DSCI 554 Final Project:
Olist Brazilian E-commerce Visual Analysis

Lukas Stein[1], Vinjit Regulagedda[1], Amit Birajdar[1], Navyada Koshatwar[1], and Tejas Sujit Bharambe[1]

1 University of Southern California, Los Angeles CA 90089, USA
{lukasste, sregulag, abirajda, koshatwa, tbharamb}@usc.edu

Abstract. This paper outlines the steps towards and importance of creating a data visualization platform for Brazilian E-Commerce giant Olist. The data acquired is from the Kaggle datastore, where Olist has released information regarding orders, payments, customers, and ratings. This data is given to generate insights into the e-commerce industry over a time period of August 2016 to October 2018. After critical data cleaning and formatting, we have prepared a series of important data frames helpful for depicting the most important information as visualizations. We have created several critical visualizations of orders, region-based analyses, and payments of orders, customers, and sellers. These data and infographics are also helpful in creating a Machine Learning model to forecast the rating of the sold item. After testing several basic methods, we utilized the gradient-boosting technique (XGBoost) which resulted in the best RMSE score of 1.171. These techniques and a Machine Learning Model can give several insights to Olist teams, to rework and create better strategies to get an edge over their competition.

Keywords: E-Commerce · Machine Learning · Data Visualization.

1 Introduction

1.1 Motivation

Olist[5], the biggest department store in Brazilian marketplaces, links small businesses from all over Brazil to channels with a single contract. These business owners can use Olist logistics partners to sell their goods through the Olist Store and send them straight to customers. This is a rich dataset and can be used for exploring the Brazilian markets and generating numerous types of visualizations pertaining to orders, payments, sales, and ratings. Furthermore, this dataset can also be used for various machine-learning methods like clustering, sales prediction, delivery performance, feature engineering, NLP, and much more.

1.2 Scope

This platform is created for the sales and marketing teams of the OLIST e-commerce portal. The purpose of this platform is to serve them with the essential information needed to drive business decisions that could generate higher revenues and, eventually, profits. The platform provides users with the ability to analyze the following:

- Order trends and growth M-o-M with an ability to deep-dive by the status of the orders and the day of the orders. Regional-level analysis to analyze the number of orders and their price, delivery time, delay time, and freight value across regions of Brazil. Payment analysis to visualize the trends in payments bifurcated by payment method and the number of installments. The density of e-commerce across regions is based on counts of cities, customers, and vendors plotted on a map to compare raw data against the relative area. A Machine learning model to predict the ratings of items based on several features, which can be useful for the sales and marketing teams at Olist to improve their recommendation system for users and to generate high revenues.

1.3 Contributions

Our work aims to visualize the sales-related metrics and enhance the revenues for Olist with a particular focus on improving the user-item recommendations for the existing users and how it could help the sales and marketing team to price the products accurately to boost the ratings. We have also added geographic data to get a better picture of how the sales are performing across various parts of Brazil.

1.4 Introduction to each section

In section 2, we present existing analysis and visualization work related to Brazilian E-Commerce, such as algorithms that are currently being developed to improve customer profiles and sales data predictions. In section 3, we present the dataset used by providing descriptive statistics, describing the database schema, and pre-processing the data. In section 4, we explain how we came up with a design for the system, and the process we used to build it. In section 5, we showcase the main aspects of the website we have built. In section 6, we present the outcome of our analyses and discuss the results of the machine learning model. Finally, section 7 concludes the paper and reiterates important findings.

2 Related Work

Currently, the recommendation system is a major aspect of research that is being done in the e-commerce space. It is essential for the organization to understand the customer and be able to cater to his needs better. This has proved to be a way to retain existing customers and get new ones on board, which evidently relates to higher revenues and profit margins. Apart from the recommendation systems, organizations have also started focusing on optimizing their supply chain and reducing the time and waste required to complete a product to the user cycle by adopting a lean methodology. Several projects have been performing feature engineering-related work by combining this data set with the other available datasets on Brazilian e-commerce as well as economy-related government data. With this, their goal has been to increase the breadth of the data which can be used to further align their analysis of this data with that of a bigger economic picture of Brazil. With our analysis, we add to the research that is currently taking place and add to their perspectives widening the opportunity horizon.

3 Data

3.1 About the Data

The dataset contains details on 100k orders between 2016 and 2018 on Brazilian marketplaces. Some of the key features include customer location, product qualities, order status, pricing, payment, and freight performance. The features we used were mainly related to the price, order purchase timestamps, region, payment type, freight value, and delta delivery (difference between estimated and actual delivery time). Since the data is real, it was anonymized to protect the identity of companies. The dataset was provided by Olist and can be found here.

3.2 Data Preprocessing

The data was imported and formatted using pandas. Geo-location data that maps lat-long coordinates of cities to their respective Brazilian states were also included in the Olist dataset. The other necessary geospatial data, a geojson map of brazil splitting up states into separate features, is the basis for the SVG map and the draggable elements on the slippy map. Data exploration began with a pair plot to get an overview of the data, then creating boxplots for the continuous variables, and finally creating histograms for the categorical variables. The review score was examined as a histogram – it revealed a left-skewed distribution. This was important to note, as it means our dataset is unbalanced.

