Final Deliverable

Find Patterns in Customer Satisfaction Based on Product Reviews

Project Team A7: Alice, Ana, Keshuo, Mekhal, Zheming

I. Problem Statement

As e-commerce shopping becomes a more widely-used method of retail shopping in today's world, the importance of customer reviews and satisfaction has also become more valuable in determining the continued success of a product. Customer reviews have a significant influence in determining whether other potential buyers will purchase the product or not. For instance, factors such as the number of reviews, review scores and quality of reviews play on a product may persuade the buyer in choosing one competitor's product over the other. In order for a company to remain competitive within its market segment, it is vital to monitor product reviews and scores to quickly identify any potential issues that may be leading to customer dissatisfaction. We are analyzing the online shopping dataset of Olist, the largest departmental store in Brazil in order to figure out how the review text and sentiment relate to customer satisfaction to improve future customers' user experiences.

II. The Data

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

The dataset was downloaded from Kaggle, and originally provided by Olist which is the largest department store in Brazilian marketplaces. It contains 9 csv files with over 50 columns in total, most of them have about 100,000 rows but some do not.

Before go in to this dataset, we need to keep in mind:

- An order might have multiple items.
- Each item might be fulfilled by a distinct seller.
- All text identifying stores and partners were replaced by the names of Game of Thrones great houses.

1. Data Cleaning

We filled the missing, empty, and duplicated values and we used binary or ordinal encoding to transform categorical data into numeric values. We added a column 'review_grade' that categorized customer ratings into three categories:

- negative: -1 for review score of 1 and 2;
- neutral: 0 for review score of 3;
- positive: 1 for review score of 4 and 5.

We also added a column 'review_label' that classified the ratings as :

- unsatisfied: for review score of 1 and 2;
- neutral: for review score of 3;
- satisfied: for review score of 4 and 5.

Next, we grouped similar products into common categories and finally, we combined all these datasets into one by merging 'order_id', 'customer_id', 'seller_id', 'product_id' and 'product_category_name'.

2. Data Modeling

For modeling, we added another dataset that contained all the Holidays over the time period of our original dataset so that we could find the impact of holidays on shipping time. We then split the data to train/validation/test dataset for supervised machine learning with the ratio of 6/2/2. Next, we created dummy variables for the column 'customer_state' and then we scaled the data using 'MinMaxScaler' and did PCA on it.

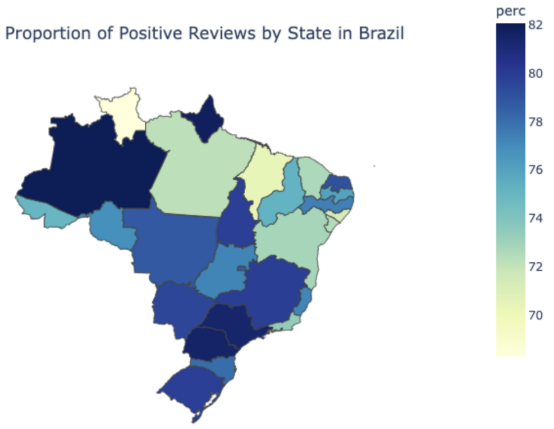
III. Methodology

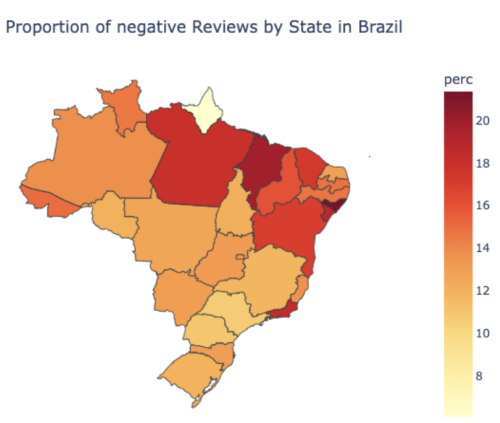
1. Exploratory Data Analysis

a. *Are there geographical patterns (state-level) that correlate with high and low review scores? What does the geographic distribution of reviews look like?*

We created a choropleth map comparing review scores across Brazil. We grouped the scores into two bins: satisfied review scores (scored: 4 or 5) and unsatisfied scores (scored: 1, 2 or 3) by state. We found the following:

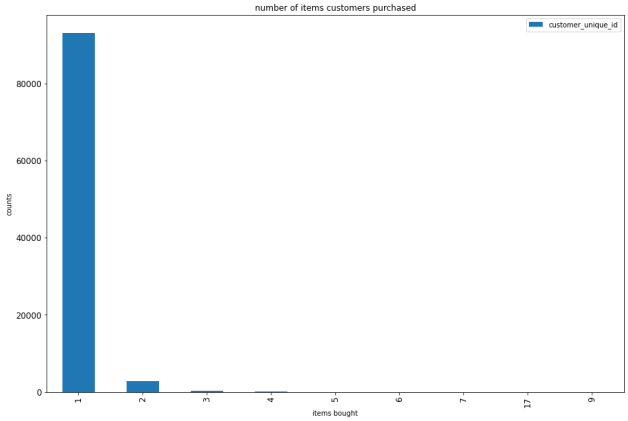
- States AP, AM, PR, SP had a relatively high proportion of satisfied customers.
- States AL had the highest proportion of unsatisfied customers.





