svm

October 1, 2023

0.1 DataCo SMART SUPPLY CHAIN FOR BIG DATA ANALYSIS

A DataSet of Supply Chains used by the company DataCo Global was used for the analysis. Dataset of Supply Chain , which allows the use of Machine Learning Algorithms and R Software.

Areas of important registered activities : Provisioning , Production , Sales , Commercial Distribution.It also allows the correlation of Structured Data with Unstructured Data for knowledge generation.

Type Data :
Structured Data : DataCoSupplyChainDataset.csv
Unstructured Data : tokenized_access_logs.csv (Clickstream)

Types of Products : Clothing , Sports , and Electronic Supplies

Additionally it is attached in another file called DescriptionDataCoSupplyChain.csv, the description of each of the variables of the DataCoSupplyChainDatasetc.csv.

0.1.1 Goal

The goal of this analysis is to predict whether the package delivery gonna be late or not (variable Late_delivery_risk)

[1]:
```
import pandas as pd
import matplotlib.pyplot as plt
import numpy as np
from scipy import stats
import seaborn as sns

from imblearn.over_sampling import SMOTE
from sklearn.model_selection import train_test_split
from sklearn import svm
from sklearn.metrics import accuracy_score, f1_score, recall_score,␣
 ↪precision_score
from sklearn.model_selection import GridSearchCV
import lime
from lime import lime_tabular

import logging
import os
import warnings
```

```
from urllib.parse import urlparse
import mlflow
import mlflow.sklearn
from mlflow.models import infer_signature

pd.set_option('display.max_columns', 100)
pd.set_option('display.max_colwidth', None)
logging.basicConfig(level=logging.WARN)
logger = logging.getLogger(__name__)
```

[2]:
```
data_raw = pd.read_csv("D:/Kuliah/semester_3/kecerdasan_buatan/Github/
 ↪Artificial_Inteligence/Pertemuan_4/DataCoSupplyChainDataset.csv",␣
 ↪encoding="ISO-8859-1")
data_unstructured = pd.read_csv("D:/Kuliah/semester_3/kecerdasan_buatan/Github/
 ↪Artificial_Inteligence/Pertemuan_4/tokenized_access_logs.csv",␣
 ↪encoding="ISO-8859-1")
data_desc = pd.read_csv("D:/Kuliah/semester_3/kecerdasan_buatan/Github/
 ↪Artificial_Inteligence/Pertemuan_4/DescriptionDataCoSupplyChain.csv",␣
 ↪encoding="ISO-8859-1")
```

[3]:
```
# Column Description
data_desc
```

[3]:
```
                                 FIELDS  \
0                                  Type
1   Days for shipping (real)
2   Days for shipment (scheduled)
3                      Benefit per order
4                     Sales per customer
5                        Delivery Status
6   Late_delivery_risk
7                            Category Id
8                          Category Name
9                          Customer City
10                      Customer Country
11                        Customer Email
12                        Customer Fname
13                           Customer Id
14                        Customer Lname
15                     Customer Password
16                      Customer Segment
17                        Customer State
18                       Customer Street
19                      Customer Zipcode
20                         Department Id
21                       Department Name
22                              Latitude
```

```
23                      Longitude
24                         Market
25                     Order City
26                  Order Country
27              Order Customer Id
28        order date (DateOrders)
29                       Order Id
30         Order Item Cardprod Id
31            Order Item Discount
32  Order Item Discount Rate
33                  Order Item Id
34  Order Item Product Price
35        Order Item Profit Ratio
36            Order Item Quantity
37                          Sales
38             Order Item Total  
39         Order Profit Per Order
40                   Order Region
41                    Order State
42                   Order Status
43                Product Card Id
44            Product Category Id
45            Product Description
46                  Product Image
47                   Product Name
48                  Product Price
49                 Product Status
50  Shipping date (DateOrders)
51                  Shipping Mode

                                                         DESCRIPTION
0
:  Type of transaction made
1
:  Actual shipping days of the purchased product
2
:  Days of scheduled delivery of the purchased product
3
:  Earnings per order placed
4
:  Total sales per customer made per customer
5
:  Delivery status of orders: Advance shipping , Late delivery , Shipping
canceled , Shipping on time
6
:  Categorical variable that indicates if sending is late (1), it is not late
(0).
```

```
7
:  Product category code
8
:  Description of the product category
9
:  City where the customer made the purchase
10
:  Country where the customer made the purchase
11
:  Customer's email
12
:  Customer name
13
:  Customer ID
14
:  Customer lastname
15
:  Masked customer key
16
:  Types of Customers: Consumer , Corporate , Home Office
17
:  State to which the store where the purchase is registered belongs
18
:  Street to which the store where the purchase is registered belongs
19
:  Customer Zipcode
20
:  Department code of store
21
:  Department name of store
22
:  Latitude corresponding to location of store
23
:  Longitude corresponding to location of store
24
:  Market to where the order is delivered : Africa , Europe , LATAM , Pacific
Asia , USCA
25
:  Destination city of the order
26
:  Destination country of the order
27
:  Customer order code
28
:  Date on which the order is made
29
:  Order code
```

30
: Product code generated through the RFID reader
31
: Order item discount value
32
: Order item discount percentage
33
: Order item code
34
: Price of products without discount
35
: Order Item Profit Ratio
36
: Number of products per order
37
: Value in sales
38
: Total amount per order
39
: Order Profit Per Order
40 : Region of the world where the order is delivered : Southeast Asia ,South Asia ,Oceania ,Eastern Asia, West Asia , West of USA , US Center , West Africa, Central Africa ,North Africa ,Western Europe ,Northern , Caribbean , South America ,East Africa ,Southern Europe , East of USA ,Canada ,Southern Africa , Central Asia , Europe , Central America, Eastern Europe , South of USA
41
: State of the region where the order is delivered
42
: Order Status : COMPLETE , PENDING , CLOSED , PENDING_PAYMENT ,CANCELED , PROCESSING ,SUSPECTED_FRAUD ,ON_HOLD ,PAYMENT_REVIEW
43
: Product code
44
: Product category code
45
: Product Description
46
: Link of visit and purchase of the product
47
: Product Name
48
: Product Price
49
: Status of the product stock :If it is 1 not available , 0 the product is available
50
: Exact date and time of shipment

```
51
:  The following shipping modes are presented : Standard Class , First Class ,
Second Class , Same Day
```

[4]:
```
data_raw.head()
```

[4]:
```
       Type  Days for shipping (real)  Days for shipment (scheduled)  \
0     DEBIT                         3                              4   
1  TRANSFER                         5                              4   
2      CASH                         4                              4   
3     DEBIT                         3                              4   
4   PAYMENT                         2                              4   

   Benefit per order  Sales per customer   Delivery Status  \
0          91.250000          314.640015  Advance shipping   
1        -249.089996          311.359985     Late delivery   
2        -247.779999          309.720001  Shipping on time   
3          22.860001          304.809998  Advance shipping   
4         134.210007          298.250000  Advance shipping   

   Late_delivery_risk  Category Id   Category Name Customer City  \
0                   0           73  Sporting Goods        Caguas   
1                   1           73  Sporting Goods        Caguas   
2                   0           73  Sporting Goods      San Jose   
3                   0           73  Sporting Goods   Los Angeles   
4                   0           73  Sporting Goods        Caguas   

  Customer Country Customer Email Customer Fname  Customer Id Customer Lname  \
0      Puerto Rico      XXXXXXXXX          Cally        20755       Holloway   
1      Puerto Rico      XXXXXXXXX          Irene        19492           Luna   
2          EE. UU.      XXXXXXXXX        Gillian        19491      Maldonado   
3          EE. UU.      XXXXXXXXX           Tana        19490           Tate   
4      Puerto Rico      XXXXXXXXX           Orli        19489      Hendricks   

  Customer Password Customer Segment Customer State           Customer Street  \
0         XXXXXXXXX         Consumer             PR  5365 Noble Nectar Island   
1         XXXXXXXXX         Consumer             PR          2679 Rustic Loop   
2         XXXXXXXXX         Consumer             CA       8510 Round Bear Gate   
3         XXXXXXXXX      Home Office             CA           3200 Amber Bend   
4         XXXXXXXXX        Corporate             PR  8671 Iron Anchor Corners   

   Customer Zipcode  Department Id Department Name   Latitude   Longitude  \
0             725.0              2         Fitness  18.251453  -66.037056   
1             725.0              2         Fitness  18.279451  -66.037064   
2           95125.0              2         Fitness  37.292233 -121.881279   
3           90027.0              2         Fitness  34.125946 -118.291016   
4             725.0              2         Fitness  18.253769  -66.037048   
```

```
         Market  Order City Order Country  Order Customer Id  \
0  Pacific Asia      Bekasi     Indonesia              20755
1  Pacific Asia     Bikaner         India              19492
2  Pacific Asia     Bikaner         India              19491
3  Pacific Asia  Townsville     Australia              19490
4  Pacific Asia  Townsville     Australia              19489

  order date (DateOrders)  Order Id  Order Item Cardprod Id  \
0         1/31/2018 22:56     77202                    1360
1         1/13/2018 12:27     75939                    1360
2         1/13/2018 12:06     75938                    1360
3         1/13/2018 11:45     75937                    1360
4         1/13/2018 11:24     75936                    1360

   Order Item Discount  Order Item Discount Rate  Order Item Id  \
0            13.110000                      0.04         180517
1            16.389999                      0.05         179254
2            18.030001                      0.06         179253
3            22.940001                      0.07         179252
4            29.500000                      0.09         179251

   Order Item Product Price  Order Item Profit Ratio  Order Item Quantity  \
0                    327.75                     0.29                    1
1                    327.75                    -0.80                    1
2                    327.75                    -0.80                    1
3                    327.75                     0.08                    1
4                    327.75                     0.45                    1

    Sales  Order Item Total  Order Profit Per Order    Order Region  \
0  327.75        314.640015               91.250000  Southeast Asia
1  327.75        311.359985             -249.089996      South Asia
2  327.75        309.720001             -247.779999      South Asia
3  327.75        304.809998               22.860001         Oceania
4  327.75        298.250000              134.210007         Oceania

       Order State     Order Status  Order Zipcode  Product Card Id  \
0  Java Occidental         COMPLETE            NaN             1360
1         Rajastán          PENDING            NaN             1360
2         Rajastán           CLOSED            NaN             1360
3       Queensland         COMPLETE            NaN             1360
4       Queensland  PENDING_PAYMENT            NaN             1360

   Product Category Id  Product Description  \
0                   73                  NaN
1                   73                  NaN
2                   73                  NaN
```

```
3                   73                  NaN
4                   73                  NaN

                                  Product Image  Product Name  Product Price  \
0  http://images.acmesports.sports/Smart+watch   Smart watch          327.75
1  http://images.acmesports.sports/Smart+watch   Smart watch          327.75
2  http://images.acmesports.sports/Smart+watch   Smart watch          327.75
3  http://images.acmesports.sports/Smart+watch   Smart watch          327.75
4  http://images.acmesports.sports/Smart+watch   Smart watch          327.75

   Product Status shipping date (DateOrders)   Shipping Mode
0               0            2/3/2018 22:56  Standard Class
1               0           1/18/2018 12:27  Standard Class
2               0           1/17/2018 12:06  Standard Class
3               0           1/16/2018 11:45  Standard Class
4               0           1/15/2018 11:24  Standard Class
```

[5]:
```
data_unstructured.head()
```

[5]:
```
                                       Product             Category  \
0     adidas Brazuca 2017 Official Match Ball  baseball & softball
1       The North Face Women's Recon Backpack   hunting & shooting
2      adidas Kids' RG III Mid Football Cleat       featured shops
3  Under Armour Men's Compression EV SL Slide          electronics
4                   Pelican Sunstream 100 Kayak         water sports

            Date Month  Hour Department              ip  \
0  9/1/2017 6:00   Sep     6    fitness     37.97.182.65
1  9/1/2017 6:00   Sep     6   fan shop     206.56.112.1
2  9/1/2017 6:00   Sep     6    apparel   215.143.180.0
3  9/1/2017 6:00   Sep     6   footwear     206.56.112.1
4  9/1/2017 6:01   Sep     6   fan shop  136.108.56.242

                                 url
0  /department/fitness/category/baseball%20&%20softball/product/adidas%20Brazuca
%202017%20Official%20Match%20Ball
1  /department/fan%20shop/category/hunting%20&%20shooting/product/The%20North%20
Face%20Women's%20Recon%20Backpack
2        /department/apparel/category/featured%20shops/product/adidas%20Kids'%20
RG%20III%20Mid%20Football%20Cleat
3        /department/footwear/category/electronics/product/Under%20Armour%20Men'
s%20Compression%20EV%20SL%20Slide
4                         /department/fan%20shop/category/water%20sports/product/
Pelican%20Sunstream%20100%20Kayak
```

[6]:
```
data_raw.isnull().sum()
```

```
[6]: Type                                  0
     Days for shipping (real)              0
     Days for shipment (scheduled)         0
     Benefit per order                     0
     Sales per customer                    0
     Delivery Status                       0
     Late_delivery_risk                    0
     Category Id                           0
     Category Name                         0
     Customer City                         0
     Customer Country                      0
     Customer Email                        0
     Customer Fname                        0
     Customer Id                           0
     Customer Lname                        8
     Customer Password                     0
     Customer Segment                      0
     Customer State                        0
     Customer Street                       0
     Customer Zipcode                      3
     Department Id                         0
     Department Name                       0
     Latitude                              0
     Longitude                             0
     Market                                0
     Order City                            0
     Order Country                         0
     Order Customer Id                     0
     order date (DateOrders)               0
     Order Id                              0
     Order Item Cardprod Id                0
     Order Item Discount                   0
     Order Item Discount Rate              0
     Order Item Id                         0
     Order Item Product Price              0
     Order Item Profit Ratio               0
     Order Item Quantity                   0
     Sales                                 0
     Order Item Total                      0
     Order Profit Per Order                0
     Order Region                          0
     Order State                           0
     Order Status                          0
     Order Zipcode                    155679
     Product Card Id                       0
     Product Category Id                   0
     Product Description              180519
     Product Image                    0
     Product Name                     0
     Product Price                    0
     Product Status                   0
     shipping date (DateOrders)       0
     Shipping Mode                    0
     dtype: int64
```

Missing data is found on Customer Zipcode, Order Zipcode, and Product Description. Since those columns most likely wasnt going in training data, this can be ignored.