- *Data Joins*: joined various tables using Excel and pandas
- *Data Cleaning*: Missing data was dropped (only a few rows), outliers removed
- *Data Transformations*: Converting MultiPolygon shapefile to Polygons and then into geoJSON.

4 Approach

4.1 Design Process

We started with defining the *Problem Statement*. Based on our problem statement, we prepared an exhaustive list of requirements in terms of data. We came up with multiple data sources and then discussed their possibles

use cases, pros, and cons. After a deliberate discussion, we finalized on using the e-commerce dataset provided by *Olist* from *Kaggle*. We then brainstormed ideas regarding the analysis that could be done using this dataset and as we came up with different ideas, we distributed the work among ourselves based on data themes like orders, payments, regional analysis, Geo analysis, and the machine learning model. Once we had a rough draft of our individual analysis, we combined it together in a web app using *Vue.js*. We then scripted a storyline to discuss how the analysis can help a user wanting to get more insight into this data and recorded a video displaying the same.

4.2 Design Considerations

- *Flexibility and Ease of use for users*: We ensured that our platform provides flexibility to users in the form of interactivity and also ease of use by segregating the appropriate information into desired tabs.
- *Layout and Effective Visualizations*: The layout was ensured to be consistent across all the tabs of the platforms and visualizations were created from the end-user perspective in terms of flow and usefulness of the information.
- *Connectivity and Reliability of Information*: The raw data was verified with quality checks and made sure the right information was shown in the visualizations.
- *Cognitive Abilities*: All of the designs were made with knowledge of human cognitive capabilities in mind. For example, humans struggle to compare areas of adjacent objects, so visuals meant for comparing areas on the map page were made to be overlaid and moved around.

5 System

5.1 Explain how you built the dashboard/infographic (technologies and methods)

The dashboard is built using the Vue-js framework along with bootstrap for styling. The Vue-js [8] framework is used to build reusable components across the dashboard. Several sections of code related to Vue-js are helpful to maintain the website like routing, state management, data, etc. We have also used several javascript libraries like:

- *d3*: d3.js[3] is useful to create several interactive charts across the website, which helped in analyzing and visualizing the data.
- *bootstrap*: Bootstrap[1] is a free CSS framework useful for mobile-first responsive web development. It contains predesigned templates for forms, buttons, div orientations, and many more.
- *Leaflet*: Leaflet[4] is a free Javascript library used for mobile-friendly geo-mapping. It is widely used in HTML5 and CSS3.
- *Vue router*: Vue Router[9] is the router for Vue.js. It is helpful for creating single-page applications based on changes in URLs.

5.2 Summary of website

- *Home Page*: This explains the story and flow of the website by giving an abstract of the analysis and model. Link to image
- *Orders on E-Commerce*: Visualized the number of orders based on their status like delivered shipped, canceled. Visualized the trend of the number of orders from 09/2016 to 10/2018. Analyzed order frequency like orders by day of the week and further deep-dived into orders by the time of the day Link to image
- *Regional-level analysis of E-Commerce*: Visualized the trend of total orders received across regions of Brazil. Envisioned the number of products bought in each state of the region and how they are related to the average price of the products across all the states of the region. Analyzed differences in estimated delivery, freight value, and delivery time for products across all the states of the regions. Link to image
- *Payment type analysis*: Analyzed the ratio between different modes of payment for the orders. Created distribution of the payment installments from 1 to up to 24 installments. Showed evolution of the payment types from Jan 2017 to Aug 2018. Link to image

– *Map*: Implemented SVG clickable map that displays data about the selected region to compare the relative density of sellers and customers between states. Implemented Leafleft slippy map using true size library to create a draggable element for each Brazilian state, allowing the area to overlay for comparison. Required data cleaning and manipulation from the GeoJSON file. Link to image
– *About*: This page contains information about the contributors to the project. Link to image

