EDA and problem-solving in the E-commerce/ supply chain domains using various ML techniques

Karan Parashar(202218004)[1], Rohan Shah(202218002)[2], Rutul Patel(202218036)[3]

·Under the guidance and mentorship of Dr. Rachit Chhaya

May 6, 2023

Abstract

Our work aims to explore and analyze a dataset related to the supply chain in the e-commerce domain using Python. It involves various exploratory data analysis techniques, such as identifying null values, handling categorical and numerical features, pairwise nature of features, correlation analysis and much more. We have also used The trends library of Google has been utilized to understand the current trend of the supply chain. In the machine learning domain, the project has applied multiple regression using the Forward propagation method and K-means clustering to obtain insights into the data. Overall, the project aims to provide a comprehensive understanding of the supply chain in the e-commerce domain, using various exploratory data analysis techniques and machine learning algorithms.

Keywords
exploratory data analysis, correlation analysis, K-means clustering

1 Introduction

Before we get to the technicalities, let us explain how EDA (Exploratory Data Analysis) can be used to gain insights from historical data and make decisions accordingly. In the supply chain domain, EDA can be used to analyze data on production rates, inventory levels, shipping times, and other key metrics to identify bottlenecks and inefficiencies in the system/s, followed by developing predictive ML algorithms for forecasting demand, optimize production and logistics processes, and improve inventory management to deadstock.

2 DataSet

We worked on a data set from DataCo Global. It comprises nearly 1.8 lac instances and fifty-three features for each in the Structured Data "DataCoSupplyChainDataset.csv". The file named "DescriptionDataCoSupplyChain.csv" is for the description of each feature in the Structured Data.

The file "tokenized_access_logs.csv" is clickstream data that refers to the record of user activity. Primarily, we worked on the Structured Data file only. It has features such as Type (mode/type of payment made), Days for shipping, customer details such as Name and location, order details like order date, order Id, and quantity and Product details such as its Category Id, Description and a few more.

Keeping the heterogeneity of Data in mind, we focused on those variables only, which helped us extract insights from them.

3 Exploratory Data Analysis

Some of the basic Python libraries we used are Pandas, NumPy, Matplotlib, Seaborn, scikit-learn and pytrends.

3.1 Data profiling

Analysing and compiling the characteristics of stored data is the process of "data profiling".

For data cleaning, feature selection, data exploration, and modelling, it is essential to understand the distinct values in each dataset feature. It makes it possible to find missing or inconsistent data. Understanding the distinctive values in each feature contributes to ensuring the accuracy and dependability of the analysis's findings. Features like 'Type', 'Late_Delivery_risk', 'Delivery Status' and 'Shipping Mode' are finitely unique in this data. Therefore, they can be handled easily in ML models by encoding.

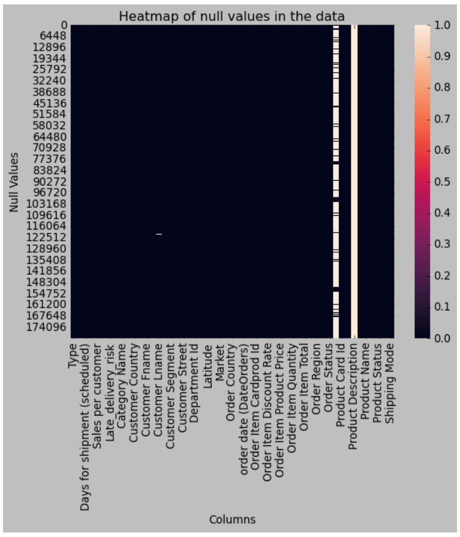

Figure 1: *This Heatmap visualizes the sparsity of missing data in the form of the white bars in between (As we can see that there are numerous missing entries for Product Description and Zipcode)*

