RGU

BSc (Hons) Artificial Intelligence and Data Science

**Module: CM2604 Machine Learning**

Coursework Report

**Module Leader: Mr. Sahan Priyanayana**

RGU Student ID : 2330906

IIT Student ID : 20230145

Student Name : P. K. S. D. C. Bandara

Contents

**No table of contents entries found.**

# Introduction

This report demonstrates the outcomes and analysis of the machine learning module coursework. Binary classification problem which predicts whether a client will subscribe to a term deposit based on historical bank marketing data is the primary goal of this project. The dataset called “Bank Marketing” from UCI machine learning repository was used as the primary dataset for the classification training and testing.

To achieve the above objective, Random Forest and Neural Network machine learning models were implemented and evaluated. This report provides a clear overview on data preprocessing, model training and evaluation.

# Corpus Preparation

Initially dataset was imported using Pandas library and separated to columns

## **Handling “Unknown” values**

Dataset did not have any null values. But it had values which were named as “unknown” to represent null values. First values that was unknown was converted to null values. All together there were 12718 rows that contains null values. All null rows were dropped except rows that contain null values in the default column.

## **Handling Duplicates**

There were 12 duplicate rows in the dataset. All duplicate rows were dropped.

## **Handling Outliers**

Outliers was identified by examining boxplots of numeric columns. After the inspection was done outliers were identified in 3 features (age, campaign, duration). Outliers were removed respect to the 1st and 3rd quantiles of each feature.

## **Standardization**

Numeric features were transformed to a constant format across the dataset. StandardScaler (z-Score) was used in transforming data.

## **Encoding Categorical columns**

All the categorical columns were transformed into numerical format using encoding. One-hot encoding and label encoding was used in this process.

## **Handling Default column**

Default column was dropped due to the extreme imbalance. Default column was an extremely imbalanced column which has Boolean values. Because of the extreme imbalance this column will not help in resulting the final outcome in the training process. Considering those, Default column was dropped.

## **Handling Duration column**

In the text file of the dataset it is given that the attribute highly affects the output target (duration=0 -> output =0). Such scenario is not suitable for realistic predictive model. Dataset contained 4 rows that duration was equal to 0. Those rows were removed.

## **Handling the target variable imbalance**

Target variable y also had a high imbalance which was 88.73% and 11.26% this might cause a biasness towards one class. Oversampleing method SMOT was used in this scenario to balance the dataset.

# Solution Methodology

Well known supervised learning algorithms (***Random Forest*** & ***Neural networks***) were used for this project solution. Both of these models has their own strengths in these kind of classification problems.

## **Random Forest**

Random Forest is made out of many decision trees where the mean output of each tree is considered as the prediction. Such model is suitable for this kind of project for several reasons.

* Handling overfitting
* Nonlinear approach
* High predictive accuracy with minimal hyper parameter tuning

## **Neural Network**

Neural network is a combination of neurons which has a weight, bias and an activation function. In this specific problem rectified linear unit and Sigmoid functions are used as activation functions of hidden layers and output layer.

* Effective for High-Dimensional Data
* Ability to Model Complex Patterns
* Normalization Benefits

# Evaluation Methodology

After the model training, evaluation is one of the most important tasks that has to follow. Metrics such as accuracy, precision, recall and f1-Score is used in evaluating the project