Machine Learning Based Implementation of Online Retail Analytics Using Elastic Cloud

**AUTHORS**

Rahul Subramaniam  
*Information Science and Engineering  
R.V College of Engineering*Bangalore, India  
rahulsubramaniam0510@gmail.com

Kunal Bhandari  
*Information Science and Engineering  
R.V College of Engineering*Bangalore, India  
kunalbhandari15@gmail.com

Arnab Jana  
Information Science and Engineering  
R.V College of EngineeringBangalore, India  
arnabfromjec05@gmail.com

***Abstract***—*In the highly competitive online retail sector, making decisions intuitively or by simply viewing selling history to improve overall sales is insufficient. Most of the online stores of today demonstrate a reliance on data analytics to make business, customer centric, operational as well as strategic decisions. This given work performs various kinds of data analysis on the data of online retail stores. The work also provisions the methodology to performs data analysis on the same. The dataset for this project was obtained by developing a web crawler and scraper that extracted relevant details with respect to every store item for the given retail stores. This unstructured data was then catalogued, and appended to an already existing database to create a custom dataset. The categorization process was followed by a graphical exploratory analysis, that would help data analysts, business analysts to make product specific discussions. Post this, Predictive analysis was done on the data, by performing Product based clustering and Customer based Clustering. This portion of the project made use of various Natural Language Processing techniques, and defines the basic logic of a Retail Recommender system. The code was run in the Amazon cloud (Amazon Machine Image in the Amazon Elastic cloud) to ameliorate the execution speed. The project also verifies the correctness of the segmentation by the application of existing Machine Learning algorithms on the product choices that are made by the customers. The accuracy of the various models are then compared to conclude which algorithm is most suited for prediction. A detailed analysis of this kind can help provide inferences and details to organizations to enhance their product sales and profits.*

Keywords— Clustering, Exploratory Analysis, Machine Learning, Predictive analysis, Stemming.

1. **Introduction**

With the advent of Interconnected Networks, the retail sector saw itself migrating from the traditional shop based model to the online item delivery based model. Such a model offers benefits to both, the vendor as well as the buyer. The vendor is able to provide doorstep delivery of goods, which is more convenient to the customer as well. The vendor on the other hand, benefits from the fact that he or she does not need to invest much capital in the development of retail outlets/stores. A minimal cost warehouse would suffice.

Yet another benefit of online retail stores is that in the due process of a customer transaction, the details of the deal can be stored with ease in contrast to manual entries as in the case of the shop-based models. The amount of data collected varies from store to store, but is generally perceived to be enormous in its nature. Several inferences can be obtained from the data of previous transactions to help improve overall sales and profits, and to minimize production costs. However, when dealing with high volume data, attempting to perform analysis manually may prove to be tedious and long drawn.

Leveraging the power of machine learning and big data is a viable option to perform analysis on the retail store data, and to perform necessary predictions that will benefit the organization. It can help identify which place records maximum sales, when to increase server/load capacity to meet demands, which commodity is bought the most, which commodity has maximum contribution to sales and many other such inferences. It can also help recommend to the customer which item he or she would be interested in, based on his or her previous purchases, similarity of items etc. The recommendation can be performed by clustering various objects into groups, and then suggesting items based on buying patterns and the cluster in which the original item under study is. The given work used both Supervised and Unsupervised Machine Learning algorithms to perform appropriate data analytic processes.

Models of this kind can turn out to be an efficient and cost-effective solution to optimize the marketing and sales efforts made at an online retail store. The model in this project can further be improved by incorporation of Hadoop ecosystem to handle greater amount of data and to make the model more scalable in nature.

1. **Literature Survey**

With the diversification of products, the number of retail stores (online as well as offline) have increased greatly in the 21st century. This increase has lead online retail analytics to gain a lot of traction in the recent times. A lot of research is being done in this field, in order to improve both, customer experience as well as to vendor experience. This has resulted in manifold approaches that have been worked upon in the field of retail analytics.

The first approach to retail analytics dealt with an Agent based paradigm to facilitate the use of Big Data Analytics in the field of retail. The paradigm exploits characteristics of an agent such as autonomy, proactivity and intelligence in the performance of data analytics processes. The work also reviewed the situation background and analyzes the applications, properties and challenges Big Data Analytics faces in the field of retail.

