E-Commerce Analysis of Consumer Behavior

Taya Stewart, Raymond Delgado, Navaneeth Visagan, Frank Chen, Ninelia Talverdi
CIS 5200 - Group 3
Advisor: Dr. Jongwook Woo
California State University, Los Angeles
tstewar2@calstatela.edu, rdelga40@calstatela.edu, nvisaga@calstatela.edu, fchen26@calstatela.edu, ntalver2@calstatela.edu

Abstract: This paper serves as an analysis of e-commerce consumer centric behaviors. In reviewing, relevant data, statistics, and producing particular product research we were able to find key insights in consumer behavior that will enable sellers to make better informed and strategic business decisions. The data source is a Comma Separated Values (CSV) file type packed in zip format downloaded from kaggle.com. The dataset was collected from online stores that logged users' interactions during October 2019. The total file size is 5.28 GB of relevant consumer data. Using text analysis, we found a series of the customers' shopping events, patterns by factors such as day and hour, product category and brand. Finally, we utilized business intelligence software to visualize insights and provide analysis from the dataset.

1. Introduction

E-Commerce sales are growing exponentially and have become preferable to in store sales for a substantial sector of consumers. Consumers favor e-commerce sales because it saves them money and time when comparing different features among a broad range of products. The new challenge raised by the developing market is to use unique insights and analysis to drive a competitive advantage and to improve profits. However, this is a conflict to the consumers interests as they want low cost products. To balance the interests of the company and better service customers, it is important to investigate a panoramic view of the market to develop a comprehensive market strategy and make efficient management decisions.

Our team of data analysts utilized Big Data technologies such as Hadoop and IBM Cognos Analytics, SAP Predictive Analytics, Tableau and Power BI for visualization to analyze the large amount of e-commerce data. In the process, we revealed valuable insights regarding user behavior that have a substantial impact on ecommerce strategy.

2. Related Work

A research study funded by Oracle Cloud Innovation Accelerator attempted to predict customer ratings on Amazon products by developing models in Oracle Big Data and Azure Cloud Computing services in conjunction with Spark ML and AzureML architecture [1]. The report was able to review customer behavior using 15 attributes to analyze the ecommerce industry and predict within reasonable confidence the rating of an Amazon product. The methodology for the massive 7 million records was based on big data mining, machine learning, and predictive algorithm technologies.

With the increase of customers choosing to shop online, a report published in the 5th International Conference on Cloud Computing and Big Data Analytics reviewed online transaction systems records on consumer data to predict customer's buying preferences. The model used is able to predict with 88.51% accuracy the purchasing behavior of users within this dataset [2]. While consumers are incentivized by numerous factors, these studies were able to quantify consumer behavior and provide key insights to e-commerce producers interested in developing a consumer-centric business model based on big data analysis.

3. System Specification

To work on this data file, we used services offered by Oracle cloud cluster with 3 nodes. We worked on Hadoop, but put our focus on several data analysis and visualization software. The details are given below:

Cluster Version – Oracle Big Data Compute Edition
Number of Nodes – 3
Memory size – 160 GB
# of OCPUs – 8
CPU speed – 2.20 GHz
HDFS capacity – 802 GB

4. Architecture Workflow

We downloaded the data file from kaggle.com, extracted the csv file from the package and uploaded the file to the HDFS, utilized the Hive query language (HiveQL) to create an external table and view which convert fields, based on the data. Then we executed the Hive queries to analyze the data and got the result file in csv format for data analysis and visualizations. At the end, we found these insights could be used for business decisions.

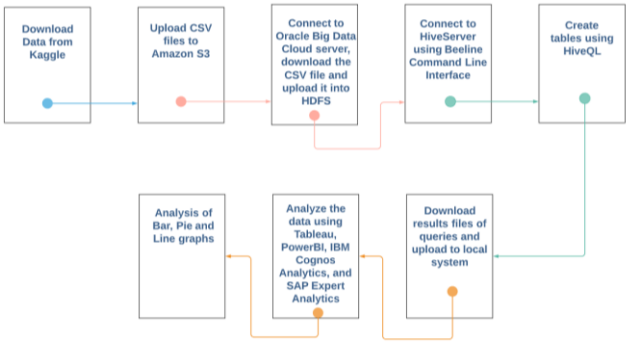
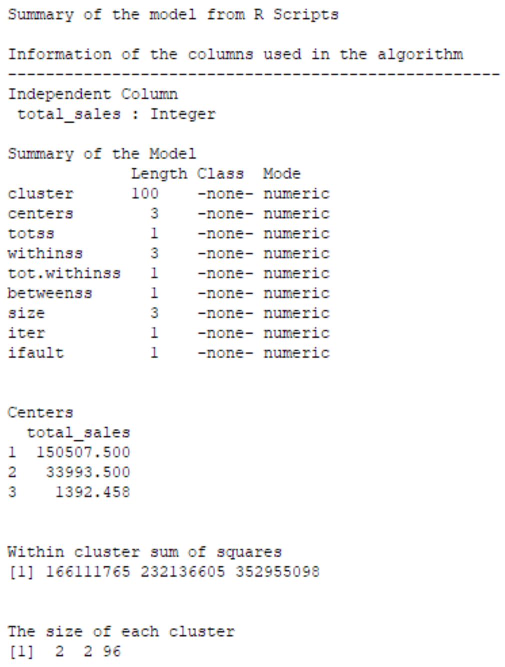