[7]:
```
data_raw.head()
```

[7]:
```
       Type  Days for shipping (real)  Days for shipment (scheduled)  \
0     DEBIT                         3                              4
1  TRANSFER                         5                              4
2      CASH                         4                              4
3     DEBIT                         3                              4
4   PAYMENT                         2                              4

   Benefit per order  Sales per customer   Delivery Status  \
0          91.250000          314.640015  Advance shipping
1        -249.089996          311.359985     Late delivery
2        -247.779999          309.720001  Shipping on time
3          22.860001          304.809998  Advance shipping
4         134.210007          298.250000  Advance shipping

   Late_delivery_risk  Category Id   Category Name Customer City  \
0                   0           73  Sporting Goods        Caguas
1                   1           73  Sporting Goods        Caguas
2                   0           73  Sporting Goods      San Jose
3                   0           73  Sporting Goods   Los Angeles
4                   0           73  Sporting Goods        Caguas

  Customer Country Customer Email Customer Fname  Customer Id Customer Lname  \
0      Puerto Rico      XXXXXXXXX          Cally        20755       Holloway
1      Puerto Rico      XXXXXXXXX          Irene        19492           Luna
2         EE. UU.       XXXXXXXXX        Gillian        19491       Maldonado
3         EE. UU.       XXXXXXXXX           Tana        19490           Tate
4      Puerto Rico      XXXXXXXXX           Orli        19489      Hendricks

  Customer Password Customer Segment Customer State           Customer Street  \
0         XXXXXXXXX         Consumer             PR  5365 Noble Nectar Island
1         XXXXXXXXX         Consumer             PR          2679 Rustic Loop
2         XXXXXXXXX         Consumer             CA        8510 Round Bear Gate
3         XXXXXXXXX      Home Office             CA           3200 Amber Bend
4         XXXXXXXXX        Corporate             PR  8671 Iron Anchor Corners
```

```
   Customer Zipcode  Department Id Department Name   Latitude   Longitude  \
0             725.0              2         Fitness  18.251453  -66.037056
1             725.0              2         Fitness  18.279451  -66.037064
2           95125.0              2         Fitness  37.292233 -121.881279
3           90027.0              2         Fitness  34.125946 -118.291016
4             725.0              2         Fitness  18.253769  -66.037048

         Market  Order City Order Country  Order Customer Id  \
0  Pacific Asia      Bekasi     Indonesia              20755
1  Pacific Asia     Bikaner         India              19492
2  Pacific Asia     Bikaner         India              19491
3  Pacific Asia  Townsville     Australia              19490
4  Pacific Asia  Townsville     Australia              19489

  order date (DateOrders)  Order Id  Order Item Cardprod Id  \
0         1/31/2018 22:56     77202                    1360
1         1/13/2018 12:27     75939                    1360
2         1/13/2018 12:06     75938                    1360
3         1/13/2018 11:45     75937                    1360
4         1/13/2018 11:24     75936                    1360

   Order Item Discount  Order Item Discount Rate  Order Item Id  \
0            13.110000                      0.04         180517
1            16.389999                      0.05         179254
2            18.030001                      0.06         179253
3            22.940001                      0.07         179252
4            29.500000                      0.09         179251

   Order Item Product Price  Order Item Profit Ratio  Order Item Quantity  \
0                    327.75                     0.29                    1
1                    327.75                    -0.80                    1
2                    327.75                    -0.80                    1
3                    327.75                     0.08                    1
4                    327.75                     0.45                    1

    Sales  Order Item Total  Order Profit Per Order    Order Region  \
0  327.75        314.640015               91.250000  Southeast Asia
1  327.75        311.359985             -249.089996      South Asia
2  327.75        309.720001             -247.779999      South Asia
3  327.75        304.809998               22.860001         Oceania
4  327.75        298.250000              134.210007         Oceania

       Order State     Order Status  Order Zipcode  Product Card Id  \
0  Java Occidental         COMPLETE            NaN             1360
1         Rajastán          PENDING            NaN             1360
2         Rajastán           CLOSED            NaN             1360
3       Queensland         COMPLETE            NaN             1360
```

```
4       Queensland  PENDING_PAYMENT            NaN             1360

   Product Category Id  Product Description  \
0                   73                  NaN
1                   73                  NaN
2                   73                  NaN
3                   73                  NaN
4                   73                  NaN

                                  Product Image  Product Name  Product Price  \
0  http://images.acmesports.sports/Smart+watch   Smart watch         327.75
1  http://images.acmesports.sports/Smart+watch   Smart watch         327.75
2  http://images.acmesports.sports/Smart+watch   Smart watch         327.75
3  http://images.acmesports.sports/Smart+watch   Smart watch         327.75
4  http://images.acmesports.sports/Smart+watch   Smart watch         327.75

   Product Status shipping date (DateOrders)   Shipping Mode
0               0            2/3/2018 22:56  Standard Class
1               0           1/18/2018 12:27  Standard Class
2               0           1/17/2018 12:06  Standard Class
3               0           1/16/2018 11:45  Standard Class
4               0           1/15/2018 11:24  Standard Class
```

[8]:
```
data_raw.dtypes
```

[8]:
```
Type                                 object
Days for shipping (real)              int64
Days for shipment (scheduled)         int64
Benefit per order                   float64
Sales per customer                  float64
Delivery Status                      object
Late_delivery_risk                    int64
Category Id                           int64
Category Name                        object
Customer City                        object
Customer Country                     object
Customer Email                       object
Customer Fname                       object
Customer Id                           int64
Customer Lname                       object
Customer Password                    object
Customer Segment                     object
Customer State                       object
Customer Street                      object
Customer Zipcode                    float64
Department Id                         int64
Department Name                      object
```

```
Latitude                         float64
Longitude                        float64
Market                            object
Order City                        object
Order Country                     object
Order Customer Id                  int64
order date (DateOrders)           object
Order Id                           int64
Order Item Cardprod Id             int64
Order Item Discount              float64
Order Item Discount Rate         float64
Order Item Id                      int64
Order Item Product Price         float64
Order Item Profit Ratio          float64
Order Item Quantity                int64
Sales                            float64
Order Item Total                 float64
Order Profit Per Order           float64
Order Region                      object
Order State                       object
Order Status                      object
Order Zipcode                    float64
Product Card Id                    int64
Product Category Id                int64
Product Description              float64
Product Image                     object
Product Name                      object
Product Price                    float64
Product Status                     int64
shipping date (DateOrders)        object
Shipping Mode                     object
dtype: object
```

[9]:
```
col_to_object = ["Customer Id", "Customer Zipcode", "Department Id", "Order␣
↪Customer Id", "Order Id", "Order Item Cardprod Id",
                 "Order Item Id", "Product Card Id", "Product Category Id",␣
↪"Product Status", "Late_delivery_risk", "Category Id",
                 "Latitude", "Longitude", "Order Zipcode", "Product␣
↪Description"]
col_to_date =  ["order date (DateOrders)", "shipping date (DateOrders)"]

data_1 = data_raw
data_1[col_to_object] = data_1[col_to_object].astype(str)
data_1[col_to_date] = data_1[col_to_date].apply(pd.to_datetime, format='%m/%d/
↪%Y %H:%M')
```

0.1.2 Descriptive Statistics

```
[10]: print(data_1.describe(include='all'))
```

```
C:\Users\PC\AppData\Local\Temp\ipykernel_7520\3281304271.py:1: FutureWarning:
Treating datetime data as categorical rather than numeric in `.describe` is
deprecated and will be removed in a future version of pandas. Specify
`datetime_is_numeric=True` to silence this warning and adopt the future behavior
now.
  print(data_1.describe(include='all'))
```

```
          Type  Days for shipping (real)  Days for shipment (scheduled)  \
count   180519             180519.000000                  180519.000000
unique       4                       NaN                            NaN
top      DEBIT                       NaN                            NaN
freq     69295                       NaN                            NaN
first      NaN                       NaN                            NaN
last       NaN                       NaN                            NaN
mean       NaN                  3.497654                       2.931847
std        NaN                  1.623722                       1.374449
min        NaN                  0.000000                       0.000000
25%        NaN                  2.000000                       2.000000
50%        NaN                  3.000000                       4.000000
75%        NaN                  5.000000                       4.000000
max        NaN                  6.000000                       4.000000

        Benefit per order  Sales per customer Delivery Status  \
count       180519.000000       180519.000000          180519
unique                NaN                 NaN               4
top                   NaN                 NaN   Late delivery
freq                  NaN                 NaN           98977
first                 NaN                 NaN             NaN
last                  NaN                 NaN             NaN
mean            21.974989          183.107609             NaN
std            104.433526          120.043670             NaN
min          -4274.979980            7.490000             NaN
25%              7.000000          104.379997             NaN
50%             31.520000          163.990005             NaN
75%             64.800003          247.399994             NaN
max            911.799988         1939.989990             NaN

       Late_delivery_risk Category Id Category Name Customer City  \
count              180519      180519        180519        180519
unique                  2          51            50           563
top                     1          17        Cleats        Caguas
freq                98977       24551         24551         66770
first                 NaN         NaN           NaN           NaN
last                  NaN         NaN           NaN           NaN
mean                  NaN         NaN           NaN           NaN
```

```
std                     NaN         NaN           NaN           NaN
min                     NaN         NaN           NaN           NaN
25%                     NaN         NaN           NaN           NaN
50%                     NaN         NaN           NaN           NaN
75%                     NaN         NaN           NaN           NaN
max                     NaN         NaN           NaN           NaN

       Customer Country Customer Email Customer Fname Customer Id  \
count            180519         180519         180519      180519
unique                2              1            782       20652
top            EE. UU.      XXXXXXXXX           Mary        5654
freq             111146         180519          65150          47
first               NaN            NaN            NaN         NaN
last                NaN            NaN            NaN         NaN
mean                NaN            NaN            NaN         NaN
std                 NaN            NaN            NaN         NaN
min                 NaN            NaN            NaN         NaN
25%                 NaN            NaN            NaN         NaN
50%                 NaN            NaN            NaN         NaN
75%                 NaN            NaN            NaN         NaN
max                 NaN            NaN            NaN         NaN

       Customer Lname Customer Password Customer Segment Customer State  \
count          180511            180519           180519         180519
unique           1109                 1                3             46
top             Smith         XXXXXXXXX         Consumer             PR
freq            64104            180519            93504          69373
first             NaN               NaN              NaN            NaN
last              NaN               NaN              NaN            NaN
mean              NaN               NaN              NaN            NaN
std               NaN               NaN              NaN            NaN
min               NaN               NaN              NaN            NaN
25%               NaN               NaN              NaN            NaN
50%               NaN               NaN              NaN            NaN
75%               NaN               NaN              NaN            NaN
max               NaN               NaN              NaN            NaN

                Customer Street Customer Zipcode Department Id  \
count                    180519           180519        180519
unique                     7458              996            11
top     9126 Wishing Expressway            725.0             7
freq                        122            66770         66861
first                       NaN              NaN           NaN
last                        NaN              NaN           NaN
mean                        NaN              NaN           NaN
std                         NaN              NaN           NaN
min                         NaN              NaN           NaN
25%                         NaN              NaN           NaN
```

```
50%                        NaN              NaN           NaN
75%                        NaN              NaN           NaN
max                        NaN              NaN           NaN

       Department Name    Latitude    Longitude  Market     Order City  \
count           180519      180519       180519  180519         180519
unique              11       11250         4487       5           3597
top           Fan Shop  18.2275734  -66.3706131   LATAM  Santo Domingo
freq             66861         417         3821   51594           2211
first              NaN         NaN          NaN     NaN            NaN
last               NaN         NaN          NaN     NaN            NaN
mean               NaN         NaN          NaN     NaN            NaN
std                NaN         NaN          NaN     NaN            NaN
min                NaN         NaN          NaN     NaN            NaN
25%                NaN         NaN          NaN     NaN            NaN
50%                NaN         NaN          NaN     NaN            NaN
75%                NaN         NaN          NaN     NaN            NaN
max                NaN         NaN          NaN     NaN            NaN

         Order Country Order Customer Id order date (DateOrders) Order Id  \
count           180519            180519                  180519   180519
unique             164             20652                   65752    65752
top     Estados Unidos              5654     2016-12-14 12:29:00    48880
freq             24840                47                       5        5
first              NaN               NaN     2015-01-01 00:00:00      NaN
last               NaN               NaN     2018-01-31 23:38:00      NaN
mean               NaN               NaN                     NaN      NaN
std                NaN               NaN                     NaN      NaN
min                NaN               NaN                     NaN      NaN
25%                NaN               NaN                     NaN      NaN
50%                NaN               NaN                     NaN      NaN
75%                NaN               NaN                     NaN      NaN
max                NaN               NaN                     NaN      NaN

       Order Item Cardprod Id  Order Item Discount  Order Item Discount Rate  \
count                  180519        180519.000000             180519.000000
unique                    118                  NaN                       NaN
top                       365                  NaN                       NaN
freq                    24515                  NaN                       NaN
first                     NaN                  NaN                       NaN
last                      NaN                  NaN                       NaN
mean                      NaN            20.664741                  0.101668
std                       NaN            21.800901                  0.070415
min                       NaN             0.000000                  0.000000
25%                       NaN             5.400000                  0.040000
50%                       NaN            14.000000                  0.100000
75%                       NaN            29.990000                  0.160000
max                       NaN           500.000000                  0.250000
```

```
       Order Item Id  Order Item Product Price  Order Item Profit Ratio  \
count         180519             180519.000000            180519.000000
unique        180519                       NaN                      NaN
top           180517                       NaN                      NaN
freq               1                       NaN                      NaN
first            NaN                       NaN                      NaN
last             NaN                       NaN                      NaN
mean             NaN                141.232550                 0.120647
std              NaN                139.732492                 0.466796
min              NaN                  9.990000                -2.750000
25%              NaN                 50.000000                 0.080000
50%              NaN                 59.990002                 0.270000
75%              NaN                199.990005                 0.360000
max              NaN               1999.989990                 0.500000

        Order Item Quantity          Sales  Order Item Total  \
count         180519.000000  180519.000000     180519.000000
unique                  NaN            NaN               NaN
top                     NaN            NaN               NaN
freq                    NaN            NaN               NaN
first                   NaN            NaN               NaN
last                    NaN            NaN               NaN
mean               2.127638     203.772096        183.107609
std                1.453451     132.273077        120.043670
min                1.000000       9.990000          7.490000
25%                1.000000     119.980003        104.379997
50%                1.000000     199.919998        163.990005
75%                3.000000     299.950012        247.399994
max                5.000000    1999.989990       1939.989990

        Order Profit Per Order     Order Region Order State Order Status  \
count            180519.000000           180519      180519       180519
unique                     NaN               23        1089            9
top                        NaN  Central America  Inglaterra     COMPLETE
freq                       NaN            28341        6722        59491
first                      NaN              NaN         NaN          NaN
last                       NaN              NaN         NaN          NaN
mean                 21.974989              NaN         NaN          NaN
std                 104.433526              NaN         NaN          NaN
min              -4274.979980              NaN         NaN          NaN
25%                   7.000000              NaN         NaN          NaN
50%                  31.520000              NaN         NaN          NaN
75%                  64.800003              NaN         NaN          NaN
max                 911.799988              NaN         NaN          NaN

       Order Zipcode Product Card Id Product Category Id Product Description  \
count         180519          180519              180519              180519
```

```
unique         610          118              51                 1
top            nan          365              17               nan
freq        155679        24515           24551            180519
first          NaN          NaN             NaN               NaN
last           NaN          NaN             NaN               NaN
mean           NaN          NaN             NaN               NaN
std            NaN          NaN             NaN               NaN
min            NaN          NaN             NaN               NaN
25%            NaN          NaN             NaN               NaN
50%            NaN          NaN             NaN               NaN
75%            NaN          NaN             NaN               NaN
max            NaN          NaN             NaN               NaN

                                                    Product Image  \
count                                                      180519
unique                                                        118
top     http://images.acmesports.sports/Perfect+Fitness+Perfect+Rip+Deck
freq                                                        24515
first                                                         NaN
last                                                          NaN
mean                                                          NaN
std                                                           NaN
min                                                           NaN
25%                                                           NaN
50%                                                           NaN
75%                                                           NaN
max                                                           NaN

                            Product Name  Product Price Product Status  \
count                             180519  180519.000000         180519
unique                               118            NaN              1
top     Perfect Fitness Perfect Rip Deck            NaN              0
freq                               24515            NaN         180519
first                                NaN            NaN            NaN
last                                 NaN            NaN            NaN
mean                                 NaN     141.232550            NaN
std                                  NaN     139.732492            NaN
min                                  NaN       9.990000            NaN
25%                                  NaN      50.000000            NaN
50%                                  NaN      59.990002            NaN
75%                                  NaN     199.990005            NaN
max                                  NaN    1999.989990            NaN

       shipping date (DateOrders)   Shipping Mode
count                      180519          180519
unique                      63701               4
top           2016-01-05 05:58:00  Standard Class
freq                           10          107752
```

```
first         2015-01-03 00:00:00            NaN
last          2018-02-06 22:14:00            NaN
mean                          NaN            NaN
std                           NaN            NaN
min                           NaN            NaN
25%                           NaN            NaN
50%                           NaN            NaN
75%                           NaN            NaN
max                           NaN            NaN
```

```
C:\Users\PC\AppData\Local\Temp\ipykernel_7520\3281304271.py:1: FutureWarning:
Treating datetime data as categorical rather than numeric in `.describe` is
deprecated and will be removed in a future version of pandas. Specify
`datetime_is_numeric=True` to silence this warning and adopt the future behavior
now.
  print(data_1.describe(include='all'))
```

