Analyzing Factors Affecting Review Scores in Brazilian E-commerce Marketplaces: A Big Data Project

Authors:
Yramaia Salviano 20220033
Antonina Filatova 20221104
Marta Manevska 20220056
Ludovico Toscano 20220044

1. Abstract

This report presents the findings of a big data project conducted on a dataset obtained from Kaggle, focusing on orders made at multiple marketplaces in Brazil. The objective was to develop a machine learning model, specifically linear regression, to predict the review score. The research question addressed in this project was to identify the factors influencing the review scores of purchases.
The analysis revealed that customers tend to provide predominantly positive reviews. Primary factors were found to impact the review scores: customer payment with multiple installments, freight value, customer region, and payment type. This project is an example of utilizing big data and machine learning techniques in understanding customer behavior and predicting review scores, however there are a number of limitations that have to be considered and resolved in further research.

2. Introduction & Literature Review

2.1. Background and context of the project

E-commerce has grown substantially over the past years and has become increasingly important in our daily life, especially under the influence of COVID-19 recently (Hasanat et al., 2020). Considering this context, this project has the objective to study data about Olist, which is a Brazilian e-commerce company that connects businesses to different online sales channels and logistic solutions. They also provide services such as client attraction by developing promotional campaigns and orders management.

The analysis was made using a collection of datasets that have information about more than 100.000 orders made between 2016 and 2018.

2.2. Research question and objectives

Considering the importance of the reviews for the sales performance, the research question proposed to be answered in this project is: "*what factors impact the review score of the purchases*?".

To answer the question, the following objectives will be pursued:

a. Business understanding to identify the variables more relevant with the objective of this project
b. Understand what are the main factors that have impact to the review score
c. Predict the review score;
d. Propose actions that improve the review score of the products

2.3. Brief overview of the dataset

According to the description of the dataset available on Kaggle, the collection of dataset used to develop the project refers to orders made at multiple marketplaces in Brazil. Its features allows viewing an order from multiple dimensions: from order status, price, payment and freight performance to customer location, product attributes and finally reviews written by customers. In the item 3.1 the datasets used to develop the project will be described.

2.4. Overview of big data in marketing

In the realm of marketing, Big Data has emerged as a powerful tool that empowers marketers to gain a deeper understanding of their customers. By harnessing vast amounts of data collected from diverse sources,

including social media platforms, customer feedback channels, and website analytics, marketers can extract valuable insights into customer behavior, preferences, and demographics.

The benefits of using big data in marketing includes:

- Effective predictive modeling: marketers use this information to develop targeted marketing campaigns that are more effective at driving sales.
- Better personalization: this personalization can lead to increased customer engagement and loyalty.
- Optimizing marketing spend: Targeting only valuable customers allows companies to get the most out of their marketing efforts.
- Reducing customer churn: marketers can reduce customer churn by targeting these customers with personalized offers and messages.
- Improving customer experience: companies can improve customer experience and loyalty by making changes based on this information.

2.5. Brief description of key concepts and theories related to the research question

The significance of online reviews in e-commerce success is grounded in two key concepts: social influence and word of mouth.

Social influence, as highlighted in a study by Sridhar and Srinivasan (2012), is the tendency of human beings to follow the actions of others when making decisions and placing weight on those actions to assume “the correct decision” In the context of online product ratings, the ratings provided by other consumers serve as moderators, influencing the impact of positive and negative product experiences, product performance on the reviewer's own product rating.

On the other hand, word of mouth refers to consumer communication about a product, service, or company, where the sources are perceived as independent of commercial influence (Litvin et al., 2008). These interpersonal exchanges offer insights into the consumption of a particular product or service that surpass formal advertising and go beyond company messages, involuntarily shaping individual decision-making (Brown et al., 2007).

These two concepts, social influence and word of mouth, are closely interconnected in their impact on consumer behavior and collectively contribute to the understanding of the significance of online reviews in the e-commerce success.

2.6. Brief review of related studies (academic or non-academic) and their findings

In the context of online shopping, consumer reliance on product reviews for information is growing. Unlike official product information provided by sellers, reviews are generated by other consumers who have already purchased the product through online shopping websites. Understanding the factors that impact these reviews is crucial, as well as having strategies to encourage customers to provide their review scores and comments after each purchase.