Download Data from Kaggle
Upload CSV files to Amazon S3
Connect to Oracle Big Data Cloud server, download the CSV file and upload it into HDFS
Connect to HiveServer using Beeline Command Line Interface
Create tables using HiveQL
Download results files of queries and upload to local system
Analyze the data using Tableau, PowerBI, IBM Cognos Analytics, and SAP Expert Analytics
Analysis of Bar, Pie and Line graphs

Figure 1 - Architecture Flowchart

5. Analysis and Visualization

After cleaning the original data set, parsing the primary key into core components, and developing queries, the data was visualized using business intelligence software. While the queries output tabulated rows of data, it needed to be further synthesized into understandable imagery. From our depictions, there were key insights on consumer trends, behaviors, and preferences.

5.1 K-Means Analysis



```
Summary of the model from R Scripts

Information of the columns used in the algorithm
-----------------------------------------------------
Independent Column
 total_sales : Integer

Summary of the Model
             Length Class  Mode
cluster      100    -none- numeric
centers        3    -none- numeric
totss          1    -none- numeric
withinss       3    -none- numeric
tot.withinss   1    -none- numeric
betweenss      1    -none- numeric
size           3    -none- numeric
iter           1    -none- numeric
ifault         1    -none- numeric


Centers
  total_sales
1  150507.500
2   33993.500
3    1392.458


Within cluster sum of squares
[1] 166111765 232136605 352955098


The size of each cluster
[1]  2  2 96
```

Figure 2 - K-Means Analysis

SAP Expert Analytics provides R-Studio integration. Operating this feature, a K-Means analysis developed clusters that segregated products into definitive and similar groups.

5.2 K-Means Distribution

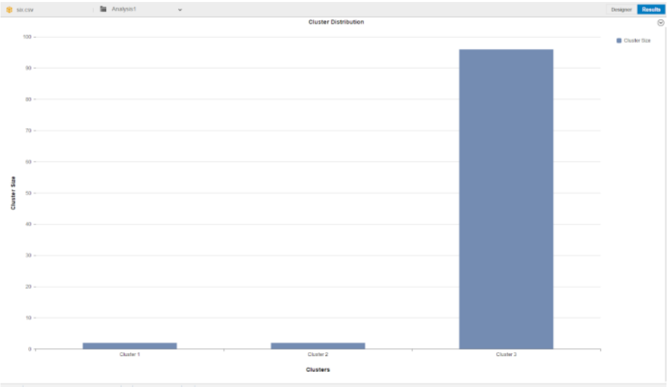


Figure 3 - K-Means Distribution

Utilizing a query of primary category, brands, and total sales, within SAP Expert Analytics, we performed a K-Means Analysis to formulate 3 clusters. It was found that most of the sales are on the low end in cluster 3. Most products are centered around 1,300 sales and total sales peak at 16,500 total sales which should be employed to determine production needs so that the brands in our data set do not over or under produce.

5.3 Forecast of Sales Revenue by Day

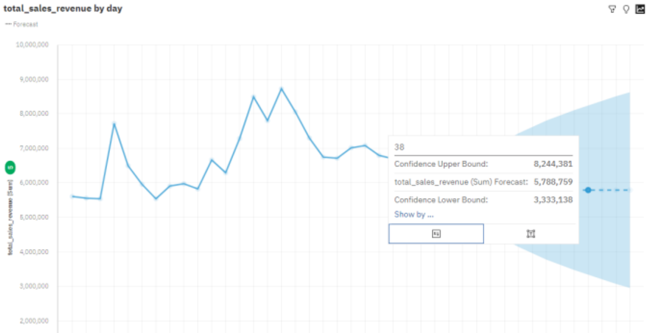


Figure 4 - Forecast of Sales Revenue by Day

Using IBM Cognos, we developed a forecast of the total revenue using October 2019 data to estimate the total revenue for an additional 10 days. The resulting forecast gives both an upper bound and lower bound estimate with 95% confidence. From this, we can determine industry trends that producers can use to establish price, seasonality, and competitive strategy.