Since SVM is relatively demanding algorithm in case of using this large dataset, lets do stratified sampling based on Late_delivery_risk

[11]:
```
data_1 = data_1.groupby('Late_delivery_risk', group_keys=False).apply(lambda x:␣
↪x.sample(frac=0.1))
```

0.1.3 Bar Chart

[12]:
```
data_cat = data_1.select_dtypes(include=['object'])
data_cat.head()
```

[12]:
```
            Type    Delivery Status Late_delivery_risk Category Id  \
131982      CASH   Advance shipping                  0          46
122089  TRANSFER  Shipping canceled                  0          29
59741      DEBIT   Advance shipping                  0          43
146707     DEBIT   Shipping on time                  0          45
79280      DEBIT   Shipping on time                  0          48

               Category Name Customer City Customer Country Customer Email  \
131982  Indoor/Outdoor Games     Opelousas          EE. UU.      XXXXXXXXX
122089          Shop By Sport        Caguas      Puerto Rico      XXXXXXXXX
59741       Camping & Hiking      Wheeling          EE. UU.      XXXXXXXXX
146707                Fishing        Caguas      Puerto Rico      XXXXXXXXX
79280           Water Sports      Cerritos          EE. UU.      XXXXXXXXX

       Customer Fname Customer Id Customer Lname Customer Password  \
131982           Mary          38          Smith         XXXXXXXXX
122089       Nicholas       12002          Davis         XXXXXXXXX
59741            Mary        3477          Smith         XXXXXXXXX
146707           Kyle       10253          Smith         XXXXXXXXX
79280          Teresa       10142          Smith         XXXXXXXXX
```

```
       Customer Segment Customer State       Customer Street Customer Zipcode  \
131982         Consumer             LA     2805 Crystal Moor          70570.0
122089        Corporate             PR   3977 Old Dale Point            725.0
59741         Corporate             WV   1397 Colonial Point          26003.0
146707         Consumer             PR      1023 Honey Grove            725.0
79280          Consumer             CA  3504 Dusty View Loop          90703.0

       Department Id Department Name     Latitude     Longitude        Market  \
131982             7        Fan Shop  30.24161911  -97.89144135  Pacific Asia
122089             5            Golf  18.28413773  -66.37057495         LATAM
59741              7        Fan Shop   40.0639801  -80.72141266        Europe
146707             7        Fan Shop  18.21958351   -66.3706131        Africa
79280              7        Fan Shop  33.80993271  -118.0119705          USCA

           Order City   Order Country Order Customer Id Order Id  \
131982        Yakarta       Indonesia                38    27562
122089    Villa Nueva       Guatemala             12002     9109
59741   Villefontaine         Francia              3477    12481
146707      Quelimane      Mozambique             10253    46243
79280     Springfield  Estados Unidos             10142    34938

       Order Item Cardprod Id Order Item Id     Order Region  \
131982                   1014         68994   Southeast Asia
122089                    627         22722  Central America
59741                     957         31224   Western Europe
146707                   1004        115597      East Africa
79280                    1073         87289      East of USA

                  Order State     Order Status Order Zipcode Product Card Id  \
131982                Yakarta           CLOSED           nan            1014
122089              Guatemala  SUSPECTED_FRAUD           nan             627
59741   Auvernia-Ródano-Alpes         COMPLETE           nan             957
146707               Zambezia         COMPLETE           nan            1004
79280                    Ohio         COMPLETE       45503.0            1073

       Product Category Id Product Description  \
131982                  46                 nan
122089                  29                 nan
59741                   43                 nan
146707                  45                 nan
79280                   48                 nan

 Product Image  \
131982
http://images.acmesports.sports/O%27Brien+Men%27s+Neoprene+Life+Vest
122089  http://images.acmesports.sports/Under+Armour+Girls%27+Toddler+Spine+Surg
e+Running+Shoe
```

```
59741   http://images.acmesports.sports/Diamondback+Women%27s+Serene+Classic+Com
fort+Bike+2014
146707
http://images.acmesports.sports/Field+%26+Stream+Sportsman+16+Gun+Fire+Safe
79280
http://images.acmesports.sports/Pelican+Sunstream+100+Kayak

                                        Product Name Product Status  \
131982               O'Brien Men's Neoprene Life Vest               0
122089  Under Armour Girls' Toddler Spine Surge Runni               0
59741   Diamondback Women's Serene Classic Comfort Bi               0
146707      Field & Stream Sportsman 16 Gun Fire Safe               0
79280                        Pelican Sunstream 100 Kayak               0

         Shipping Mode
131982  Standard Class
122089     First Class
59741   Standard Class
146707  Standard Class
79280   Standard Class
```

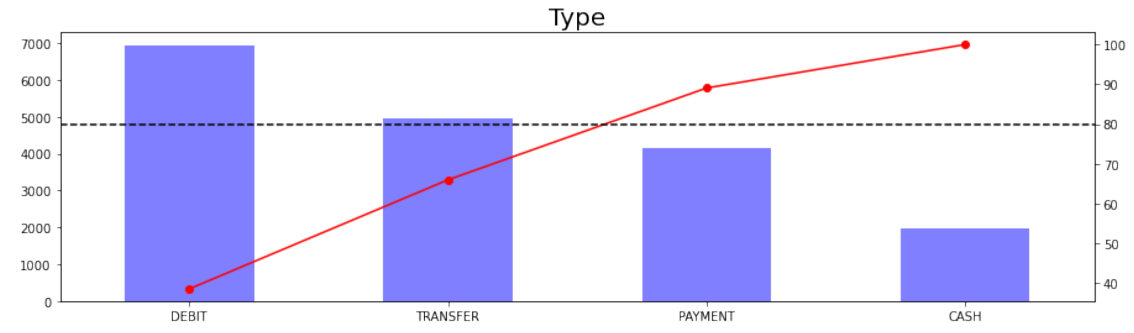
[13]:
```
data_cat.nunique()
```

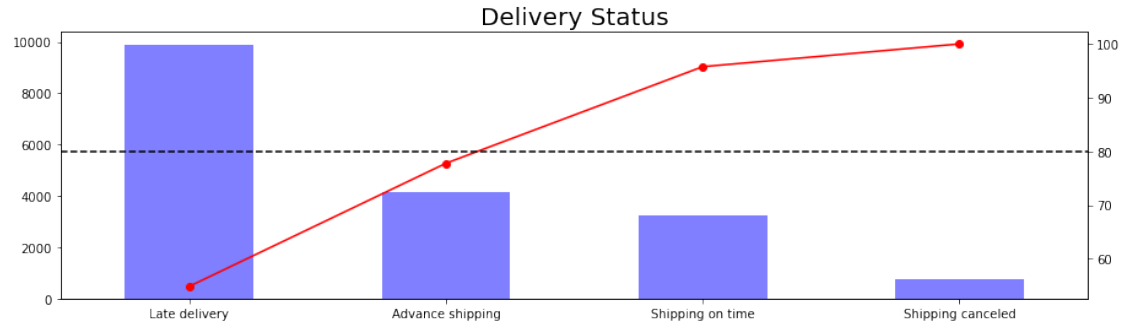
[13]:
```
Type                        4
Delivery Status             4
Late_delivery_risk          2
Category Id                51
Category Name              50
Customer City             558
Customer Country            2
Customer Email              1
Customer Fname            597
Customer Id              9564
Customer Lname           1036
Customer Password           1
Customer Segment            3
Customer State             44
Customer Street          5851
Customer Zipcode          988
Department Id              11
Department Name            11
Latitude                 6556
Longitude                3360
Market                      5
Order City               2769
Order Country             146
Order Customer Id        9564
Order Id                   15872
Order Item Cardprod Id       118
Order Item Id              18052
Order Region                  23
Order State                  912
Order Status                   9
Order Zipcode                453
Product Card Id              118
Product Category Id           51
Product Description            1
Product Image                118
Product Name                 118
Product Status                 1
Shipping Mode                  4
dtype: int64
```

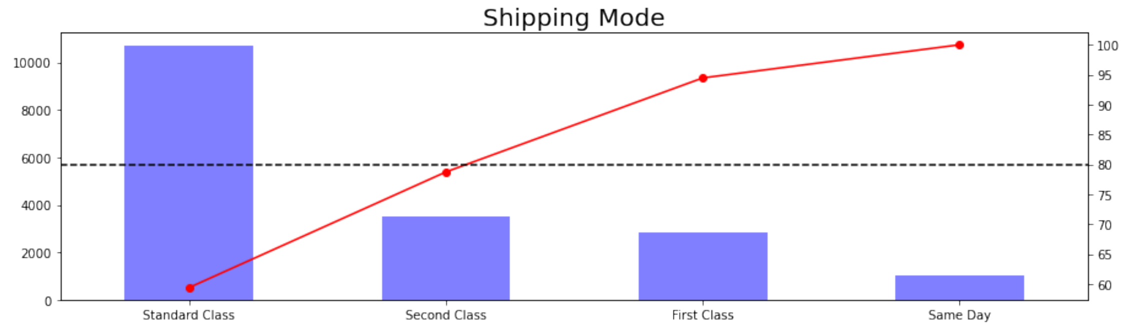
[14]:
```
retained_cat_col = ["Type", "Delivery Status", "Shipping Mode",␣
 ↪"Late_delivery_risk"]
data_cat = data_cat[retained_cat_col]
data_cat.head()
```

