1. Handling Time-stamps

March 31, 2022

[1]:

```
# Stages of Data Cleaning
# 1. Parse date/time (if time-stamps present)
# 2. Drop un-necessary columns (for machine learning) like ID, etc.
# 3. Feature Engineering (derive new features from existing features)
# 4. Data imputation: Fill the missing values (with mode, median, mean, etc.)␣
 ↪OR Drop rows with missing values
# 3. Handling class imbalance: oversampling e.g. SMOTE, ADASYN
# 5. One Hot Encoding of categorical features
# 6. Normalization (0-1) / Standarization (mean=0, SD=1)
# 7. PDimensionality Reduction: Feature Transformation (PCA/t-SNE) or Feature␣
 ↪selection (chiq-square test, RFE, etc. )
# After data cleaning we can perform:
# Data Modeling (Machine Learning + Regularization) with Hyper-parameter Tuning␣
 ↪(Grid search)
# Model Evaluation (Accuracy, Precision, F1 score, Confusion Matrix, AUC) and␣
 ↪Visualization
```

[2]:

```
import datetime
import pandas as pd
```

[3]:

```
df = pd.read_excel("online_retail.xlsx", sheet_name='data',␣
 ↪parse_dates=['InvoiceDate'])
df.info() # this will help us know the datetime columns which need to be parsed␣
 ↪
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 541909 entries, 0 to 541908
Data columns (total 8 columns):
 #   Column       Non-Null Count   Dtype
---  ------       --------------   -----
 0   InvoiceNo    541909 non-null  int64
 1   StockCode    541909 non-null  object
 2   Description  540455 non-null  object
 3   Quantity     541909 non-null  int64
 4   InvoiceDate  541909 non-null  datetime64[ns]
 5   UnitPrice    541909 non-null  float64
 6   CustomerID   406829 non-null  float64
 7   Country      541909 non-null  object
dtypes: datetime64[ns](1), float64(2), int64(2), object(3)
memory usage: 33.1+ MB
```

[4]:

```
df.head(2).transpose()
```

[4]:

```
                                             0                    1
InvoiceNo                               536365               536365
StockCode                               85123A                71053
Description  WHITE HANGING HEART T-LIGHT HOLDER  WHITE METAL LANTERN
Quantity                                     6                    6
InvoiceDate                2010-12-01 08:26:00  2010-12-01 08:26:00
UnitPrice                                 2.55                 3.39
CustomerID                             17850.0              17850.0
Country                         United Kingdom       United Kingdom
```

[5]:

```
print("Shape of dataset: ", df.shape)
```

```
Shape of dataset:  (541909, 8)
```

[6]:

```
total_customers=len(df['CustomerID'].unique())
print("Total customers: ", total_customers)
```

```
Total customers:  4373
```

[7]:

```
df['Date']=[d.date() for d in df['InvoiceDate']]
df.head(2).transpose()
```

[7]:

```
                                             0                    1
InvoiceNo                               536365               536365
StockCode                               85123A                71053
Description  WHITE HANGING HEART T-LIGHT HOLDER  WHITE METAL LANTERN
Quantity                                     6                    6
InvoiceDate                2010-12-01 08:26:00  2010-12-01 08:26:00
UnitPrice                                 2.55                 3.39
CustomerID                             17850.0              17850.0
Country                         United Kingdom       United Kingdom
Date                                2010-12-01           2010-12-01
```

[8]:

```
x=df['Date'].value_counts()
print("Total transactions per date:\n",x)
```

```
Total transactions per date:
 2011-12-05    5331
2011-12-08    4940
2011-11-29    4313
2011-11-16    4195
2011-11-11    4089
```

```
              …
2011-03-13     537
2010-12-19     522
2011-05-01     452
2010-12-22     291
2011-02-06     279
Name: Date, Length: 305, dtype: int64
```

[9]:

```python
df['year'] = df['InvoiceDate'].dt.year.astype(int)
df['month'] = df['InvoiceDate'].dt.month.astype(int)
df['day'] = df['InvoiceDate'].dt.day.astype(int)
df['day_name'] = df['InvoiceDate'].dt.weekday
df['hour'] = df['InvoiceDate'].dt.hour.astype(int)
df['minute'] = df['InvoiceDate'].dt.minute.astype(int)
print(df.head(2).transpose())
```

```
                                              0                    1
InvoiceNo                                536365               536365
StockCode                                85123A                71053
Description  WHITE HANGING HEART T-LIGHT HOLDER  WHITE METAL LANTERN
Quantity                                      6                    6
InvoiceDate                 2010-12-01 08:26:00  2010-12-01 08:26:00
UnitPrice                                  2.55                 3.39
CustomerID                              17850.0              17850.0
Country                          United Kingdom       United Kingdom
Date                                 2010-12-01           2010-12-01
year                                       2010                 2010
month                                        12                   12
day                                           1                    1
day_name                                      2                    2
hour                                          8                    8
minute                                       26                   26
```

[10]:

```python
# transactions of last 30 days
date_cutoff = df['InvoiceDate'].max() - datetime.timedelta(30, 0, 0)
df['active'] = (df['InvoiceDate'] > date_cutoff).astype(int)
print(df.head(2).transpose())
```

```
                                              0                    1
InvoiceNo                                536365               536365
StockCode                                85123A                71053
Description  WHITE HANGING HEART T-LIGHT HOLDER  WHITE METAL LANTERN
Quantity                                      6                    6
InvoiceDate                 2010-12-01 08:26:00  2010-12-01 08:26:00
UnitPrice                                  2.55                 3.39
CustomerID                              17850.0              17850.0
Country                          United Kingdom       United Kingdom
Date                                 2010-12-01           2010-12-01
```

```
year                                           2010                 2010
month                                            12                   12
day                                               1                    1
day_name                                          2                    2
hour                                              8                    8
minute                                           26                   26
active                                            0                    0
```

```
[11]: df = df.drop(["Country", "UnitPrice", "StockCode", "Description"], axis=1)
      df = df.drop(["year", "month", "day", "hour", "minute"], axis=1)
      df.head(5).transpose()
```

```
[11]:                               0                    1                    2  \
      InvoiceNo                536365               536365               536365
      Quantity                      6                    6                    8
      InvoiceDate  2010-12-01 08:26:00  2010-12-01 08:26:00  2010-12-01 08:26:00
      CustomerID              17850.0              17850.0              17850.0
      Date                 2010-12-01           2010-12-01           2010-12-01
      day_name                      2                    2                    2
      active                        0                    0                    0

                                    3                    4
      InvoiceNo                536365               536365
      Quantity                      6                    6
      InvoiceDate  2010-12-01 08:26:00  2010-12-01 08:26:00
      CustomerID              17850.0              17850.0
      Date                 2010-12-01           2010-12-01
      day_name                      2                    2
      active                        0                    0
```

```
[ ]:
```