5.4 Trend of Total Unique Buyers by Hour

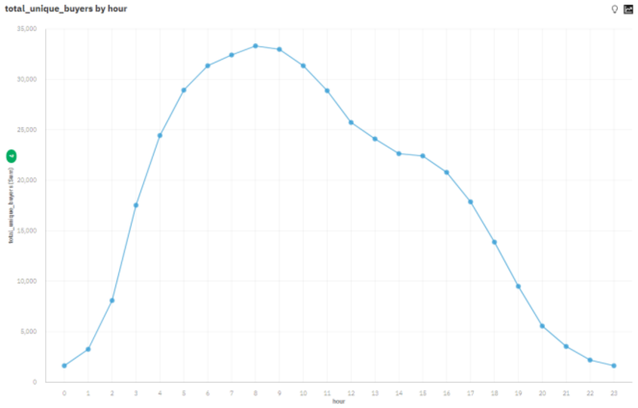


Figure 5 - Trend of Total Unique Buyers by Hour

Additionally, IBM Cognos was used to develop a trend line of total unique buyers by hour. We are able to see that there is a peak around 8:00. This is when sellers should send out advertisements, discounts, and incentives that will increase the amount of buyers' engagement.

5.5 Active Users by Hour

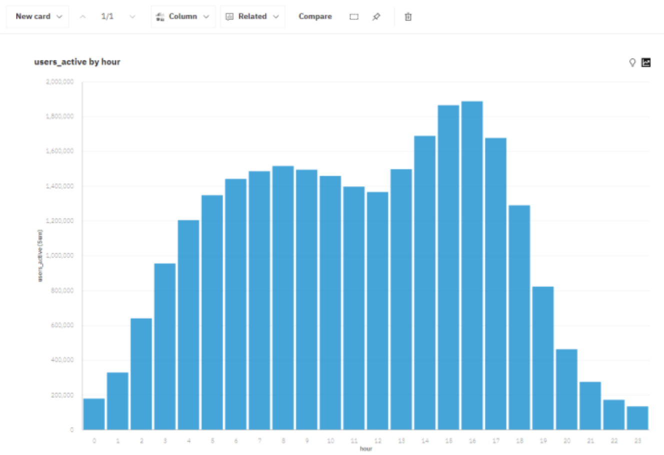


Figure 6 - Active Users by Hour

In IBM Cognos, using queried data, we created a column chart that depicts active users by hour. This is different from unique users in that it includes all users instead of new buyers. From this we see that the busiest time is between hour 16 and 15, as well as the most inactive hour, hour 23. Given this data is focused on ecommerce, it can be used to determine load balancing, labor needs, and similar to total unique users we can use this to increase user engagement, especially on high grossing products.

5.6 Most Popular Categories by View

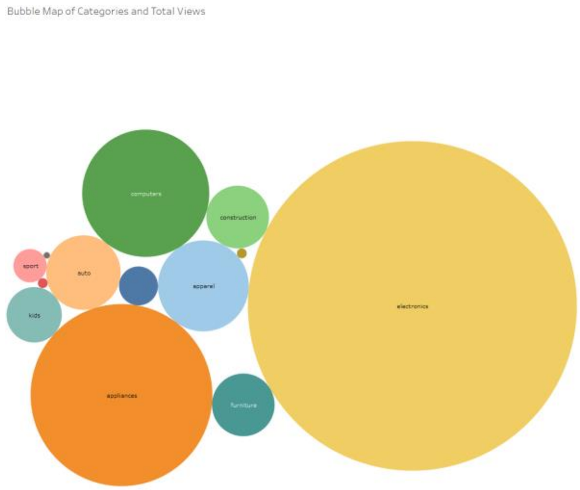


Figure 7 - Most Popular Categories by View

Using Tableau and the queried data, we created a bubble chart to showcase the popularity of certain categories of shopping. The popularity of a category, in this instance, is measured by views. Using that metric, the most viewed and subsequently the most popular category is electronics followed by appliances and computers. Businesses can use this analysis, to help prioritize supply planning based on the demand/popularity of the categories to make sure their customers have products available.