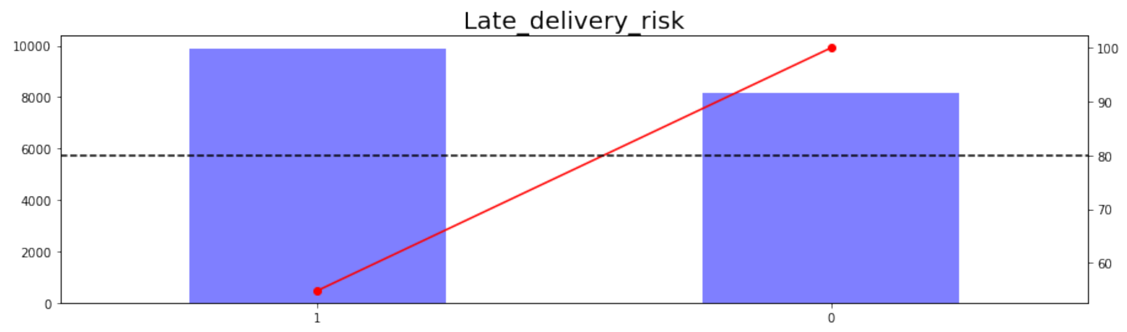
[14]:
```
            Type    Delivery Status   Shipping Mode Late_delivery_risk
131982      CASH   Advance shipping  Standard Class                  0
122089  TRANSFER  Shipping canceled     First Class                  0
59741      DEBIT   Advance shipping  Standard Class                  0
146707     DEBIT   Shipping on time  Standard Class                  0
79280      DEBIT   Shipping on time  Standard Class                  0
```

[15]:
```
for column in data_cat:
    plt.figure(figsize=(15,4))

    # Calculate value counts and sort by descending order
    value_counts = data_cat[column].value_counts().sort_values(ascending=False)

    # Create bar chart
    value_counts.plot(kind='bar', color='blue', alpha=0.5)

    # Calculate cumulative sums and convert to percentage of total
    cumulative_sums = value_counts.cumsum() / value_counts.sum() * 100

    # Create Pareto line
    cumulative_sums.plot(kind='line', marker='o', color='red', secondary_y=True)

    # Add dotted line at 80%
    plt.axhline(y=80, color='k', linestyle='--')

    plt.title(column, fontdict={'fontsize': 20})
plt.show()
```


Type
7000
6000
5000
4000
3000
2000
1000
0
DEBIT
TRANSFER
PAYMENT
CASH
100
90
80
70
60
50
40


Delivery Status
10000
8000
6000
4000
2000
0
Late delivery
Advance shipping
Shipping on time
Shipping canceled
100
90
80
70
60


Shipping Mode
10000
8000
6000
4000
2000
0
Standard Class
Second Class
First Class
Same Day
100
95
90
85
80
75
70
65
60



From the visualization

a. about 80% type of transaction made consist of DEBIT, TRANSFER, and PAYMENT

b. about 80% delivery status consist of Late Delivery, Advance Shipping, and Shipping on Time

c. about 80% shipping mode used were Standard Class and Second Class

d. Late delivery risk, which is the label we want to predict were almost equal in occurence

TODO-> bisa eksplorasi gimana statistiknya kalau Late delivery risk nya 0 dan 1

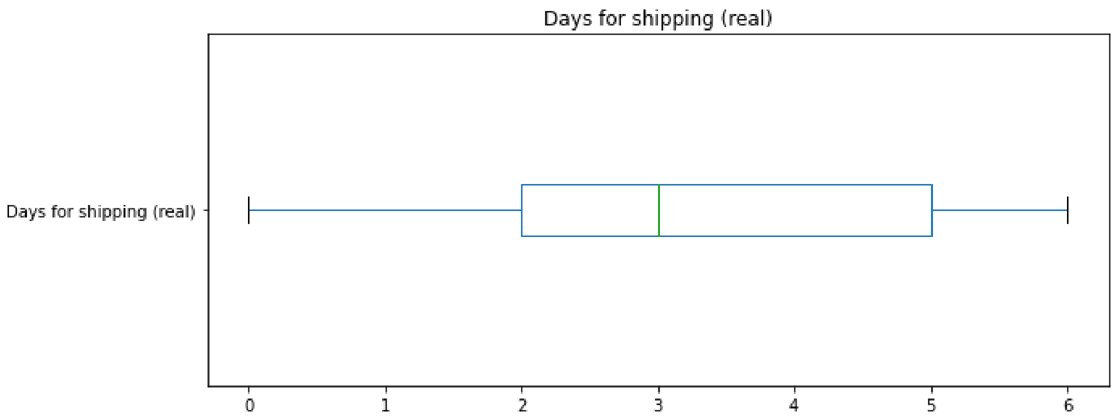
1 Box Plot

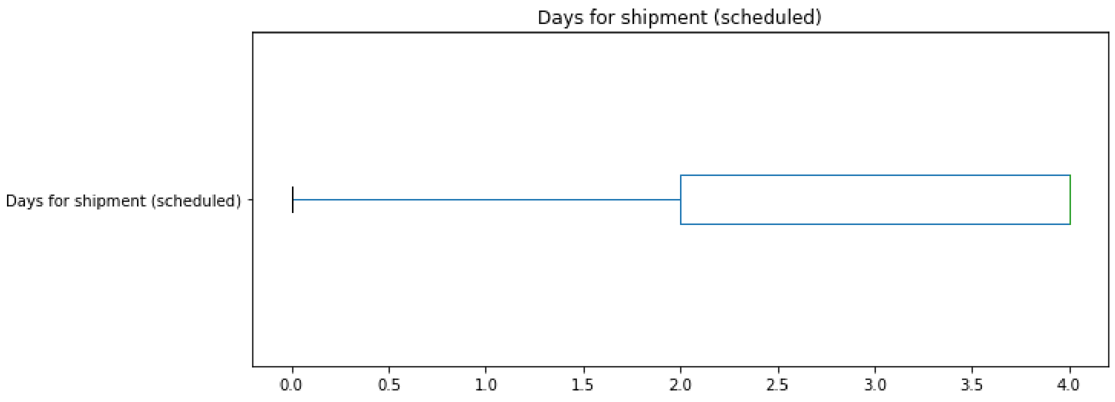
[16]:
```
data_num = data_1.select_dtypes(exclude=['object', 'datetime64[ns]'])
data_num.head()
```

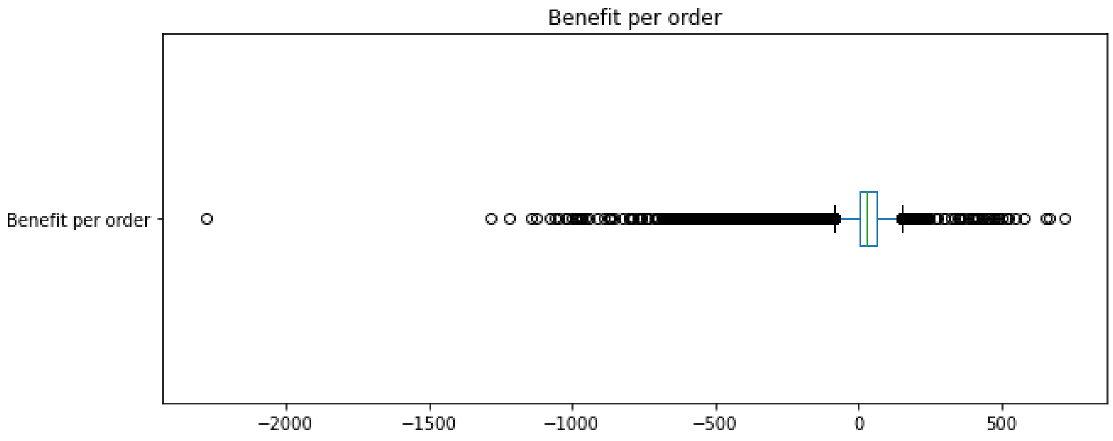
[16]:
```
        Days for shipping (real)  Days for shipment (scheduled)  \
131982                         3                              4
122089                         2                              1
59741                          2                              4
146707                         4                              4
79280                          4                              4

        Benefit per order  Sales per customer  Order Item Discount  \
131982         -47.139999          173.929993                25.99
122089          77.470001          188.949997                11.00
59741          119.839996          254.979996                45.00
146707         151.309998          387.980011                12.00
79280           46.400002          159.990005                40.00

        Order Item Discount Rate  Order Item Product Price  \
131982                      0.13                 49.980000
122089                      0.06                 39.990002
59741                       0.15                299.980011
146707                      0.03                399.980011
```

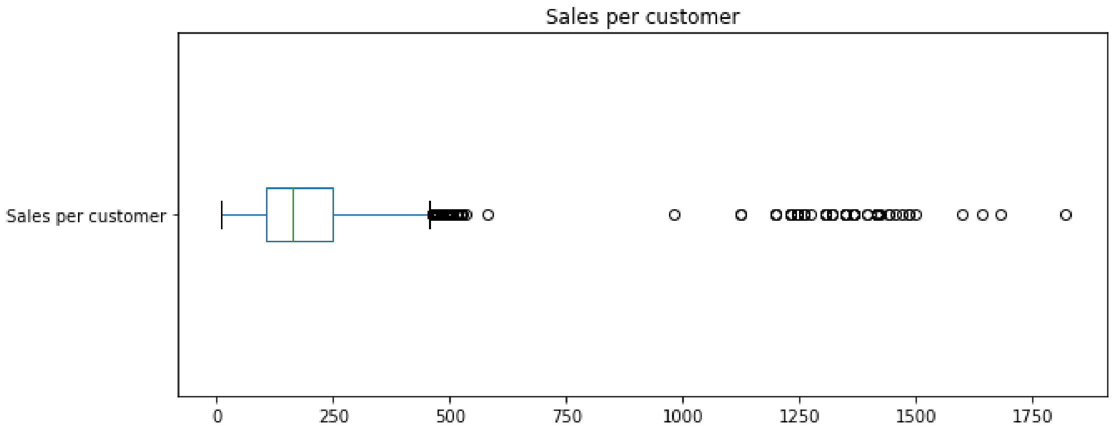
```
79280                      0.20            199.990005

        Order Item Profit Ratio  Order Item Quantity       Sales  \
131982                    -0.27                    4  199.919998
122089                     0.41                    5  199.949997
59741                      0.47                    1  299.980011
146707                     0.39                    1  399.980011
79280                      0.29                    1  199.990005

        Order Item Total  Order Profit Per Order  Product Price
131982        173.929993              -47.139999      49.980000
122089        188.949997               77.470001      39.990002
59741         254.979996              119.839996     299.980011
146707        387.980011              151.309998     399.980011
79280         159.990005               46.400002     199.990005
```

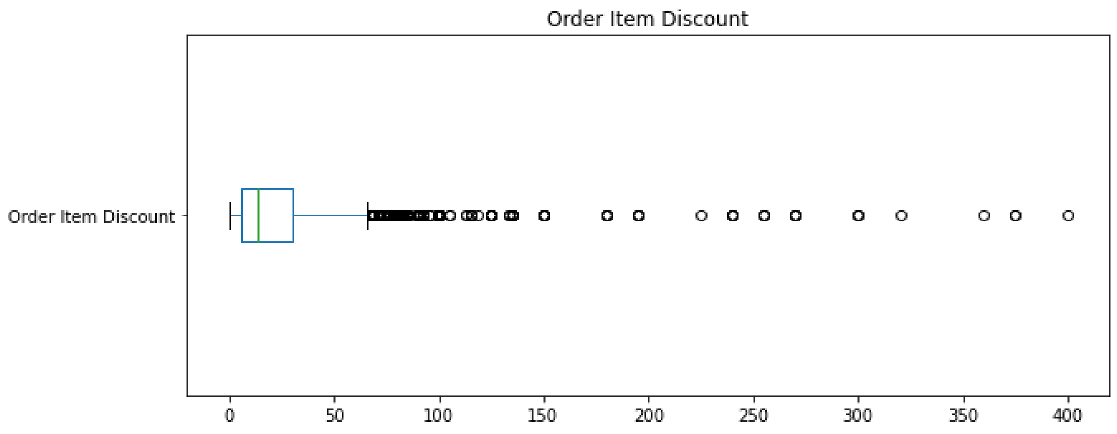
[17]:
```
for column in data_num:
    plt.figure(figsize=(10,4))
    data_num.boxplot([column], vert=False, grid=False)
    plt.title(column)
    plt.show()
```