One article by Orenstein (2023) explores the importance of product reviews to online sales. The author mentions that a study from the company BrightLocal found that 93% of consumers check online reviews prior to making a purchase and that 85% of consumers trust online reviews as much as personal recommendations. In addition, another study by the Spiegel Research Center indicated that when compared to products without evaluations, those with reviews have a 270% higher chance of being bought. The relevance of reviews to sales is relevant and must be monitored and used to find useful insights to the e-commerce business.

The paper of Lei, Quian and Zhao (2016) proposes a recommendation model that incorporates sentiment information from social users' reviews. By combining user sentiment similarity, interpersonal sentiment influence, and item reputation similarity, the model achieves accurate rating predictions. Experimental results demonstrate significant improvements over existing approaches, and future work can explore linguistic rules, sentiment dictionaries, and hybrid factorization models for further enhancements.

These studies collectively underscore the significance of factors such as product characteristics, consumer characteristics, review text content, and review ratings in impacting product ratings and sales in e-commerce.

2.7. Outline of the report

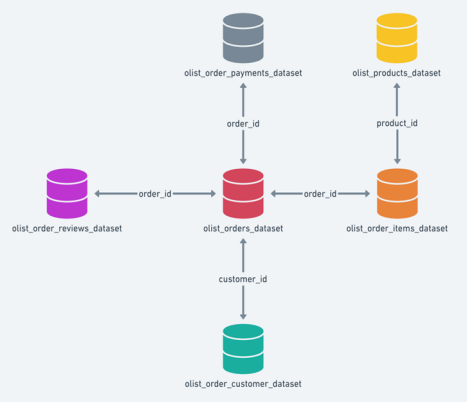
This report will begin by discussing the methodology employed to explore and process the data, including the selection of an appropriate model to address the research question and an assessment of its performance. Following this, the results obtained from the implementation of the machine learning model will be presented and interpreted. Lastly, the report will present the findings of the study in detail, accompanied by recommendations and an exploration of the study's limitations. This structure ensures a logical flow and comprehensive coverage of the research process and outcomes.

3. *Methodology*

3.1. Description of the dataset and its source

The collection of dataset used to develop the project contains data of 100.000 orders from 2016 to 2018 made at multiple marketplaces in Brazil. The files can be accessed on Kaggle via the link provided in the references section. From the total of datasets available, the ones described in the image below were the ones used for the analysis and modeling. The dataset consists of 9 SQL database files, related to each other by common keys.

There is a detailed description of the dataset variables in the appendix.



Customers Dataset: information about the customer and its location. Each order is assigned to a unique customer_id. This means that the same customer will get different ids for different orders. The purpose of having a customerunique_id on the dataset is to allow you to identify customers that made repurchases at the store. Otherwise, you would find that each order had a different customer associated with it it.

Order Items Dataset: data about the items purchased within each order.

Payments Dataset: data about the orders payment options.

Order Reviews Dataset: data about the reviews made by the customers

Order Dataset: the core dataset. From each order you might find all other information.

Products Dataset: data about the products sold by Olist.

3.2. Data pre-processing techniques used

Missing values

Data preprocessing was started with exploring missing values. First missing value check was run on each csv independently.

- In the Review table the missing values were found in comment_title and comment_message columns. The first one is about the title that the customer left, in this case is in portuguese. Second one is a message about the review. After a customer purchases the product from Olist Store a seller gets notified to fulfill that order. Once the customer receives the product, or the estimated delivery date is due, the customer gets a satisfaction survey by email where he can give a note for the purchase experience and write down some comments. For example, on Amazon, leaving a review for a product is not mandatory. It is completely optional for customers to provide feedback or write a review. Moreover, review content is not relevant for our objectives, so these rows were left as they are.

- In the Product table there were 610 observations missing in product_category_name, product_name_length, product_description_length, product_photos_qty and 2 observations missing for product dimensions variables. It doesn't seem possible to sell products on the marketplace without indicating any information and description, neither dimensions, which are necessary for shipping. These observations are considered wrong and shall be removed further.

- In the Order table missing values are found order_delivered_carrier_date (1783 missing) which is order posting timestamp (When it was handled to the logistic partner) and order_delivered_customer_date which shows the actual order delivery date to the customer. (2965 missing). Explanation of this is the status of the order: most missing values were canceled or in the process of shipping meaning they have other status than 'delivered'. Only 10 observations with missing values were delivered, this is

considered as a mistake and further would be removed. Moreover, for the task of review score prediction, only delivered orders matter.