Another perspective to online retail analytics dealt with Clickstream Analytics on the Amazon users’ simulated monthly traffic. It adopted a sequential frequent item set detection methodology to perform clickstream analytics. A composite dataset that simulated the monthly traffic of an Amazon U.S online retail shop was analyzed in order to demonstrate the efficient functioning of the analytical process.

A third approach dealt with the use of deep learning to perform Retail Analytics. It presented an automated approach to obtain information relevant to analytics using deep learning and image processing. The approach combined face detection, tracking, best view estimation, repeat customer identification, blacklisted customer warnings and facial sentimental classification. The work states that this approach achieved satisfactory results.

Recommendation of item category in the field of retail Analytics was also performed using an approach known as the Random Walk Model. This approach dealt with combining users’ behavior with the objective knowledge of items. It employed an improved graph based algorithm to predict items and compute a customer score. The paper claims that the proposed method can find possible purchases for users, as well as find potential customers for goods.

Yet another analytical approach proposed a Color Expansion method exclusively for Clothing Recommendation systems. It is based on the principle of similarity of colors, which is further combined with RGB and HSV color space. This is further computed to obtain color expansion. The work states that under experimental conditions, a single base color can expand to 300 similar colors, and these results can be stored in the color similarity. The paper advocates the simplicity of the color expansion calculation, stating that it can avoid the weakness of RGB color space. It also states that this method conforms to the rule of visual similarity judgement, which is adapted to the experts recommendation.

1. **Data Description**

The data for this analytical project was initially a single csv file that was previously acquired. However, due to the paucity of data to perform a detailed analysis on, the project retrieved data from the web with the help of crawlers and scrapers. The below diagram shows the flow of events ultimately leading to the dataset.

The process of obtaining and defining the data was primarily composed of three essential steps: Data Sourcing, Web scraping, Mapping the scraped data to the acquired structured data.

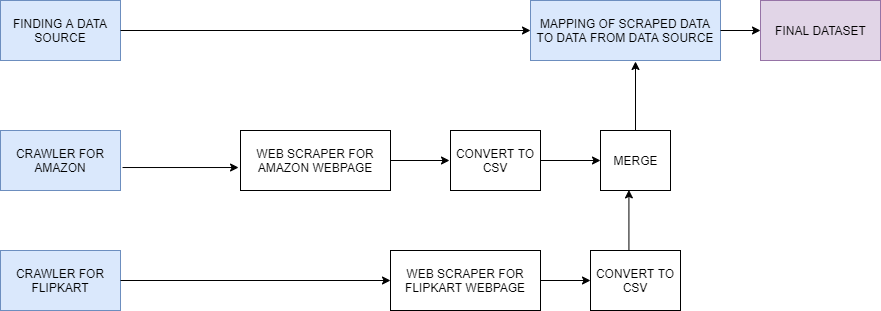


Fig 1. Flow diagram of the data description step

**3.1 Data Source**

The dataset was fetched from the Kaggle E-Commerce dataset. The dataset is 43 MB in size. It contains a single file known as data.csv. The dataset was a transnational dataset that contained all the transactions from 1st December, 2010 to 9th December, 2011 for a United Kingdom based online retail store. The dataset had a total of eight columns. These included Invoice number, Stock code, Description, Quantity, Invoice data, Unit price, Customer ID, Country.

There was no operation on this data until the scraped data was merged with it. The bifurcation to training and testing set also took place after this data was merged with the scraped.

**3.2 Scraped Data**

The Data was scraped from selected categories that the Flipkart and Amazon website offered. The Scrapy framework was used to firstly crawl through all the pages in the selected category for a given website, and then scrape the contents of that particular page. The contents retrieved were then filtered, and stored in a dictionary. Individual dictionaries for every category were then merged in order to create a set of set of final contents.

The final contents stored in the dictionary contained the Product description, and the Product unit price. The dictionary was then converted to a csv file. The programming for the entire crawling-scraping framework was done using the Scrapy Shell, and with relevant commands that run on the Scrapy shell. The categories from which the data was scraped were: books, computers, electronics, food, luggage, men’s fashion, women’s fashion, shoes, phones, watches, common items (default items on the Flipkart front page).

