E-commerce Sales: Prediction and Recommendation System

Summer 2021 Big Data Final Project

Contributors : Sahil Chitnis(ssc9983), Phoebe Zhou(yz6729)

Content

1) Introduction
2) Problem Statement (i.e. Need Analysis)
3) Solution Briefly (ie Concept Definition)
4) Dataset Introduction
5) Technology Stack
6) Detailed Solution
7) Shortcomings
8) Conclusion
9) Gratitude
10) References

Introduction

The growth and evolution in consumer demand, amalgamated with technological innovations, continue's to drive growth in global e-commerce sales. According to Statista, a company specializing in market and consumer data, the number of people buying goods and services online is around 2.14 billion in 2021, skyrocketing up from 1.66 billion global buyers in 2016.

Also, what's even more spectacular is the fact that the industry is forecasted to 2 folds in size within the next 2-3 years and growing from 3.53 trillion USD in retail e-commerce sales in 2019, to a whopping 6.54 trillion USD by 2022-2023.

And thus one can safely say that the key drivers of success over the next 10 years will be those who are focused on building a deep understanding of and improved relationship to the empowered consumer, and the only way to understand consumer behavior is to measure, analyze and predict using existing data to take effective and meaningful decisions that will make digital shopping a fun experience for the consumer as well as drive in the customers for the sellers i.e. a win - win situation.

Problem Statement (ie Need Analysis)

As of 2021, almost all the commodities such as clothes, electronics, everyday use groceries as well as services such as house cleaning, plumbing services etc. are sold online on E-commerce websites. It is safe to say that we humans have understood its vitality in our lives and can consider like our best friend. And like any best friend E-commerce websites too have taken up the task to use existing transaction Big Data (sometimes in size of GB's) and run Big Data technologies on them to analyze the data quickly and make well informed decisions to improve user experience, predict sales to make sure the companies are ready for heavy days as well as recommend products to us, making it a more personal and thus enriched experience to use these websites. Recommendation systems have started gaining momentum in companies such as Amazon with its recommendation engine that recommends a personalized list of products as soon as we login or Spotify with its list of songs that we are most likely to listen to it.

Solution Briefly (i.e. Concept Definition)

Based on the results of the previous need analysis phase we can try to outline below functionalities that will add great value to our software system.

- **Month-wise Sales analysis**: Helps to analyze months with max/min sales which helps to prepare better accordingly (i.e. heavy sales month - increase supply, beef up staff)
- **Hour-wise Daily Sales analysis**: Helps to analyze hourly sales which helps in better strategizing daily plans.
- **Map based (i.e. Geospatial) analysis of sales distribution:** Very visual representation of data, helps to understand areas (as local as zip code level) in Brazil which require additional marketing to improve sales
- **Spending Trend:** Helps to understand the most / least used of payment modes by customer
- **Product Review based analysis:** Helps to understand which products are most liked / hated by customer there by help to strategize product development
- **Predicting Sales for next FY:** With Sales Forecast for next year, the management can better plan inventory and logistics for the future well in advance.
- **Customer Segmentation:** dividing a company's customers into groups that reflect similarity among customers in each group, to maximize the value of each customer to the business.
- **Recommendation System:** The objective of recommender systems is to provide recommendations based on recorded information on the users' preferences. In this way, to increase the revenue of companies.

Dataset Introduction

We have used the Brazilian e-commerce public dataset of orders made at **Olist** Store. This dataset can be found on Kaggle here : **https://www.kaggle.com/olistbr/brazilian-ecommerce.**

The dataset has information of 100k orders from 2016 to 2018 made at multiple marketplaces in Brazil. Its features allow 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. It also features a geolocation dataset that relates Brazilian zip codes to latitude / longitude coordinates.

Olist is a Brazilian departmental store (marketplace) that operates in e-commerce segment but is not an e-commerce itself. It offers a marketplace solution (of e-commerce segment) to shopkeepers of all sizes (and for most segments) to increase their sales whether they have online presence or not.

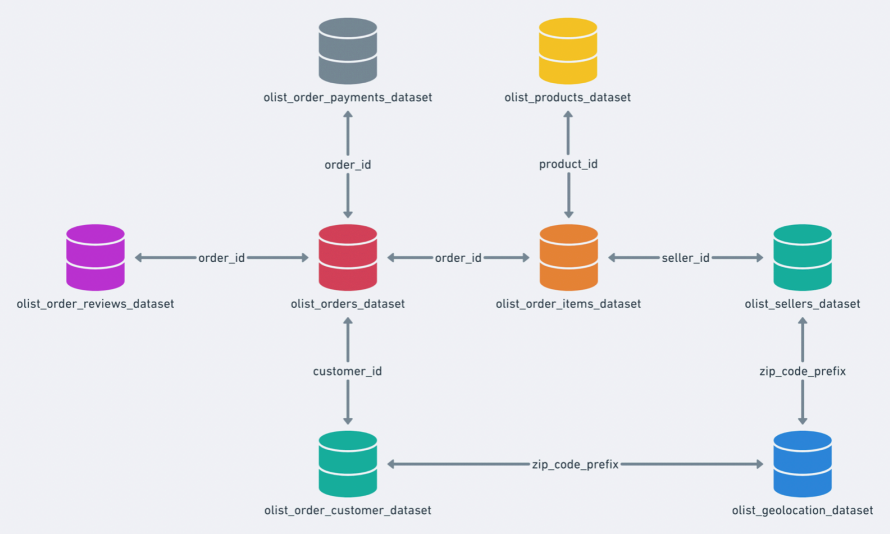

Figure 1

Technology Stack

We used multiple tools and technologies in this project. Specifically, we use Hadoop and Spark to store csv file, then we mostly used PySpark for data processing. We also built predictive model and recommendation system by utilizing Spark MLlib. Moreover, to achieve better visualization and get valuable business insight, we used Matplotlib and ArcGIS.