- We observed that there are two rows with payment number of payment_installments = 0. That is considered as a mistake and further would be removed.

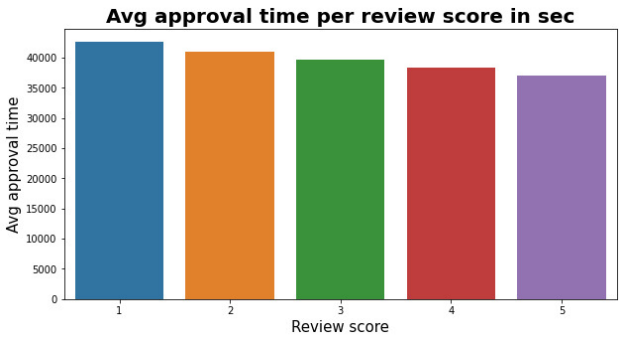
After having merged all csv tables in Databricks we obtain some missing values and discrepancies. Important note: we counted unique order IDs for each table: orders, payments, items and reviews and we saw that these numbers don't match. This happened because some tables had less unique order IDs, so while merging, mismatched cases appeared.
Missing values are less than 1% of the data, as next all missing values were deleted.

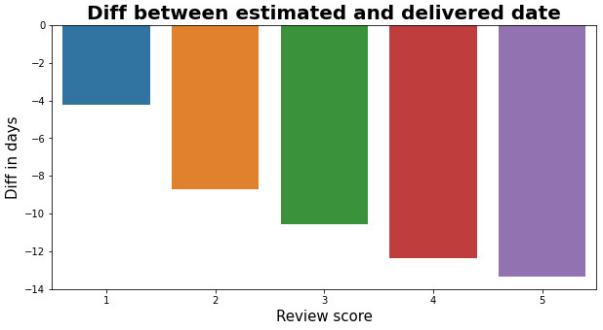
Multiple items orders cleaning
It was indicated on Kaggle by the dataset owner that there exist duplicated in review dataset, these observations are wrong.
There is another problem related to attribution of reviews to orders. The way the data is structured is that reviews are linked to the product by order_id not by product_id and this brings a problem in case when an order has more than one item: it is not clear to which product the review refers to. In the dataset there are 29602 orders that include more than one item, precisely 25574 orders with the same product bought more than once and 4028 that include different products. In both cases it creates a disbalance in data. We believe that 3 reviews of one customer that bought the same product 3 times don't have the same "significance" for the analysis as 3 reviews of 3 different customers for that product. So we decide to treat these cases as duplicates and only leave one observation. In the case of different products in one order we choose to remove all the observations since it's not clear to which product the review score belongs to and this data is misleading. As a result the dataset is reduced to 98126 rows.

Feature engineering

- Regions. Dataset includes 4190 unique cities and 27 states, which makes geographical data highly fragmented. A more advanced feature for geographical location of customers and sellers could be beneficial for prediction of the score, but in this study we choose a more basic solution. For the purpose of data analysis and modeling we decided to match regions to Brazilian states. We included 5 unique regions to the dataset which are north, northeast, midwest, southeast and south.

- Time_to_be approved and estimated_vs_delivered. We assume that the time of order delivery impacts customer experience and satisfaction, and therefore can influence the review score. New variables were created to analyze the timing aspect: time_to_be approved (as difference between the purchase timestamp and approved timestamp) and estimated_vs_delivered (as difference between estimate and real delivery time). Plots below show that there is a relation between delivery time and review score: when the average delivery time is less, the score is higher. Note: the graph of estimated_vs_delivered shows negative variables because marketplaces tend to set the estimated delivery date much later than expected (presumably to avoid the risk of customer complaints). It can be seen that the bigger the gap, the sooner the delivery happens, which corresponds to higher review score.





- Payment_sequential: a customer may pay an order with more than one payment method. If he does so, a sequence is created to accommodate all payments. More than 99000 (96%) observations have 1 payment_sequential meaning that the same payment method was used (even in case of multiple installments). The majority of multiple payment_sequential cases are paid with vouchers. This variable will not be taken into consideration as it accounts for less than 4% of data.

- Payment_value: this variable shows the value of a transaction. In case an order has one item that the payment_value = price + freight_value. Previously the data was cleaned so that the orders with more than one item always have the same product. In this case payment_value is multiplied by the number of items. Since the number of items in one order is not relevant for the objective, instead of keeping the number of items we just kept the same order more than once.

