National College of Ireland

Project Submission Sheet – 2020/2021

School of Computing

Student Name:	Tej Rup Sai Munagala
Student ID:	19196628
Programme:	Msc Data Analytics **Year:2020**
Module:	Domain Application of Predictive Analysis
Lecturer:	Vikas Sahni
Submission Due Date:	28/06/20
Project Title:	**A predictive analytics of customer review data.**
Word Count:	2519

I hereby certify that the information contained in this (my submission) is information pertaining to research I conducted for this project. All information other than my own contribution will be fully referenced and listed in the relevant bibliography section at the rear of the project.

ALL internet material must be referenced in the bibliography section. Students are encouraged to use the Harvard Referencing Standard supplied by the Library. To use other author's written or electronic work is illegal (plagiarism) and may result in disciplinary action. Students may be required to undergo a viva (oral examination) if there is suspicion about the validity of their submitted work.

Signature: Tej Rup Sai Munagala

Date: 27/06/2020

PLEASE READ THE FOLLOWING INSTRUCTIONS:

1.Please attach a completed copy of this sheet to each project (including multiple copies).
2.**You must ensure that you retain a HARD COPY of ALL projects**, both for your own reference and in case a project is lost or mislaid. It is not sufficient to keep a copy on computer.

Office Use Only	
Signature:	
Date:	
Penalty Applied (if applicable):	

A PREDICTIVE ANALYTICS OF CUSTOMER REVIEW DATA

Tej Rup Sai Munagla
Master of Science in Data Analytics
National College of Ireland
Dublin, Ireland
x19196628@student.ncirl.ie

***Abstract*— This paper discusses about how predictive algorithms played an important role in domain application of predictive analytics. The dataset selected for this research is "Women's E-commerce clothing reviews". This information about the customers reviews were used to classify the type of customer and recommend the appropriate products according to customer behavior. Ethical concerns are discussed about how to deal with the consumer data and some preliminary steps to inspect the flow of information within a supply chain. Preprocessing techniques like resampling, cleaning unnecessary data and transformations are applied by using exploratory data analysis (EDA). Several machine learning models are used for sentimental analysis of the customer review and few recommendation algorithms like naïve Bayes, Support vector machine (SVM), random forest used to recommend products to the customers.**

Keywords*— *Predictive algorithms, Analytics, E-commerce, classification, recommendation, Ethics, Preprocess techniques, EDA, sentiment analysis, SVM, naïve Bayes, Random Forest.

I. INTRODUCTION

The evolution of technology has brought many changes in the world, it brought a revolutionary change in business by completely virtualization of total business process [1]. This evolution led to development of E-commerce where it provides cutting edge technology to many businesses. The term E-commerce stands for Electronic commerce and it's defined as 'all electronically mediated information exchange between an organization and its external stakeholders' [2].

E-commerce has changed many elements of business environment and the operations between the organization and people. Online shopping is one of the advantageous outcomes out of E-commerce [3], it is a process where the consumers buy products or services from the internet. It is said that there will be a skyrocket in online expenditure about 4 trillion mark in 2020 around the world [4]. In Europe, the digital marketing by using AI has increased the sales from 15% in 2014 to 19% in 2017 [5].

Currently, AI has a prominent role in most of the online shopping website. It generates lot of revenue for the E-commerce business by learning the online activities of consumers, product ratings and product sales. AI can reduce the overstock and out-of-stock for the retailers, Retalon stated that 'AI can lower the cost up to 25-40 percent on Inventory Management and increases the sales turnover by 350 percent' [6].

Presently, the most influential factors that leading for the online purchases are by the product reviews and product recommendations. Online reviews have a prominent role in making a consumer decision whether to purchase a certain product or not. Positive reviews on the websites are like a promotional material that lead to increase in sales [7]. Secondly, recommendation systems suggest the best product to the consumers based on their preferences and helps the E-business to filter out the customers based on their accurate estimation approach [8].

There are many emerging machine learning algorithms that are being used for the text analysis and for the recommendations. Algorithms for sentiment classification based on logistic regression, Support Vector Machine, Neural Network, and the recommendation algorithm like naïve Bayesian (NB), Pearson r correlation.