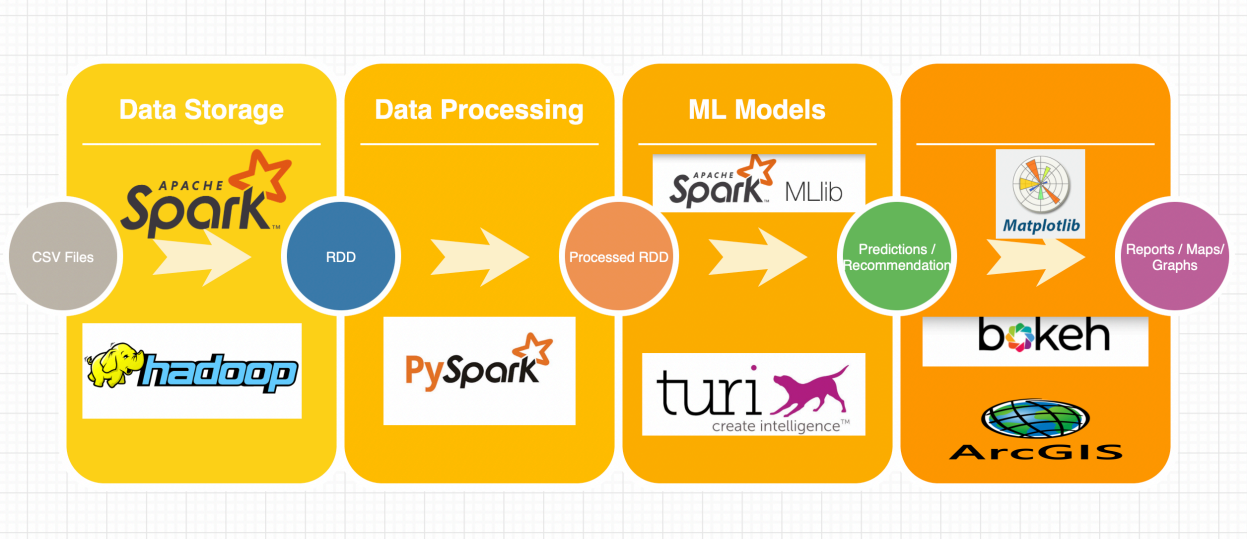

Figure 2

Detailed Solution

A) **Sales Prediction**:

Future Sales Prediction an e-commerce company is one of the most vital aspects of strategic planning. It helps the company, looking into the pseudo future, to plan better, especially if bad days are anticipated. We wanted to analyze how internal and external factors of the e-commerce company in one of the biggest countries in the western hemisphere can affect their Annual Sales in the future.

Linear Regression: We have used linear regression ML model to predict sales. Linear regression is used for evaluating trends and sales estimate, analyzing the impact of price changes, assessment of risk in financial services and insurance domain.

Linear regression is a statistical model used to predict the relationship between independent and dependent variables. As the name suggest, we use a straight-line equation (i.e. linear):

$$\mathbf{Y = (M * X) + C}$$

Where ,

Y= dependent variable (In our case, it is the predicted sale at some “Hour-Day-Month” for the next FY)

X= independent variable (In our case it were a bunch of features that explain/contibute the Y such as Sales, Hour, Day, Month, freight_value, product_description_length, product_weight_g, payement_value, payment_installments , review_score, product_photos_qty)

We followed a linear modeling approach to find the relationship between X above (i.e. predictors) denoted as X and dependent variable, Y above (i.e. target). We 1st cleaned the data and used PySpark to aggregate all the date with valid connections, following which we ran a feature selector, using RFE ie Recursive Feature Elimination, to select top 10 features that would best describe our model. Post this we used a VectorAssembler that combined our 10 raw features into a a single list. We then used a Normalizer to transform our data set into a vector with unit norm, which helped to standardize our input and improve the behavior of learning algorithms. We then split our data into 70% training data and remaining 30% test data.

We used the Linear regression technique from spark ML library to find the best fit line for the training as well as test data. The best fit line was found minimizing the distance between all data points and its distance to the regression line (by calculating the error (sum of squares error), we found the minimize distance), i.e., the distance between the points and the line was minimum. This was done in a recursive method with 20 iterations.

Our model generated a R2 (goodness-of-fit) of around 54%, which explains the % of variance of Y (dependent variable) from X (independent variable) ie how well the model fits the data and was a good enough to predict the future sales

