eCommerce Product Recommendation and Pricing Algorithm

José Pérez and Telmo Linacisoro

Visual Analytics 2024-25,
Universitat Pompeu Fabra (Barcelona)
josemaria.perez02@estudiant.upf.edu, telmomaximlian.linacisoro01@estudiant.upf.edu

1. Problem Statement

The goal is to assist eCommerce sellers in selecting trending products and setting optimal prices. This involves four aspects: constructing a useful dataset, exploring and understanding the dataset, developing two (machine learning) algorithms to identify current trends and predict pricing, and finally, creating a user-friendly interface to take advantage of these insights.

2. Dataset Overview

The dataset is obtained from Brazilian eCommerce public information from 2016 to 2018 made at Olist Store across multiple marketplaces. It includes features such as order status, price, payment, freight performance, customer location, product attributes and customer reviews. An additional geolocation dataset links Brazilian zip codes to latitude/longitude coordinates. The data is real but anonymized, with references in reviews replaced by names of Game of Thrones' houses. As shown in Fig. I, we merged individual CSV files and processed the dataset (approximately 120.000 entries with 23 columns) by handling null values, duplicates, outliers, renaming, formatting and encoding.

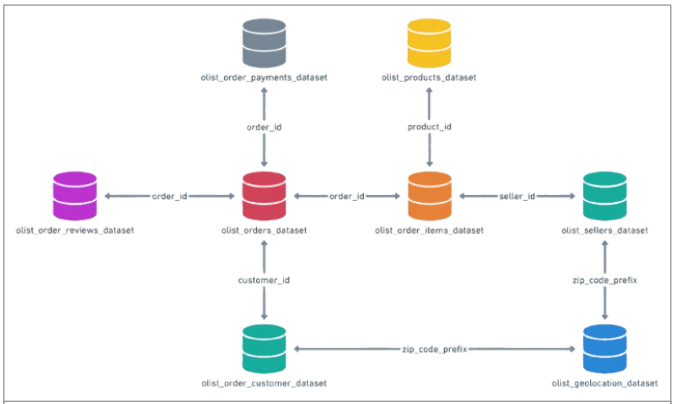

Fig. I Data Scheme

The processed data was stored in both, encoded and unencoded formats for categorical columns, for later use in machine learning models and Streamlit. In the notebook, we do include boxplots and histograms for each numerical variable, though this Exploratory Data Analysis is further developed in Streamlit, incorporating a range of relevant plots and diving deep into specific details of the data. The main insights to be extracted from the data are the following:

a. Price distribution showcases an economy of sales centered around products ranging from $20 to $60, creating a right skewed distribution with a particularly short right tail, with the most extreme observations landing at about $300, after removing outliers. Most common sales are furniture, beauty and phones, followed by sports, presents and computer-related products.
b. Correlation between variables confirm one's a priori hypothesis, with product size, weight and shipping (freight) variables being the ones more correlated to price. However, location, number of payments and product description length do present correlation with price as well.
c. Geographical insights outline an economy centered around Sao Paulo, both in terms of sellers and customers. This is consistent with the socio-economic situation in Brazil, especially when we take into account that the dataset belongs to an online commerce.

3. Business Questions and Objectives

In order to make the most of our tool, we settled our key questions to be the following:

a. Which products are trending?
b. What features are most important in determining whether a product is trending?
c. What are the optimal price points for products based on their attributes and market conditions?
d. What factors determine the best price point for a product?

We aimed to define a project that could later be developed into software applicable to any eCommerce platform, allowing companies in the sector to address their specific challenges using our algorithms. Our goal is to design this platform as a B2B SaaS solution, which can be integrated into a company's existing suite. This approach allows for potential upselling opportunities for our software products. Initially, we focused on providing a high-quality first interaction rather than incorporating a lot of unnecessary features from the outset, prioritizing the user's ability to extract clear insights.

4. Methodology

Ever since we began the project, we have been aware of the relevance of having a dataset that has been created and transformed in a way that makes logical sense and is aligned with our goals. We were particularly careful when carrying out the process of merging the different datasets to create the master database, as we knew that a mistake in this section could lead to the rows and features being altered in a way that could eliminate all relevant relationships between them.

