DSA Project Report: Analyzing Brazilian E-Commerce Dataset (Olist)

DSA Project Report: Analyzing Brazilian E-Commerce Dataset (Olist)

Project Objective

The main objective of this project is to extract insights and build predictive models using the Olist Brazilian E-Commerce dataset. Specifically, the focus is on:

- Predicting product price based on physical and logistical attributes.
- Predicting customer satisfaction.
- Predicting delivery performance.

Step-by-Step Methodology

1. Data Understanding & Loading

The dataset comprises multiple CSV files linked by key identifiers. These include information on customers, sellers, orders, products, payments, reviews, and geolocations. The initial step involves:

- Importing Python libraries
- Reading CSVs into DataFrames
- Understanding key files like orders, reviews, products, etc.

2. Data Merging & Preprocessing

To create a comprehensive dataset:

- Merged tables on keys
- Converted dates

- Calculated delivery metrics

- Handled missing data

- Engineered features like product volume, is_late, is_satisfied

- Encoded categorical variables

3. Exploratory Data Analysis (EDA)

Explored key trends:

- Product pricing distribution

- Freight and delivery delays

- Satisfaction trends and delivery impact

- Correlations using plots and heatmaps

4. Price Prediction (Regression Task)

Goal: Predict product price

- Model: Linear Regression

- Metrics: R = 19.6%, RMSE = R$52.21, MAPE = 89.59%

- Key drivers: weight, freight value

- Limitations: lacks brand, category, and promotional info

5. Customer Satisfaction Prediction (Classification Task)

Goal: Classify satisfaction

- Model: Logistic Regression

- Metrics: Accuracy = 78.4%, AUC-ROC = 0.596

- Issue: Severe class imbalance, recall = 0 for dissatisfied customers

- Recommendation: Use SMOTE or cost-sensitive learning

6. Delivery Performance Prediction (Classification Task)

Goal: Predict late delivery

- Model: Logistic Regression
- Metrics: Accuracy = 96.7%, AUC-ROC = 0.859, F1 = 98.3%
- Key drivers: delivery time, distance, freight value

7. Model Evaluation & Cross-Validation

- Used 5-fold CV
- Stable performance for delivery model
- Satisfaction model remained weak

8. Business Recommendations

For Sellers:

- Optimize weight/volume
- Improve delivery speed

For Olist:

- Offer pricing tools
- Alert system for delays

For Logistics:

- Use delivery models for planning
- Dynamic routing

9. Regional Context: Brazils E-Commerce

- Distance is a major factor

- Strong logistics despite geography

- Delivery models succeed; satisfaction needs better features

Conclusion

Integrated data can generate powerful insights. Delivery prediction is strong; satisfaction and pricing models need richer features.