Data Science - Bank Marketing Campaign

1. Group Information

Group Name: Datalux

Group Members: 3

Name	Email	Country	Uni/Company
Huu Thien Nguyen	nguyenhuuthien27296@gmail.com	Sweden	Skövde University
Asmaa Alqurashi	asmaa.idk@gmail.com	Saudi Arabia	Taif University
Deepak Rawat	deepakrawat68@gmail.com	Ireland	Dublin Business School

Specialisation: Data Science

Submitted to: Data Glacier canvas platform

Internship Batch: LISUM09

2. Data cleansing and transformation have been done on the data

2.1. Thien:

2.1.1. Age - Remove outliers based on quantile:

An outlier is an observation “that appears to deviate markedly from other members of the sample in which it occurs”. Note: we focus on univariate outliers, those found when looking at a distribution of values in a single dimension (Age feature).

The outliers detection procedure was based on the graphical pandas profiling library. For each feature, we drew histograms and one-way plots of the logarithms of the unit values, using each to detect the presence of gross outliers for further investigations.

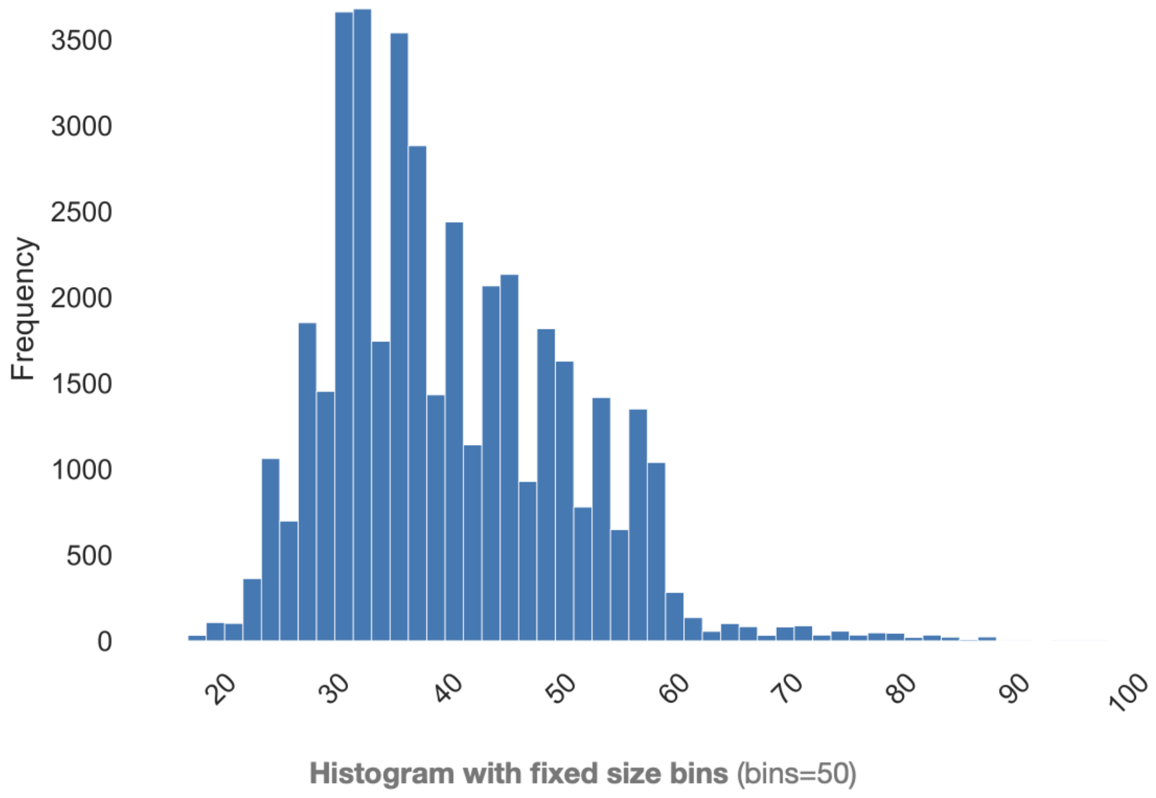


Figure 1: Age histogram visualisation

We must specify a threshold for deciding whether each observation is ‘too extreme’ (outlier or not?). Common ‘thumb-rule’ thresholds: an observation is considered an outlier if it is more than 2.5, 3, or 3.5 standard deviations far from the mean of the distribution. In our scenario, the 1% and 99% were applied to maintain the integrity of the dataset (not so many rows were affected)