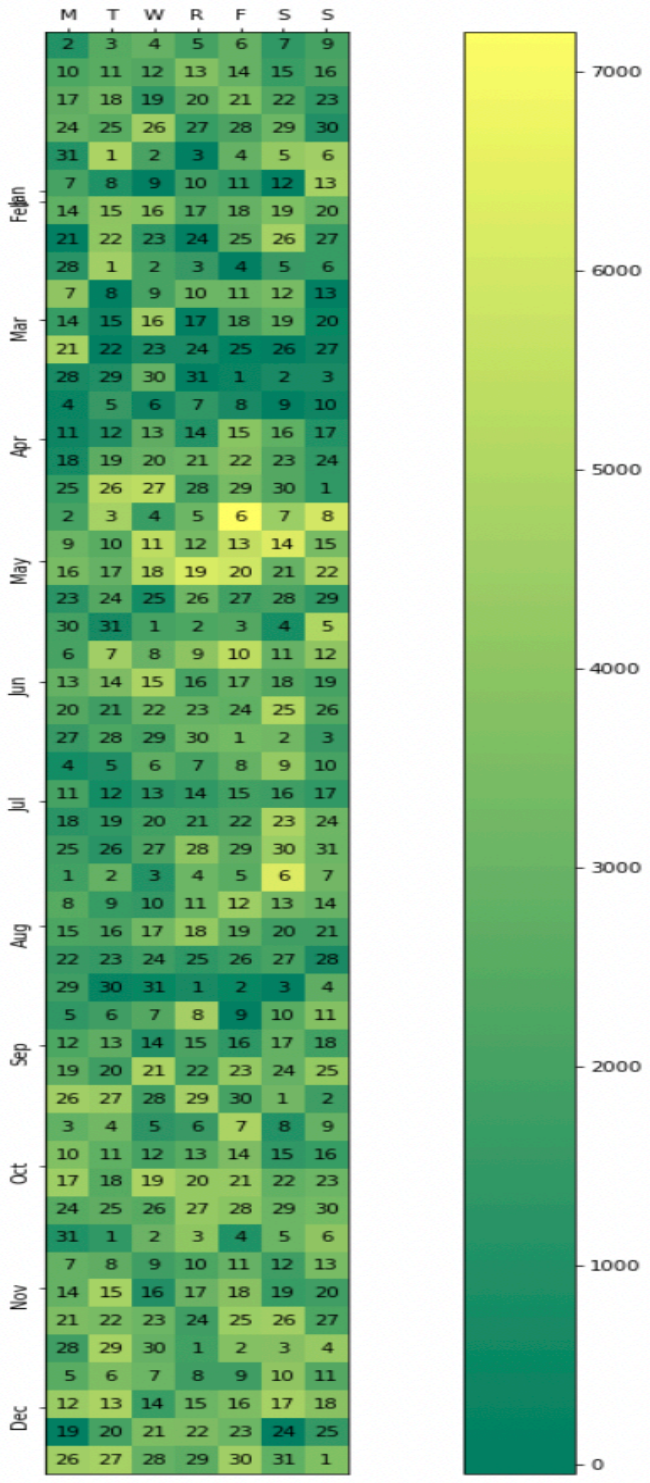
M T W R F S S
Jan
Feb
Mar
Apr
May
Jun
Jul
Aug
Sep
Oct
Nov
Dec
7000
6000
5000
4000
3000
2000
1000
0

Figure 3

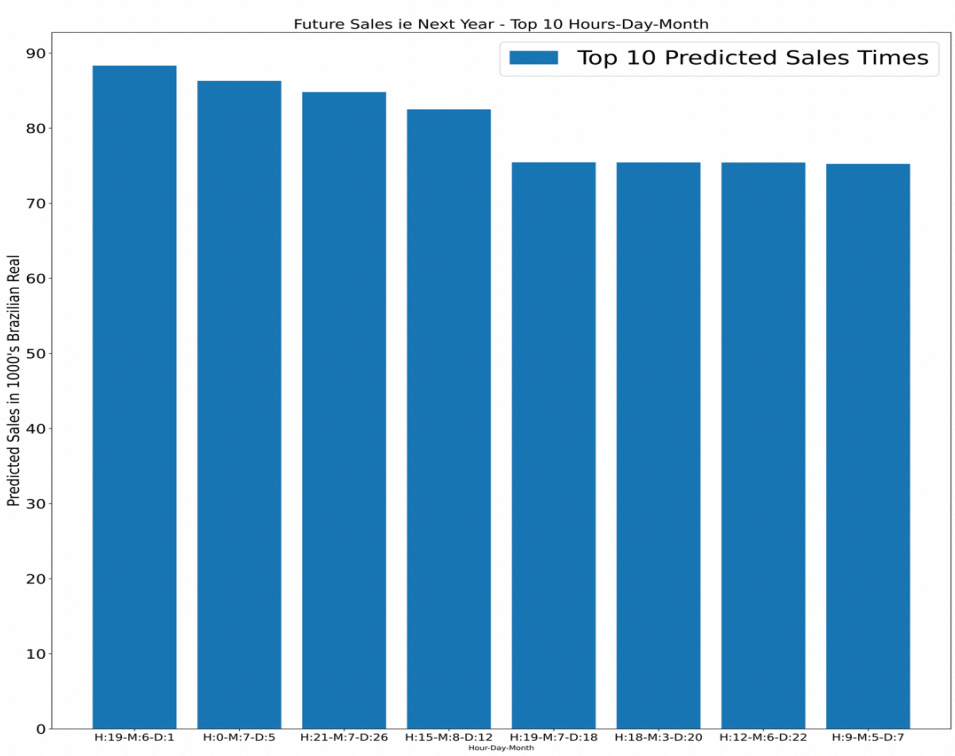

Figure 4

1) On using a Linear Regression ML model, we can see the above results, the **Calender** above gives us a more detailed view of which all days are going to be busy (Ex April 8th) and need extra preparation to keep up to the demand. As well as analyze what's wrong with days of Low sales.

2) Also above we can see the bar graph which shows the top 10 times in the following year with highest number of sales. This will help management to better prepare for these 10 days with product stock.

