Assignment 4 - Day 7

Data Preprocessing

Atharva Ramgirkar 19BCE0114

In [ ]:

Table of Content

In [ ]:

Importing Libraries

In [1]:
```
import pandas as pd
import numpy as np
```

In [2]:
```
from sklearn.preprocessing import LabelEncoder
from sklearn.preprocessing import OneHotEncoder
from sklearn.model_selection import train_test_split
from sklearn.preprocessing import StandardScaler
from sklearn.compose import ColumnTransformer
```

Getting the Data

In [211]:
```
df = pd.read_csv("bank.csv")
```

In [5]:
```
# Making copy of original data
df_ori = df.copy()
```

Data Overview

In [4]:
```
df.head()
```

Out[4]:

	age	job	marital	education	default	balance	housing	loan	contact	day	month	duration	cam
0	59	admin.	married	secondary	no	2343	yes	no	unknown	5	may	1042	
1	56	admin.	married	secondary	no	45	no	no	unknown	5	may	1467	
2	41	technician	married	secondary	no	1270	yes	no	unknown	5	may	1389	
3	55	services	married	secondary	no	2476	yes	no	unknown	5	may	579	
4	54	admin.	married	tertiary	no	184	no	no	unknown	5	may	673	

In [6]:
```
df.isnull().sum()
```

Out[6]:
```
age          0
job          0
marital      0
education    0
default      0
balance      0
housing      0
loan         0
contact      0
day          0
month        0
duration     0
campaign     0
pdays        0
previous     0
poutcome     0
deposit      0
dtype: int64
```

Encoding

Binary Encoding

In [8]:
```
df['default'].value_counts()
```

Out[8]:
```
no     10994
yes      168
Name: default, dtype: int64
```

In [9]:
```
df['housing'].value_counts()
```

Out[9]:
```
no     5881
yes    5281
Name: housing, dtype: int64
```

In [10]:
```
df['loan'].value_counts()
```

Out[10]:
```
no     9702
yes    1460
Name: loan, dtype: int64
```

In [222]:
```
df['default'] = np.where(df['default']=="yes",1,0)
df['housing'] = np.where(df['housing']=="yes",1,0)
df['loan'] = np.where(df['loan']=="yes",1,0)
```

In [223]:
```
df.head()
```

Out[223]:

	age	job	marital	education	default	balance	housing	loan	contact	day	month	duration	cam
0	59	admin.	married	secondary	0	2343	1	0	unknown	5	may	1042	
1	56	admin.	married	secondary	0	45	0	0	unknown	5	may	1467	
2	41	technician	married	secondary	0	1270	1	0	unknown	5	may	1389	
3	55	services	married	secondary	0	2476	1	0	unknown	5	may	579	
4	54	admin.	married	tertiary	0	184	0	0	unknown	5	may	673	

Multilabel Encoding

In [18]:
```
list(df['job'].unique())
```

Out[18]:
```
['admin.',
 'technician',
 'services',
 'management',
 'retired',
 'blue-collar',
 'unemployed',
 'entrepreneur',
 'housemaid',
 'unknown',
 'self-employed',
 'student']
```

In [24]:
```
for i in list(df['job'].unique()):
    df['job_'+i] = np.where(df['job']==i,1,0)
```

In [28]:
```
df.drop(columns=['job'],inplace=True)
```

In [29]:
```
df.head()
```

Out[29]:

	age	marital	education	default	balance	housing	loan	contact	day	month	...	job_services	job_man
0	59	married	secondary	0	2343	1	0	unknown	5	may	...	0	
1	56	married	secondary	0	45	0	0	unknown	5	may	...	0	
2	41	married	secondary	0	1270	1	0	unknown	5	may	...	0	
3	55	married	secondary	0	2476	1	0	unknown	5	may	...	1	
4	54	married	tertiary	0	184	0	0	unknown	5	may	...	0	

5 rows × 28 columns

In [30]:
```
for i in list(df['marital'].unique()):
    df['marital_'+i] = np.where(df['marital']==i,1,0)
df.drop(columns=['marital'],inplace=True)

for i in list(df['education'].unique()):
    df['education_'+i] = np.where(df['education']==i,1,0)
df.drop(columns=['education'],inplace=True)

for i in list(df['contact'].unique()):
    df['contact_'+i] = np.where(df['contact']==i,1,0)
df.drop(columns=['contact'],inplace=True)

for i in list(df['month'].unique()):
    df['month_'+i] = np.where(df['month']==i,1,0)
df.drop(columns=['month'],inplace=True)

for i in list(df['poutcome'].unique()):
    df['poutcome_'+i] = np.where(df['poutcome']==i,1,0)
df.drop(columns=['poutcome'],inplace=True)
```

