E-commerce Product Recommendation based on User Ratings and Reviews

K L C D Senarath

2021

E-commerce Product Recommendation based on User Ratings and Reviews

A dissertation submitted for the Degree of Master of Business Analytics

K L C D Senarath

University of Colombo School of Computing

2021

Declaration

The thesis is my original work and has not been submitted previously for a degree at this or any other university/institute.

To the best of my knowledge it does not contain any material published or written by another person, except as acknowledged in the text.

Student Name: K.L.C. D Senarath

Registration Number:2018/BA/031

Index Number: 18880315

K.L.C.D Senarath

_____

Signature: Date: 20/09/2021

This is to certify that this thesis is based on the work of Miss K.L.C.D Senarath under my supervision. The thesis has been prepared according to the format stipulated and is of acceptable standard.

Certified by:

Supervisor Name: Dr.L.N.C De Silva

_____

Signature: Date:20/09/2021

I would like to dedicate this thesis to...

ACKNOWLEDGEMENTS

I would like to express special thanks & gratitude to my supervisor, Dr. L.N.C De Silva who gave me a lot of ideas and help to work on this project on the topic of "E-commerce Product Recommendation Based on User Ratings and Reviews", which led me into doing a lot of Research which diversified my knowledge to a huge extent for which I am thankful.

Also, I would like to thank my parents who supported me a lot in finalizing this project within the limited time frame.

K.L.C.D Senarath

2018/BA/031

ABSTRACT

The fast-growing retail industry amidst the current pandemic is eCommerce. It is becoming popular due to emerging information technologies, rapid growth, and the high adapting rate of e-banking services. In addition, benefits such as low cost, flexibility, speed in the buying process, availability of comprehensive descriptions of products, convenience due to no geographical limitations, availability of reviews, etc., are a few psychological factors that impact the consumer behavior towards eCommerce. Furthermore, millions of products are available in retail e-commerce, and users post reviews every minute. Therefore, customer reviews and ratings are crucial factors nowadays, and it affects the customer's buying behavior.

Those reviews ultimately increase sales by giving the consumers the information they need to choose to buy the product. People are always more eager to purchase products that others have already recommended. However, due to a large number of products and customer reviews available, it has become tedious to understand the actual quality of the product. Therefore, it is tough to make a goodchoice whether to buy the products. Hence, it is vital to analyze the customer ratings, reviews, and recommendations to assist consumers' decision-making process.

In order to derive valuable insights from a large set of reviews with ratings, this study has been conducted using four supervised machine learning techniques, Linear SVC, Decision Tree, K-Nearest Neighbor(KNN), and Naive Bayes on fashion products from Amazon. Exploratory data analysis was applied to the dataset containing more than 10,000 records and 16 attributes for feature selection and handling missing values. Customer reviews were pre-processed using natural language processing techniques and classified data into three classes called Good, Moderate, and Not Recommended product based on rating score and reviews. The pre-processed data set is divided into training and testing, and the modelwas trained using different algorithms. The detailed output is generated using a confusion matrix and a classification report. The accuracies and prediction time have then been compared to identify the best fit.

The results showed that the Linear SVC approach performed better than other algorithms. Accuracy, precision, recall,F1-score, and confusion matrix are used as performance measures. The KNN algorithm was applied against different k values and observed that the K-Nearest Neighbor

algorithm classifier was further improved when the value of k was increased. Several statistical analyses such as Pearson correlation, OLS regression, etc. was carried out on the attributes, rating, number of reviews, and price to identify their relationships and impact. Based on the study, most popular product categories, manufacturers, and products were identified. Hence, these statistics and findings would help sellers, marketers to make betterdecisions to improve their revenue.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

CHAPTER 1

INTRODUCTION

1.1 Overview

The fast-growing retail industry amidst the current pandemic is eCommerce. Electronic commerce is defined as buying and selling goods and services over the internet. Due to the massive usage of the internet, marketers, and customers are highly influenced through eCommerce, and it is yet another way of boosting business practices. Mitra Abhijit (2013) suggests that e-Commerce has unleashed another revolution, changing how businesses buy and sell products and services (Mitra, 2013). In addition to that, with the current COVID pandemic, people tend to use eCommerce to get their daily needs done. As a result, businesses tend to sell their products through eCommerce, which is a trend for almost every product. Business-to-Business (B2B), Business-to-Consumer (B2C), Business-to-Government (B2G), Consumer-to-Consumer (C2C), and Mobile Commerce (m-commerce) are the main types of e-commerce.