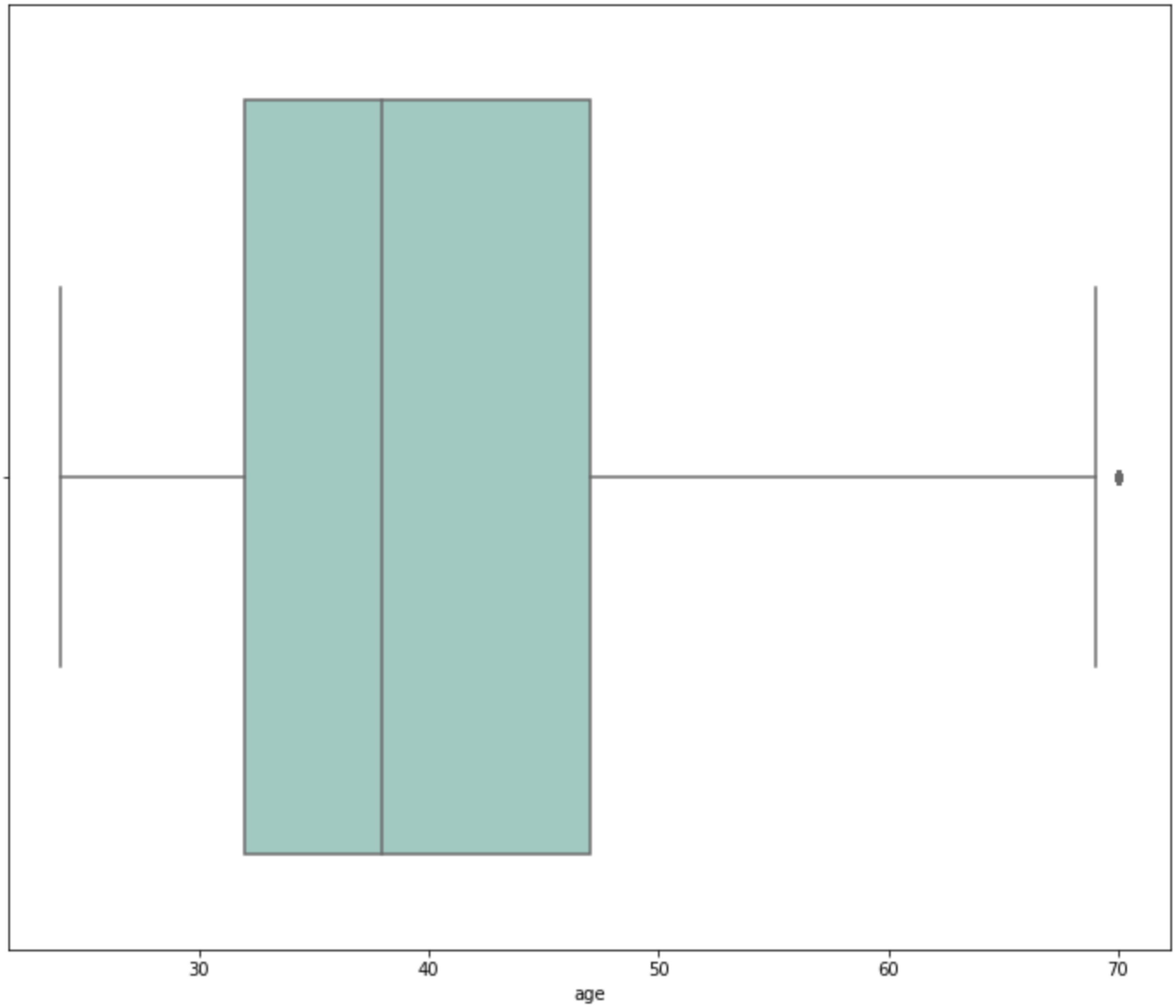
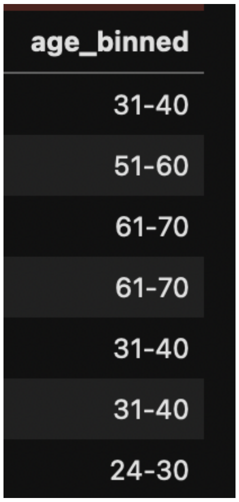


Figure 2: Age boxplot after the removal of the outliers

The age range after removal was between 24 and 70. The dataset has 40161 rows remaining, which is an acceptable amount.