B) **Geospatial Analysis**

1) Using geographical co-ordinates (latitude, longitude) for each customer who purchased this a product online in Brazil this year and a list of all zip codes for Brazil, we could holoviews, geoviews and ARCGIS libraries to draw a graph of BRAZIL with area wise representation of customer count.
2) Sao Paulo, the capital of Brazil, and its surrounding capital region showed the highest customer density, as expected from any capital region of a big country.
3) The northern part of Brazil however shows less sparsely populated sales, partly because it has more rural areas with less exposure to digital shopping. The company can adopt strategies focused towards improving sales in rural parts.

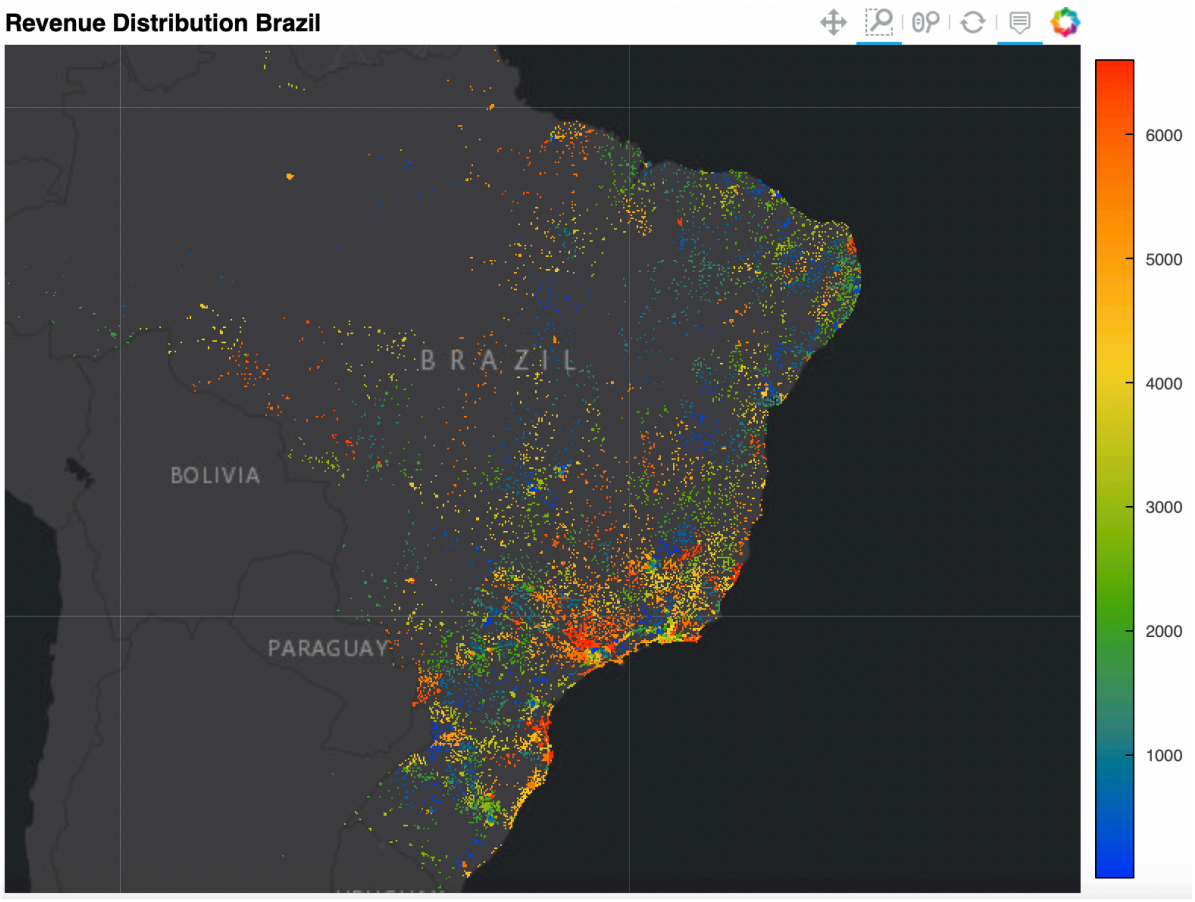

Figure 5

C) **Monthwise / Hourwise Sales Analysis**

1) Pyspark RDD based inner joins on Product, Order, Customer, Order_Info yields a nice aggregation of sales data for each product purchased by each customer in single/multiple orders with a purchase timestamp, grouping (PySpark GroupBy/ReduceByKey) on which yields a monthwise and hour wise sales analysis. Helpful to find out pain points for sellers.

