Week 9: Data Cleansing and Transformation

Group Name: Data Visionaries

Name: Abdukhakimov Asatilla

Email: tremendous54@gmail.com

Country: Uzbekistan

Company: Data Glacier

Specialization: Data Science

Problem Description

ABC Bank wants to sell its term deposit product to customers. Before launching the product, the bank aims to develop a model to understand whether a particular customer will buy their product or not, based on the customer's past interaction with the bank or other financial institutions.

Data Cleansing and Transformation

Import libraries

[1]:
```
import pandas as pd
import numpy as np
import matplotlib.pyplot as plt
import seaborn as sns
import warnings
warnings.filterwarnings("ignore")
```

Let’s read the dataset

[2]:
```
df = pd.read_csv('bank-full.csv', sep=';')
```

[3]:
```
df.head()
```

[3]:

	age	job	marital	education	default	balance	housing	loan	contact	day	month	duration	campaign	pdays	previous	poutcome	y
0	58	management	married	tertiary	no	2143	yes	no	unknown	5	may	261	1	-1	0	unknown	no
1	44	technician	single	secondary	no	29	yes	no	unknown	5	may	151	1	-1	0	unknown	no
2	33	entrepreneur	married	secondary	no	2	yes	yes	unknown	5	may	76	1	-1	0	unknown	no
3	47	blue-collar	married	unknown	no	1506	yes	no	unknown	5	may	92	1	-1	0	unknown	no
4	33	unknown	single	unknown	no	1	no	no	unknown	5	may	198	1	-1	0	unknown	no

[4]:
```
df.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 45211 entries, 0 to 45210
Data columns (total 17 columns):
 #   Column     Non-Null Count  Dtype
---  ------     --------------  -----
 0   age        45211 non-null  int64
 1   job        45211 non-null  object
 2   marital    45211 non-null  object
 3   education  45211 non-null  object
 4   default    45211 non-null  object
 5   balance    45211 non-null  int64
 6   housing    45211 non-null  object
 7   loan       45211 non-null  object
 8   contact    45211 non-null  object
 9   day        45211 non-null  int64
 10  month      45211 non-null  object
 11  duration   45211 non-null  int64
 12  campaign   45211 non-null  int64
 13  pdays      45211 non-null  int64
 14  previous   45211 non-null  int64
 15  poutcome   45211 non-null  object
 16  y          45211 non-null  object
dtypes: int64(7), object(10)
memory usage: 5.9+ MB
```

[5]:
```
df.describe()
```

[5]:

	age	balance	day	duration	campaign	pdays	previous
count	45211.000000	45211.000000	45211.000000	45211.000000	45211.000000	45211.000000	45211.000000
mean	40.936210	1362.272058	15.806419	258.163080	2.763841	40.197828	0.580323
std	10.618762	3044.765829	8.322476	257.527812	3.098021	100.128746	2.303441
min	18.000000	-8019.000000	1.000000	0.000000	1.000000	-1.000000	0.000000
25%	33.000000	72.000000	8.000000	103.000000	1.000000	-1.000000	0.000000
50%	39.000000	448.000000	16.000000	180.000000	2.000000	-1.000000	0.000000
75%	48.000000	1428.000000	21.000000	319.000000	3.000000	-1.000000	0.000000
max	95.000000	102127.000000	31.000000	4918.000000	63.000000	871.000000	275.000000

Check for missing values

[6]:
```
missing_values = df.isnull().sum()
missing_values = missing_values[missing_values > 0]
print("\nMissing Values:")
print(missing_values if not missing_values.empty else "No missing values")
```

```
Missing Values:
No missing values
```

Observations:

1. No missing values are present in the dataset. ¶

2. Potential outliers:

- balance has a wide range (-8019 to 102127), suggesting possible outliers.
- duration has a max value of 4918, which seems extreme.
- campaign has a max value of 63, which is significantly higher than the 75th percentile.
- pdays has a lot of -1 values, which might indicate "never contacted before."

Next Steps:

Handle outliers in balance, duration, and campaign using:

Winsorization (capping extreme values).

Transformation to reduce skewness.

Address pdays == -1 as a special category rather than treating it as a numerical value.

[7]:
```
# Handle outliers using Winsorization (capping extreme values at the 99th percentile)
for col in ['balance', 'duration', 'campaign']:
    upper_limit = df[col].quantile(0.99)
    lower_limit = df[col].quantile(0.01)
    df[col] = np.clip(df[col], lower_limit, upper_limit)

# Transform 'balance' using log transformation to reduce skewness (adding 1 to avoid log(0))
df['balance'] = np.log1p(df['balance'])

# Treat 'pdays' == -1 as a special category
df['pdays_category'] = np.where(df['pdays'] == -1, 'never_contacted', 'contacted_before')

df.drop(columns=['pdays'], inplace=True)

df.describe()
```

[7]:

	age	balance	day	duration	campaign	previous
count	45211.000000	4.149500e+04	45211.000000	45211.000000	45211.000000	45211.000000
mean	40.936210	-inf	15.806419	254.375993	2.691403	0.580323
std	10.618762	NaN	8.322476	234.779511	2.585740	2.303441
min	18.000000	-inf	1.000000	11.000000	1.000000	0.000000
25%	33.000000	4.983607e+00	8.000000	103.000000	1.000000	0.000000
50%	39.000000	6.295266e+00	16.000000	180.000000	2.000000	0.000000
75%	48.000000	7.374629e+00	21.000000	319.000000	3.000000	0.000000
max	95.000000	9.485385e+00	31.000000	1269.000000	16.000000	275.000000

[8]:
```
df.head()
```

[8]:

	age	job	marital	education	default	balance	housing	loan	contact	day	month	duration	campaign	previous	poutcome	y	pdays_category
0	58	management	married	tertiary	no	7.670429	yes	no	unknown	5	may	261	1	0	unknown	no	never_contacted
1	44	technician	single	secondary	no	3.401197	yes	no	unknown	5	may	151	1	0	unknown	no	never_contacted
2	33	entrepreneur	married	secondary	no	1.098612	yes	yes	unknown	5	may	76	1	0	unknown	no	never_contacted
3	47	blue-collar	married	unknown	no	7.317876	yes	no	unknown	5	may	92	1	0	unknown	no	never_contacted
4	33	unknown	single	unknown	no	0.693147	no	no	unknown	5	may	198	1	0	unknown	no	never_contacted

Github Link:

https://github.com/asat94/Data-Glacier-Internship/tree/main/Week%209%20Data%20Cleansing%20and%20Transformation