6 Results

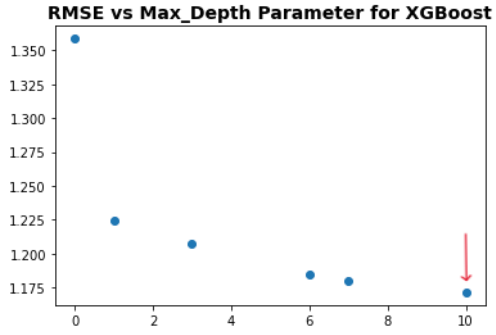

Fig. 1. RMSE vs Max_Depth of XGBoost

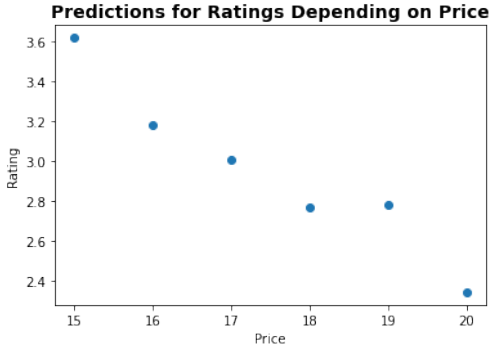

Fig. 2. Rating prediction based on price

– The model in this project focused on predicting ratings based on information from the order. We have used all the features as described in the data section.
– We built and fitted linear regression, logistic regression, random forest regression, and XGBoost regression models. The XGBoost regressor performed best. XGBoost stands for extreme gradient boosting. A simple decision tree regressor uses a path of "decisions" based on the features to predict a label. A gradient-boosted decision tree like XGBoost combines decisions from several weaker trees to output a prediction.
– Using cross-validation, we tuned the max depth parameter on XGBoost. We found that a max depth of 10 reduced the RMSE to 1.171. Given more time, it would be interesting to see what other parameters could make significant increases in scores. The model could also likely be improved with additional data on features like demographics, shipment methods, brand popularity, etc.
– The model is useful for a marketing team, as review scores can be leveraged to target marketing or improve sales. The notebook includes a use case to visualize the improvement in rating at each dollar decrease. Using this, the team can make an informed decision on how to leverage price reduction to boost ratings.
– The model and related infographics have been presented in the Notebook.

7 Conclusion

As a part of our analysis on the Olist E-commerce dataset, we mainly focused on inferring the following:

– Trends related to orders and purchases made by the user: number of orders showed a steady increase in 2017 and plateaued in 2018 after which we see a substantial drop. We also infer that purchases were mostly made during the first half of the week and around the second half of the day (afternoon and night). The ratio of canceled orders to the total shipped or delivered orders were low which is an encouraging sign for the e-commerce market.
– Analyzing Regional level trends in orders: *Sudeste* region contributed to the most orders among all the regions but, even if the number of orders were most in this region, the average price of order was least when compared to other regions. This gives an insight into the population density and the purchasing power of this population.

- Understanding payment behavior of users: It is evident from our analysis that customers most often prefer paying using their credit card. Also, most payments were made as a one-time payment instead of a monthly installment.
- Analyzing data based on geographical properties to understand the density of the population in relation to the geographical area of a region.
- Predicting the rating for items based on features used in the analysis which could be used by marketing and sales teams at Olist to better understand their product and rework on their strategies to better place their products and get an edge over their competition.

7.1 List your contributions

- Lukas Stein: Responsible for the maps page. Contributed to sections 1 and 7 of the paper and where map and map data is mentioned. Recorded map page demonstration for video.
- Vinjit Regulagedda: Responsible for the payment-type analysis page. Administrated Vue pages. Contributed to section 5 of the paper and other sections relevant to his data analysis. Participated in a live recording of the demo video.
- Amit Birajdar: Responsible for the E-Commerce exploratory analysis page based on orders data. Contributed to the sections 2, 3, and 7 of the paper and where related to his visualizations. Participated in a live recording of the demo video. Edited, merged, and uploaded the video to YouTube.
- Navyada Koshatwar: Responsible for the machine learning model. Contributed to sections 3 and where applicable for the ML part of the project. Recorded a one-minute overview of the machine learning model.
- Tejas Sujit Bharambe: Responsible for Regional E-commerce analysis page. Contributed to sections 1 and 4 and parts that relate to his visualizations. Participated in live recording of demo video.

7.2 Explain what you would do next or differently

We would expand our dataset by finding additional related data for a better overview and to further our analysis. The ML model would improve with additional data and features. Additionally, topoJSON instead of geoJSON for the SVG maps would provide a quicker load time and better experience for the page. Over time, we could also continue experimenting with other parameters to improve the RMSE and generate better predictions and richer analyses.
https://www.overleaf.com/project/6382bfbc87d785672405dec3

References

1. Bootstrap. `https://getbootstrap.com/`
2. Bostock, M.: Data-driven documents - d3.js (2014)
3. d3js. `http://d3js.org`, last accessed 10 Aug 2019
4. Leaflet. `https://leafletjs.com/`
5. Olist. `https://olist.com/pt-br/`
6. Satyanarayan, A., Moritz, D., Wongsuphasawat, K., Heer, J.: Vega-lite: A grammar of interactive graphics. IEEE transactions on visualization and computer graphics **23**(1), 341–350 (2016)
7. VanderPlas, J., Granger, B.E., Heer, J., Moritz, D., Wongsuphasawat, K., Satyanarayan, A., Lees, E., Timofeev, I., Welsh, B., Sievert, S.: Altair: Interactive statistical visualizations for python. Journal of open source software **3**(32), 1057 (2018)
8. Vue-js. `https://vuejs.org/`
9. Vue router. `https://router.vuejs.org/`
10. Wickham, H.: ggplot2: elegant graphics for data analysis. Springer (2016)

[2,7,10], [2,10,6,3].