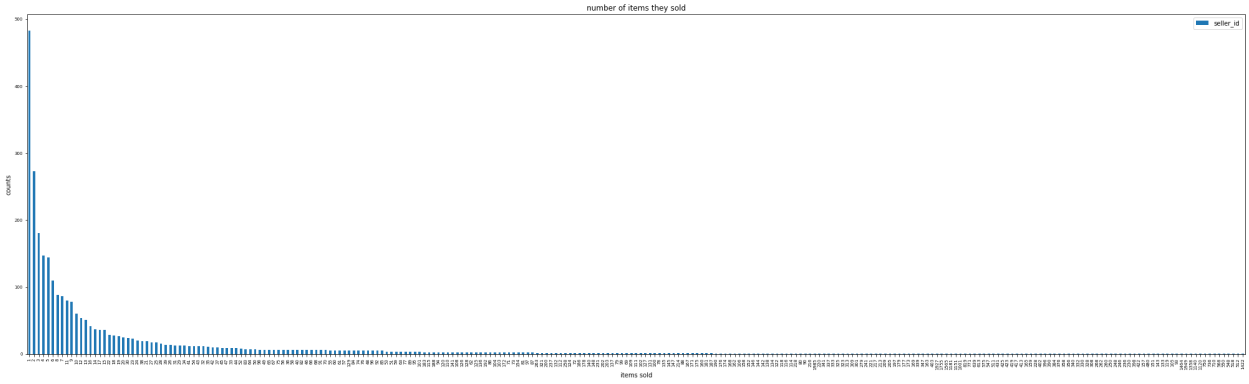
b. How many items do customers frequently purchase on this ecommerce site?

As shown in our findings below, customers mostly purchased 1 item on the site. However, there was one outlier customer who purchased 17 items.



c. How many items do sellers usually sell on this site?

There were 483 sellers who only sold 1 item, while 1 seller sold 1,985 items. Only 13 sellers sold more than 1000 items on the site. It also suggested that detect and improve customer sentiment are important for Olist to attract and retain users.



d. What is the average review score for each product category?

The product category with the highest average review score is Book.

e. Display the review scores for the top and worst sellers

Sellers with best review scores:

	seller_id	review_score
2399	d263fa444c1504a75cbca5cc465f592a	5.0
453	28872dc528e978a639754bc8c2ce5a4c	5.0
2607	e5def42655b7490edac5a56fe8e9e603	5.0
1370	761681a821d8275bc79f552116d06869	5.0
1365	75745ef7bc7d4f3ea3380f6f5303f514	5.0

Sellers with lowest review scores:

	seller_id	review_score
2484	da2782c804606d2a5d8e1760dbb3e7ec	1.0
2542	df683dfda87bf71ac3fc63063fba369d	1.0
636	391bbd13b6452244774beff1824006ed	1.0
2738	f114dca2828bf718548db175ebe2cfcd	1.0
658	3a52d63a8f9daf5a28f3626d7eb9bd28	1.0

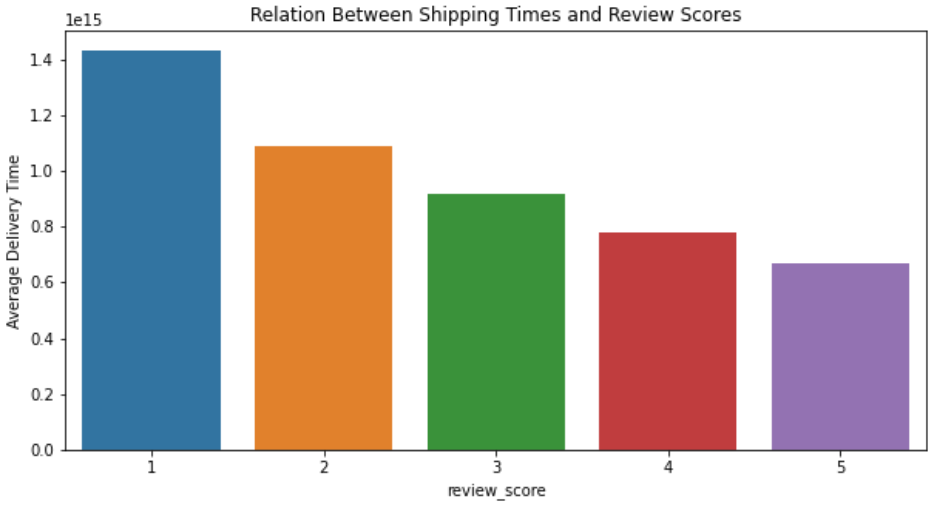
f. What are the top 5 selling items by frequency?