3.2 Correlation Heatmap

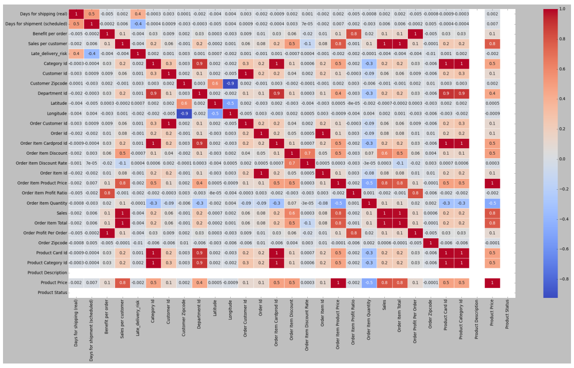

Figure 2: *This correlation Heatmap visualizes the strength of dependency amongst features in terms of correlation coefficient* r *(*$r \in [-1, 1]$*)*

As we can see, features like ‘Product Price’ and ‘Sales per Customer’, ‘Days for shipment (scheduled)’ and ‘Days for shipping (real)’ and ‘Days for shipping (real)’ and ‘Late_delivery_risk’ are positively correlated while ‘Days for shipment (scheduled)’ and ‘Late_delivery_risk’ are negatively correlated, which justifies our intuition of the effect-cause relationship between these feature pairs.

While there exists a spurious correlation between features such as ‘Order Item Quantity’ and ‘Category Id’.

3.3 CountPlot and Piechart

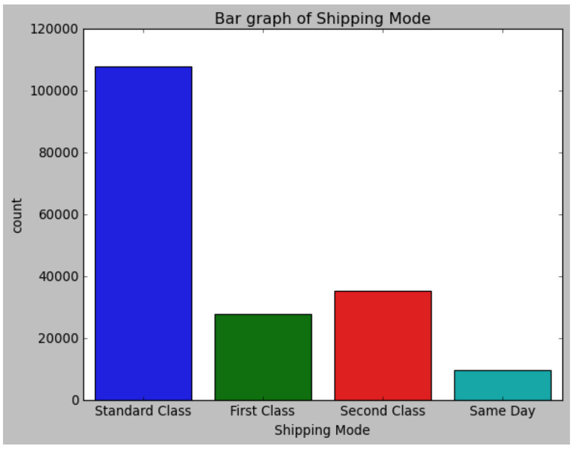

Figure 3: *Majority of the customers prefer to place order/s opting for standard mode of delivery as it’s more affordable than others followed by Second Class.*

One of the explanations for why the Second Class mode is the second most popular could be that less a difference between the price of Second and First Class delivery, with significant differences in the service. This reflects the mindset of privileged consumers who are willing to pay a bit extra to meet their needs.

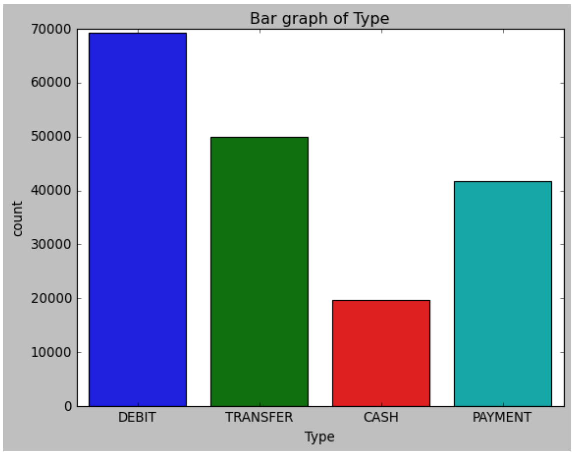

Figure 4: *This countplot represents that the customers are more likely to use Debit card for payments and cash is the least used payment option)*

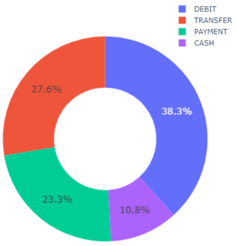

Figure 5: *Piechart for the same*

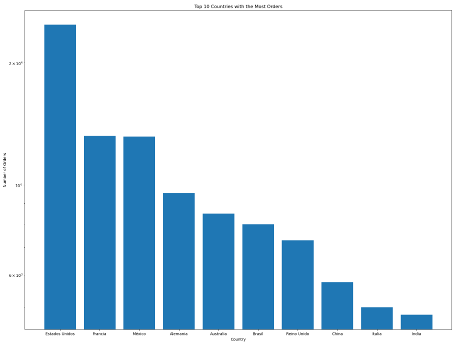

Figure 6: *Majority of the orders were placed from Estados Unidos (USA), followed by France and Mexico*

