

MA429  
ALGORITHMIC TECHNIQUES FOR DATA MINING

SUMMATIVE PROJECT

**WE need a Title**

Candidates  
*NEED CANDIDATE NUMS*  
*NEED CANDIDATE NUMS*  
*43058*

May 3, 2020

---

# Contents

|          |                                   |          |
|----------|-----------------------------------|----------|
| <b>1</b> | <b>Introduction</b>               | <b>3</b> |
| <b>2</b> | <b>Data Set</b>                   | <b>3</b> |
| 2.1      | Data Selection . . . . .          | 3        |
| 2.2      | Data Description . . . . .        | 3        |
| <b>3</b> | <b>Pre-processing</b>             | <b>4</b> |
| 3.1      | Selecting Features . . . . .      | 4        |
| 3.2      | Missing Values . . . . .          | 4        |
| <b>4</b> | <b>Data Mining Methods</b>        | <b>4</b> |
| 4.1      | Random Forest . . . . .           | 4        |
| 4.2      | Logistic Regression . . . . .     | 4        |
| 4.3      | Support Vector Machines . . . . . | 4        |
| 4.4      | Neural Networks . . . . .         | 4        |
| <b>5</b> | <b>Summary</b>                    | <b>4</b> |
| <b>6</b> | <b>Limitations</b>                | <b>4</b> |
| <b>7</b> | <b>Conclusions</b>                | <b>4</b> |
| <b>A</b> | <b>Table of Features</b>          | <b>6</b> |

---

## 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 deposits through phone campaigns [2]. 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 resilient 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 therefore 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 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.

---

|          |                            |
|----------|----------------------------|
| <b>3</b> | <b>Pre-processing</b>      |
| 3.1      | Selecting Features         |
| 3.2      | Missing Values             |
| <b>4</b> | <b>Data Mining Methods</b> |
| 4.1      | Random Forest              |
| 4.2      | Logistic Regression        |
| 4.3      | Support Vector Machines    |
| 4.4      | Neural Networks            |
| <b>5</b> | <b>Summary</b>             |
| <b>6</b> | <b>Limitations</b>         |
| <b>7</b> | <b>Conclusions</b>         |

---

## 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] 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 |