AMAZON SALES DOCUMENTATION

By Hassan Eltinay

AGENDA

- Exploratory Data Analysis
- Data Preprocessing
- Data Visualization & Insights
- Predictive Modeling

1. EXPLORATORY DATA ANALYSIS

Data Inspection

Data Source: CSV file

Dataset Overview: 24 columns and 128974 rows

Data types: bool(1), float64(2), int64(2), object(19)

It was evident that the dataset indeed needs a lot of data-preprocessing

2. DATA PRE-PROCESSING

Removal of irrelevant Columns

- Unnamed: 22

Why? its an unnamed column and its purpose is unknown

- fulfilled-by

Why? since we are not interested in fulllfilment methods and our focus is on overall sales trends and distribtions

- B2B

Why? since we arent interested in B2B vs. B2C trends

- promotion ids

Why? promotion effectivness is irrelevant to us

- courier status

Why? how courier performance affects sales is not important

- fulfilment

Why? who fulfills the orders is not relevant to us since understanding fulfilment impact is not in our agenda

Identifying and Dealing with Missing Values

- 6 Columns included Standard missing values (null values), and 1 Column included Outliers

- The 6 columns are currency, amount, ship-city, ship-state, ship postal-code and ship-country , Amount column is the one with the outliers

- Since all missing values represent only 6% of the dataset , we can simply drop all of them without affecting the quality of the dataset

Data Type Conversion

- Date column needed conversion from “object” to “datetime64[ns]”

- The rest of the columns data types were OK to proceed with the analysis

Identifying Outliers in Numerical Columns

- Used a boxplot on the Amount column to show the outliers

- The idea of how to deal with outliers is that outliers represent any value above the upper T-shaped whisker, which is qual to (Q3+1.5*IQR), or below the lower T-shaped whisker, which is equal to (Q1-1.5*IQR)
- Hence, we calculate the ist quartile Q1, 3rd quartile Q3 and the interquartile range IQR

Treatment of Outliers

- Instead of removing them, we replace the outliers with the median value of our Amount column, because removing them will result in the loss of considerable chunk of our dataset, affecting its quality

3. DATA VISUALIZATION & INSIGHTS

Monthly Sales trends

- The line plot implemented using seaborn revealed that April had the most sales, followed closely by May and then June
- March had a relatively large fall in sales

Top-Selling Categories

- Bar chart using matplotlib was used to extract the following insights:
- Set and Kurta categories had by far the most sales, proving to be the most popular categories
- Western Dress and Top categories followed, with the two having a close number of sales
- Ethnic Dress, Blouse and Bottom had low sales, with Saree and Dupatta having really, really low sales
- With Saree and Dupatta having really, really low sales, they proved to be least popular categories

Regional Sales Distribution

- Mahrashtra is the state with the highest sales, followed closely by Karnataka
- Sikkim and Nagaland are amongst the states with the lowest sales

Recommendations

- Marketing campaigns should be focused on the months of April and June
- More manufacturing of top selling products like Set and Kurta
- Focus on customer satisfaction in states that have high sales, and more marketing states and cities that have low sales

Dashboard Development

- I had some technical issues that have to do with my PC during installation of dash and other necessary libraries and packages to develop the dashboard using python, but I didnt allow it to stop me and opted to develop the dashboard using Microsoft PowerBI. I believe that either way, a tool is just a means to an end, and shouldnt be the most critical factor to reach the endgoal

- The dashboard includes KPIs, such as Average Order Value, which is a crucial e-commerce metric
- By analyzing trends in AOV over time or comparing it across different customer segments (e.g., new vs. returning customers), you can gain insights into customer spending habits and identify opportunities to influence them.
- AOV can vary by product category. Analyzing AOV alongside product categories can reveal which categories drive higher spending per order. This can inform product placement strategies or upsell/cross-sell recommendations.

4. PREDICTIVE MODELING

Logistic Regression Classification Model

- The code separates the features (relevant columns) into a DataFrame named X and the target variable ("Status") into a Series named y.
- Since Logistic Regression works best with numerical data, the code uses LabelEncoder from sklearn.preprocessing to convert categorical features in X into numerical labels. This allows the model to understand the relationships between these features and the target variable.
- The code uses train_test_split from sklearn.model_selection to split the data into training and testing sets. The training set (80% by default) is used to train the model, and the testing set (50%) is used to evaluate its performance on unseen data.

Model Training

- The code creates a Logistic Regression model using LogisticRegression from sklearn.linear_model.

- Fitting the Model: The fit method trains the model on the training data (X_train and y_train). During training, the model learns the relationships between the features and the target variable, allowing it to make predictions for new data points.
- The code uses the trained model (model) to predict the order status ("fulfilled" or "cancelled/pending") for the unseen data points in the testing set (X_test). The predictions are stored in the y_pred variable.

Model Evaluation

- The code calculates the accuracy score using accuracy_score from sklearn.metrics
- Accuracy represents the percentage of predictions the model got correct on the testing set
- The model scored a humble 63% accuracy score, 40% Precision and 63% Recall. I feel it makes sense to have such numbers since there isnt really a strong relationship/correlation betwen the input features present in the dataset and our target variablel the order status

THANK YOU