After loading, preprocessing and ensuring the correctness of the final dataset, we tried several machine learning models, including LinearRegression, RidgeRegression and RandomForestClassifier. We carried out long grid searches with each of the models to extract their maximum performance in our context. Ultimately, we selected LightGBM as our final model due to its superior performance, which is analysed in the following section. LightGBM is a gradient-boosting framework that uses decision tree algorithms. These decision trees are built sequentially, with each successive tree focusing on errors from the previous ones using a weight scheme. Since we have two ML models, we will go into detail about each of them separately.

For the price prediction task, we implemented LightGBM with hyperparameter tuning using GridSearchCV. This allowed us to optimize the model's performance by adjusting the hyperparameters. The model was trained using input features related to product specifications and other relevant numerical data and, most importantly, targeting the price of the product. This was the model in which we carried out most of the model testing and the one which has the more specific set of hyperparameters. Once we saw the performance we were able to extract from LightGBM, we decided to try to apply the same approach for the second task.

To detect trending products, we began by performing feature engineering through data aggregation across product categories. Specifically, we calculated the total number of products ordered for each category to serve as the target variable. Additionally, we computed the mean and standard deviation of review scores for each product category. Lastly, we applied the LightGBM model. It is important to note that this task is relatively simple for a machine learning model and is not designed to predict future trends. Instead, the goal is to provide users with a quick estimate of where a specific product stands in terms of current popularity or trends.

The last statement of the paragraph above is consistent with the fact that we focused all our efforts on explainability in allowing the user to understand the price point selected by our price algorithm, as we consider this to be the cornerstone of our application. To do so, we combined multiple local and global explainability plots such as: summary plots, bar plots, force plots, decision plots and others.

The last step was to integrate all of the components into a single, practical web application using Streamlit. This became more complex that what one would have anticipated, since we wanted to create a proper site with some advanced features, while Streamlit is more aimed towards the creation of simpler dashboards. Through loads of investigation, we ended up finding workarounds for most of the features we wanted to introduce, such as tabs within a specific page to have easy access to the different sections within that page or displaying some SHAP plots which were not compatible with Streamlit at first.

5. Conclusion

Our models achieved strong performance, with a test RMSE of 21.94 for price prediction and 1.05 for predicting count of sold products per category (what we define as trends). The model for trending categories performed exceptionally well because it ended up using features like review scores and order counts to identify trending products accurately in a task that, as we said before, does not explicitly require of machine learning, but given its accessibility, its ability to perform in this context and simplicity of usage we identified it as our best option.

The price prediction model, while more complex due to pricing volatility, helps get optimal pricing based on the general market conditions and captures 81% of the variability of prices. When looking at further development, it could be interesting to construct a pricing algorithm not only based on setting the appropriate price point for a product but also on maximising revenue.

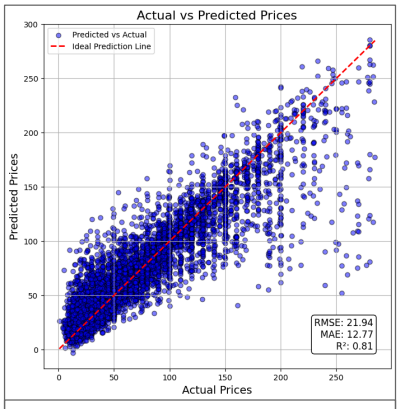

Fig. II Pricing Model Performance

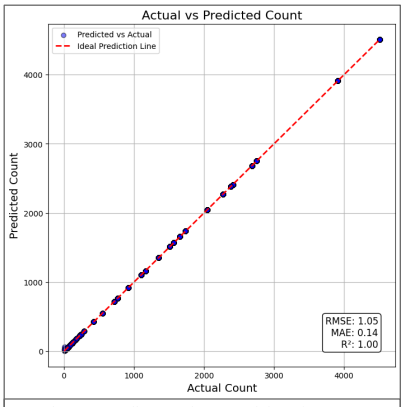

Fig. III Trending Products Model Performance

The SHAP plots and overall explainability worked exactly as intended. They provided insights that were both expected, such as the importance of a product's weight in price prediction, and less obvious ones, like the impact of the product's name length, description length or the number of pictures in its listing. This information, combined with the transparency of explaining test samples and, more importantly, each user's interaction with our algorithm, allowed us to deliver a highly interpretable solution. Additionally, incorporating current trends into the analysis further enhanced the utility and relevance of the results, leaving us very satisfied with the final outcome, as we believe our integration of machine learning models with intuitive visualizations could impact the decision-making process for eCommerce sellers, enabling them to make data-driven pricing and trend-based decisions easily.