For this research I have used a data related to women's E-commerce clothing reviews was acquired from Kaggle repository which has 11 columns and 23468 attributes.

II. OBJECTIVE

The main goal for this project is to analyze the various customers and product behaviors. For this predictive model are being used to achieve the following problems:

- To discover the sentiment analysis of various customers reviews on the website
- To recommend the products based on the customer preferences.

This project is all about recommending products to the customers and to find the sentiments of the customer reviews.

III. ETHICAL CONCERNS

The women's clothing E-commerce reviews dataset has been used for this project. As this is a real time customer dataset all the customer names are anonymized with certain number i.e. most of the customer names are replaced with some random digits.

The dataset also is also included with company details, so they are replaced by the type of department name and with division of which product it comes under. The dataset is acquired for the research purpose from an open source Kaggle repository. Hence, it is not violating any ethical data laws.

As the dataset is all about the consumer it has many ethical complications about the data is being shared in the supply chain market. All the customers details are being processed through different stages of supply chains, there would be the risk of misuse of the data.

On top on that online retailer uses the machine learning algorithms to predict various activities of the customers like recommending certain type of product, classifying the type of customers etc. however, there are many misuses of these predictions like based on a pregnant teenager purchase history

and sent a congratulatory letter to their parents that they aren't aware of it [9].

Access to the activities of consumer review data and categorizing the individuals based on their sentiment analysis like alcoholic can be disrespectful for objectifying them as a slender category [9]. Moreover, we have no clue which measures are considered for categorizing those areas? It appears that these answers are yet less detailed.

Consumer datasets contains lot of sensitive information like card details, personal details, and lot of sensitive information. Firms should always inspect the flow of information within the supply chain by giving access to verified members, confirmed trusted recipients and unverified recipients etc. [9]

Predictive algorithms should always follow inclusivity and stakeholder awareness related to the potential ethical risks, issues while designing an AI algorithm. Having bias while designing the algorithm causes lot of problems, so the people behind the predictive models should always be cautious while designing them [10].

IV. Planned Strategy

To achieve a profitable revenue out of the competitive world of E-commerce, Online E-business need to give more importance in many aspects of business to consumer (B2C). Among them the most three important factors are Human-computer interaction (HCI), behavioral, and consumerist orientations [11].

Fig. 1 – Business model.

Human-computer interaction (HCI): E-business should primarily focus to create a user interface that is easy, efficient, and pleasant to use. Factors such as visual attractiveness of website, web design, ease of navigation, information content are the most influencing factors for a pleas able shopping behavior.

Behavioral approach: E-business must focus on the trust elements of online shopping. There should always have a strong sense of trust between the online merchant and consumer because all the transactions take place virtually. So, it is important to have enough confidence on sellers.

Consumer characteristics: Online shopping behavior also depends on the consumer characteristics such as their personality, demographics, and their profiles. The friendliness of website and customer comfort level always have a prominent effect on sales.

These are the most three influential factors that effects for an online business to develop, by investing enough time on these can bring a dramatic change in the world of E-commerce.

V. Exploratory Data Analysis (EDA)

Exploratory Data Analysis (EDA) is a process of visualizing the dataset and finding the insights out of the given data. By going through the insights, we can summarize what type of machine learning models can be applied for the given type of dataset.

EDA is the first step of pre-processing the data, following are few insights that are drawn out from the women's E-commerce dataset:

The Fig. 2 tells that the age group 10-20 gave less rating. It is obvious, teenagers generally do not care about online shopping and reviews. However, for the age group of 30-40 gave high number of 5-star rating as compared to all other age groups. This is only the group who gave a greater number of rating and reviews when compared to other age groups. Finally, when we look at age group of 70 and above did not care about online shopping.

