Data Wrangling

Capstone Project -1

Prediction of telemarketing conversion probability using bank user dataset

1. **What kind of cleaning steps did you perform?**

Data exploration —

- **Columns : bank_data.head()**
- **Unique values in a column: bank_data.info()**
- **Describe: bank_data.describe()**

There are 4521 different rows in the dataset

- **Duplicates: len(bank_data[bank_data.duplicated()])**

2. **How did you deal with missing values, if any?**

Dealing with missing values:

— Quantifying missing values per column, filling & dropping missing values.

bank_data.isnull().sum()

bank_data.isna().sum()

No missing values in the bank_dataset

3. **Were there outliers, and how did you handle them?**

To find Outliers:

```
sns.boxplot(x=bank_data['age'])
sns.boxplot(x=bank_data['balance'])
sns.boxplot(x=bank_data['day'])
sns.boxplot(x=bank_data['duration'])
sns.boxplot(x=bank_data['campaign'])
sns.boxplot(x=bank_data['pdays'])
sns.boxplot(x=bank_data['previous'])
```

To remove Outlier:

Outliers found in all the numerical attribute (age, balance, duration, campaign, Pdays,previous), **except day attribute.**

```
sns.boxplot(x=bank_data['age'])

q3 = bank_data['age'].quantile(0.75)
q3
q1 = bank_data["age"].quantile(0.25)
q1
iqr = q3-q1
iqr
upper_limit= q3+(1.5*iqr)
upper_limit
lower_limit= q1-(1.5*iqr)
lower_limit
bank_data.loc[bank_data['age'] < (q1 - 1.5 * iqr),['age']] = q1 - 1.5 * iqr
bank_data.loc[bank_data['age'] <= (q1 - 1.5 * iqr),['age']]
bank_data.loc[bank_data['age'] > (q3 + 1.5 * iqr),['age']] = q3 + 1.5 * iqr
```

```
bank_data.loc[bank_data['age'] >= (q3 + 1.5 * iqr),['age']]
```

After removed outliers:

```
sns.boxplot(x=bank_data['age'])
```

Similarly for all the attributes except days which is not having outliers.