Benefits such as low cost, flexibility, speed in the buying process, availability of comprehensive description of products, convenience due to no geographical limitations, availability of reviews, etc., are a few psychological factors that impact the behavior of consumers towards eCommerce. On the other end, the producers can reap many benefits due to the ability to analyze consumer behavior that is not possible in physical stores. However, perceived risks such as financial loss due to product performance, quality, delivery issues, and psychological factors like trust and security are a few concerns in eCommerce (Moshref et al., 2012). Therefore, these risks and psychological factors play a more significant impact on determining consumer behavior towards eCommerce. In that regard, customer reviews and ratings are crucial as the customers look into these in prior buying through these sites (Hinckley, 2015).

Customer reviews and ratings are very crucial factors affecting online shopping (Hinckley, 2015). Every day millions of reviews are posted for different products. Those reviews ultimately increase sales by giving the consumers the information, they need to choose to buy the product. People are always more eager to purchase products that others have already recommended. At present, retail e-commerce sales are increasing (statistica, 2021), and many new products are introduced daily. Therefore, customers have to rely on product reviews to make up their minds to make better purchase decisions. So it is vital to analyze ratings and reviews. It can also help businesses extend

sales, and improve the product by understanding customer needs, reducing the risks due to financial losses.

Every day millions of reviews are posted by the buyers of the product. Positive reviews give the consumer confidence in buying the relevant product. People are always more eager to purchase products that others have already recommended. At present, retail e-commerce sales are increasing (statistica, 2021), and as a result, many new businesses emerge, introducing various products daily into the online market. Hence, customers rely on product reviews to make up their minds to make better purchase decisions. Therefore, it is vital to analyze ratings and reviews to better options for consumers to make actionable decisions. It can also help businesses build trust in consumers, thereby improving product quality and extending sales with a better understanding of customer reviews. This, in turn, will cater towards reducing the risks of business failures and financial losses.

Lack of customer service is one limitation in retail e-commerce. Hence, assisting the customers in finding the right product is a significant issue. Analyzing the product reviews posted by thousands of consumers on a specific product can be helpful in one way to support customer service. The proposed study was conducted to analyze the feasibility of recommending the products for consumers and future buyers. This was done by analyzing online reviews, ratings, prices, and other features important in making consumer decisions when purchasing the products. In return, the business can be increasing online sales by attracting more customers. For this study, Amazon, one of the most popular e-commerce sites that people use every day for online purchases, is being used.

1.2 Motivation

At present, product reviews and ratings play an essential role in the growth of e-commerce. Consumers post positive as well as negative reviews and rate the products accordingly. A customer buying a specific product usually goes through consumer reviews before purchasing the product. However, going through the massive amount of consumer reviews is not an easy task. Customer reviews and ratings are also essential to identify the market reactions to a specific product. As e-commerce continues to grow, businesses should survive in the new economic world by properly balancing business practices and information technologies.

There are several factors behind the rapid growth of eCommerce, such as the increase of dynamic business relations, the elimination of price differences, the need to cooperate in new product development (Bell, 2021). In addition, the increasing use of information technologies and the widespread of e-banking increased the use of e-commerce. Apart from that, the recent pandemic

has created a vast impact on eCommerce. Many, even those used only for physical stores, are currently adapting and converting to such technologies.

Although the companies like Amazon, Walmart are not involved in physical assets, their commercial activities are in the electronic environment. However, with the current context to sustain retail eCommerce, they need more competitive features to attract more consumers. Therefore, ratings and reviews have a direct impact on their businesses and increase sales. As for consumers, reviews and ratings provide recommendations and help them make decisions about products. On the other end, reviews assist online retailers by providing critical feedback about their products. In addition to that, going through the feedback, the producers can increase the product quality and other features to attract more consumers. Furthermore, they can also adjust the product prices to meet customer needs and increase their sales.

The advantages and benefits in customer feedback and identifying the feature required the retail eCommerce to sustain more consumers motivated to analyze the customer reviews and feedback further. For this study the Amazon fashion products dataset was selected due to Amazon is one of the giant retail companies available in the world and most customers around the world access this site to purchase products. So this dataset provides more accurate data to get the most accurate results. Also this dataset contains data related to amazon fashion products like toys, accessories etc. Since researchers were not much focused on this area, this study also will be help to get an idea about customer preferences related to these kind of fashion products.