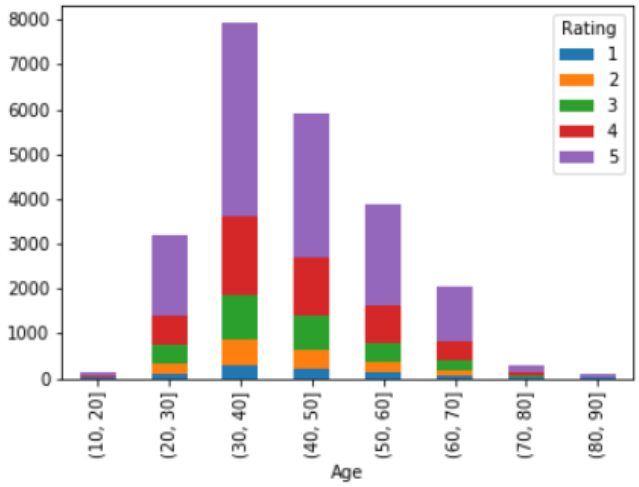

Fig. 2 – Plot between the Age group and Ratings

The Fig. 3 tells about the what departments of clothes are chosen with a specified age group. The females from 20-70 age are more active and brought numerous clothes from online. By looking at the plot we can conclude that females are more focused on tops and dresses but not much interested on bottoms. Most people are not concerned about the trend department.

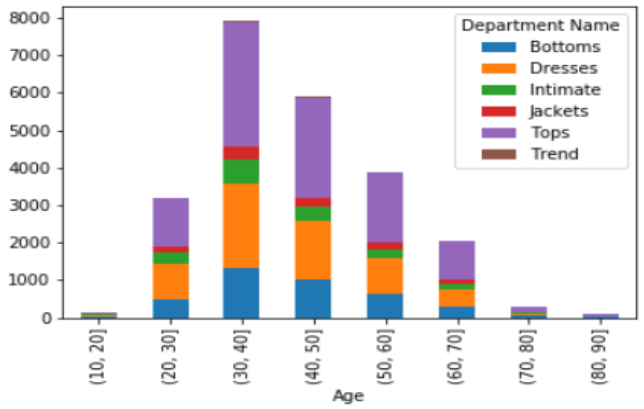

Fig. 3 – Plot between the Age group and Department type.

The Fig. 4 tells about the 3 divisions which are general, general petite, and intimates. The general division products are most sold out as compared to general petite and intimates. There are around 14k products are sold in general division, 8k products in general petite division and around 1600 products are sold in intimates division. Most people recommend intimates and general petite.

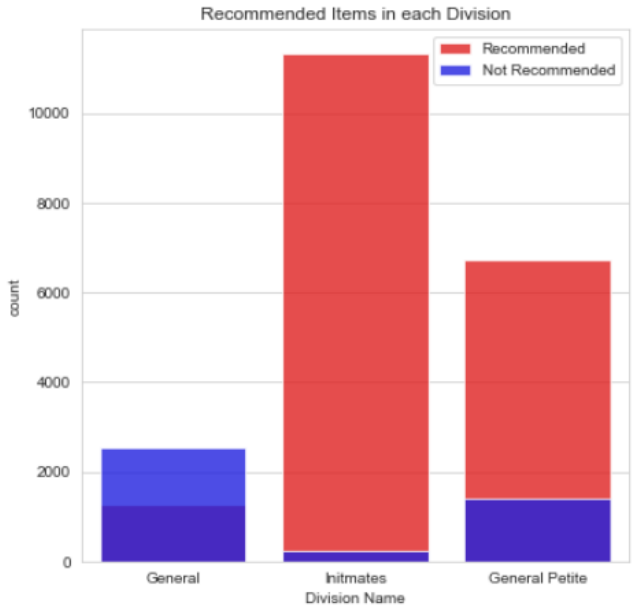

Fig. 4 – Recommended item in each division.

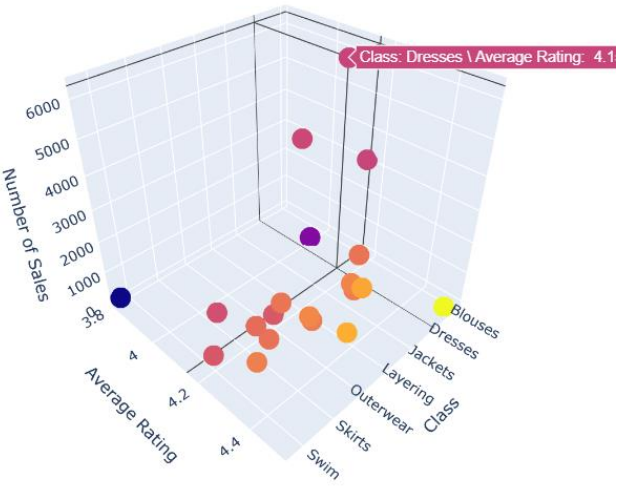

Fig. 5 – 3D plot between Average rating & classes & number of reviews.