The top 5 items by selling frequency are highlighted in order in the table below:

	product_category_name_english	product_id	Count
22715	furniture	aca2eb7d00ea1a7b8ebd4e68314663af	517
26017	housewares	422879e10f46682990de24d770e7f83d	486
22238	furniture	99a4788cb24856965c36a24e339b6058	472
25860	housewares	389d119b48cf3043d311335e499d9c6b	389
25815	housewares	368c6c730842d78016ad823897a372db	388

g. Are faster shipping times associated with higher review scores?

As shown in the plot below, on average, items with faster shipping times had high review scores.



2. Modeling

a. Supervised Machine Learning

Model	AUC	Accuracy
Random Forest	✔ 0.65	✔ 0.81
Bagging	0.62	0.81
Logistic Regression	0.59	0.81

MLP (Neural Networks)	0.55	0.76
K Neighbors	0.59	0.70
Decision Tree	0.58	0.68

Among all models we tested, Random Forest has the highest accuracy and auc scores when run on the validation data. It has 0.81 accuracy score and 0.65 auc score on the test dataset, suggesting that we are not overfitting this model.

3. Unsupervised Machine Learning

a. Hierarchical clustering:

Our team conducted a Hierarchical clustering analysis and found that the optimal cluster number by hierarchical clustering is **5 or 6.** At time of drafting, the optimal number is 5, and its silhouette score is **0.43.** We also drew a graph about the distribution of labels among clusters under hclust.

b. K-Means clustering:

Our team also conducted a K-Means clustering analysis to understand how many natural clusters are formed within the dataset. From the inertia and silhouette plots, we found that we would be minimizing inertia while maximizing silhouette scores at k=4 or k=7. At k=4 clusters we would receive the highest silhouette score of 0.397, and this fits with our business needs as we are looking to cluster customers' sentiments into a few main categories.

4. Text Analysis

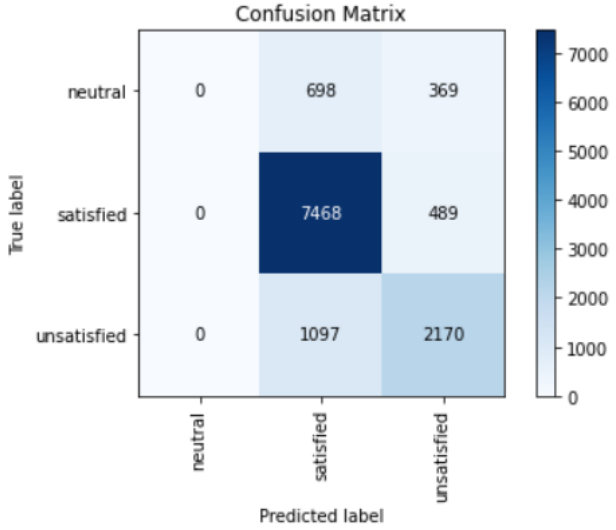
a. Word cloud

We used about 40,000 reviews to generate a word cloud. From the word cloud above, customers mainly expressed their opinion and experiences on the product and the delivery.

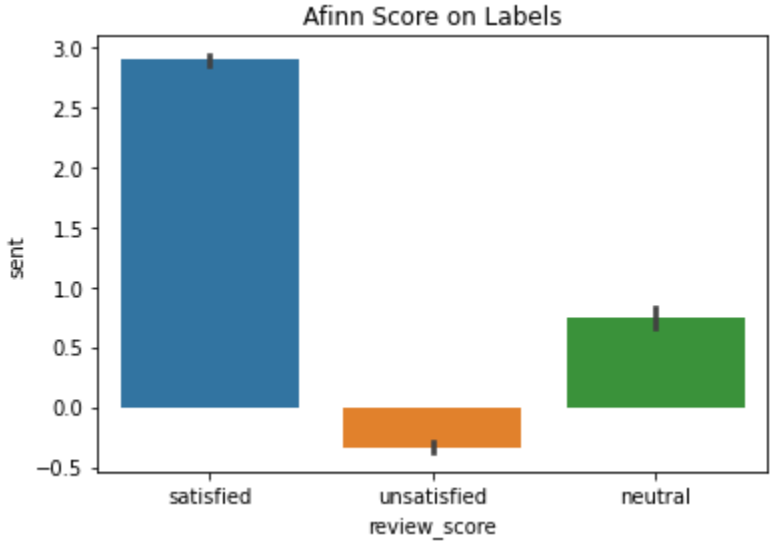
b. CountVectorizer + Decision Tree

We then tried to use CountVectorizer to get token count and use the decision tree model to fit and got an accuracy score of 0.78.