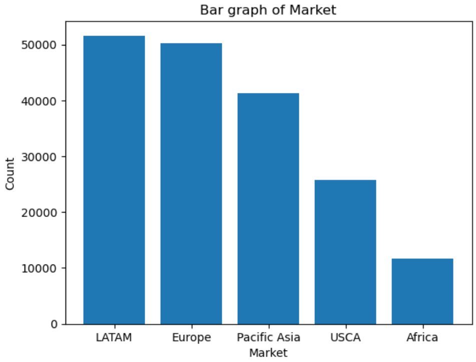

Figure 7: *LATAM contributes to the market the most*

4 Current Trend

We have used the *pytrends* library to fetch Google Trends data for the search term "E-commerce" (the same can be done for any keyword).

Here, we generate region-wise search stats using *trends.interest_by_region()* method. For instance, the latest (as of 29/04/2023) popularity of the term "E-commerce" across the globe is as follows:

geoName	E-commerce
Madagascar	100
Zimbabwe	62
Kenya	54
Senegal	52
Somalia	44
Nepal	42
Morocco	34
Nigeria	32
Tanzania	31
Philippines	31
Zambia	19

Table 1: Relative number of searches for the keyword (country-wise)

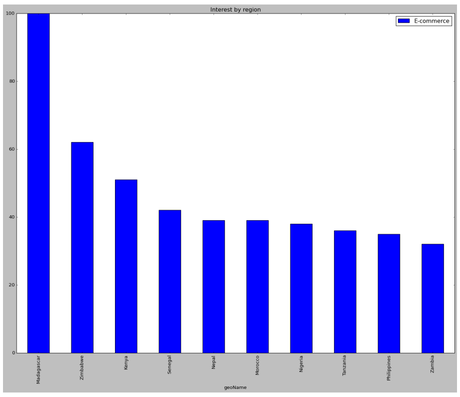

Figure 8: *Bargraph of the above result*

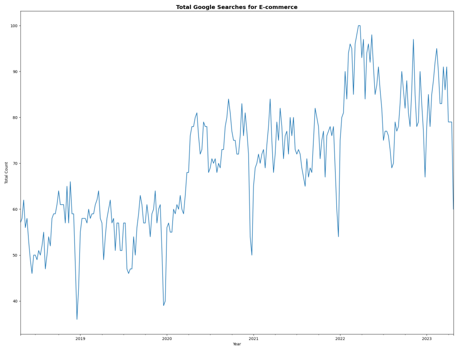

Figure 9: *Time series trend for the word "E-commerce" on Google Search Engine.*

5 Machine Learning

Our prime aim is to develop a very basic ML model to predict ‘Sales per customer’ (target variable) by selecting appropriate feature/s that minimize the test error (RMSE)

5.1 Multiple Regression using Forward Propagation

Initially, to avoid model complexity, we implemented Linear Regression for all the potential features that could contribute to determining the target variable - Y.

Forward Propagation is an ML paradigm that starts with only two determining features to build a model for univariate/bivariate data and then keeps adding more features. In the end, based on the graph of Predicted Y vs Actual Y determining the best choice of features.

Firstly, converting the categorical features into numeric, using *.cat.codes* and splitting the data in the ratio of 3:2 for the training and testing dataset. These are the plot for Actual Y vs Predicted Y value of ‘Sales per customer’ for bi, tri and multivariate data

