Mathematics in Machine Learning

Bank Marketing Dataset Analysis

Matteo Merlo
s287576

June 8, 2022

Politecnico di Torino

Abstract

In this project I introduced possible approaches for preparing the dataset and building a classification model for bank marketing data obtained from [1]. Specifically, methodology consists of preprocessing the data by cleaning, encoding, scaling, and balancing, and comparing built classification models, namely Decision Tree, Random Forest, Logistic Regression, Support Vector Machines, and AdaBoost by predefined metrics, such as accuracy and f1-score. The approaches achieve overall adequate results.

1 Introduction

Bank sector is one of the demanding fields of machine learning applications. The dataset used in this project specifically gathered by a Portuguese banking institution to improve the marketing campaigns which at a time was undergone by the phone calls. At the end of each call customer's answer to accepting bank term deposit (yes or no) is registered. Purpose of the machine learning is to predict this outcome so that there is no need to waste time on old manners.

2 Overview of the data

3 Data exploration

3.1 Missing Values and Data Types

First thing we can do is to see if the data has any missing values. However it has no null or NAN values, it has some instances that has 'unknown' in one or more of its features. Features of 'job', 'marital', 'education', 'default', 'housing', 'loan' have 330, 80, 1731, 8597, 990, 990 'unknown's respectively.

Moreover, information about columns' types are extracted as follows:

- Binary columns: ['default', 'housing', 'loan', 'y']
- Categorical columns: ['job', 'marital', 'education', 'contact', 'month', 'day of week', 'poutcome']
- Numerical columns: ['age', 'duration', 'campaign', 'pdays', 'previous', 'emp.var.rate', 'cons.price.idx', 'cons.conf.idx', 'euribor3m', 'nr.employed']

3.2 Data distribution

References