**3.3 Merging the Data**

On account of the unavailability of Customer Identities and Invoice Identities, the scraped data could not be directly merged with the dataset. Thus the team had to assign Customer credentials already present in the dataset to the scraped data. Firstly, a set of all the Customer Identity Numbers present in the dataset was created. This was followed by obtaining all the Invoices (Invoice number, and the Invoice date) that a given Customer has. A nested dictionary data structure was created Customer Identity Number as the key used for the outer dictionary. The inner dictionary contained details regarding the country of transaction and the set of invoices (Invoice number and Invoice date) associated with the given Customer Identity Number. Post this, a random Customer Identity was selected using the python random function. For this given Customer Identity Number, a random Invoice Identity Number was selected (using a similar custom made random function). The first item in the scraped data file was picked up, the product description and unit price was extracted and was appended to the inner dictionary. The range of values that the entries in the ‘Quantity’ column (in the dataset) assumed was calculated, and a random value within this given range was used as the quantity of product purchased in the given invoice.

The main purpose to Scrape data from different Data sourced despite having to randomly generate data is to demonstrate the volume of data that online retail stores generally have. The code for the merging process was run in the Amazon Elastic Cloud (EC2-AMI) to attain a greater execution speed. The final dataset generated was of size roughly 1.50 Gigabytes.

1. **Methodology**

This covers the technique and flow of events that was used to perform the cluster prediction process. The prediction methodology itself was composed of three integral steps: product cluster creation, customer clustering and the final prediction process. The code was run in the Amazon Elastic Cloud (EC2-Amazon Machine Image) to attain a greater execution speed.

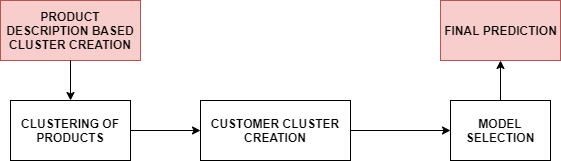


Fig 2. Flow diagram of the Methodology step

**4.1 Product Cluster Creation**

The Product Clustering was entirely based on the individual product descriptions. This given section leveraged the functions of Natural Language Processing to extract relevant keywords from the product descriptions. Firstly, the proper and common nouns present in the Product Description were extracted using Parts of Speech tagging and the Word Tokenizer functions available as a part of the Natural Language Processing Toolkit. For each of these given nouns, the root word was extracted using the Snowball Stemming technique. The individual nouns were then aggregated using the given set of names. This was followed by the step of counting the number of times each given root appears in the Dataframe. These roots were then stored in a single list.

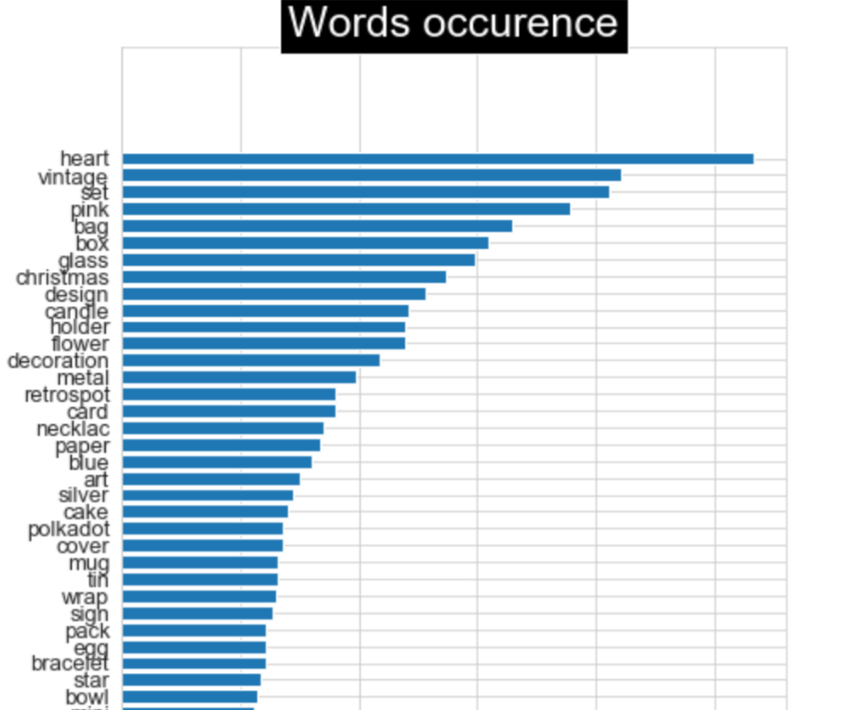


Fig 3. Diagram showing few of the most common root words extracted.