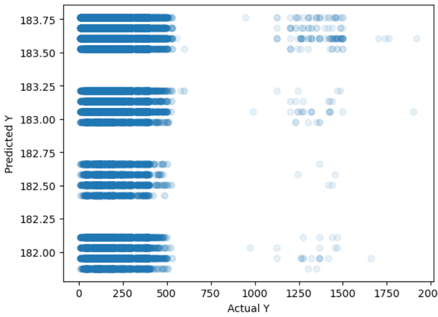

Figure 10: *X = [‘Type’, ‘Shipping Mode’] (RMSE1 = 119.831)*

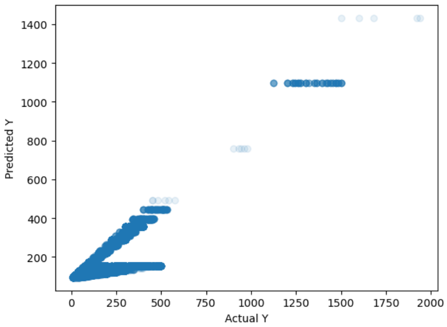

Figure 11: *X = [‘Type’, ‘Shipping Mode’, ‘Order Item Product Price’] (RMSE2 = 74.484)*

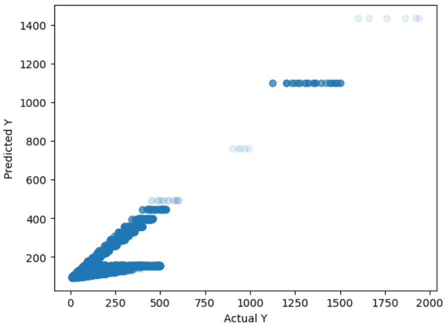

Figure 12: *X = [‘Type’, ‘Shipping Mode’, ‘Order Item Product Price’, ‘Customer Segment’] (RMSE3 = 74.841)*

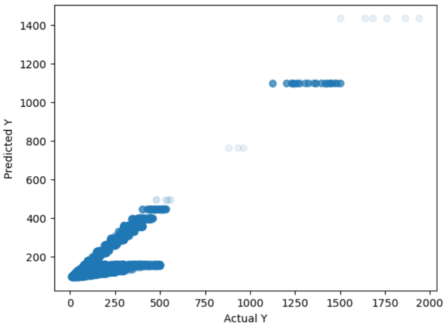

Figure 13: *X = [‘Type’, ‘Shipping Mode’, ‘Order Item Product Price’, ‘Customer Segment’, ‘Market’] (RMSE4 = 75.195)*

We get the least Test RMSE error for four features X = [‘Type’,‘Shipping Mode’,‘Order Item Product Price’].

Insights: The target variable (‘Sales per customer’), is well determined from the mentioned four features for the basic linear regression model. The same procedure can be done for other target variables using other types/forms of regression methods. Starting from a set of potential/possible determining features { Xi}, reducing each feature one by one, the set of features corresponding to *argmin(RMSE)* are considered the determining variables.

6 Conclusion

Factors such as 'Product Price', 'Shipping Mode' and 'Mode of Payment' in any consumer purchase can highly determine the number of Sales the individual contributes to the Net Sales of the company, Globally, LATAM countries (Brazil, Mexico, Argentina, Colombia, Chile etc.) has the highest market. We have also observed the price skimming strategy in the pricing of the products/services in which slight differences in quality or features, at prices that are not too far apart. Due to their perception that the more expensive product offers greater value for money than the less expensive option, customers may be more tempted to choose it.

7 Further Studies

Keeping the non-linearity of the data in mind, we will implement LASSO regression for feature selection to get a more accurate model by plotting the regularization path.

8 Acknowledgements

We sincerely thank Dr. Rachit Chhaya for his advice and mentoring during our project work. We are really grateful that he served as our mentor and value his effort to help us progress as students. We also thank our batchmates for their passive participation in seeding ideas.

9 References

- https://www.geeksforgeeks.org/exploratory-data-analysis-in-python/
  Accessed on January 30, 2023.
- https://www.microsoft.com/en-us/research/uploads/prod/2006/01/Bishop-Pattern-Recognition-and-Machine-Learning-2006.pdf
  Accessed on February 09, 2023.
- https://hastie.su.domains/Papers/ESLII.pdf
  Accessed on February 28, 2023.
- https://scikit-learn.org/stable/
  Accessed on March 16, 2023.