- Order_status: In the data set there are shipped products, canceled products, unavailable products and so on, but this data is not important since we are working only with the delivered products and our goal is analyzing the factor of the review score on our purchases.

- Average_product_score: feature engineered as an average score for a specific product.

3.3. Explanation of the chosen algorithms and models.

As the task of predicting the review score is a classification issue, in this case specifically, multicategorical classification, the chosen ML algorithm is logistic regression. Logistic regression is a relatively simple algorithm, and moreover, it provides the explanation of the target variable.

3.4. Performance metrics used to evaluate the models.

The metrics used to evaluate the model are:

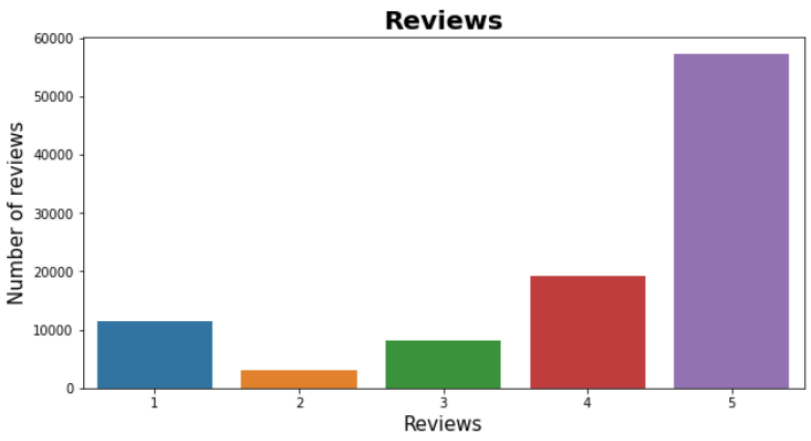
- Accuracy
- F1 score
- True positive rate
- Weighted precision
- Weighted recall

Cross-validation graph would be helpful to evaluate if the model is overfitting or underfitting, however, in this project it was not possible to plot the cross-validation graph, so it stays as a limitation of the study.

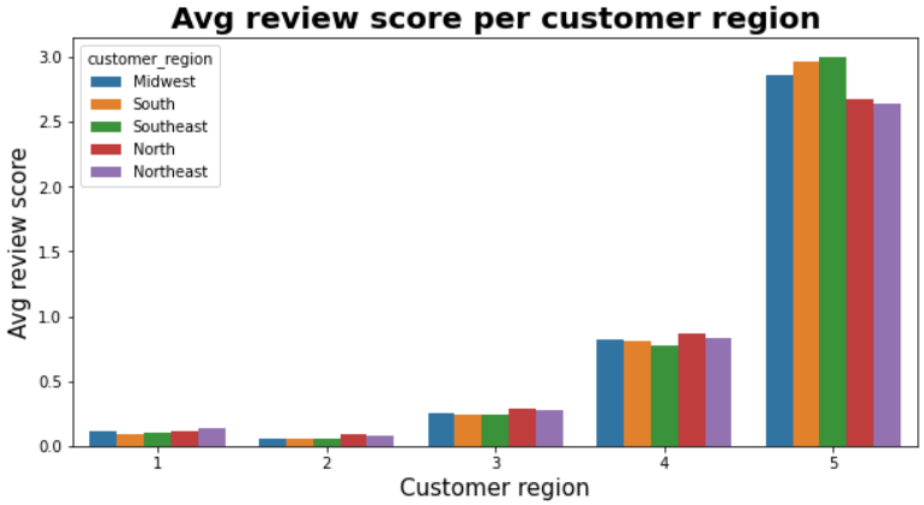
4. Results & Discussion

4.1. Data exploration and visualization

In the Data Exploration part we focus on Review score, since this is the target variable and with EDA we can start getting some insights on customers patterns in giving reviews.