	precision	recall	f1-score	support
neutral	0.00	0.00	0.00	106
satisfied	0.81	0.94	0.87	795
unsatisfied	0.72	0.66	0.69	326
accuracy			0.78	1229
macro avg	0.51	0.53	0.52	1229
weighted avg	0.71	0.78	0.74	1229

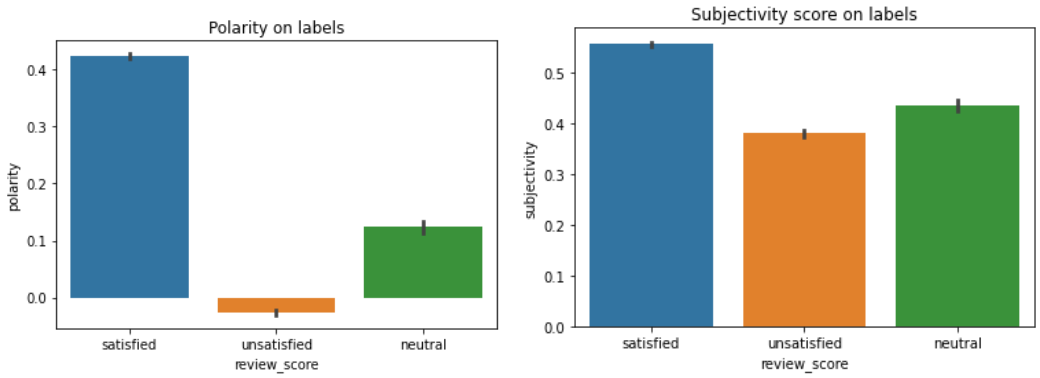


c. Afinn Score



Besides using running tests on text and scores, we then want to conduct sentiment analysis to see if customers' review scores aligned with their text sentiment. Afinn score is one mature model to realize it. The average afinn scores for each label fit what we expected.

d. Polarity and Subjectivity Score



Polarity score distribution among labels are similar to that from afinn score.Subjectivity scores show that customers are less subjective for unsatisfied reviews. This is questionable, so we further checked the sentiment score with some text messages. It turned out that the pre-set model does not fit the real sentiment very well.

e. Named Entity Recognition

In order to improve sentiment detection, we used label studio to conduct named entity recognition, manually assigning certain keywords to positive, neutral and negative. Because of the time limit, we only sampled 100 reviews to do NER. The project we used is called BA820 A7 in the label studio. To test this model, we tested a basic methodology of using the sign of each review's sum sentiment score using our customized model.

	precision	recall	f1-score	support
neutral	0.12	0.28	0.17	1067
satisfied	0.69	0.85	0.76	7957
unsatisfied	0.75	0.00	0.00	3267
accuracy			0.58	12291
macro avg	0.52	0.38	0.31	12291
weighted avg	0.65	0.58	0.51	12291

The accuracy score is 0.58, but it's from using 100 samples to test for about 40,000 reviews. It's relatively decent. Further work on annotation can prominently increase the accuracy score.

IV. Limitations

1. We cannot take the impact of products into consideration, because one order may have multiple products.
2. Supervised Machine Learning: the accuracy and auc scores still have some potential as the translation function may not recognize typo in the original Portuguese.
3. Our current clustering results only suggest that more clusters can be applied other than just using our current scoring scale to segment. Further exploration of customer segmentation on unsatisfied reviews can be done.
4. Named Entity Recognition step only used 100 samples because of the time limit. More effort can foreseeably improve the accuracy of scores.

V. Conclusion

Based on our analysis results, we found that the best model is Random Forest as it gave us the highest AUC and accuracy scores. From our unsupervised ML results, we found with HCluster that the optimal number would be 5-6 clusters, whereas in our KMeans analysis 4 clusters was ideal. Lastly, through our text analysis we found that while tokenization will help to predict labels correctly it unfortunately, cannot capture the true sentiment accurately. Additionally, Named Entity Recognition (NER) can be helpful to understand the sentiment of the reviews.

VI. Recommendations

Based on our results we propose the following business recommendations for the eCommerce company:

- To improve user experience, the website can segment customers into 4-6 groups or conduct customer segmentation only for unsatisfied reviews and develop targeted features or review schemes
- For customer sentiment prediction, we recommend using supervised machine learning methods (Random Forest).
- If the company want to use the review text to detect incorrect review scores, invest time on Named Entity Recognition (NER) and try to combine it with other models, then the machine learning model performance can be improved