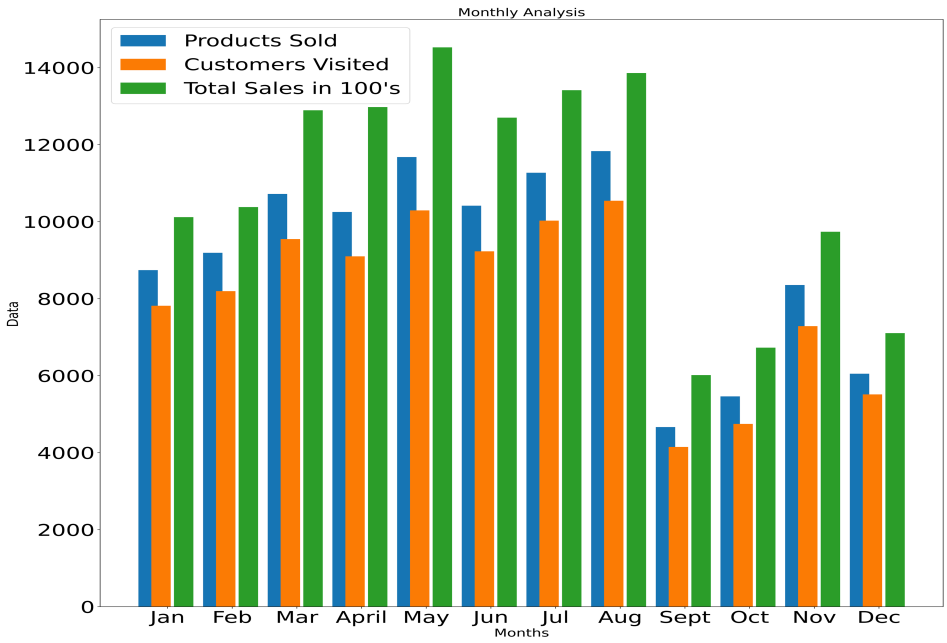

Figure 6

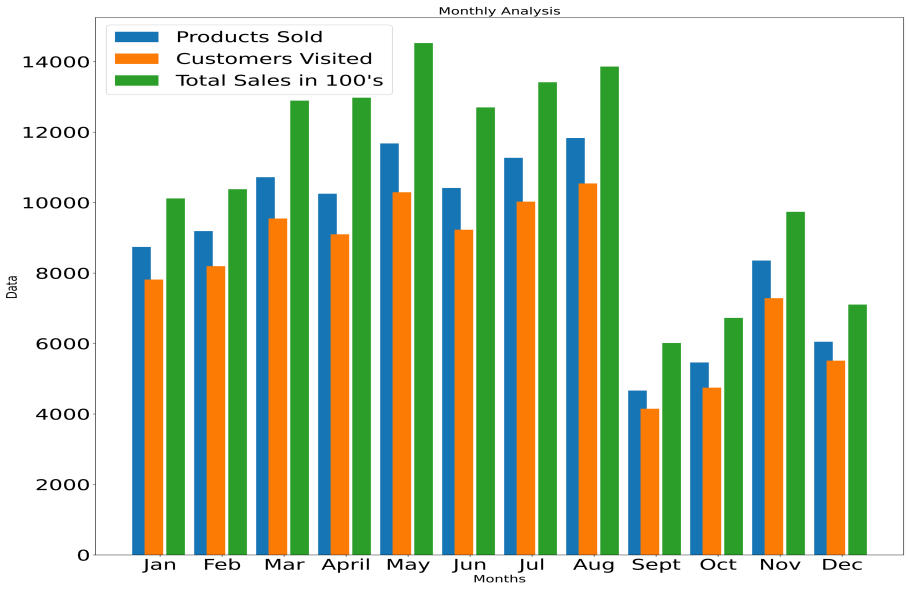

Figure 7

D) **Review Based Analysis**

1) Pyspark RDD based inner joins on Product, Order, Review, Order_Info yields a nice aggregation of average review scores data for each product purchased by each customer in single/multiple orders and grouping (PySpark GroupBy/ReduceByKey) on product id and then to sort to find the top 10 products with highest average review score helps pick the top 10 liked products by customers and similar least 10 liked products.

   a. It can be seen that health beauty is the most liked product followed by bed_bath_table and coming in at number 3 is housewares. The sellers can thus understand to keep high stocks of these products and to keep them as benchmarks.

b. Least 10 reviewed products: It shows that flowers, tables_printing_image, fashion_children_clothes are the worst products based on customer reviews. Sellers should introduce strategy to improve their quality.