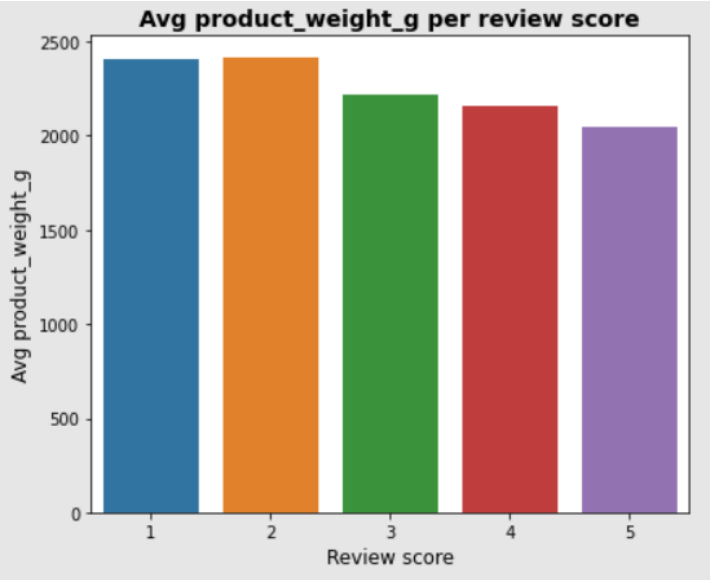


- Looking at the distribution of the review score, we can clearly say that the dataset is imbalanced, the majority of ratings are 5, and the second most frequent review score is 4, then follows 1. Ratings 2 and 3 are least frequent. It can be said that customers overall tend to rate products either positively (which happens in the majority of cases) or explicitly negatively. Rarely do customers put “medium” scores.



- Regions & scores. In the table below, we can observe some difference between the regions regarding average score. For instance, customers from North and Northeast rate 5 less often, and put 3 and 4 more often than the rest of the sample. On the contrary, customers from Southeast and South are more likely to give 5 stars to products.

- Prices & scores. On the graphs below we can observe that products with higher freight value are more likely to receive lower ratings, this can be related to longer distances or higher weights. Interesting trend can be seen with the product price: it seems that the cheapest products are more likely to get an average score like 3, while the most expensive are more often rated 1 or 5.



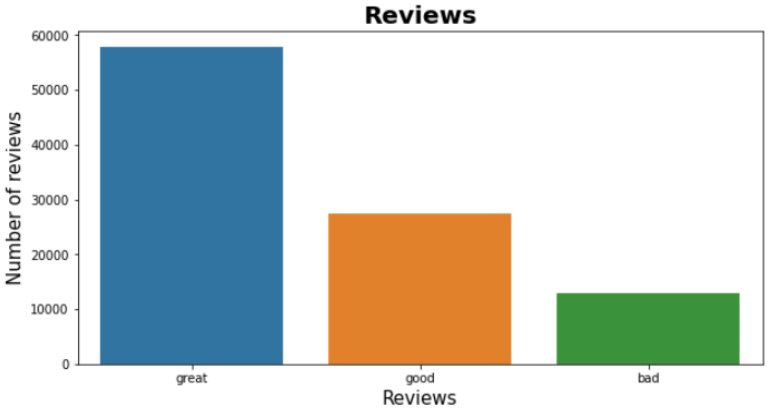
- Product weight & scores. From the graph above we can see a descending trend on weight of the product, as the score goes up: more heavy products are more likely to get lower review scores, which can be related to a more complicated delivery for or a higher price of the products with bigger weight and dimensions.

4.2. Description of the ML models' performance

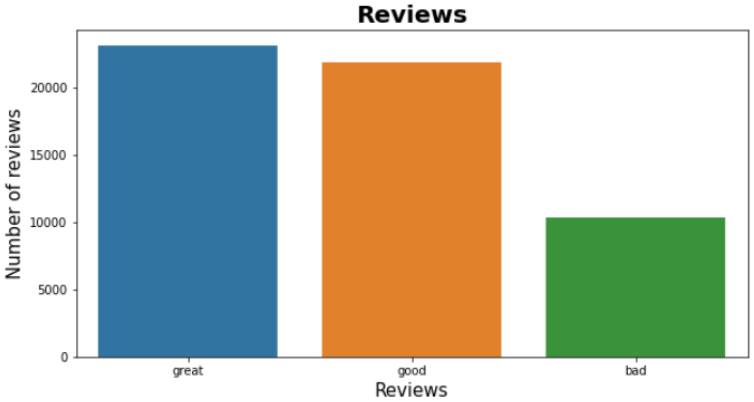
The feature selection was made with the objective of having a good performance of the model. All the variables that were related to the freight value, such as origin and destination, price, product dimension and weight and description were dropped, since the sign that they might give to the model is already in the freight_value. Other engineered features apart from more_than_one_installment, and estimated_vs_delivery were also disconsidered because of their low impact on the performance of the model. Features like photo quantity and length of the metadata were not selected either as we consider that their impact is included in the average score of the product.
First run of the model brings the following results (image below). The evaluation metrics are quite low and most importantly, true positive rate is immensely skewed for label 5, with 0.97 versus 0.57, 0.014, 0.003. 0.0 for other labels.

```
Metric: accuracy  =  0.6343253159396657
Metric: f1  =  0.5238178921038421
Metric: weightedPrecision  =  0.5025947069061916
Metric: weightedRecall  =  0.6343253159396657
truePositiveRateByLabel for label 1  =  0.5781637717121588
truePositiveRateByLabel for label 2  =  0.0
truePositiveRateByLabel for label 3  =  0.0030902348578491965
truePositiveRateByLabel for label 4  =  0.014952307295694767
truePositiveRateByLabel for label 5  =  0.9710082215491129
```