5.7 Event Types

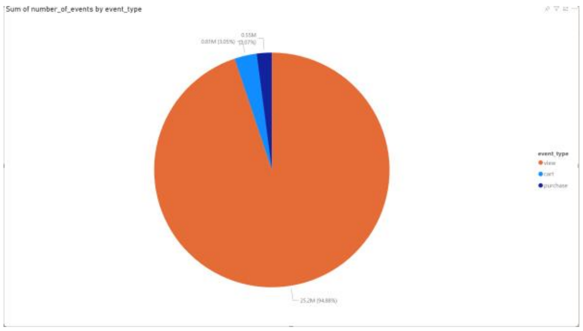


Figure 8 - Event Types

Figure 8 illustrates a pie chart using Power BI. The queried data shows how many times a certain event occurred. These events are labeled as view, cart, purchase. View refers to when a customer views the product on the site, cart when a product is in their virtual cart, and purchase when a customer buys the product. An overwhelming majority is a view event, customers browsing the online store. However only 3% actually move an item to their cart, and even less purchase an item, at 2%. This is a valuable metric to understand showcase. Businesses need to figure out how to convert view and cart events into purchases or else start losing profits.

5.8 Customers with the Most Purchases

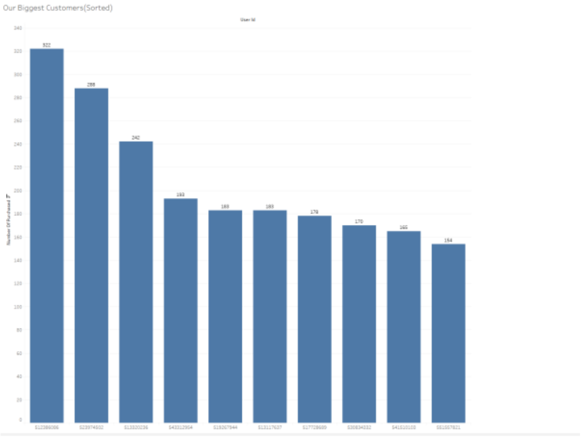


Figure 9 - Customers with the Most Purchases

Many businesses have loyal customers. Businesses have been seen to profit from rewarding their loyal customers. Figure 9 uses Power BI to provide the most loyal customers to the e-commerce business. Here, they can find loyal customers to provide benefits to and help maintain loyalty to the business.

5.9 Active Users by Weekday

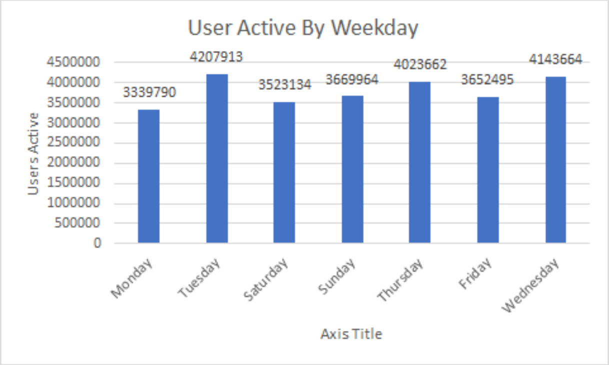


Figure 10 - Active Users by Weekday

We used Microsoft Excel to develop a column chart that depicts the total active users by weekday. The first finding that we discovered is that there is variety in user activity:

- Days with the most users active: Tuesday, Wednesday, and Thursday
- Days with the least users active: Monday, and Sunday.

The main insights that we gained from user activities are:

- Most users are active during the middle of the week.
- The e-commerce platform experiences the smallest amount of traffic during the weekend (Friday--Sunday) and Monday

Nonetheless, the e-commerce platform still experiences a large amount of traffic--the average number of active users is about 3,800,000.

6. Conclusion

Applying big data analytics principles and business intelligence software, it is evident that consumers are swayed toward products that are lower cost and accessible. There is a trend of unique and consistent consumers visiting ecommerce websites to shop within a predictable time frame, consistent revenue growth, and comparable user engagement for certain products. Sellers may apply the resulting data analysis to determine competitive sales and pricing per unit, reliable inventory, and to increase user engagement. During busy online hours, we recommend that the companies increase advertisements, incentives, and sales to match the consumers behavior.

References

[1] Woo, J. and Mishra, M. (2020). *Predicting the ratings of Amazon products using Big Data*. Wiley Online Library. Retrieved from https://doi.org/10.1002/widm.1400.

[2] X. Dou, "Online Purchase Behavior Prediction and Analysis Using Ensemble Learning," 2020 IEEE 5th International Conference on Cloud Computing and Big Data Analytics (ICCCBDA), Chengdu, China, 2020, pp. 532-536, doi: 10.1109/ICCCBDA49378.2020.9095554.

[3] GitHub URL: neltf. “Neltf/CSULA_CIS_5200.” *GitHub*, 2020, github.com/neltf/CSULA_CIS_5200.

[4] Data source URL: Kechinov, Michael. “ECommerce Behavior Data from Multi Category Store.” *Kaggle*, 9 Dec. 2019, www.kaggle.com/mkechinov/ecommerce-behavior-data-from-multi-category-store.