2.1.2. Age - Binning value:

Since age is a numerical feature, the binning technique was implemented to create a new categorical, which can provide a huge insight for the model in the future. The binning range was selected as 24-30, 31-40, 41-50, 51-60, and 61-70. In the modelling step, the feature importance will be measured and the binning range will get adjusted based on the results.



age_binned
31-40
51-60
61-70
61-70
31-40
31-40
24-30

Figure 3: age_binned feature

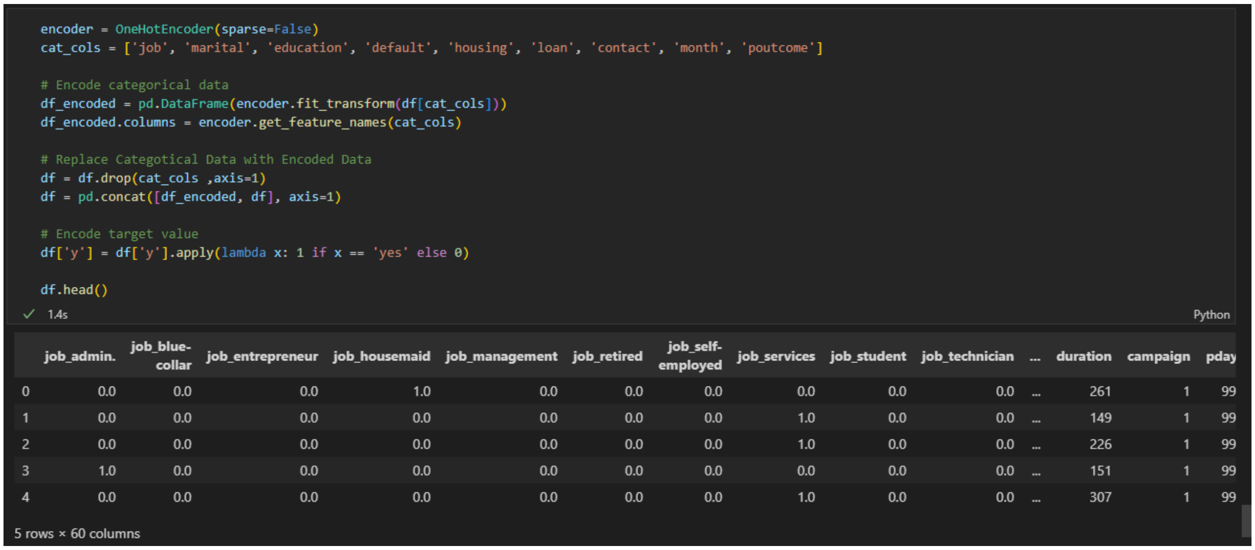
2.2. Asmaa:

2.2.1 Encode Categorical Data

Pre-process categorical data from words to the numeric value to use it in the model. To do this we will use OneHotEncoder() provided by sklearn.

2.2.2 Encode Target Column

Encode the target column yes = 1 and no = 0



```python
encoder = OneHotEncoder(sparse=False)
cat_cols = ['job', 'marital', 'education', 'default', 'housing', 'loan', 'contact', 'month', 'poutcome']

# Encode categorical data
df_encoded = pd.DataFrame(encoder.fit_transform(df[cat_cols]))
df_encoded.columns = encoder.get_feature_names(cat_cols)

# Replace Categotical Data with Encoded Data
df = df.drop(cat_cols ,axis=1)
df = pd.concat([df_encoded, df], axis=1)

# Encode target value
df['y'] = df['y'].apply(lambda x: 1 if x == 'yes' else 0)

df.head()
```

	job_admin.	job_blue-collar	job_entrepreneur	job_housemaid	job_management	job_retired	job_self-employed	job_services	job_student	job_technician	...	duration	campaign	pday
0	0.0	0.0	0.0	1.0	0.0	0.0	0.0	0.0	0.0	0.0	...	261	1	99
1	0.0	0.0	0.0	0.0	0.0	0.0	0.0	1.0	0.0	0.0	...	149	1	99
2	0.0	0.0	0.0	0.0	0.0	0.0	0.0	1.0	0.0	0.0	...	226	1	99
3	1.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	...	151	1	99
4	0.0	0.0	0.0	0.0	0.0	0.0	0.0	1.0	0.0	0.0	...	307	1	99

5 rows × 60 columns

Figure 4: Encode categorical data

2.3. Deepak

2.3.1 Data Transformation

Feature Scaling

It is a Data Pre Processing step that is used with independent data characteristics or variables. In general, it assists in normalising the data within a specific range. It occasionally benefits in accelerating algorithmic calculations.