The explanation of the result is the imbalance of the data for the target variable. As mentioned before, the dataset is strongly imbalanced to positive reviews. Two steps were taken to balance the data, first is to merge reviews into 3 categories instead of 5, making “bad” (stands for scores 1 and 2), “good” (stands for scores 3 and 4) and “great” (stands for score 5) labels for the reviews.



With this binning the majority of data still stays in “great” reviews, so the second step is undersampling this category. Graph below shows the distribution after the manipulation.



After running the model on the modified data, we see the improvement of the results. F1 score, weighted precision, weighted recall went up, and importantly, true positive rate is now more balanced, even though the label 0 (great) still has a higher rate than other labels.

```
Metric: accuracy  =  0.6281084386465552
Metric: f1  =  0.639348763369143
Metric: weightedPrecision  =  0.6676417850046765
Metric: weightedRecall  =  0.6281084386465552
truePositiveRateByLabel for label 0  =  0.683686715707486
truePositiveRateByLabel for label 1  =  0.571766418046207
truePositiveRateByLabel for label 2  =  0.4988335925349922
```

Images below show which factors logistic regression model has attributed to the review score, in absolute numbers and as positive or negative coefficients. From the output it is not clear to which category bad, good or great these coefficients relate, this is a limitation of the project.

Name	Weight
more_than_one_ins...	62.382786
freight_value	62.24596
customer_region_vec	61.644405
payment_type_vec	61.62089
average_product_s...	34.027622
estimated_vs_deli...	1.5876719

Name	Weight
customer_region_vec	7.555803
payment_type_vec	7.377288
freight_value	7.280585
more_than_one_ins...	7.208983
average_product_s...	1.6371744
estimated_vs_deli...	0.41306403

Name	Weight
more_than_one_ins...	55.173805
freight_value	54.965374
payment_type_vec	54.243603
customer_region_vec	54.0886
average_product_s...	35.6648
estimated_vs_deli...	2.000736

Name	Weight
estimated_vs_deli...	-1.5876719
average_product_s...	-34.027622
payment_type_vec	-61.62089
customer_region_vec	-61.644405
freight_value	-62.24596
more_than_one_ins...	-62.382786

Name	Weight
customer_region_vec	7.555803
payment_type_vec	7.377288
freight_value	7.280585
more_than_one_ins...	7.208983
estimated_vs_deli...	-0.41306403
average_product_s...	-1.6371744

Name	Weight
more_than_one_ins...	55.173805
freight_value	54.965374
payment_type_vec	54.243603
customer_region_vec	54.0886
average_product_s...	35.6648
estimated_vs_deli...	2.000736

4.3. Comparison of the results with existing studies

Comparing the results obtained in this study with the paper of Lei, Quian and Zhao (2016) about rating reduction it was possible to identify that not all the variables that we used for the model make sense with the results obtained by the authors, but there are some related:

1. Freight value: the article briefly touched upon shipping as an important factor for customer satisfaction. The authors expand on it by stating that customers appreciate cost-effective or free shipping, as it adds value to their purchase. However, if the freight value is considered high or unreasonable, it can result in customer dissatisfaction and potentially negative reviews.
2. Customer region: the article did not specifically discuss regional factors. However, the authors point out that certain factors such as delivery time, availability of local support or services, and regional preferences can vary across different locations. If a product aligns well with the specific needs or preferences of customers in a particular region, it can contribute to more positive reviews from that region.
3. Payment type: authors mentioned payment options briefly. If customers encounter difficulties, limited options, or security concerns during the payment process, it can lead to dissatisfaction.