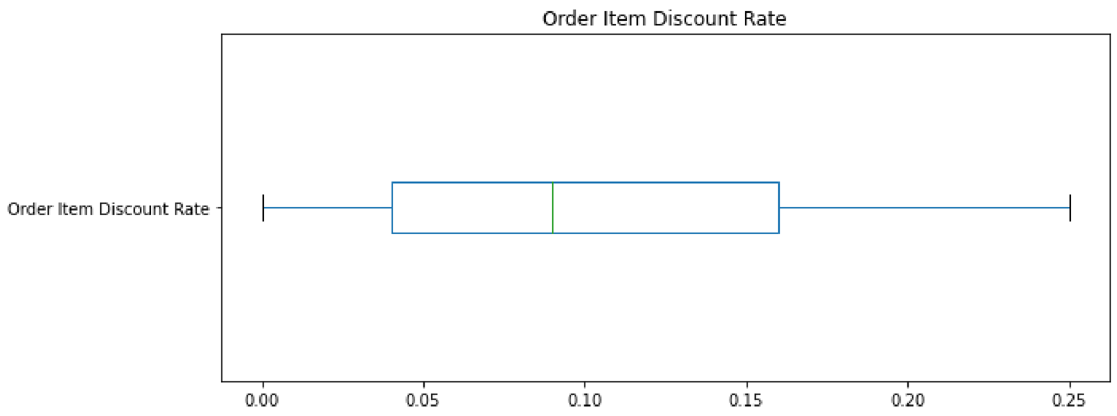


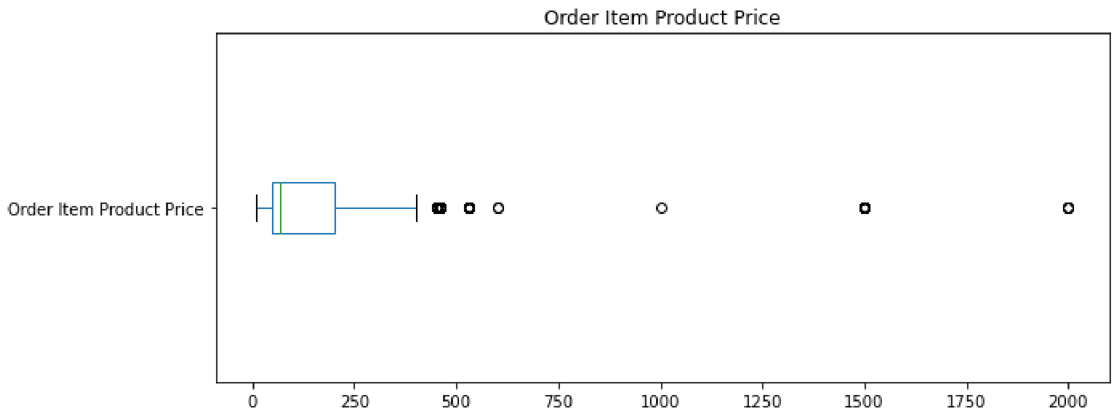

Days for shipment (scheduled)
Days for shipment (scheduled)
0.0
0.5
1.0
1.5
2.0
2.5
3.0
3.5
4.0

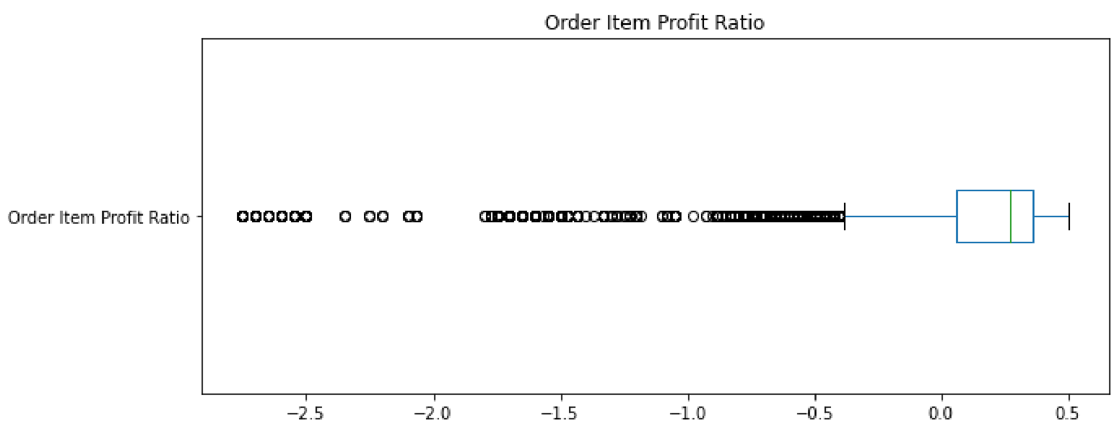

Benefit per order
Benefit per order
−2000
−1500
−1000
−500
0
500

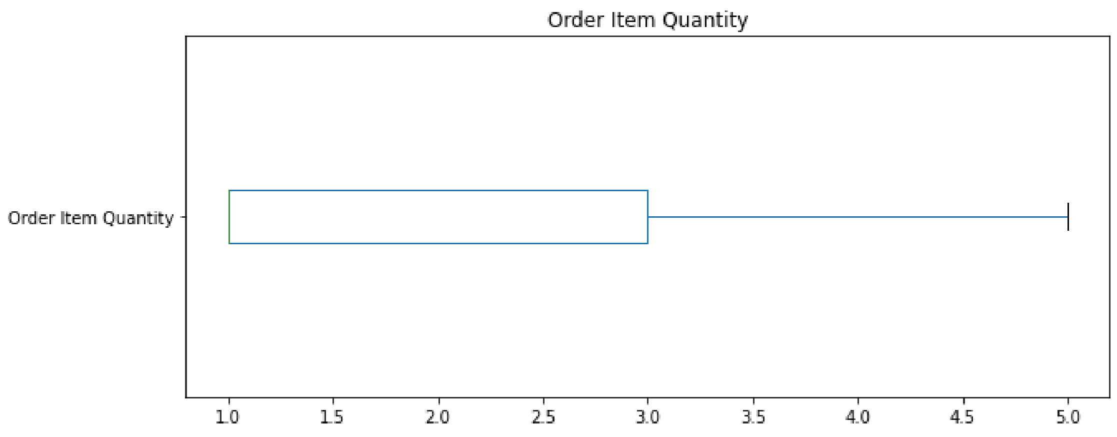

Sales per customer
Sales per customer
0
250
500
750
1000
1250
1500
1750

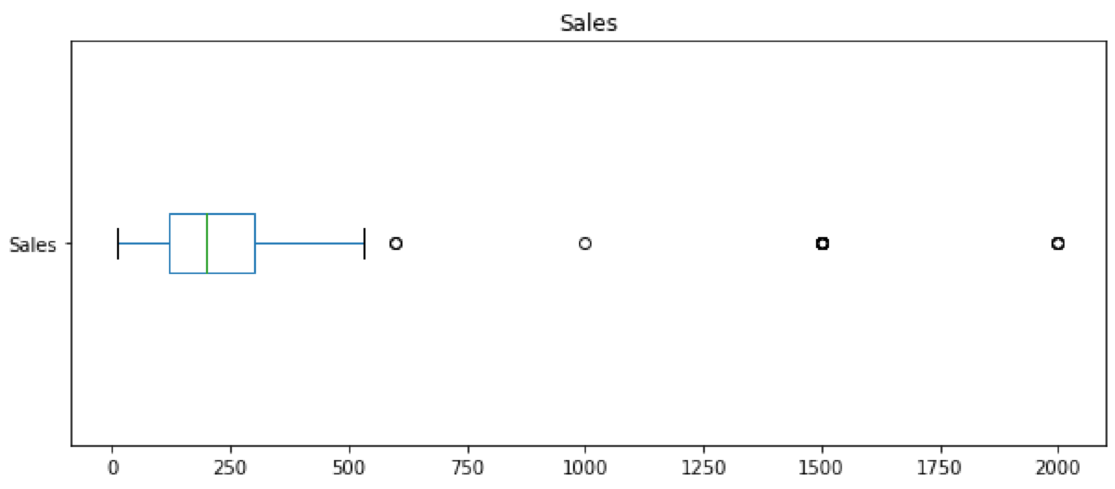

Order Item Discount
Order Item Discount
0
50
100
150
200
250
300
350
400

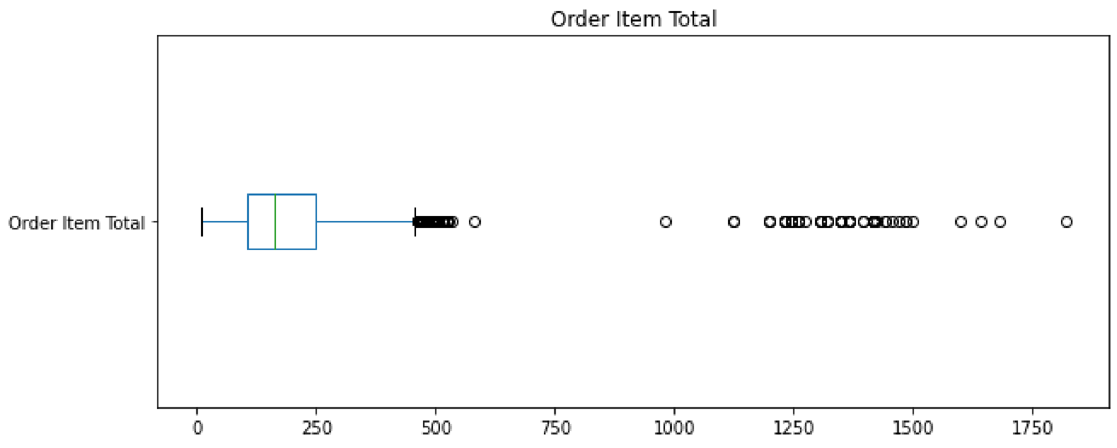

Order Item Discount Rate
Order Item Discount Rate
0.00
0.05
0.10
0.15
0.20
0.25

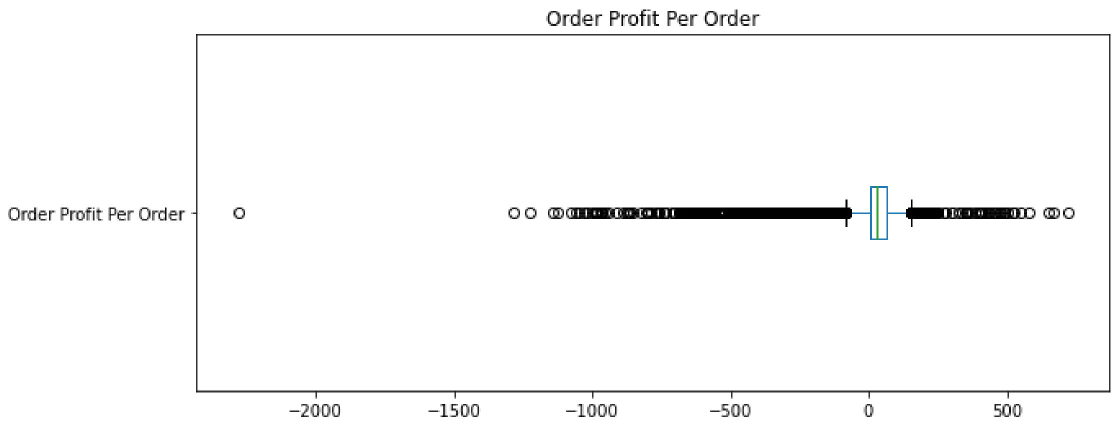

Order Item Product Price
Order Item Product Price
0
250
500
750
1000
1250
1500
1750
2000

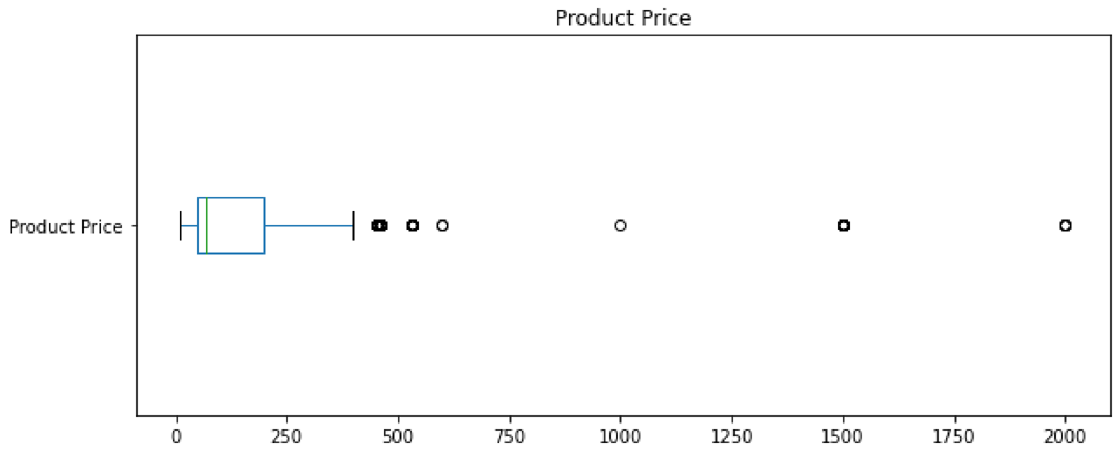

Order Item Profit Ratio
Order Item Profit Ratio
−2.5
−2.0
−1.5
−1.0
−0.5
0.0
0.5


Order Item Quantity
Order Item Quantity
1.0
1.5
2.0
2.5
3.0
3.5
4.0
4.5
5.0


Sales
Sales
0
250
500
750
1000
1250
1500
1750
2000







From boxplot created above, several point can be derived:

a. Outliers exist in Benefit per order, Sales per customer, Order Item Discount, Order Item Product Price, Order Item Profit Ratio, Sales, Order Item Total, Order Profit Per Order, Product Price