However, it was observed that among all the root words extracted, not all of them were relevant to online retail store analytics. The team believed that having colors in the set of root words did not serve the purpose. Moreover, it was observed that the length of some of the words in the most common list was less than two. Hence these kind of words were removed from the list. To further narrow down the redundancy, only words with occurrences more than 13 times were considered for the clustering purposes. After this, binary encoding of values was done using One Hot encoding technique.

To finally group these words, the K-Means clustering algorithm was used. The number of clusters that must be used is based on the silhouette coefficient parameter, that is calculated for a given number of clusters.

(1)

Here SC is the Silhouette Coefficient, while b is the mean nearest cluster distance and a is the intra cluster distance. It was observed that for this analytical project, the silhouette score for the number of cluster equal to 5 was greater than the other values, and also showed uniform cluster distribution. Hence, a total of five clusters was considered for further computation.



Fig 4. Diagram showing word cloud that demonstrates the various clusters.

**4.2 Customer Clustering**

Each of the products present in the dataset was allotted one of the five clusters. The dataset was structured such that the Customer Identity Number was the Primary Key, and hence each customer had purchased multiple products that belonged to multiple categories (or clusters). Thus, for every given Customer in the dataset, the sum of the prices for all the products, the mean value of the product price, the maximum price of all the commodities the customer bought, the minimum price of all the commodities the customer bought was calculated. Furthermore, for each Customer in the dataset, the percentage of products the Customer has bought in each category was calculated.

Any String values present in the dataset were temporarily dropped, to create a complete numeric dataset. On this given dataset, K-Means clustering algorithm was applied to classify the customers into clusters. The value of the number of clusters chosen for this clustering process was eleven. This was due to the high silhouette score and uniform distribution achieved using eleven clusters.

**4.3 Model Selection**

The given problem in hand was essentially a Classification problem. For any classification problem, a plenitude of machine learning algorithms are available and thus to choose the best among them, it is necessary to evaluate them on the same data based on a suitable parameter.

Here we chose the accuracy score metric to perform the computation. It was used on account of its simplicity and straightforwardness in demonstrating how accurate the given algorithm is. On running the predefined classification algorithms on the preprocessed data, the following accuracy results were obtained, as performed by the sklearn library. The training and testing set split ratio was 80% to 20 %. This was used due to the availability of great volume of data to demonstrate how accurate the model is, and hence the team believed that data would be better used when provided to the training set.

1. Algorithm and Corresponding Accuracy Scores

|  |  |
| --- | --- |
| **Algorithm** | **Score** |
| Support Vector Classifier | 0.8431 |
| Decision Tree Classifier | 0.8408 |
| K Neighbors Classifier | 0.8016 |
| Logistic Regression | 0.8743 |

The models developed by the individual algorithms were trained on the training dataset and then test data was used for final prediction of which category the Customer belonged to (they were generally used to view how accurate the clustering was). Since on evaluation, Logistic Regression algorithm was found to possess the highest accuracy, the model was considered for the prediction process.

1. **Results and Discussions**

The given model, apart from being instrumental in the working of a Retail Recommendation System, can also be used to perform different kinds of Analysis. This section primarily deals with explaining how the project was used to perform different kinds of exploratory analysis. Though the inferences obtained from this analysis cannot be extrapolated to the real world scenarios due to the random generation of certain fields, the work demonstrates how analytics can be used to obtain inferences to help organizations make better business decisions. The below graphs were plotted using the Matplotlib, an Open Source Python Library.

* 1. **Customer Analysis**

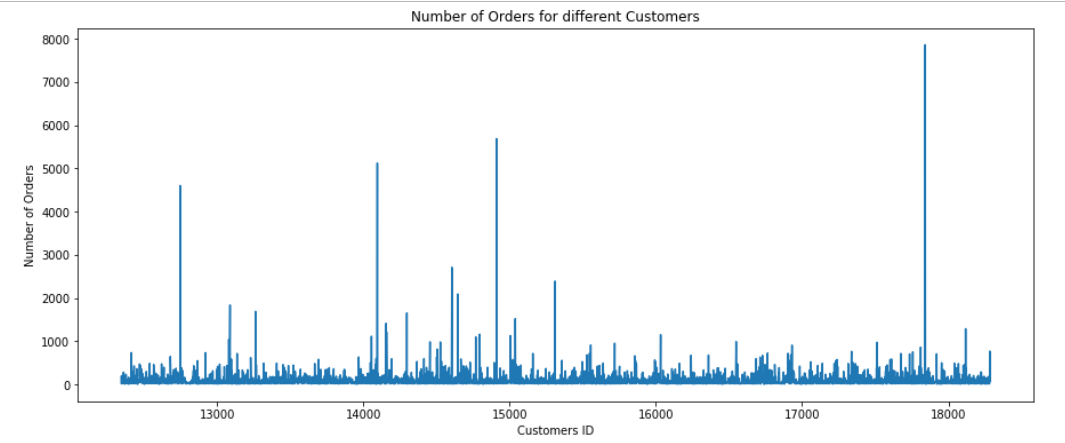


Fig 5. Diagram showing Exploratory Analysis for the entire set of Customers (based on Orders)