4.4. Interpretation and discussion of the findings.

- Freight value: Customers may appreciate cost-effective or free shipping, as it adds value to their purchase. However, if the freight value is considered high or unreasonable, customers may feel dissatisfied, leading to a potentially negative review.
- Customer region: Certain factors like average delivery time per region, availability of local support or services, and regional preferences may vary across different locations. If the product aligns well with the specific needs or preferences of customers in a particular region, it can contribute to more positive reviews from that region
- Payment type: It can be assumed that different payment options provide customers more convenience, security, and ease of use. If customers encounter difficulties, limited options, or security concerns during the payment process, it may lead to dissatisfaction.
- Difference between estimated and real time of delivery: Timeliness of delivery seems to have an impact on customer satisfaction. If the actual delivery time exceeds the estimated time significantly, it may lead to frustration and disappointment.

5. Conclusion

5.1. Summary of the findings

Overall, it is observed that customers are more inclined to leave great or good reviews. There key factors that have the greatest impact on review scores are: customer payment with multiple installments, freight value, customer region, average review of the product, difference between estimated and real time of delivery and payment type. The study suggests that depending on these factors, customers are either highly satisfied or dissatisfied overall.

5.2. Implications of the results for marketing

- Olist can negotiate favorable shipping rates with logistics partners on behalf of its clients. By securing cost-effective shipping solutions, Olist helps its clients offer competitive or discounted freight values to customers. This contributes to a positive perception of value and a seamless shopping experience.
- Negotiating with logistics partners the contracts to make sure the delivery timings are respected and delivery time is predicted precisely.
- Regional trends and preferences in Brazil can be analyzed to help sellers tailor their product offerings and services to specific regions. Olist can assist its clients in optimizing their inventory, providing localized support, and ensuring efficient delivery.
- By sharing valuable insights from collecting and analyzing customer reviews and feedback with its clients, Olist will help them to identify areas for improvement and take necessary actions to enhance the customer experience.

5.3. Limitations of the study

- The original dataset has the imbalance, and it is the main limitation of the current study.It was leveraged with the undersampling technique. Using the SMOTE function could improve model results. Another way would be using multiple binary logistic regression for each score, in order to avoid the issue of imbalanced data.
- The cross validation hasn't been run for the model, this is a significant limitation of the projects, since there is no clarity if the model is overfitting, underfitting or not.
- From the output of logistic regression it is not clear to which category bad, good or great the coefficients relate, so there is no interpretation which factors are more important for positive and negative reviews. Using the binary classification for each score of review could solve this issue.
- The only customer data provided is customer unique number and customer zip code. In this project zip code data wasn't employed which could potentially bring some insights as well as obtaining more behavioral information.
- In the dataset there is no information about the products, which leaves a gray zone in understanding how products are ranked. Moreover, the categories provided in the dataset don't have hard borders and are overlapping, so they couldn't be used as a feature for the algorithm. Regrouping the categories in less categories with distinct structure could bring more insights.

5.4. Suggestions for future research

To enrich the understanding of the factors that explain customer review scores there can be done further research about logistics and distance part of the business. In the current dataset there are only zip codes and cities from customers and sellers, it could be insightful to include some additional information about the cities, for instance size of population, linkage to other cities, transportation development, etc.

5.5. Conclusion and overall contribution of the study to the field

The model can be used as a starting point and be improved to be used by e-commerce platforms to predict the score provided to customers to understand how to design actions that can be implemented in advance to promote the best purchase experience possible.

6. References

1. Kaggle. (n.d.). Brazilian E-commerce Dataset. Retrieved from https://www.kaggle.com/datasets/olistbr/brazilian-ecommerce

2. Orenstein, S. (2023, April 18). The Importance of Product Reviews for Ecommerce Sales. Retrieved from: https://www.locate2u.com/ecommerce/the-importance-of-product-reviews-for-ecommerce-sales/

3. Lei, X., Qian, X., & Zhao, G. (2016). Rating Prediction Based on Social Sentiment From Textual Reviews. IEEE Xplore Digital Library. Retrieved from https://ieeexplore.ieee.org/document/7484319

4. Sridhar, S., & Srinivasan, R. (2012). Social Influence Effects in Online Product Ratings. *Journal of Marketing*, *76*(5), 70–88. https://doi.org/10.1509/jm.10.0377

5. Predik Data Driven https://predikdata.com/big-data-in-marketing-role-applications-benefits/#:~:text=Big%20Data%20enables%20marketers%20to,Optimize%20pricing%20decisions

