STW7082CEM: Big Data Management and Data Visualization

Submitted To:
Siddhartha Neupane

Submitted by:
Sandeep Shrestha

Project Proposal on Binary Classification of Bank Marketing Data Using Pyspark.

Module: STW7082CEM: Big Data Management and Data Visualization

Submitted By: Sandeep Shrestha

Softwarica College ID: 230127

Introduction

The data relates to a Portuguese financial institution's direct marketing initiatives. The Bank Marketing Dataset, which is being made accessible here, covers a broad spectrum of attributes and variables meant to provide insight into customer demographics, financial practices, and the effectiveness of marketing campaigns. The primary objective of this dataset is to provide analysis aimed at enhancing operational efficiency. Predicting whether a consumer will open a term deposit is the aim of classification.

Objective

This dataset is a useful tool for comprehending transactional dynamics, consumer behavior, and important for making wise decisions in the banking industry. The objective of categorization is to forecast whether the customer will sign up for a term deposit.

Dataset Name

bank.csvs

Data Sources

The data has been obtained from Kaggle where we can find numerous data sets.

Dataset Link

https://www.kaggle.com/code/palmer0/binary-classification-with-pyspark-and-mllib/notebook#Machine-Learning-with-PySpark-and-MLlib:-Solving-a-Binary-Classification-Problem

Dataset Description

The chosen "bank.csv" file has marketing campaign data in it. The dataset has the following fields: campaign, pdays, previpous, poutcome, deposit, age, job, education, balance, housing, loan, contact, day, month, duration, and campaign. The dataset

includes the following datatypes: date, float, string, and integer. There are 11163 rows and 17 columns in all.