DELIVERABLE WEEK – 10

Group Name	: Walther Rathenau Team
Specialization	: Data Science

Team members:

1. Name	: Mohammed Maqsood
Email	: mohammedmaqsood48@gmail.com
Country	: Germany
College	: Otto von Guericke University
Specialization	: Chemical and Energy Engineering

2. Name	: Davangam Sreedhar Saikiran
Email	: saikiran.davangams@gmail.com
Country	: Germany
College	: Otto von Guericke University
Specialization	: Chemical and Energy Engineering

Problem description

ABC Bank wants to sell its term deposit product to customers and before launching the product they want to develop a model which help them in understanding whether a particular customer will buy their product or not (based on customer's past interaction with bank or other Financial Institution).

Data Cleaning and Transformation

The dataset was checked for missing values, duplicates, outliers, and skewness. No missing values or duplicates were found. Summary statistics such as the mean, standard deviation, distribution, and kurtosis, skewness were checked. Many machine learning algorithms do not understand categorical data. Hence, they must be converted into integer values. The unknown values are handled and columns with categorical data are changed to Boolean or integer values. The dataset is cleaned and any ML model can use it for training and prediction.

Github repo link:

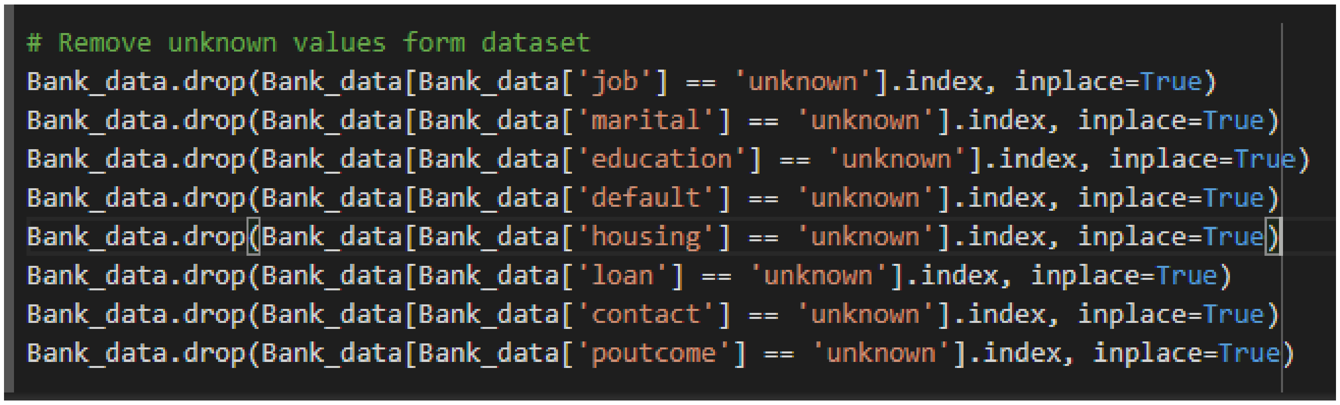
https://github.com/Maqsood8/Group-Project-Bank-Marketing.git

Results and approaches

- EDA was performed for continuous and categorical variables
- Methods for data cleaning and transformation
  - Step 1: Removed unknown values from the Dataset

    Rows (Job, Marital, Education, Default, Housing, and Loan) with "unknown" valuesare removed from the Dataset using drop function as shown in Figure 1.

Figure 1 – Python Code – Unknown Values



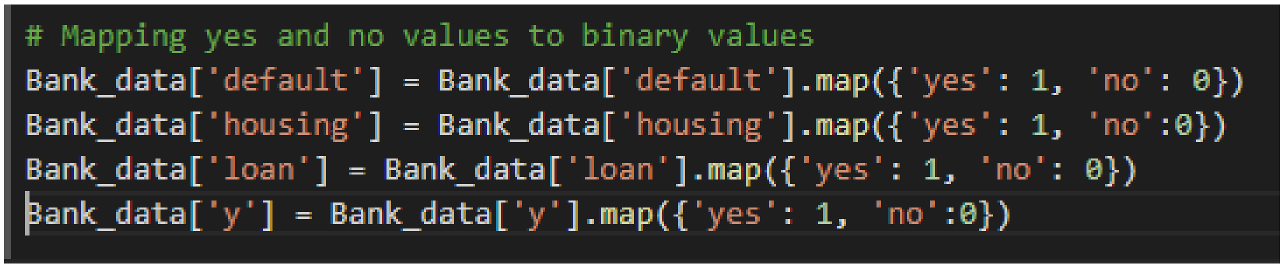
```
# Remove unknown values form dataset
Bank_data.drop(Bank_data[Bank_data['job'] == 'unknown'].index, inplace=True)
Bank_data.drop(Bank_data[Bank_data['marital'] == 'unknown'].index, inplace=True)
Bank_data.drop(Bank_data[Bank_data['education'] == 'unknown'].index, inplace=True)
Bank_data.drop(Bank_data[Bank_data['default'] == 'unknown'].index, inplace=True)
Bank_data.drop(Bank_data[Bank_data['housing'] == 'unknown'].index, inplace=True)
Bank_data.drop(Bank_data[Bank_data['loan'] == 'unknown'].index, inplace=True)
Bank_data.drop(Bank_data[Bank_data['contact'] == 'unknown'].index, inplace=True)
Bank_data.drop(Bank_data[Bank_data['poutcome'] == 'unknown'].index, inplace=True)
```

Figure 1 - Before Transformation

- Step 2: Binning of outliers

  Rows (Default, Housing, y, and Loan) which has only two values are replaced with 1s and 0s using map function as shown in figure 2.