7. Appendices

7.1. Additional visualizations and tables

Description of the dataset variables

Variable	Definition	Dataset
customer_id	key to the orders dataset. Each order has a unique customer_id.	olist_customers_dataset
customer_unique_id	unique identifier of a customer.	olist_customers_dataset
customer_zip_code_prefix	first five digits of customer zip code	olist_customers_dataset
customer_city	customer city name	olist_customers_dataset
customer_state	customer state	olist_customers_dataset
geolocation_zip_code_prefix	first 5 digits of zip code	olist_geolocation_dataset
geolocation_lat	latitude	olist_geolocation_dataset
geolocation_lng	longitude	olist_geolocation_dataset
geolocation_city	city name	olist_geolocation_dataset
geolocation_state	state	olist_geolocation_dataset
order_item_id	sequential number identifying number of items included in the same order.	olist_order_items_dataset
product_id	product unique identifier	olist_order_items_dataset
seller_id	seller unique identifier	olist_order_items_dataset
shipping_limit_date	Shows the seller shipping limit date for handling the order over to the logistic partner.	olist_order_items_dataset

price	item price	olist_order_items_dataset
freight_value	item freight value item (if an order has more than one item the freight value is splitted between items) *Example* The order_id = has 3 items (same product). Each item has the freight calculated according to its measures and weight. To get the total freight value for each order you just have to sum. The total order_item value is: 21.33 * 3 = 63.99 The total freight value is: 15.10 * 3 = 45.30 The total order value (product + freight) is: 45.30 + 63.99 = 109.29	olist_order_items_dataset
order_id	unique identifier of an order.	olist_order_payments_dataset
payment_sequential	a customer may pay an order with more than one payment method. If he does so, a sequence will be created to accommodate all payments.	olist_order_payments_dataset
payment_type	method of payment chosen by the customer.	olist_order_payments_dataset
payment_installments	number of installments chosen by the customer.	olist_order_payments_dataset
payment_value	transaction value *(check the explanation of the freight value variable)*	olist_order_payments_dataset
review_id	unique review identifier	olist_order_reviews_dataset
order_id	unique order identifier	olist_order_reviews_dataset
review_score	Note ranging from 1 to 5 given by the customer on a satisfaction survey.	olist_order_reviews_dataset
review_comment_title	Comment title from the review left by the customer, in Portuguese.	olist_order_reviews_dataset
review_comment_message	Comment message from the review left by the customer, in Portuguese.	olist_order_reviews_dataset
review_creation_date	Shows the date in which the satisfaction survey was sent to the customer.	olist_order_reviews_dataset
review_answer_timestamp	Shows satisfaction survey answer timestamp.	olist_order_reviews_dataset

order_id	unique identifier of the order.	olist_orders_dataset
customer_id	key to the customer dataset. Each order has a unique customer_id.	olist_orders_dataset
order_status	Reference to the order status (delivered, shipped, etc).	olist_orders_dataset
order_purchase_timestamp	Shows the purchase timestamp.	olist_orders_dataset
order_approved_at	Shows the payment approval timestamp.	olist_orders_dataset
order_delivered_carrier_date	Shows the order posting timestamp. When it was handled to the logistic partner.	olist_orders_dataset
order_delivered_customer_date	Shows the actual order delivery date to the customer.	olist_orders_dataset
order_estimated_delivery_date	Shows the estimated delivery date that was informed to customer at the purchase moment.	olist_orders_dataset
product_id	product unique identifier	olist_products_dataset
product_category_name	root category of product, in Portuguese.	olist_products_dataset
product_name_lenght	number of characters extracted from the product name.	olist_products_dataset
product_description_lenght	number of characters extracted from the product description.	olist_products_dataset
product_photos_qty	number of product published photos	olist_products_dataset
product_weight_g	product weight measured in grams.	olist_products_dataset
product_length_cm	product length measured in centimeters.	olist_products_dataset
product_height_cm	product height measured in centimeters.	olist_products_dataset
product_width_cm	product width measured in centimeters.	olist_products_dataset
seller_id	seller unique identifier	olist_sellers_dataset
seller_zip_code_prefix	first 5 digits of seller zip code	olist_sellers_dataset
seller_city	seller city name	olist_sellers_dataset
seller_state	seller state	olist_sellers_dataset
product_category_name	category name in Portuguese	product_category_name_translation
product_category_name_english	category name in English	product_category_name_translation

Correlation matrix