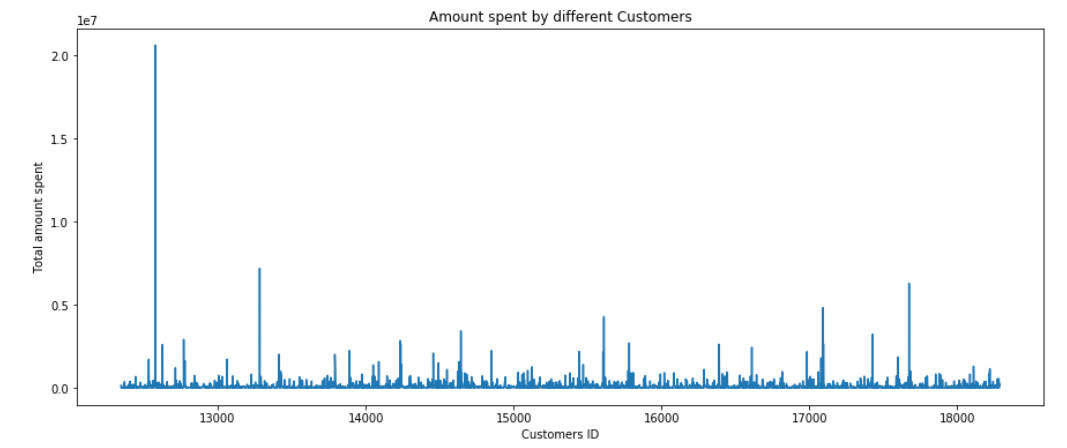


Fig 6. Diagram showing Exploratory Analysis for the entire set of Customers (based on Amount)

In Figure 5, it can be observed that most Customers have ordered less than thousand items, while a few customers have ordered more than 4000 items over the span of roughly 1 year. The Maximum number of items a Customer has ordered almost touches 8000. Customers with orders of such a magnitude are generally individuals who are happy with the products/services the online retail portal provides. Retaining such motivated customers could greatly benefit the business since such customers are responsible for Word of Mouth Marketing, raising the Customer Satisfaction Index and are representative of a set of customers who will purchase goods on a nearly daily basis. Providing incentives such as Loyalty programs, Long term customers discount could help maintain (if not raise) their satisfaction level.

In Figure 6, it can be observed that most customers have a spending of less than 5 million currency units, while a few customers tend to spend more than 5 million. Furthermore, one Customer spends 20 million currency units in online purchases. Such individuals have direct contributions to the organizations’ Sales numbers, and hence it is of paramount importance that such customers be retained. Incentives such as loyalty programs, personalized services could help retain such customers, and ensure steady consistent Sales figures.

Further analysis of Customer data with respect to products can provide information on number of customers that buy a given product. Such information is valuable to knowing how to improve the overall value of the product in the Market.