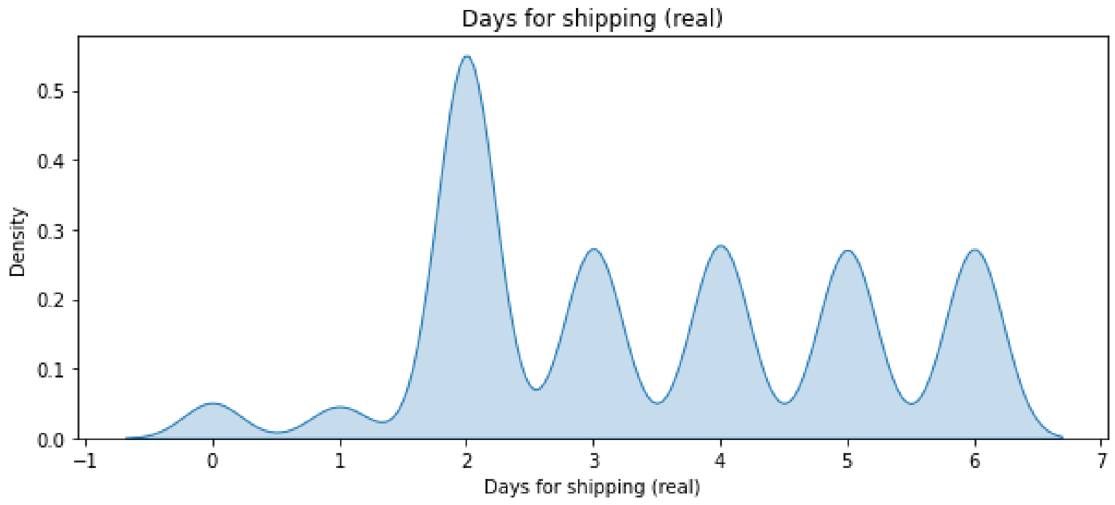
1.0.1 Data Preprocessing

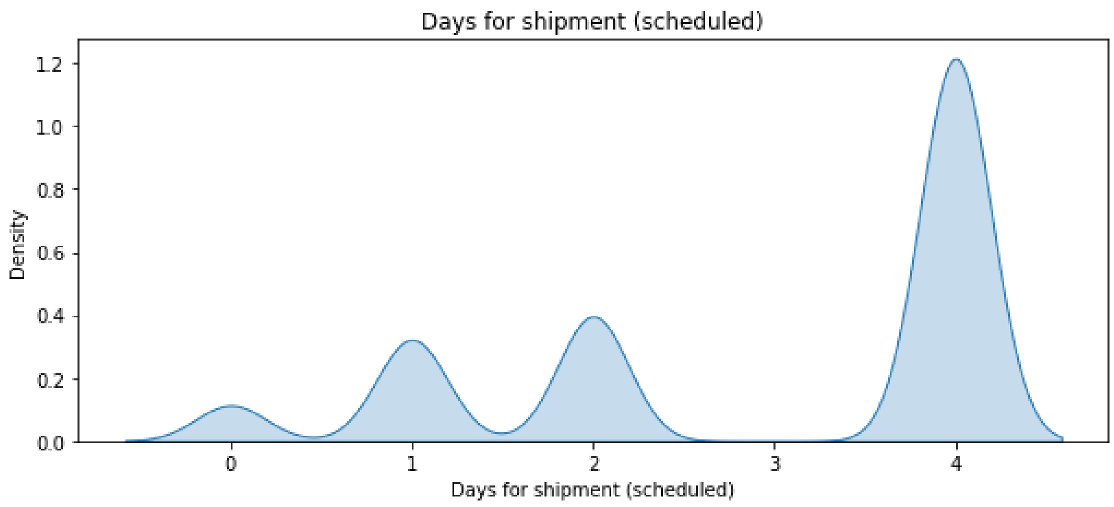
According to prior search, SVM most likely does not have prior assumptions, therefore we can proceed with current data and make preparation for training, test, and validation data. The special treatment is we gonna use stratified sampling accross data.

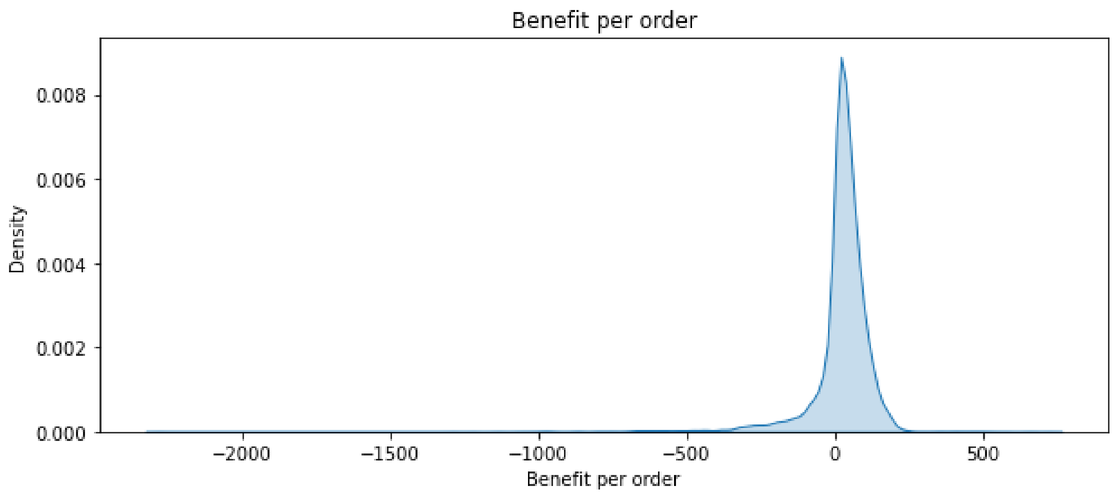
Source: https://stackoverflow.com/questions/35422072/major-assumptions-of-machine-learning-classifiers-lg-svm-and-decision-trees

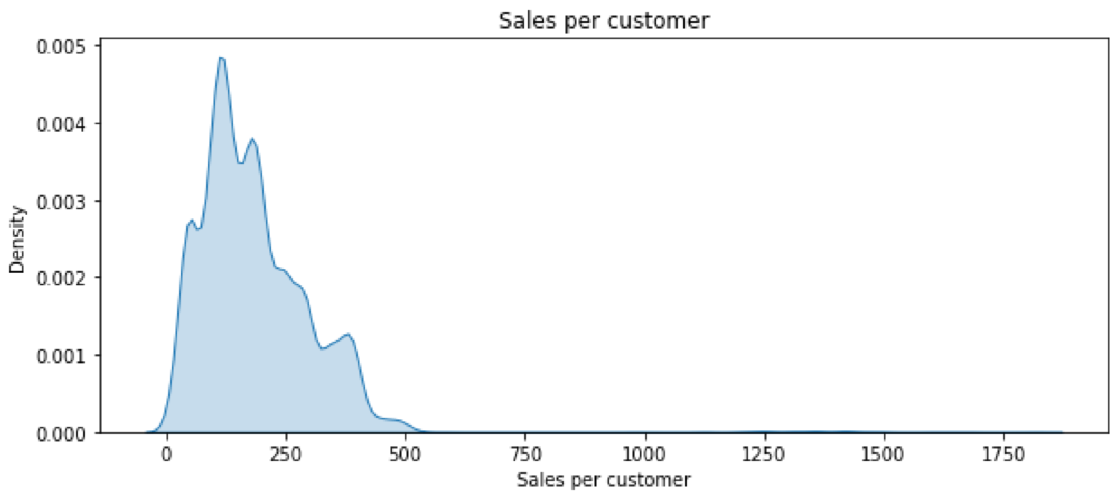
[18]:
```
rand_seed = 123
np.random.seed(rand_seed)
```

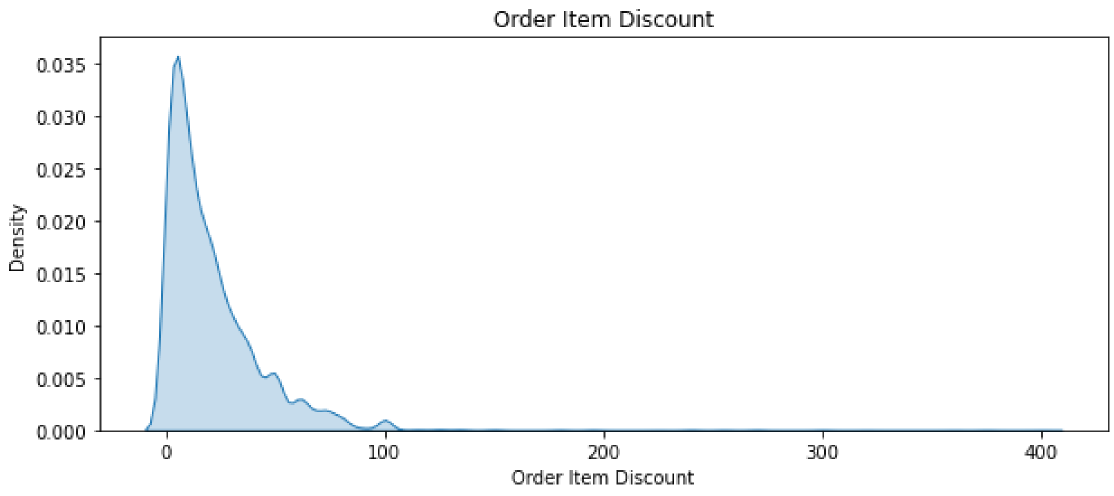
[19]:
```
for column in data_num:
    plt.figure(figsize=(10,4))
    sns.kdeplot(data_num[column], fill=True)
    plt.title(column)
    plt.show()
```

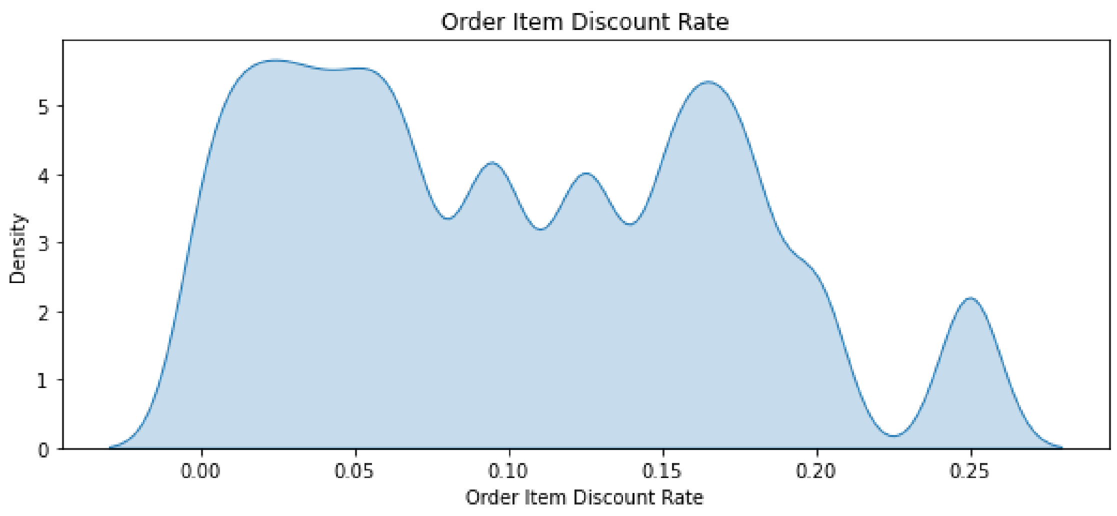


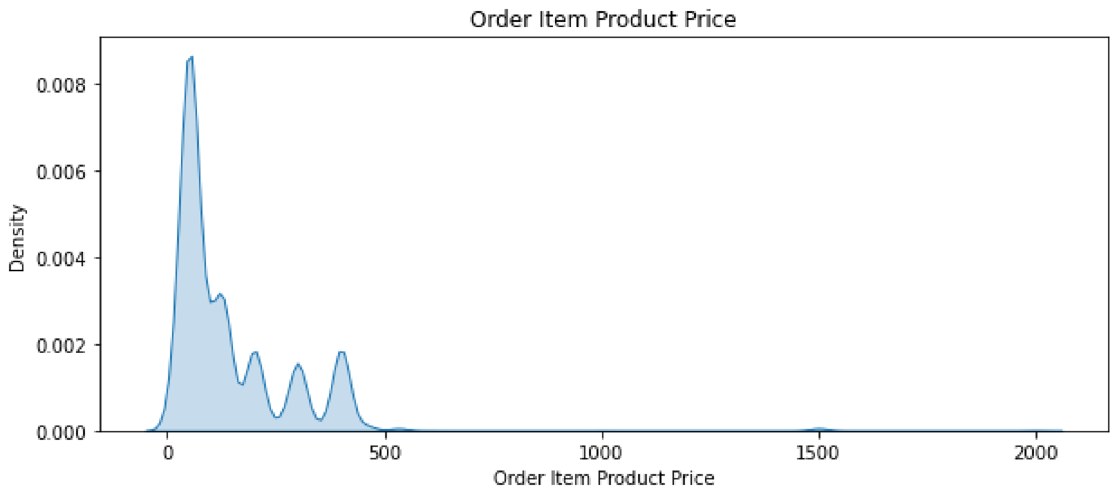

Days for shipment (scheduled)
Density
1.2
1.0
0.8
0.6
0.4
0.2
0.0
0
1
2
3
4
Days for shipment (scheduled)

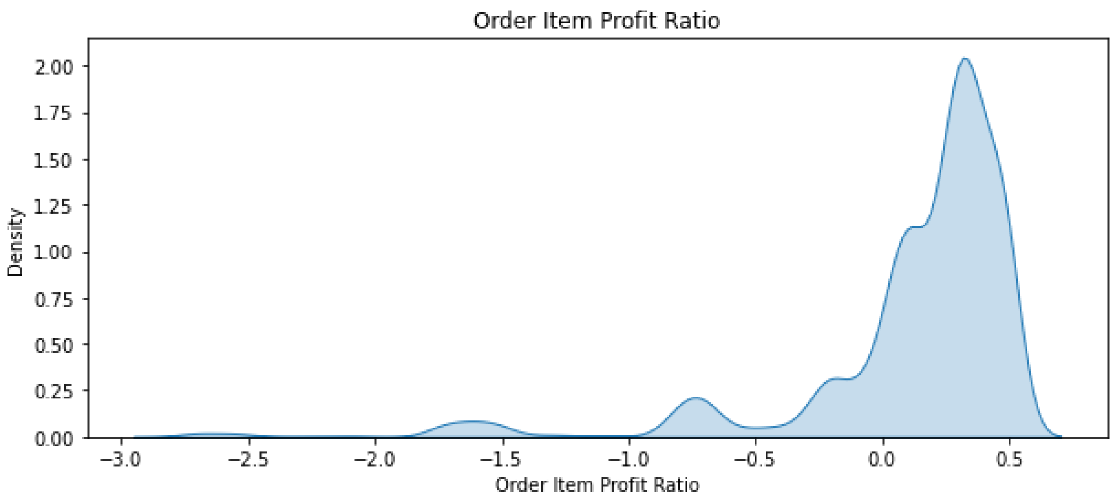

Benefit per order
Density
0.008
0.006
0.004
0.002
0.000
−2000
−1500
−1000
−500
0
500
Benefit per order