Feature scaling can be mainly done in three ways -

Min-Max scaling/Normalisation - In min-max we subtract the minimum value in the dataset with all the values and then divide this by the range of the dataset(maximum-minimum). In this case, your dataset will lie between 0 and 1 in all cases

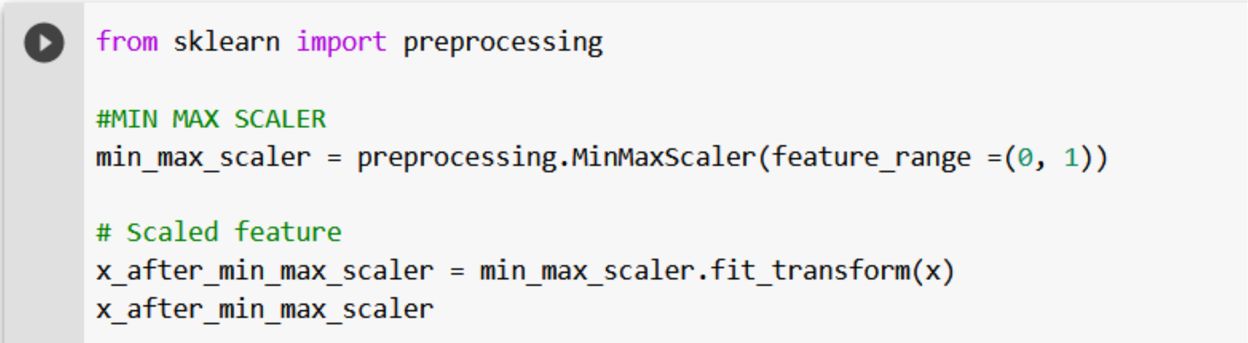
$$X_{new} = \frac{X - X_{min}}{X_{max} - X_{min}}$$

Standardisation - Standardisation involves rescaling the features such that they have the properties of a standard normal distribution with a mean of zero and a standard deviation of one.

$$X_{new} = \frac{X - X_{mean}}{\sigma}$$

Feature scaling can be applied to the columns which contain numeric values. These columns are - 'duration', 'campaign', 'pdays', 'previous', 'emp.var.rate', 'cons.price.idx', 'cons.conf.idx', 'euribor3m', 'nr.employed'.

Min-max Scaler

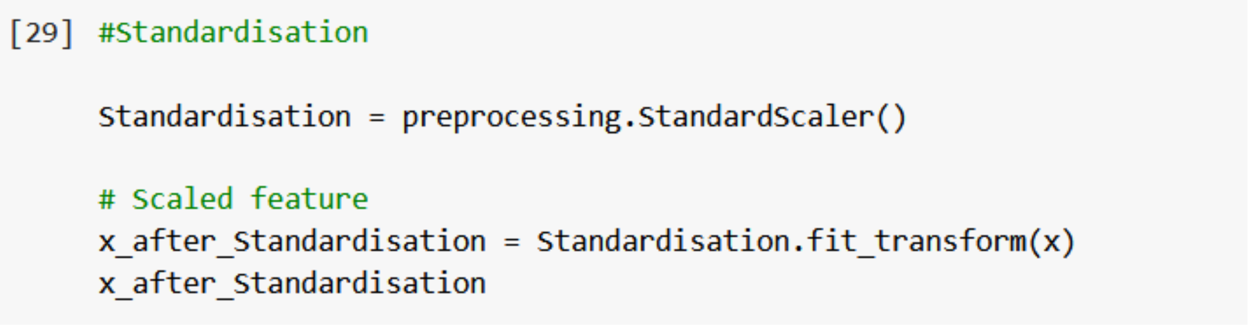


```
from sklearn import preprocessing

#MIN MAX SCALER
min_max_scaler = preprocessing.MinMaxScaler(feature_range =(0, 1))

# Scaled feature
x_after_min_max_scaler = min_max_scaler.fit_transform(x)
x_after_min_max_scaler
```

Standardisation



```
#Standardisation

Standardisation = preprocessing.StandardScaler()

# Scaled feature
x_after_Standardisation = Standardisation.fit_transform(x)
x_after_Standardisation
```

3. GitHub link

The link for GitHub: https://github.com/AndrewNguyen27296/DataGlacier