**5.2 Analysis of Number of Orders**

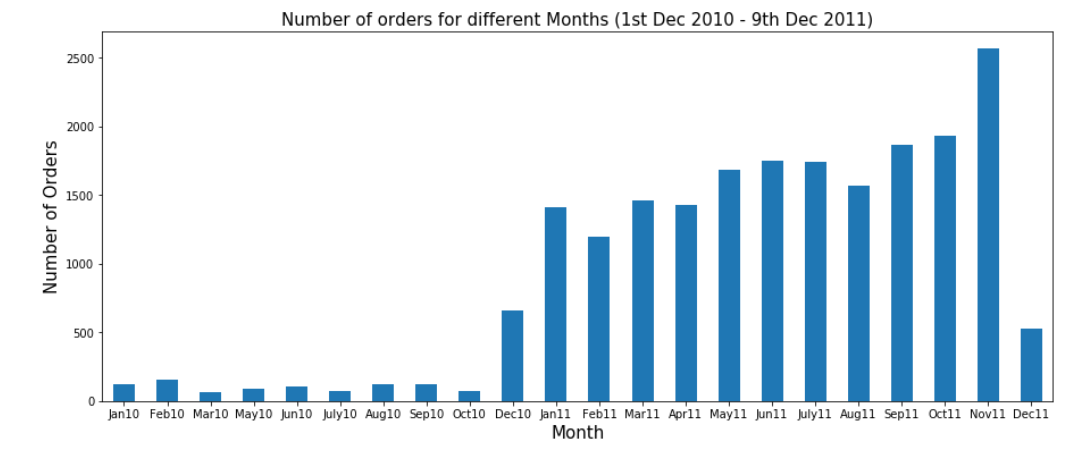


Fig 7. Diagram showing Exploratory Analysis for month wise number of orders

In Figure 7, it can be observed that most months have sold more than thousand orders, while a few of them have around 500 orders. The maximum number of orders was recorded in the month of November 2011, totalling to 2500 orders. Generally, the factors affecting number of order of the months relate to the seasonality. During a festive season, it is observed that the sale of items is more. There is greater sale of specific items on the onset of a season. For example, during the month of Winter, there is greater sale of sweaters, jackets, gloves etc. Graphical analysis of this kind (i.e. orders vs months) could provide information to the organization with respect to Inventory Management. The organization can stock up greater amount of inventory before months where number of orders is expected to be higher. Likewise, the organization could lower the amount of inventory during months where the sale of goods is low. Such actions could lead to low interest costs, and better working capital management.

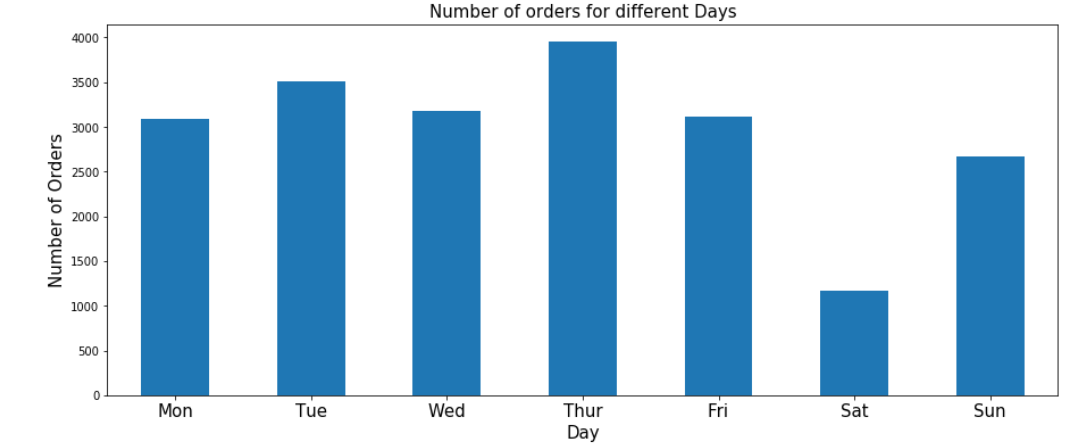


Fig 8. Diagram showing Exploratory Analysis for day wise number of orders

In Figure 8, it can be observed that the number of orders during a Thursday is above the general, while the number of orders during a Saturday is below average. An organization could use this kind of information to know when to increase server/load capacity so as to ensure it is able to cope up with the incoming traffic.



Fig 9. Diagram showing Exploratory Analysis for hour wise number of orders

From Figure 9, it can be obtained that the number of orders during a day follow a Normal Distribution pattern, with below 500 orders during till seven in the morning. The number of orders the peaks during noon (which is obvious since most people are active during this time period), and then again regresses to below 500 after 6 o’clock in the evening. Inferences from this graph could help corporations decide at what time they should launch application updates, perform testing operations etc. to ensure minimal customer interference at the given time.