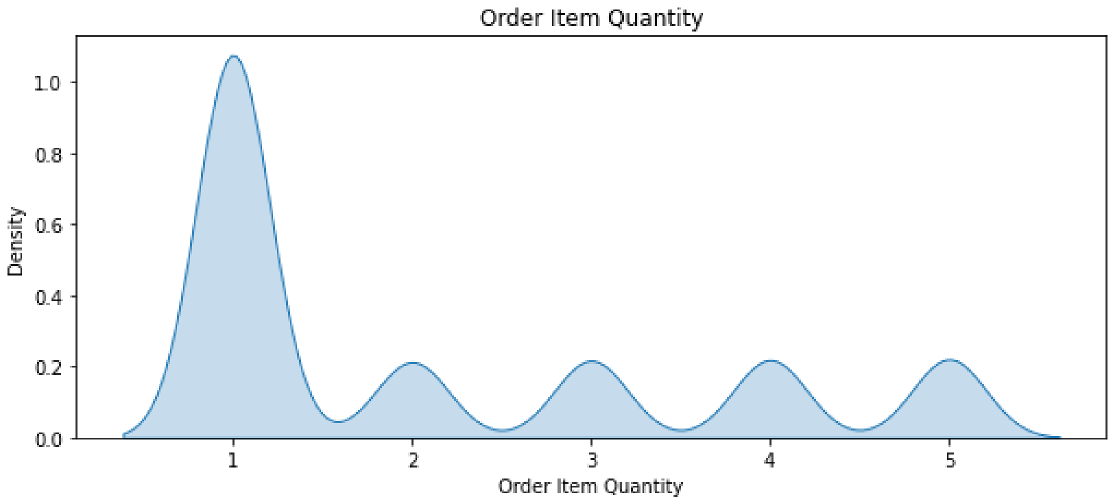

Sales per customer
Density
0.005
0.004
0.003
0.002
0.001
0.000
0
250
500
750
1000
1250
1500
1750
Sales per customer

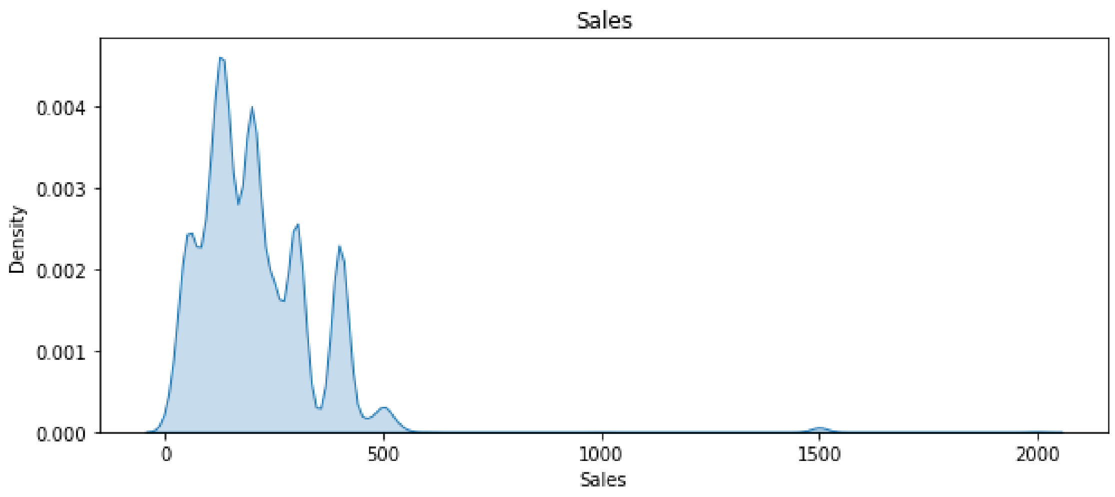

Order Item Discount
Density
0.035
0.030
0.025
0.020
0.015
0.010
0.005
0.000
0
100
200
300
400
Order Item Discount

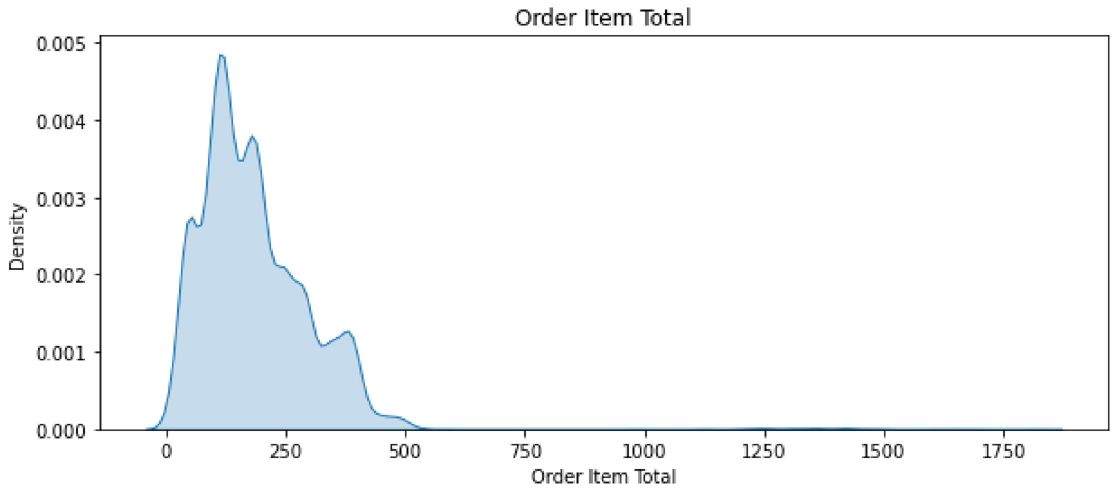

Order Item Discount Rate
Density
0
1
2
3
4
5
0.00
0.05
0.10
0.15
0.20
0.25
Order Item Discount Rate

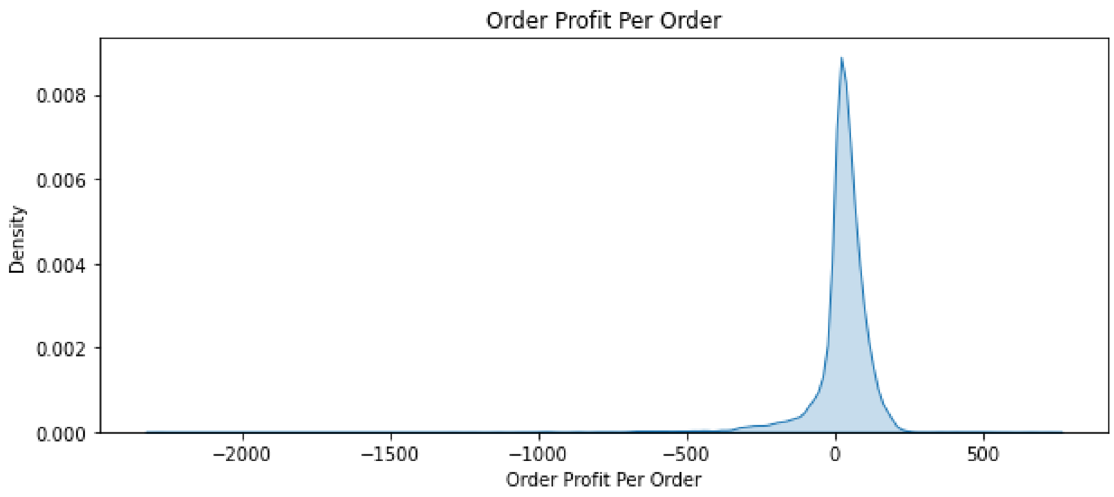

Order Item Product Price
Density
0.000
0.002
0.004
0.006
0.008
0
500
1000
1500
2000
Order Item Product Price

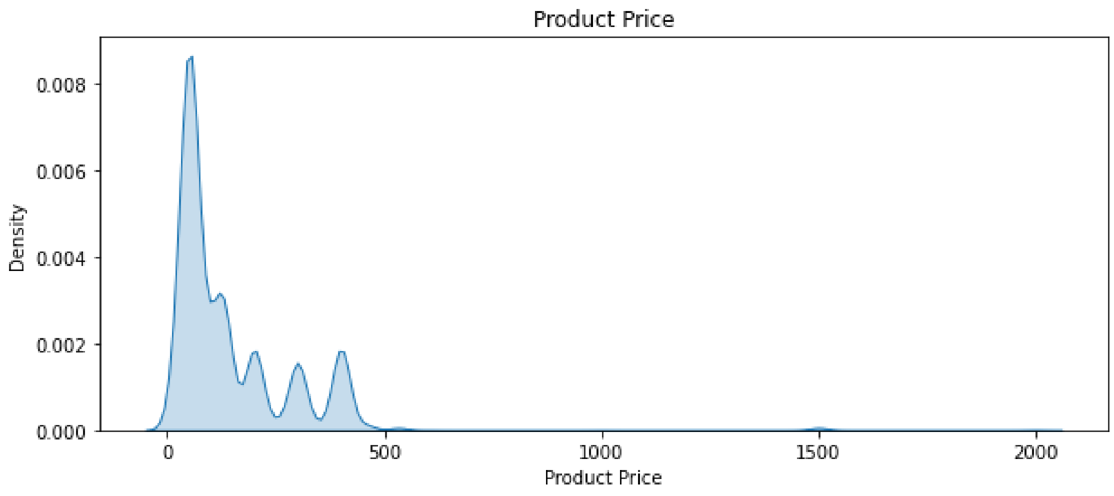

Order Item Profit Ratio
Density
2.00
1.75
1.50
1.25
1.00
0.75
0.50
0.25
0.00
−3.0
−2.5
−2.0
−1.5
−1.0
−0.5
0.0
0.5
Order Item Profit Ratio


Order Item Quantity
Density
1.0
0.8
0.6
0.4
0.2
0.0
1
2
3
4
5
Order Item Quantity


Sales
Density
0.000
0.001
0.002
0.003
0.004
0
500
1000
1500
2000
Sales


Order Item Total
Density
0.000
0.001
0.002
0.003
0.004
0.005
0
250
500
750
1000
1250
1500
1750
Order Item Total





[20]:
```
# Measure the skewness of each column
skewness = data_num.apply(lambda x: x.skew())
print(skewness)

# Normalize skewed data
#for column in data_num:
#    if np.min(data_num[column]) > 0: # Box-Cox Transformation requires␣
 ↪strictly positive data
#        if data_num[column].skew() > 0.5 or data_num[column].skew() < -0.5: #␣
 ↪check for skewness
#            data_num[column], _ = stats.boxcox(data_num[column]) # apply␣
 ↪Box-Cox transformation
```

```
Days for shipping (real)         0.086850
Days for shipment (scheduled)   -0.717514
Benefit per order               -3.978533
Sales per customer               3.139386
Order Item Discount              3.451388
Order Item Discount Rate         0.349407
Order Item Product Price         3.549491
Order Item Profit Ratio         -2.867483
Order Item Quantity              0.886592
Sales                            3.276199
Order Item Total                 3.139386
Order Profit Per Order          -3.978533
Product Price                    3.549491
dtype: float64
```

[21]:
```
#data_num.apply(lambda x: x.skew())
```

[22]:
```
# Separate features and target
X = data_num.drop(["Days for shipping (real)","Days for shipment (scheduled)"],␣
 ↪axis=1)
y = data_cat["Late_delivery_risk"].astype(int)

# Resampling configuration
#smote = SMOTE(random_state=rand_seed)

# Perform resampling
#X, y = smote.fit_resample(X, y)
```

[23]:
```
# Assuming X is your feature matrix and y are your labels
# Generate a random sample for training, testing, and validating
X_train, X_temp, y_train, y_temp = train_test_split(X, y, test_size=0.3,␣
 ↪random_state=rand_seed, stratify=y)
X_val, X_test, y_val, y_test = train_test_split(X_temp, y_temp, test_size=0.5,␣
 ↪random_state=rand_seed, stratify=y_temp)
```

Support vector machines (SVMs) are a set of supervised learning methods used for classification, regression and outliers detection.

The advantages of support vector machines are:

- Effective in high dimensional spaces.
- Still effective in cases where number of dimensions is greater than the number of samples.
- Uses a subset of training points in the decision function (called support vectors), so it is also memory efficient.
- Versatile: different Kernel functions can be specified for the decision function. Common kernels are provided, but it is also possible to specify custom kernels.

The disadvantages of support vector machines include:

- If the number of features is much greater than the number of samples, avoid over-fitting in choosing Kernel functions and regularization term is crucial.

- SVMs do not directly provide probability estimates, these are calculated using an expensive five-fold cross-validation (see Scores and probabilities, below).

The support vector machines in scikit-learn support both dense (numpy.ndarray and convertible to that by numpy.asarray) and sparse (any scipy.sparse) sample vectors as input. However, to use an SVM to make predictions for sparse data, it must have been fit on such data. For optimal performance, use C-ordered numpy.ndarray (dense) or scipy.sparse.csr_matrix (sparse) with dtype=float64.

