorical
marital	marital status	categorical
education	education level	categorical
default	has credit in default	categorical
housing	has a housing loan	categorical
loan	has a personal loan	categorical
Current Campaign Data		
contact	communication type	numeric
month	month contacted	categorical
day_of_week	dat contacted	categorical
duration	contact duration in seconds	numeric
Previous Campaign Data		
campaign		numeric
pdays		numeric
previous		numeric
poutcome		categorical
Socio-Economic Climate Data		
emp.var.rate		numeric
cons.price.idx		numeric
cons.conf.idx		numeric
euribor3m		numeric
nr.employed		numeric
Output Variable		
y	has client subscribed	categorical