The Fig. 6 is the plot of frequency distribution for rating, recommended ID, and label. Here vast majority of reviews are highly positive, with a score of five out of five. It suggests that most of retailers are performing well in the market. The binary classification of reviews i.e. good = 1 and bad = 0. There are a greater number of good reviews rather than bad reviews. From the label part we can suggest that most of the recommended products are having a review three and more.

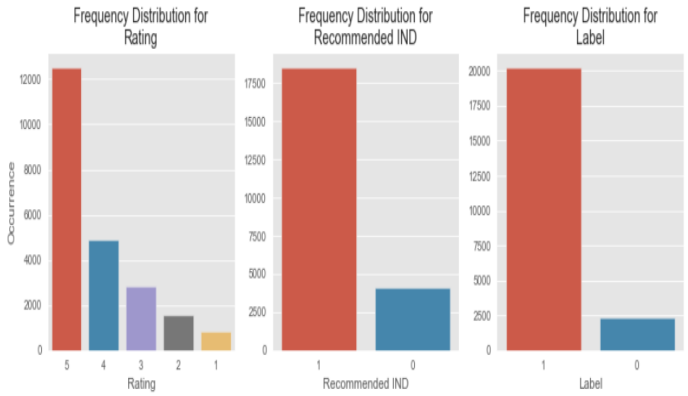

Fig. 6 – Distribution of rating, recommended IND, and labels.

The Fig. 5 is the plot between the average rating, classes, and number of classes. It is a 3-dimensional figure tells about the both the average rating and number of reviews of every class material. Among all dresses, knits and blouses are the clothes that have high review rating and number of reviews. On the other hand, swim, trend, and sleep clothes have fewer selling products with a smaller number of reviews and ratings.

The Fig 7 give details about what types of ratings are given to all clothes in the department type and in the division names. We can see most of the products have 5-star ratings, among all jackets and intimates are the highly rated clothes under department session. In the division session intimates are highly rated and all three product divisions are close by similarly rated.

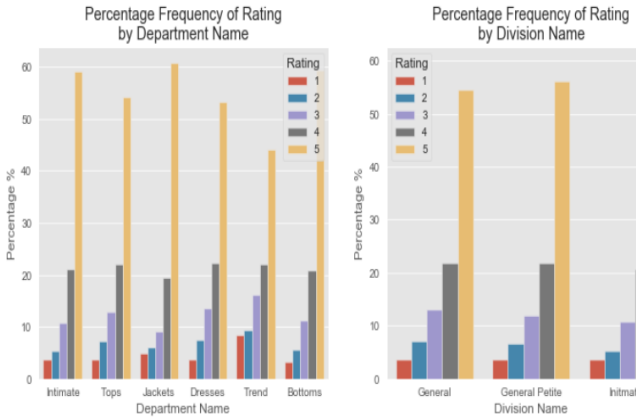

Fig. 7 – Rating by department and division name.

The Fig 8 is about the recommendations by ratings, the plot talks about the conflicting interest between the customers personal interaction with the products. Mostly the products with low rating are not recommended by the business. However, the products with high ratings are not been shown in the recommendation section because of its high number of views. Mostly the products recommended by the business are the 3-star rated products as the business wants to push their stock and shed light on these categories of products as well.

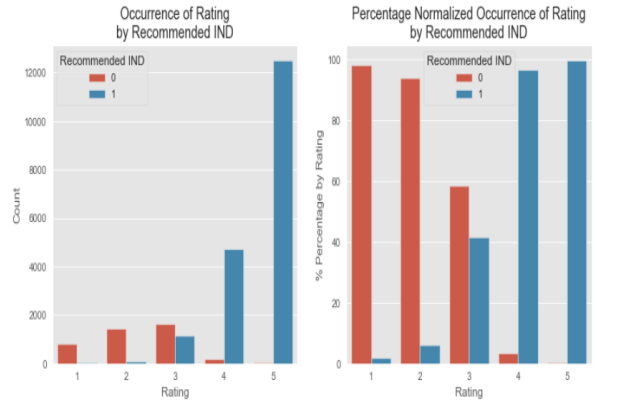

Fig. 8 – Rating by recommendations.

