Week #9 Deliverables PDF Document

Team member's details:

Group Name: Shiny Star Data Specialists

Carmelo R. Casiraro, USA, Farmingdale State University, Data Analyst

Fenil Mavani, UK, University of West London, Data Science

Nazri, London, UK, University of Greenwich, MSc Big Data and Business Intelligence

Problem description

- Implementing the solutions in Week 8 deliverables
- Click Here to View Week 8 Deliverables
- Url: https://docs.google.com/document/d/1JtSPZYQzzqRKpLK50nXwF5J-DQuR4AkAWmd0DUxa_x4/edit?usp=drivesdc

Data cleansing and transformation done on the data.

- **Prior to cleansing the columns had the following skewness:**

```
Column: age
Skewness: 0.6993
Interpretation: Moderately Positively Skewed

Column: balance
Skewness: 6.5942
Interpretation: Highly Positively Skewed

Column: day
Skewness: 0.0946
Interpretation: Approximately Symmetric (Slightly Positive)

Column: duration
Skewness: 2.7715
Interpretation: Highly Positively Skewed

Column: campaign
Skewness: 4.7423
Interpretation: Highly Positively Skewed

Column: pdays
Skewness: 2.7162
Interpretation: Highly Positively Skewed

Column: previous
Skewness: 5.8733
Interpretation: Highly Positively Skewed
['balance', 'duration', 'campaign', 'pdays', 'previous']
```

- **After cleansing the data the columns had the following skewness:**

```
Column: age
Skewness: 0.6993
Interpretation: Moderately Positively Skewed

Column: balance
Skewness: 0.1212
Interpretation: Approximately Symmetric (Slightly Positive)

Column: day
Skewness: 0.0946
Interpretation: Approximately Symmetric (Slightly Positive)

Column: duration
Skewness: -0.4677
Interpretation: Approximately Symmetric (Slightly Negative)

Column: campaign
Skewness: 0.7649
Interpretation: Moderately Positively Skewed

Column: pdays
Skewness: 1.6881
Interpretation: Highly Positively Skewed

Column: previous
Skewness: 2.2485
Interpretation: Highly Positively Skewed
['pdays', 'previous']
```

- **Checking for missing values, with code, we were able to figure out there are no null values:**

5]:
```
## Checking for missing values
import pandas as pd
df = pd.read_csv('bank.csv', sep=';')
missing_values = df.isnull().any(axis=1).sum() #it will check inside the columns of the dataframes

print (missing_values)
```

```
0
```

– However in the pdays column null values seem to be encoded by the number -1 because you can't contact someone after minus one days.

- This will be handled during construction of the model by creating a binary feature. When we make the model- whenever the model reads a -1 it will be interpreted as a false.

Try at least 2 techniques to clean the data

- **We are looking for outliers in the following columns: age, balance, duration, campaign, pdays and previous.**
- **Outliers techniques solution: for most numerical columns we impute the mean. For pdays and previous we ignore values that encode false.**

```
Outliers in column 'age':
Number of outliers: 38
Percentage of outliers: 0.84%

Outliers in column 'balance':
Number of outliers: 506
Percentage of outliers: 11.19%

Outliers in column 'day':
Number of outliers: 0
Percentage of outliers: 0.00%

Outliers in column 'duration':
Number of outliers: 330
Percentage of outliers: 7.30%

Outliers in column 'campaign':
Number of outliers: 318
Percentage of outliers: 7.03%

Outliers in column 'pdays':
Number of outliers: 7
Percentage of outliers: 0.15%

Outliers in column 'previous':
Number of outliers: 34
Percentage of outliers: 0.75%
```

- **High percentage of outliers before imputing any techniques.**

```
Outliers in column 'age':
Number of outliers: 38
Percentage of outliers: 0.84%

Outliers in column 'balance':
Number of outliers: 156
Percentage of outliers: 3.45%

Outliers in column 'day':
Number of outliers: 0
Percentage of outliers: 0.00%

Outliers in column 'duration':
Number of outliers: 34
Percentage of outliers: 0.75%

Outliers in column 'campaign':
Number of outliers: 0
Percentage of outliers: 0.00%

Outliers in column 'pdays':
Number of outliers: 119
Percentage of outliers: 2.63%

Outliers in column 'previous':
Number of outliers: 0
Percentage of outliers: 0.00%
```

- **Low percentage of outliers after we imputed the mean.**

Github Repo link

https://github.com/cralph31/Data-Glacier-Final-Group-Project-Weeks-7-12-Deliverables