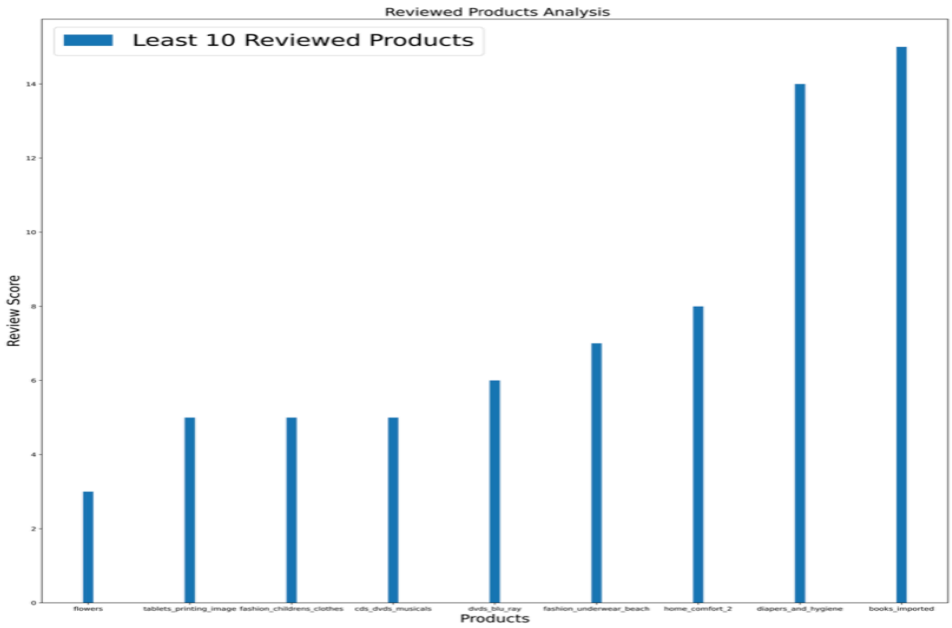

Figure 8

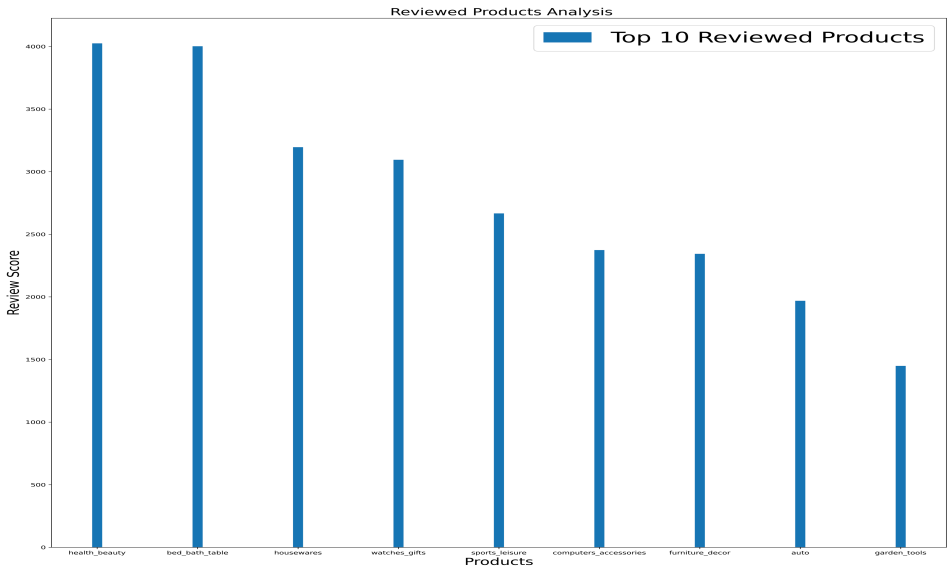

Figure 9

E) **Paying Trend Analysis**

1) Pyspark RDD based inner joins on Product, Order, Payement, Order_Info yields a nice aggregation of count of different ways which the customer has used to pay for the purchase (from 4 options being credit card, debit card, boleto, voucher) for each order and grouping (PySpark GroupBy/ReduceByKey) helps to get the count and thus the % of use of each payment type.
    a. It is quite evident from this bar graph that almost 3/4 of the customers use credit cards and a marginal section use debit card. This goes to show the popularity of credit cards which may be due to the various schemes that they have to offer. This also goes to show that the e-commerce company needs to beefup its cybersecurity to avoid being victim to credit card related frauds.

b. The next most preferred form of payement is Boleto ie payment method in Brazil regulated by FEBRABAN, short for Brazilian Federation of Banks.e

c. Vouchers are the 3rd most preferred forms of payement.

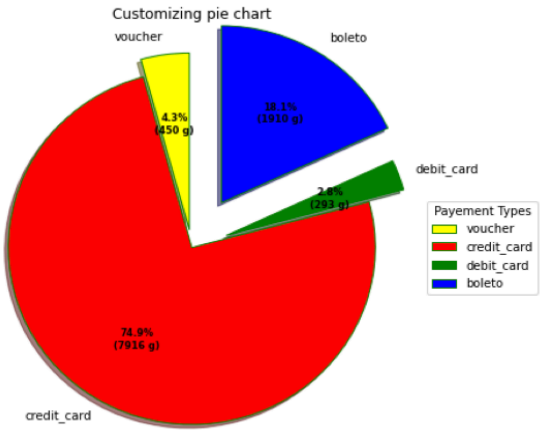

Figure 10

F) **Customer Segmentation**

Customer segmentation is very important for online retail industry. It divided the company's customers into groups that reflect similarity among them. The goal of customer segmentation is to maximize the value of each customer to the business. Now, we implemented a RFM analysis to better understand our customers so as for future marketing.

RFM stands for recency, frequency and monetary. We made data processing for our dataset and extracted 4 key features for our analysis, which are customer id, order id, payment value, and purchase time respectively. For recency, we first calculate RecencyDays, which is a measure that tells the difference in days between the 2016-09-04(which we have chosen as the day of reference) and the order date. Then we shall consider as recency the minimum RecencyDays that we

calculated in the previous step. This makes sense since the minimum RecencyDays will give us the number of days that have passed since the customer's last purchase. For frequency, it calculated as the count of purchase that customer has made. Lastly, for monetary, it is the total value of purchase that the customer has made. We utilized groupby function in Spark to get recent consumption day, shopping frequency and total revenue for each customer, generating a new data frame in Spark.

Then we assigned a score from 1 to 4 to each feature (recency, frequency, monetary). 1 indicates and 4 indicates highest score. We divided the RFM score based on statistical quartiles for 4 parts, which are 0.25,0.5,0.75 respectively. For instance, we assigned 4 to 'Monetary' when purchase amount is larger than 0.75 percentile. We assigned 3 to 'Frequency' when it is larger than 0.5 percentile but smaller than 0.75 percentile. According to the scores, we can segment customers into low value, mid value and high value customers.

- Low Value: customers are not frequent buyers and generates very low revenue.
- Mid Value: fairly frequent and generates moderate revenue.
- High Value: marketing target group with high frequency and high revenue.