The Fig. 9 tells about the average length of reviews that are given for all the products in the department section. Jackets and dresses have more length of reviews as those are most selling products and people are reviewing them most.

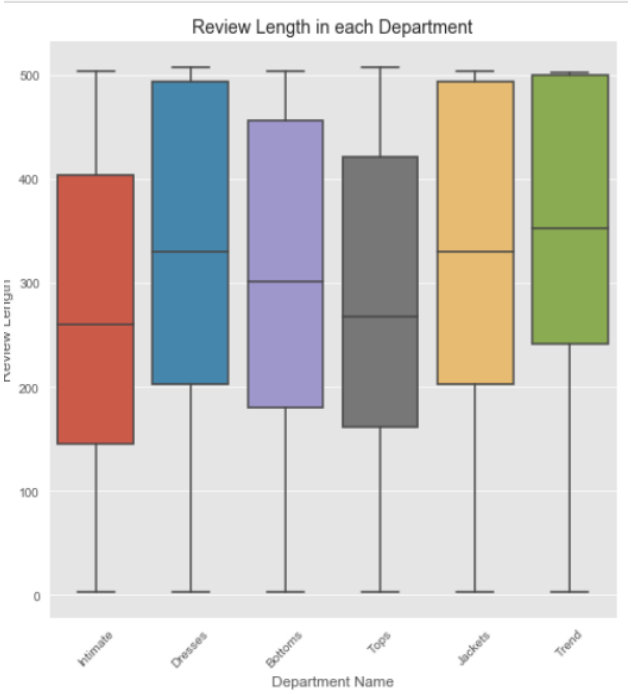

Fig. 9 – Review length in each department.

The Fig. 10 this is also about the review length of different products in the class variable. here most of the variables have the average same length of reviews i.e. around 250 letters of words. Among all dresses, out ware and trend have more length of words.

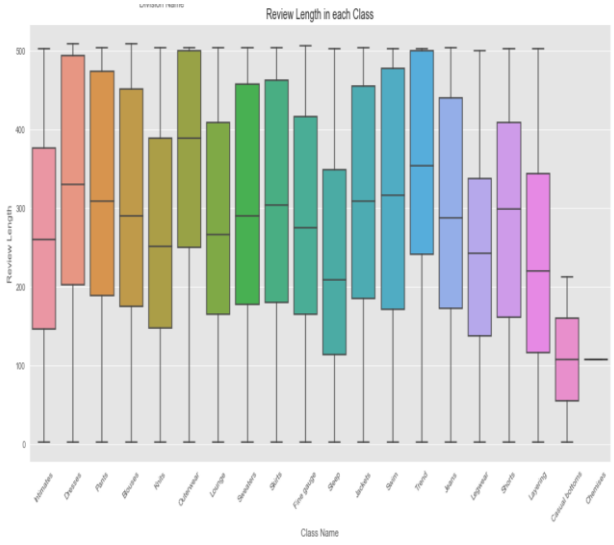

Fig. 10 – Review length in class name.

The Fig 11 is combination of the both the recommendations that are done by the business and reviews to all the categories of products that are given. Most of the products that reviewed with high ratings are not recommended, the plot show the average reviewed products like trend, bottoms are recommended most for the customers to up sale their products.

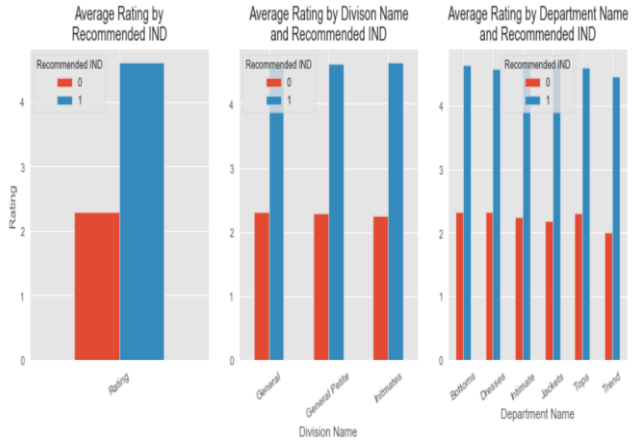

Fig. 11 – Average rating and recommendations.

