Sharhad Bashar

Data Scientist Candidate

## Data

### **Schema**

|  |  |  |
| --- | --- | --- |
| **Name** | **Type** | **Notes** |
| OrderDate | DateTime |  |
| Region | String |  |
| Rep | String |  |
| Item | String |  |
| Units | Int |  |
| UnitCost | Float |  |
| Category | Int | Types of product. E.g., supplies, furniture, etc. |
| Device | Int | Device used to place the order |
| Visits | int | Accumulated number of visits to the sales page of the corresponding item |
| FinalStatus | String |  |

### **Basic Info**

* Average size of one record is 50 bytes
* 200 B records available
* Completed : Canceled = 49:1

### **Data Sample**

|  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **OrderDate** | **Region** | **Rep** | **Item** | **Units** | **UnitCost** | **Category** | **Device** | **Visits** | **FinalStatus** |
| 1/6/20 | East | Jones | Pencil | 95 | 1.99 | 1 | 2 | 124,244 | Completed |
| 1/23/20 | Central | Kivell | Binder | 50 | 19.99 | 1 | 3 | 90,763 | Completed |
| 2/9/20 | Central | Jardine | Pencil | 36 | 4.99 | 1 | 1 | 553 | Completed |
| 2/26/20 | Central | Gill | Pen | 27 | 19.99 | 1 | 4 | 99 | Completed |
| 3/15/20 | West | Sorvino | Pencil | 56 | 2.99 | 1 | 1 | 11 | Completed |
| 4/1/20 | East | Jones | Binder | 60 | 4.99 | 1 | 1 | 242 | Completed |
| 4/18/20 | Central | Andrews | Pencil | 75 | 1.99 | 1 | 1 | 873 | Completed |
| 5/5/20 | Central | Jardine | Pencil | 90 | 4.99 | 1 | 3 | 234,633 | Canceled |
| 5/22/20 | West | Thompson | Pencil | 32 | 1.99 | 1 | 2 | 23 | Completed |
| 6/8/20 | East | Jones | Binder | 60 | 8.99 | 1 | 1 | 452 | Completed |
| 6/25/20 | Central | Morgan | Chair | 90 | 699 | 2 | 1 | 34 | Completed |
| 7/12/20 | East | Howard | Binder | 29 | 1.99 | 1 | 3 | 4 | Canceled |
| 7/29/20 | East | Parent | Binder | 81 | 19.99 | 1 | 1 | 6,987 | Completed |
| 8/15/20 | East | Jones | Pencil | 35 | 4.99 | 1 | 4 | 6 | Completed |
| 9/1/20 | Central | Smith | Desk | 2 | 125 | 2 | 2 | 13 | Completed |
| 9/29/20 | East | Parent | Shelves | 3 | 30 | 2 | 1 | 325 | Completed |

Build a model that predicts final status of incoming orders

Python

# import

Import pandas as pd

import matplotlib.pyplot as plt

From sklearn import feature\_selection

From sklearn.processing import Normalizer

From sklearn.model\_selection import train\_test\_split

From sklear.linear\_model import LogisticRegression

From sklearn.metrics import accuracy\_score, f1\_score

# code

Df = pd.read\_csv(‘orders.csv’)

df.head()

df.info()

df.describe()

Region = df[‘Region’].unique()

plt(region)

Df\_sample = df[: 100,000]

Rep = Df\_sample[‘rep’]

Y = Df\_sample[‘FinalSatatus’]

plt(rep, Y)

Corr = feature\_selection(Rep, Y)

Region\_map = {0: ‘East’, 1: North, 2: South, 3: West, 4: Central}

df[‘Region’] = df.replace(Region\_map)

X = df[[‘Region, Rep, Item, Units, UnitCost, Category, Device, Visits’]]

Y = df[‘FinalStatus]

Transformer = Normalizer.fit(X[units, unitCost, Visits])

Transformer.transform(X)

X\_train, X\_test, Y\_train, y\_test = train\_test\_split(X, y, test\_split = 0.2)

Model = LogisticRegression()

model.fit(X\_train, y\_train)

Predict = model.predict(x\_test)

model.score(predict, y\_test)

accuracy\_score(predict, y\_test)

f1\_score(predict, y\_test)

Questions:

1. How to handle large data. How to use pyspark and databricks
2. Correlation between data
3. Plot data for data analysis
4. Mapping region, we have values that are small and values that are big, for example east is 0, and central is 4. How does it effect the model? How can we fix it
5. Normalize data. What is it. Why do you do it. Real world example of normalizing
6. Diff between mean accuracy and overall accuracy
7. Precision vs recall
8. Overfitting
9. Your model does bad on blind test. Why? How do you fix it
10. Lets say completed is 1, and 0 in cancelled. Model is bad, and always predicts 1. So its 98% accurate. But if we are only evaluating cancelled, it is very bad. How to fix it