In our case, specifically, if a customer obtains 444 score, then he is identified as high value customer, while 111 identified as a low value customer. From the following table, we got Recency, Frequency, Monetary, and RFM score for each customer.

customer_id	Recency	Frequency	Monetary	R_Quartile	F_Quartile	M_Quartile	RFM_Score
4f55f44c9af248a6c...	589	1	396.97	1	1	4	114
e2b3cf60eee891a56...	67	1	171.75	4	1	3	413
d550ac7fe9688fbc8...	324	1	228.6	2	1	4	214
e154e499a4edf9f98...	508	1	105.28	1	1	2	112
6bafcf687afcf7e1c...	292	1	57.78	2	1	1	211
18f6ca10777417c93...	272	1	61.36	2	1	1	211
860ac166573be76ff...	396	1	111.02	2	1	3	213
316fc927e9216ff78...	466	1	117.94	1	1	3	113
e3c7e245a96d7fa33...	367	1	415.97	2	1	4	214
4ba1bde676ff918af...	165	1	117.26	4	1	3	413

Table 1

In order to have a more vivid intuition, we want to find the number of customers in each segmentation. Therefore, we visualized the number of customers per segment in bar chart.

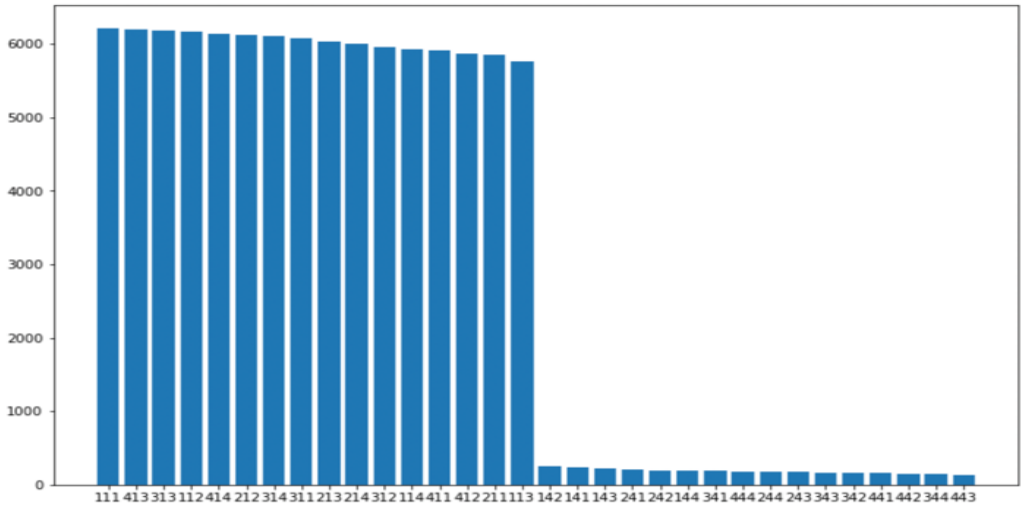

Figure 11

RFM Analysis

Surprisingly, the low value customers (111) compose the largest composition, then high-spending-new-customer takes the second largest percentage. Usually, we would ignore low value customers and target on high value customers to conduct various marketing campaigns. However, in this case, considering the large number, we need to dive it deeper. First, we'd better create customer

profile for these low-value customers, both demographic (age, sex, geography, etc) and behavior analysis, to understand their preference. Then based on our conclusion, we can conduct specific marketing penetration for these customers with the aim to increase their frequency and revenue generated per consumption. What's more, considering Olist is an Ecommerce website, we need to think about more strategies to improve retention rate and conversion rate for customers. Lastly, we can also implement competitive analysis, to figure out contribution margin and profit points of the competitors of Olist. As for high-spending-new customers, we need to target on them, with more appropriate marketing strategies. They are absolutely the group that we do not want to lose.

G) **Recommendation System**

During the last two decades, with the rise of some tech giants, like Amazon, Netflix, etc, recommendation system takes more and more place in our lives. Recommendation system are of critical importance to some industry and their profit. For instance, in 2009, Netflix awarded a $1 million prize to a developer team for an algorithm that increased the accuracy of the company's recommendation engine by 10 percent. In Ecommerce filed, it is very important too.

Looking at our datasets, we have 5 different csv files, which are customer, review, payment, order and product. We really wanted to figure out, how to recommend most appropriate and favorite product to each customer, so as to increase their purchase and revenue for the company. Therefore, we decided to build a recommendation engine in Spark by using collaborative filtering.

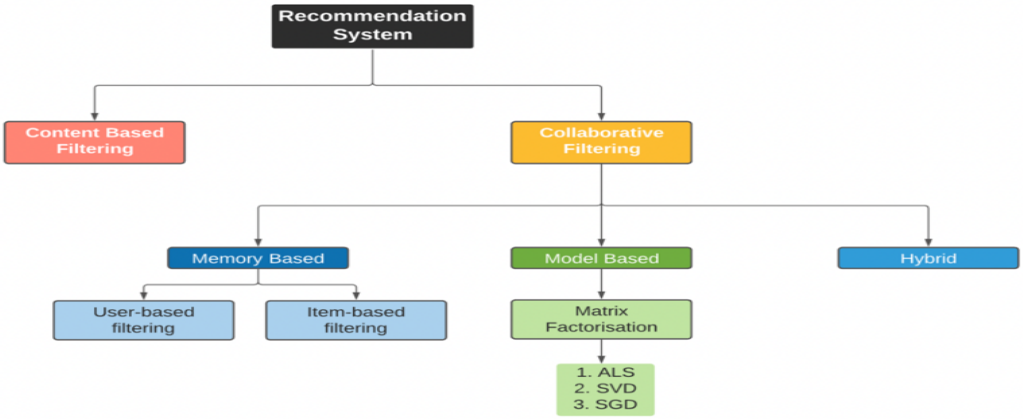

Figure 12

For recommendation system, there are two kinds, one is content based filtering, one is collaborative filtering. They are very different. A content-based recommendation system recommends items to a user by taking similarities of items. Collaborative filtering aggregates the past behaviors of all users. Let's say you know a friend who has the same taste as you because you both love French dress, then you might like purchasing other cloths that your friend has bought but you haven't.

