
# Week #8 Deliverables PDF Document

## Team member's details:

Group Name: Shiny Star Data Specialists

Carmelo R. Casiraro, cralph31@gmail.com, USA, Farmingdale State University, Data Analyst

Fenil Mavani, fenilmavani5757@gmail.com, UK, University of West London, Data Science

Nazri, nazrianwar10@gmail.com, London, UK, University of Greenwich, Data Science

## Problem description

- **What features exist in the data that would make training a model difficult?**
- **Some models are more restrictive with skewness than others.**

# Data understanding

# What type of data you have got for analysis?

***Column Descriptions***

- age (Integer): Age of the customer ● job (String): Type of job.
- marital (String): Marital status.
- education (String): Education level.
- default (Boolean): Has credit in default?
- balance (Integer): Account balance.
- housing (Boolean): Has housing loan?
- loan (Boolean): Has personal loan?
- contact (String): Contact communication type.
- day (Integer): Last contact day of the month.
- month (String): Last contact month of the year.
- duration (Integer): Last contact duration (in seconds).
- campaign (Integer): Number of contacts performed during this campaign for this client.
- pdays (Integer): Number of days that passed after the client was last contacted from a previous campaign.
- previous (Integer): Number of contacts performed before this campaign for this client.
- poutcome (Integer): Outcome of the previous marketing campaign.
- y (Boolean): Has the client subscribed to a term deposit?

**What are the problems in the data ( number of NA values, outliers , skewed etc)?**

1) Booleans are represented as yes/no strings

2) Null values are represented by the string 'unknown'

3) Unknown values in pdays seem to be represented by -1- we don't know what this means?

4) There are no NA values.

5) Data is semicolon-delimited, not comma-delimited

6) The distribution of labels is skewed heavily to one side, the model may be biased toward that result

7) There are a significant number of outliers in the numerical columns

8) Several numerical columns are heavily positively skewed

**What approaches you are trying to apply on your data set to overcome problems like NA value, outlier etc and why?**

- If we are using a logistic regression model, than we should handle skewness with other methods.
- If we are using a decision tree, then it is not as important.
- To overcome the boolean problem, when reading the file- need to add logic to parse the yes/no into boolean true or false.
- Possible solution, use the mean of each column or look at other similar data to predict to a estimated value for the column- (Technique: K nearest neighbors)if time permits we can implement this technique.
- The argument `sep=";"` must be passed in each call to pd.read_csv()

- **Pre-process the test and training data to remove "no" rows (i.e. take a random sample) until the output distribution is about 50% yes and 50% no**
- **We are shooting for 1-5% outliers so the training model is generated more accurately.**
- **How do we get in that range?**

  **-> Creating and applying a log transformation- replacing values. It would preserve information and reduce space in values.**

# Github Repo link

Data-Science-Intern-at-Data-Glacier/Week8 Project- Bank Marketing Campaign at main · NazriJasmal/Data-Science-Intern-at-Data-Glacier (github.com)