VI. Applicable Techniques

The main objective of this project is to do the sentiment analysis of the customer reviews of the women's E-commerce website and to recommend the products based on sentiments of the products. So, the target variables in the dataset are reviews and product category. Following are the steps involved in preprocessing the dataset and applying the predictive models accordingly:

- Initially check the format of all data types and then convert the datatypes of few variables for the preprocessing.
- Check for the null values in all the variables, if there are null values replace them with appropriate technique.
- According to the type of the algorithms that is being applied do the log transformations on the target variables.
- Rebalance the dataset.
- Check for the multi collinearity in the dataset, if there is higher collinearity between the variables remove them as these variables effects the model.
- Remove the variables that are redundant these variables will affect overall performance of the model.

To evaluate the precision and accuracy of the model, few techniques are used to check the overall performance of the model:

- Confusion matrix
- Classification accuracy
- F1 score

Most of these steps will follow above order during the preprocessing, however some of the steps are repeated and again due to unexpected constraints.

VII. References

[1]R. Singh and J. Kaur, "Service Quality of Online Shopping Portals: A Review of Literature," Journal of General Management Research, vol. 5, no. 1, pp. 45–63, Jan. 2018.

[2]V. Kaushik, A. Khare, R. Boardman, and M. B. Cano, "Why do online retailers succeed? The identification and prioritization of success factors for Indian fashion retailers," Electronic Commerce Research and Applications, vol. 39, p. 100906, Jan. 2020, doi: 10.1016/j.elerap.2019.100906.

[3]R. Singh and J. Kaur, "Service Quality of Online Shopping Portals: A Review of Literature," Journal of General Management Research, vol. 5, no. 1, pp. 45–63, Jan. 2018.

[4]"Online Shopping Statistics You Need to Know in 2020," OptinMonster, Nov. 06, 2019. https://optinmonster.com/online-shopping-statistics/ (accessed Jun. 26, 2020).

[5]Z. Ruan and K. Siau, "Digital Marketing in the Artificial Intelligence and Machine Learning Age," AMCIS 2019 Proceedings, Aug. 2019, [Online]. Available: https://aisel.aisnet.org/amcis2019/treo/treos/96.

[6]L. Parker, "The AI Revolution is Now: Artificial Intelligence (AI) is capturing and scanning reams of sales data for predictive analytics, tracking consumer behaviour and changing retail as we know it," Earnshaw's Review, vol. 104, no. 2, pp. 8–30, Feb. 2020.

[7]"Featuring Mistakes: The Persuasive Impact of Purchase Mistakes in Online Re...: Discovery Service for NCI Library..." http://eds.a.ebscohost.com/eds/pdfviewer/pdfviewer?vid=0&sid=1a96b241-beed-4fbf-aa46-25ddb6a030f3%40sessionmgr4008 (accessed Jun. 26, 2020).

[8]S. H. Choi, S. Kang, and Y. J. Jeon, "Personalized recommendation system based on product specification values," Expert Systems with Applications, vol. 31, no. 3, pp. 607–616, Oct. 2006, doi: 10.1016/j.eswa.2005.09.074.

[9]K. E. Martin, "Ethical Issues in the Big Data Industry," in Strategic Information Management, 5th ed., R. D. Galliers, D. E. Leidner, and B. Simeonova, Eds. Routledge, 2020, pp. 450–471.

[10]D. J. Yates, G. J. J. Gulati, and J. W. Weiss, "Understanding the Impact of Policy, Regulation and Governance on Mobile Broadband Diffusion," in 2013 46th Hawaii InternationalConference on System Sciences, Wailea, HI, USA, Jan. 2013, pp. 2852–2861, doi: 10.1109/HICSS.2013.583.

[11]"Examining Online Purchase Intentions in B2C E-Commerce: Testing an Integrated Model - ProQuest." https://search.proquest.com/docview/215884472/fulltext/805AEFE7A234D03PQ/1?accountid=103381 (accessed Jun. 27, 2020)