In collaborative filtering, matrix factorization is used to solve sparse data problem, which works by decomposing matrix. Alternating Least Square (ALS) is a matrix factorization algorithm and it runs itself in a parallel fashion. It is doing a pretty good job at solving scalability and sparseness of the Ratings data, and it's simple and scales well to very large datasets. We used ALS in Spark library.

We extracted 4 features, which are *customerId, orderId, productId and review score* respectively. Customerid and orderid are both exhibited in hex but ALS algorithm only accepts integer, so we utilized StringIndexer to convert them to index, then established pipeline to transform the whole

spark dataset. After data preparation, we built recommendation system by splitting train and test data (0.8:0.2), tuning hyperparameters with cross validation, finally, we evaluated accuracy of our model by performing out-of-sample test, achieving 1.04 RMSE. From the following plot, we generated top 10 recommendation product for each customer.

Customer_id	**Recommendations**
00012a2ce6f8dcda20d059ce98491703	Housewares, cool_stuff, stationery, toys, furniture_decor, baby, computers_accessories, perfumery, toys, toys
000161a058600d5901f007fab4c27140	Housewares, home_appliances, sports_leisure, sports_leisure, watches_gifts, housewares, auto, health_beauty, furniture_decor, fashion_shoes
0001fd6190edaaf884bcaf3d49edf079	pet_shop, sports_leisure, baby, furniture_decor, computers_accessories, toys, computers_accessories, construction_tools_construction
0002414f95344307404f0ace7a26f1d5	computers_accessories, home_construction, home_construction, sports_leisure, auto, watches_gifts, garden_tools, baby, health_beauty, furniture_decor
0004164d20a9e969af783496f3408652	housewares, computers_accessories, perfumery, market_place, bed_bath_table, air_conditioning, telephony, sports_leisure
00046a560d407e99b969756e0b10f282	toys, health_beauty, computers_accessories, perfumery, auto, ('bed_bath_table', arts_and_craftmanship, garden_tools, auto, furniture_decor
00050bf6e01e69d5c0fd612f1bcfb69c	home_appliances, luggage_accessories, books_general_interest, telephony, health_beauty, health_beauty, furniture_decor, home_appliances, computers_accessories, telephony,
000598caf2ef4117407665ac33275130	sports_leisure, pet_shop, sports_leisure, small_appliances, fashion_bags_accessories, , auto, baby, bed_bath_table, garden_tools, bed_bath_table
00062b33cb9f6fe976afdcff967ea74d	furniture_bedroom, furniture_decor, sports_leisure, garden_tools, telephony, sports_leisure, auto, sports_leisure, computers_accessories, auto
00066ccbe787a588c52bd5ff404590e3	health_beauty, bed_bath_table, home_appliances, sports_leisure, luggage_accessories, furniture_decor, toys, housewares, stationery

Table 2

Shortcomings

There are several shortcomings of our analysis. Firstly, for price forecast, we used linear regression, but accuracy is not very high, only 54 R2. if we have time, we can use feature selection, cross validation and implement other advanced supervised model to improve the accuracy of our prediction model. Secondly, we can use better visualizations tools such as TABLEAU or d3. Thirdly, for customer segmentation part, we lack competitive analysis. For deeper analysis, we should research in competitors of Olist, analyzing their contribution margin and profit point. Lastly, we should also improve the accuracy of our recommendation system. Now it is only with 1.04 RMSE.

Conclusion

To conclusion, we used PySpark, Spark ML algorithms, and ArcGIS to build predictive model and recommendation system, so as to conduct pricing, marketing and sales strategies.

According to our predictive analysis, we suggest the company make adequate preparation for large amounts of sales on April 6th, April 8th, May 19th and June 15th in the next year. We suggest the company to conduct maintenance work during 2 am to 4 am considering the traffic. Based on our geospatial analysis, we advise the company to target at Sao Paulo area given the largest revenue, meanwhile, it's better to develop rural area, which are very potential area. According to customer segmentation analysis, we suggest that the company can target at high spending new customers, create customer profiles for low value customers and dive it deeper to figure out the reason for the large composition of low value customers. In this way, we can better develop marketing strategies and conduct marketing penetration for low value customers. Lastly, we generate recommendation product for each customer.

Gratitude

We extend want to extend sincere thanks to Professor Juan Rodriguez and TA Vishal Vanam for their teachings and guidance.

Reference

1. https://spark.apache.org/docs/latest/mllib-collaborative-filtering.html
2. https://towardsdatascience.com/building-a-recommendation-system-with-spark-ml-and-elasticsearch-abbd0fb59454
3. https://towardsdatascience.com/building-a-recommendation-engine-to-recommend-books-in-spark-f09334d47d67
4. https://developers.google.com/machine-learning/recommendation/
5. https://www.statista.com/topics/4697/e-commerce-in-brazil/