**5.3 Country based Analysis**

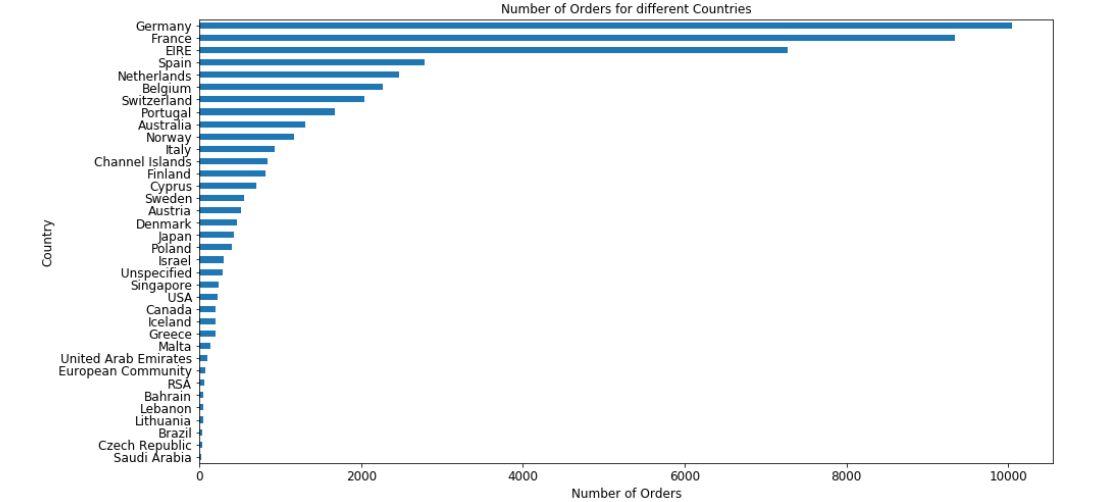


Fig 10. Diagram showing Exploratory Analysis for country wise number of orders

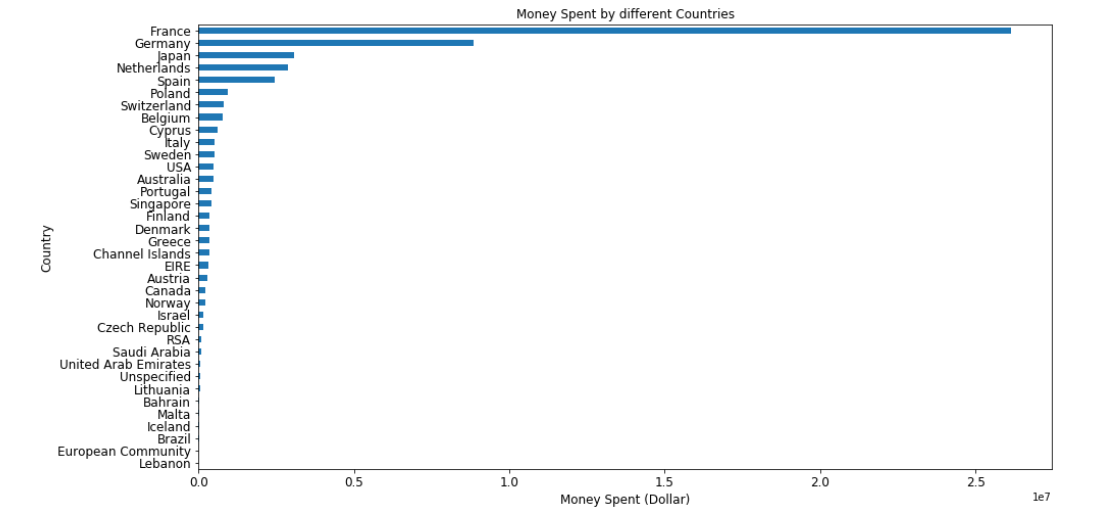


Fig 11. Diagram showing Exploratory Analysis for country wise number money spent

Figure 10 and Figure 11 provide a broad view of the sales situation in different countries. Though Germany records the highest number of customer orders, it does not record the highest amount of sales. Inferences from these kinds of graphs could suggest the company where they would like to allocate more capital and resources, would they like to discontinue services in any country, which products sell in which country, which country is more profitable for business and many other such questions.

1. **Conclusion**

This work presents that the Logistic Regression algorithm has the highest accuracy in predicting which cluster a given product belongs to, based on the accuracy score calculated for each of the four algorithms previously mentioned in this work. All the models show an accuracy of greater than eighty percent, thus proving the correctness of the model. Support Vector Classifier has been observed to be the second most preferred algorithm for category prediction. The Exploratory Analysis of the data can provide important business insights such as server capacity allotment, inventory management, profit calculation etc. The given project provides the base to develop a Customer Product Recommendation System by performing Customer Segmentation and Product Clustering.