In [31]:
```
pd.set_option('display.max_columns', None)
```

In [32]:
```
df.head()
```

Out[32]:

	age	default	balance	housing	loan	day	duration	campaign	pdays	previous	deposit	job_admin.	job
0	59	0	2343	1	0	5	1042	1	-1	0	yes	1	
1	56	0	45	0	0	5	1467	1	-1	0	yes	1	
2	41	0	1270	1	0	5	1389	1	-1	0	yes	0	
3	55	0	2476	1	0	5	579	1	-1	0	yes	0	
4	54	0	184	0	0	5	673	2	-1	0	yes	1	

In [ ]:

Features and Target Split

In [33]:
```
X = df.drop(columns=['deposit'])
```

In [34]:
```
y = df['deposit']
```

Train Test Split

In [35]:
```
X_train,X_test,y_train,y_test = train_test_split(X,
                                                 y,
                                                 test_size=0.3,
                                                 random_state=0,
                                                 stratify=y)
```

In [36]:
```
X
```

Out[36]:

	age	default	balance	housing	loan	day	duration	campaign	pdays	previous	job_admin.	job_tech
0	59	0	2343	1	0	5	1042	1	-1	0	1	
1	56	0	45	0	0	5	1467	1	-1	0	1	
2	41	0	1270	1	0	5	1389	1	-1	0	0	
3	55	0	2476	1	0	5	579	1	-1	0	0	
4	54	0	184	0	0	5	673	2	-1	0	1	
...	...	...	...	...	...	...	...	...	...	...	...	...
11157	33	0	1	1	0	20	257	1	-1	0	0	
11158	39	0	733	0	0	16	83	4	-1	0	0	
11159	32	0	29	0	0	19	156	2	-1	0	0	
11160	43	0	0	0	1	8	9	2	172	5	0	
11161	34	0	0	0	0	9	628	1	-1	0	0	

11162 rows × 48 columns

Scaling the Data

In [37]:
```
scaler = StandardScaler()
# transform data
scaled = scaler.fit_transform(X)
```

In [42]:
```
scaled_df = pd.DataFrame(scaled)
```

In [47]:
```
scaled_df.columns = X.columns
```

In [48]:
```
scaled_df
```

Out[48]:

	age	default	balance	housing	loan	day	duration	campaign	pdays	previc
0	1.491505	-0.123617	0.252525	1.052280	-0.387923	-1.265746	1.930226	-0.554168	-0.481184	-0.363
1	1.239676	-0.123617	-0.459974	-0.947616	-0.387923	-1.265746	3.154612	-0.554168	-0.481184	-0.363
2	-0.019470	-0.123617	-0.080160	1.052280	-0.387923	-1.265746	2.929901	-0.554168	-0.481184	-0.363
3	1.155733	-0.123617	0.293762	1.052280	-0.387923	-1.265746	0.596366	-0.554168	-0.481184	-0.363
4	1.071790	-0.123617	-0.416876	-0.947616	-0.387923	-1.265746	0.867171	-0.186785	-0.481184	-0.363
...	...	...	...	...	...	...	...	...	...	...
11157	-0.691015	-0.123617	-0.473616	1.052280	-0.387923	0.511650	-0.331287	-0.554168	-0.481184	-0.363
11158	-0.187357	-0.123617	-0.246658	-0.947616	-0.387923	0.040612	-0.832564	0.547981	-0.481184	-0.363
11159	-0.774958	-0.123617	-0.464934	-0.947616	-0.387923	0.396891	-0.622258	-0.186785	-0.481184	-0.363
11160	0.148416	-0.123617	-0.473926	-0.947616	2.577830	-0.909466	-1.045752	-0.186785	1.109571	1.818
11161	-0.607072	-0.123617	-0.473926	-0.947616	-0.387923	-0.790707	0.737530	-0.554168	-0.481184	-0.363

11162 rows × 48 columns

In [49]:
```
y
```

Out[49]:
```
0        yes
1        yes
2        yes
3        yes
4        yes
        ...
11157     no
11158     no
11159     no
11160     no
11161     no
Name: deposit, Length: 11162, dtype: object
```

In [ ]: