olist

# Capstone Project

Personalised product recommendation system for Olist customers using Machine Learning techniques.

Olabisi Sunmon

11th April 2023

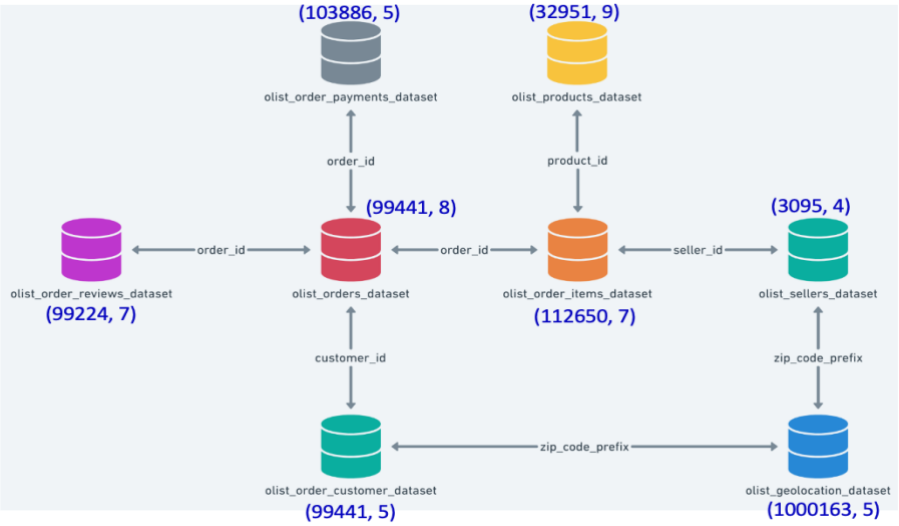
## Introduction
The goal of this project is to construct a recommendation system for Olist's customers using data from the years 2016 through 2018. The report will provide a synopsis of the approach, significant discoveries, and outcomes of the project.

## Background on the subject matter area
The e-commerce sector in Brazil has been experiencing substantial growth, thanks to increased investment and innovation from both domestic and global players. It is anticipated that the industry will continue to grow at a rapid pace, with an estimated 11.04% annual growth rate, and is projected to reach a value of US$56.9 billion by 2023. Brazil is a nation that prioritises convenience, with a preference for using smartphone apps to shop and complete transactions.

## Problem Statement
How can we create a customised product recommendation system using data analysis and machine learning techniques to help Olist customers discover new products and find relevant items for purchase, to boost revenue and customer purchase rates.

## Data Source
Olist, a Brazilian e-commerce company, gathered and saved the Brazilian E-Commerce Public Dataset on Kaggle. The data provided is authentic commercial data that has been anonymised. To maintain confidentiality, all references to the companies and partners mentioned in the review have been substituted with the names of the great houses from Game of Thrones.

The Data Schema is as follows:




There are a total of 8 datasets in the Brazilian E-Commerce Public Dataset, with an extra dataset containing English translations. All datasets are saved in CSV format.
The datasets are a complied in a combination of both Portuguese and English.
The size of the database is denoted in blue.

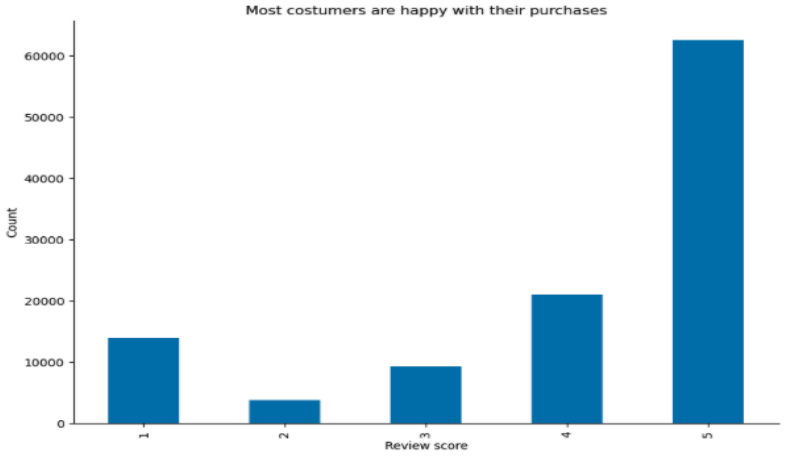
The Data can be sourced from;
https://www.kaggle.com/datasets/olistbr/brazilian-ecommerce

## Summary of Cleaning and Pre-processing
During data cleaning, I converted data types from 'object' to 'datetime' for some columns. Additionally, I removed duplicated rows from the database as they did not provide any new information. Columns were dropped when necessary to handle missing values. I merged the orders, items, customers, reviews, and product databases to create a comprehensive and relevant database for the EDA.
In the modelling process, the data frame was split into 2 by separating customers who made a single purchase (First-time Customers) and customers who made multiple purchases (Returning Customers). This is because 50% of customers have made only one purchase, this could have proposed challenges on the modelling, as the models are dependent on order history.
Due to limitations in computational power, the collaborative filtering modelling product recommendation system for returning customers was conducted on a review score stratified subset of the dataset, rather than on the entire dataset.

## Insights, modelling, and results
The goal of this project is to determine the types of products that would interest customers and recommend them accordingly. During the exploratory data analysis phase, I examined various factors such as ratings, orders, customer behaviour, and location

Dictionary
rmse;  Root mean square error

to gain valuable insights. These insights are at a high level and provide a starting point for further analysis.

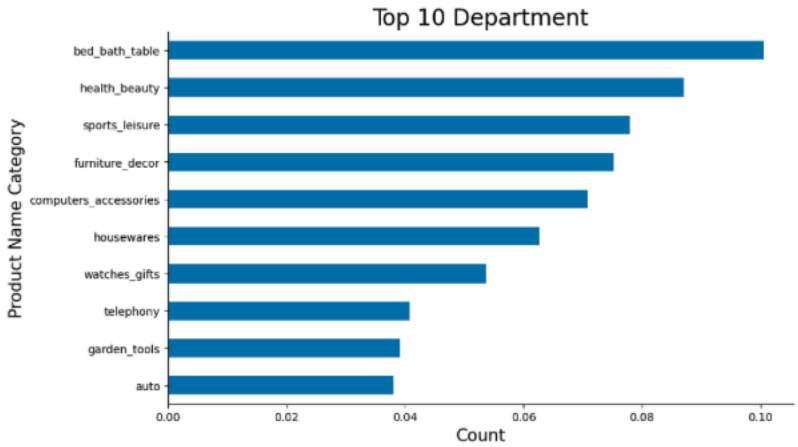



## Distribution of Ratings

While a majority of Olist orders received a 5-star rating, it would be incorrect to assume that customers are simply being generous with their ratings. Instead, based on the bar graph, it can be interpreted that customers are genuinely very satisfied with their purchases.


## Costumer State Activity

Around 70% of all product purchases come from the most heavily populated states in Brazil, namely São Paulo, Rio de Janeiro, and Minas Gerais. Meanwhile, the remaining states account for only 30% of the total product purchases.




## Most Popular Department

Out of the 71 different product categories, the most frequently purchased category is bed_bath_table, accounting for roughly 10% of all purchases. Health_beauty is a close second, making up approximately 9% of total purchases.

The purpose of this analysis was to determine whether a clustering approach should be used in the recommendation system e.g. having a system that takes state into account.

I used the market basket algorithm to devise a recommendation system that caters to both new and existing customers if they have items in their online shopping cart. Through experimenting with different levels of granularity, I found that considering all products was the optimal approach to generate more basket combinations and produce better recommendations. I recommend that Olist implements this market basket product recommendation system during the online checkout process, presented as bundle deals to incentivize customers to purchase the recommended items.

When new customers are browsing online, they should be recommended hot trending products, while returning customers should be offered products based on the FUNKSVD recommendation system. After considering different techniques such as KNNWithMeans, FunkSVD, NormalPredictor, and CoClustering, the FUNKSVD algorithm was selected as the best option for collaborative filtering.

Dictionary
rmse;  Root mean square error

| Algorithm | test_rmse | fit_time | test_time |
|---|---|---|---|
| KNNWithMeans | 1.516538 | 0.141338 | 0.006493 |
| CoClustering | 1.525092 | 0.361644 | 0.004384 |
| SVD | 1.533964 | 0.197952 | 0.006193 |
| NormalPredictor | 2.018890 | 0.002890 | 0.005964 |

Although the test_rmse score for the FunkSVD algorithm was not the lowest, it performed better on the test data due to the KNN algorithm's difficulty in handling sparse matrices. Unlike the KNN algorithm, the FunkSVD algorithm assumes that the data can be represented as a low-rank matrix, which could have resulted in a better-performing model on unseen data.

## Conclusion

Using a recommendation system that utilises machine learning can really help an e-commerce business like Olist. This system can suggest new products to customers based on their preferences, making it more likely that they'll buy something. This can increase the business's profits and make customers feel like the business understands them better which can make then a reoccurring costumer. In short, adding a recommendation system to Olist can make a big difference in both sales and customer satisfaction.

I suggest the implementation approach should be as follows.

| | Browsing | Checkout |
|---|---|---|
| New Customers | Top Trending | Market basket Recommendation |
| Returning Customers | FunkSVD recommendation system | Market basket Recommendation |

I propose dividing the online sales process into two phases: the browsing stage and the checkout stage, to provide suggestions for improving the customer experience.

FunkSVD and market basket cannot be used with new customers with an empty basket so the system should use trending products to recommend

FunkSVD predicts user preferences by analysing the feedback from similar users. Hence, implementing this technique during the browsing stage can enhance the customer experience and generate excitement.

Since market basket analysis works more effectively with a filled basket, it is recommended to apply this system at the checkout stage.

After implementing the recommendation systems, Olist should assess the model's effectiveness using various metrics such as Click-through rate and Conversion rate. As Olist's historical database grows, they should explore different levels of clustering for their recommendation systems and conduct A/B testing to determine if these models are more effective in driving sales or engagement.

## References


Data source

https://www.kaggle.com/datasets/olistbr/brazilian-ecommerce

Background on the subject matter area

https://www.researchandmarkets.com/reports/5648316/brazil-b2c-ecommerce-market-opportunities?utm_source=GNOM&utm_medium=PressRelease&utm_code=jr6rtg&utm_campaign=1837416+-+Brazil+B2C+Ecommerce+Market+Opportunities+Databook+Q1+2023+Update%3a+Launch+of+Platform+Specific+Altcoins+Presents+Opportunities&utm_exec=como322prd

https://www.jpmorgan.com/merchant-services/insights/reports/brazil-2020

https://econsultancy.com/85-of-consumers-favour-apps-over-mobile-websites/

Modelling

https://medium.com/analytics-vidhya/k-nearest-neighbors-all-you-need-to-know-1333eb5f0ed0

https://surpriselib.com

https://pythondata.com/market-basket-analysis-with-python-and-pandas/


Dictionary

rmse; Root mean square error