1. **References**

*Amazon India Official Website*. Retrieved from http://www. amazon.in

Firas D. Ahmed, Aws Naser Jaber, Mazlina Binti Abdul Majid, & Mohd Sharifuddin Ahmad. Agent based Big Data Analytics in retailing: A Case Study. *2015 4th International Conference Software Engineering and Computer Systems.* https://doi.org.10.1109/ICSECS.2015.7333085

*Flipkart official website*. Retrieved from https://www.flipkart.com

J. McAuley, C. Targett, J. Shi, & A. van den Hengel. Image-based recommendations on styles and substitutes. *SIGIR, 2015*

*Kaggle E-Commerce Data,* Retrieved from https://www.kaggle.com/carrie1/ecommerce-data

Konstantinos Xylogiannopoulos, Panagiotis Karampelas, & Reda Alhajj. Clickstream Analytics: An Experimental Analysis of the Amazon Users’ Simulated Monthly Traffic. *2018 IEEE/ACM International on Advances in Social Networks Analysis and Mining.* https://doi.org.10.1109/ICSECS.2015.7333085

*Matplotlib Plotting Library*. Retrieved from https://www.matplotlib.org.

Mohd Razilan Abdul Kadir, Fatimah Siddiqah Mohammed Nor, & Fatimah Alma Saaid. A study on customer online purchasing behavior : A Malaysian Perpsective. *2017 International Conference on Information Technology.* https://doi.org.10.1109/ICSECS.2015.7333085

Nabil Tahmidul Karim, Sanjana Jain, Jednipat Moonrinta, & Mathew N. Dailey, Mongkol Ekpanyapong. Customer and target individual face analysis for retail analytics. *2018 International Workshop on Advanced Image Technology (IWAIT).* https://doi.org.10.1109/IWAIT.2018.8369732

R. He & J. McAuley. Modeling the visual evolution of fashion trends with one-class collaborative filtering. *WWW, 2016.*

Tao Yang, Xiaoyi Man, Jiao Feng, Jie Chen, & Ran Tao. Research on Color Expansion Method in Clothing Recommendation System. *2018 Second International Conference on Data Science and Business Analytics.* https://doi.org.10.1109/ICDSBA.2018.00096

Yilong Shi, Li Jin, Jianfeng Chen, & Ke Zeng. Recommendation of item Category based on Random walk model. *2019 IEEE 4th International Conference on Big Data Analytics.* https://doi.org.10.1109/ICBDA.2019.8713258

*Seaborn official website*. Retrieved from <https://seaborn.pydata.org/introduction.html>

*Scikit-Learn documentation on Label Encoder*. Retrieved from https://scikit-learn.org/stable/modules/generated/ sklearn.preprocessing.LabelEncoder.html

**Authors Autobiography**

1. **Rahul Subramaniam**

I am a Fourth year Engineering student at R.V College of Engineering, currently pursuing my Bachelors of Education in Information Science. As a person, I believe myself to be hardworking, yet jovial by nature, I have the ability to leverage on my capabilities when the time comes. A self motivated and determined individual. With immense potential and interests, I tend to focus on my all round development. Good interpersonal skills, added with my keen interest in upcoming technologies, helps me collaborate and come up with new ideas. My sharp analytical skills give me the additional abilities to excel in my line of work.

1. **Kunal Bhandari**

A Final Year Engineering Student from R.V. College of Engineering, Bangalore, I am a cricket enthusiast. As I cherished the game with time I learnt that this game wasn't only performing the talent out but involved with a lot of analysis of about conditions, player's strengths and weakness , team potential and many other attributes . This is how my inclination towards analysis of data started. Interestingly even astrology, which is factually considered to be luck based has a lot of science behind it. This became the reason to Bachelor's degree focusing in Computer and Information Sciences. I am former software intern at Samsung Research Institute, Bangalore , mainly worked on On-Device Artificial Intelligence. This research paper Analysis of retail store is mainly of one the projects I worked on to bring out business intelligence and the value of big data and its value.

1. **Arnab Jana**

A Final year Information Science Student of RV College of Engineering Bangalore, I am really passionate about Computers, Problem Solving and Mathematics. I worked hard to get into Information Science and Engineering, after which, I got acquainted to lot of fun activities, competitions and hackathons. I participated in many hackathons and even won a few. I even used to code on Online judges like Codechef, Codeforces and participate in Online Coding Competitions organized at our college. Every new problem which I solved taught me a new concept. This was my driving force to explore new things and try out different ideas. This Research Paper Retail Store Analysis was our project to explore the field of Data Science and how it’s impacting the general trends of any multinational Online Retail Company.