Source: https://scikit-learn.org/stable/modules/svm.html

[24]:
```
# Train a Support Vector Machine

# Define the parameter grid
#param_grid = {
#    'C': [0.1, 1, 10, 100, 1000],
#    'kernel': ['linear', 'rbf', 'poly', 'sigmoid'],
#    'degree': [2, 3, 4],  # only used when kernel is 'poly'
#    'gamma': [1, 0.1, 0.01, 0.001, 0.0001],  # not used when kernel is 'linear'
#    'coef0': [0.0, 0.1, 0.5]  # only used when kernel is 'poly' or 'sigmoid'
#}


# Create a base model
#svm_base = svm.SVC(random_state=rand_seed)

#### Instantiate the grid search model
#grid_search = GridSearchCV(estimator=svm_base, param_grid=param_grid,
#                           cv=2, n_jobs=-1, verbose=2)

# Fit the grid search to the data
#grid_search.fit(X_train, y_train)

# Get the best parameters
#best_params = grid_search.best_params_

#print("Best parameters: ", best_params)
```

This hyper parameter tunning is not performed because it took a very long time to finish, even with stratified sampling it to 10% of the original data

[25]:
```
def eval_metrics(actual, pred):
    accuracy = accuracy_score(actual, pred)
    f1 = f1_score(actual, pred)
    recall = recall_score(actual, pred)
    precision = precision_score(actual, pred)
    return accuracy, f1, recall, precision
```

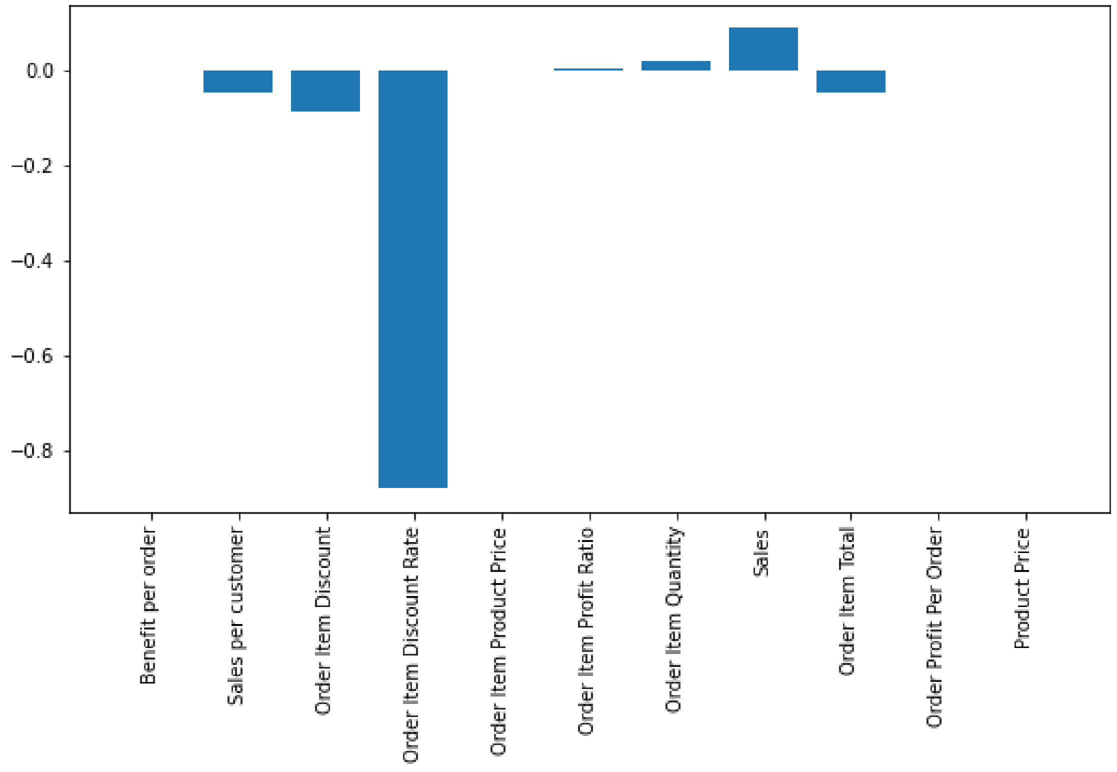
[26]:
```
mlflow.set_tracking_uri("http://localhost:5000")
mlflow.set_experiment("Order_Delivery")
```

[26]:
```
<Experiment: artifact_location='mlflow-artifacts:/102778540419101379',
creation_time=1696053492284, experiment_id='102778540419101379',
last_update_time=1696053492284, lifecycle_stage='active', name='Order_Delivery',
tags={}>
```

[27]:
```
if __name__ == "__main__":
    warnings.filterwarnings("ignore")
    rand_seed = 123
    np.random.seed(rand_seed)

    with mlflow.start_run(run_name="order_delivery_linear"):
        clf = svm.SVC(kernel="linear", random_state=rand_seed, probability=␣
 ↪True)
        clf.fit(X_train, y_train)

        # Test SVM Model on Test Data
        y_pred = clf.predict(X_test)
        (accuracy, f1, recall, precision) = eval_metrics(y_test, y_pred)

        print(f"Accuracy: {accuracy}")
        print(f"F1 Score: {f1}")
        print(f"Recall: {recall}")
        print(f"Precision: {precision}")

        mlflow.log_param("accuracy", accuracy)
        mlflow.log_param("f1 score", f1)
        mlflow.log_param("recall", recall)
        mlflow.log_param("precision", precision)

        # Test SVM Model on validation data
        y_val_pred = clf.predict(X_val)
        (val_accuracy, val_f1, val_recall, val_precision) = eval_metrics(y_val,␣
 ↪y_val_pred)

        print(f"Validation Accuracy: {val_accuracy}")
        print(f"Validation F1 Score: {val_f1}")
        print(f"Validation Recall: {val_recall}")
        print(f"Validation Precision: {val_precision}")

        mlflow.log_param("validation_accuracy", val_accuracy)
        mlflow.log_param("validation_f1 score", val_f1)
        mlflow.log_param("validation_recall", val_recall)
        mlflow.log_param("validation_precision", val_precision)
        # Assuming clf is your trained model
        try:
            # This will only work when clf is a linear model
            importance = clf.coef_[0]

            # summarize feature importance
            for i, j in enumerate(importance):
                print('Feature: %s, Score: %.5f' % (X_train.columns[i], j))

            # plot feature importance
            plt.figure(figsize=(10, 5))
            plt.bar(X_train.columns, importance)
            plt.xticks(rotation=90)  # Rotate feature names for readability
                    # Save the figure as a PNG
            if not os.path.exists("images"):
                os.mkdir("images")

            plt.savefig("feature_importance.png")
            mlflow.log_artifact("feature_importance.png")

            plt.show()
        except AttributeError:
            print("coef_ is only available when using a linear kernel")

        predictions = clf.predict(X_train)
        signature = infer_signature(X_train, predictions)
        tracking_url_type_store = urlparse(mlflow.get_tracking_uri()).scheme

        # Model registry does not work with file store
        if tracking_url_type_store != "file":
            mlflow.sklearn.log_model(clf, "model",␣
↪registered_model_name="OrderDelivery", signature=signature)
        else:
            mlflow.sklearn.log_model(clf, "model", signature=signature)
```

```
Accuracy: 0.5472673559822747
F1 Score: 0.7065581617999043
Recall: 0.9939393939393939
Precision: 0.5480876346082436
Validation Accuracy: 0.5491137370753324
Validation F1 Score: 0.7067018976699496
Validation Recall: 0.9905723905723905
Validation Precision: 0.5492905153099328
Feature: Benefit per order, Score: -0.00025
Feature: Sales per customer, Score: -0.04538
Feature: Order Item Discount, Score: -0.08485
Feature: Order Item Discount Rate, Score: -0.88036
Feature: Order Item Product Price, Score: -0.00023
```

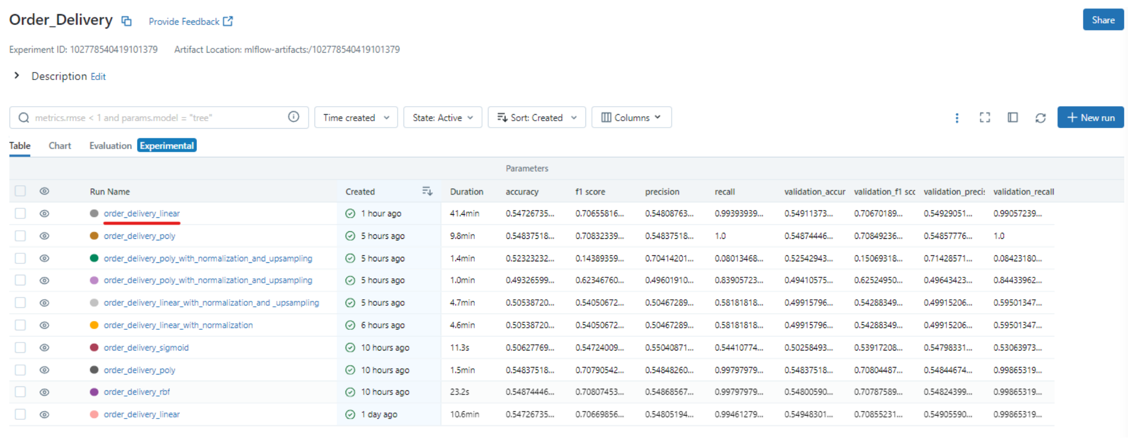
```
Feature: Order Item Profit Ratio, Score: 0.00271
Feature: Order Item Quantity, Score: 0.01784
Feature: Sales, Score: 0.08883
Feature: Order Item Total, Score: -0.04538
Feature: Order Profit Per Order, Score: -0.00025
Feature: Product Price, Score: -0.00023
```



```
Registered model 'OrderDelivery' already exists. Creating a new version of this
model…
2023/10/01 21:08:15 INFO mlflow.tracking._model_registry.client: Waiting up to
300 seconds for model version to finish creation. Model name: OrderDelivery,
version 15
Created version '15' of model 'OrderDelivery'.
```

[31]:
```
from IPython.display import Image
Image("D:
↪\Kuliah\semester_3\kecerdasan_buatan\Github\Artificial_Inteligence\Pertemuan_4\mlflow_model
↪png")
```

[31]:


Order_Delivery

Experiment ID: 102778540419101379 Artifact Location: mlflow-artifacts:/102778540419101379

Description Edit

metrics.rmse < 1 and params.model = "tree"

			Parameters							
Run Name	Created	Duration	accuracy	f1 score	precision	recall	validation_accur	validation_f1 scc	validation_preci:	validation_recall
order_delivery_linear	1 hour ago	41.4min	0.54726735...	0.70655816...	0.54808763...	0.99393939...	0.54911373...	0.70670189...	0.54929051...	0.99057239...
order_delivery_poly	5 hours ago	9.8min	0.54837518...	0.70832339...	0.54837518...	1.0	0.54874446...	0.70849236...	0.54857776...	1.0
order_delivery_poly_with_normalization_and_upsampling	5 hours ago	1.4min	0.52323232...	0.14389359...	0.70414201...	0.08013468...	0.52542943...	0.15069318...	0.71428571...	0.08423180...
order_delivery_poly_with_normalization_and_upsampling	5 hours ago	1.0min	0.49326599...	0.62346760...	0.49601910...	0.83905723...	0.49410575...	0.62524950...	0.49643423...	0.84433962...
order_delivery_linear_with_normalization_and _upsampling	5 hours ago	4.7min	0.50538720...	0.54050672...	0.50467289...	0.58181818...	0.49915796...	0.54288349...	0.49915206...	0.59501347...
order_delivery_linear_with_normalization	6 hours ago	4.6min	0.50538720...	0.54050672...	0.50467289...	0.58181818...	0.49915796...	0.54288349...	0.49915206...	0.59501347...
order_delivery_sigmoid	10 hours ago	11.3s	0.50627769...	0.54724009...	0.55040871...	0.54410774...	0.50258493...	0.53917208...	0.54798331...	0.53063973...
order_delivery_poly	10 hours ago	1.5min	0.54837518...	0.70790542...	0.54848260...	0.99797979...	0.54837518...	0.70804487...	0.54844674...	0.99865319...
order_delivery_rbf	10 hours ago	23.2s	0.54874446...	0.70807453...	0.54868567...	0.99797979...	0.54800590...	0.70787589...	0.54824399...	0.99865319...
order_delivery_linear	1 day ago	10.6min	0.54726735...	0.70669856...	0.54805194...	0.99461279...	0.54948301...	0.70855231...	0.54905590...	0.99865319...

Lets learn about how to interpret the metrics

1 - Accuracy is suitable with balanced dataset when there are an equal number of observations in each class which isn't common in real-life problems.

2 - Precision is important when the cost of false positives is high.

3 - Recall is important when the cost of false negatives is high.

4 - F1 score considers both the precision and recall.

The accuracy shows that the model is still need some experimentation, just little better from random guessing whether the delivery will be late or not. On the contrast, Recall score is high thus it is unlikely to missidentified delivery on time

[28]:
```python
# Create a LimeTabularExplainer
explainer = lime_tabular.LimeTabularExplainer(
    training_data=X_train.values,
    feature_names=X_train.columns,
    class_names=['0', '1'],
    mode='classification'
)

# Get the instance in the test set for which we want to explain the model's␣
 ↪decision
instance = X_test.iloc[1]

# Generate explanations
exp = explainer.explain_instance(
    data_row=instance,
    predict_fn=clf.predict_proba,
    num_features=5,
    top_labels=1
)
```

```
# Visualize the explanation
exp.show_in_notebook(show_table=True, show_all=True)
```

```
<IPython.core.display.HTML object>
```

```
[30]: import plotly.express as px
fig = px.scatter(data_1, x='Order Item Discount Rate', y='Sales',␣
 ↪color='Late_delivery_risk')
fig.show()
```

Based on the coefficient, Order Item Discount Rate have the highest influence (negatively) for the outcome. This might mean that the highest the discount rate is, the more likely the delivery will be on time.