Figure 2 – Python Code – Mapping

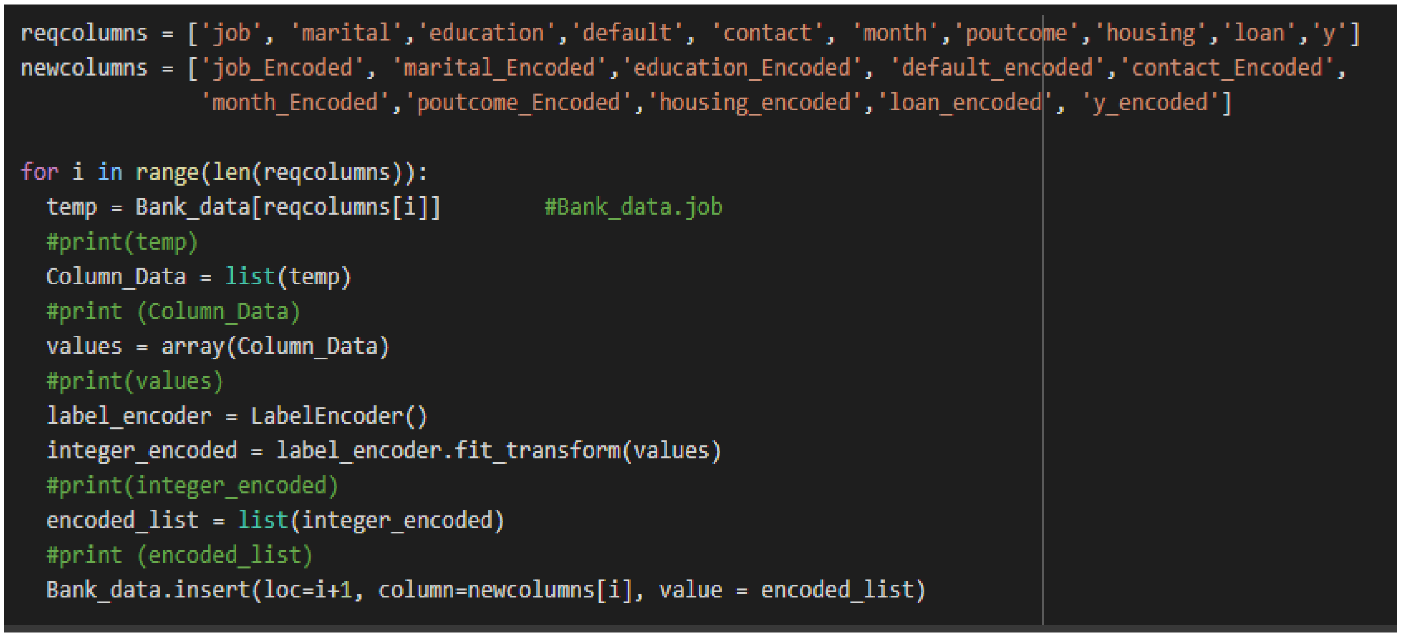


```python
# Mapping yes and no values to binary values
Bank_data['default'] = Bank_data['default'].map({'yes': 1, 'no': 0})
Bank_data['housing'] = Bank_data['housing'].map({'yes': 1, 'no':0})
Bank_data['loan'] = Bank_data['loan'].map({'yes': 1, 'no': 0})
Bank_data['y'] = Bank_data['y'].map({'yes': 1, 'no':0})
```

- Step 3: One hot encoding

  Categorical values are mapped to integer values using one hot encoding technique using scikit learn as shown in figure 3.

Figure 3 – Python Code – One Hot Encoding



```python
reqcolumns = ['job', 'marital','education','default', 'contact', 'month','poutcome','housing','loan','y']
newcolumns = ['job_Encoded', 'marital_Encoded','education_Encoded', 'default_encoded','contact_Encoded',
              'month_Encoded','poutcome_Encoded','housing_encoded','loan_encoded', 'y_encoded']

for i in range(len(reqcolumns)):
  temp = Bank_data[reqcolumns[i]]        #Bank_data.job
  #print(temp)
  Column_Data = list(temp)
  #print (Column_Data)
  values = array(Column_Data)
  #print(values)
  label_encoder = LabelEncoder()
  integer_encoded = label_encoder.fit_transform(values)
  #print(integer_encoded)
  encoded_list = list(integer_encoded)
  #print (encoded_list)
  Bank_data.insert(loc=i+1, column=newcolumns[i], value = encoded_list)
```

The dataset can now be used to train machine learning models and predict whether a particularcostumer will buy their product or not.