1.3 Objectives

This study aims to recommend products based on customer ratings and reviews and identify other features' impact on the recommendations. Therefore, the sub-objectives below were determined to achieve the main objective stated above.

- Recommend products by analyzing customer reviews and ratings:
  Analyze customer ratings/reviews and identify proper classification to recommend products.

- Identify the most popular products, product categories, and manufacturers which directly impact customers' or marketers' decisions.

- ❖ Identify the relationship between product prices, customer reviews and ratings to analyze whether there is a correlation between product price and ratings.

1.4 Scope of the study

Fashion products in Amazon are more in demand by the customers. Because Amazon, is one of the World's top fashion retailer, is in a place to grab even more market share (Forbes, 2021). Hence, this study analyzed the fashion products dataset on Amazon to recommend products based on customer reviews and ratings. In addition to that, this study identified the best products, product categories, and manufacturers and got an idea about how product price affects the ratings/reviews. The data analysis was done using the different machine learning classification algorithms like Naïve Bayes, Linear SVC, and Decision Tree. In addition to that, several statistical techniques were used to achieve the objectives mentioned above.

Python in Spyder (anaconda distribution), one of the most popular programming languages in machine learning and data science, was used to train different machine learning algorithms. In addition, some available python libraries like NLTK, Sklearn, Pandas, and Matplot are used in this study to analyze data.

The dataset is pre-crawled to select a subset of a more extensive dataset (more than 7 million fashion products) by extracting data from Amazon. There are sixteen features in the dataset. The data set consists of more than 10000 records of fashion products and each record contains sixteen features, as mentioned in Table 1.

Table 1: Features of the dataset

Feature	Type
uniq_id	String
product_name	String
manufacturer	String
price	String
number_available_in_stock	String
number_of_reviews	Integer
number_of_answered_questions	Integer
average_review_rating	String
amazon_category_and_sub_category	String
customers_who_bought_this_item_also_bought	String
Description	String
product_information	String
items_customers_buy_after_viewing_this_item	String
customer_questions_and_answers	String
customer_reviews	String
sellers	String

1.5 Structure of the dissertation

The rest of the thesis is structured as follows: Chapter 2, the Background and related work done in this area of research. In this chapter, we analyze the information available in published material like research papers, URLs, magazine articles and similar.

Chapter 3 is about the Methodology. This is followed by the systematic approach to solve the problem. In here we explained about the environment, data preprocessing, labeling under exploratory data analysis and model building.

Chapter 4 shows the overall design of the project and detailed explanation of steps taken for classification task. Chapter 5 contains the approach and proposed implementation part of this

project describing every step taken to achieve them through code snippets. Chapter 6 is about results and discussions of the study. This chapter presents the findings and the evaluation of the research and include results obtained and critical evaluation of the research work.

Finally, the findings and possible future research work are discussed in Chapter Seven. This chapter summarizes the work, discusses its findings and contributions, points out limitations of the current work, and also outlines directions for future research.

CHAPTER 2

BACKGROUND AND RELATED WORK

This chapter presents various techniques and methods used by various researchers in review and rating classification. Past literature was analyzed in this section to identify the previous studies conducted classifying user reviews. Moreover, the possibility of using machine learning approaches to classifying reviews and ratings will also be discussed in detail.

2.1 Analysis of reviews and ratings

ECommerce is more prevalent among people in the current context. Buying and selling nature through eCommerce has shown rapid growth during the last few years. According to the literature, based on the studies conducted to analyze the online product buying behavior of consumers indicated that customer reviews have a significant impact on customer reviews. Nowadays, to meet the competition, eCommerce websites allow customers to leave their opinions on various aspects. Most people who buy products through eCommerce tend to leave customer reviews and comments by all means. The product quality, delivery time, merchant's nature of the response, etc., are a few different types of comments left by the users.

On the other end, online customer reviews create new avenues in marketing as well as in communication. In the current context, everyone is eager to read online customer reviews. These reviews can drastically impact the people who search and buy the products. Studies show that 91% of people read customer reviews, and around 84% trust these customer reviews (Inc.com,2021). In addition, customers always search for more reviews to rely on before buying a product. Another study found that customers expect at least 40 reviews to rely on the star rating (15 Online Review Stats Every Marketer Should Know, 2021). Therefore, online customer reviews and ratings are essential for retail eCommerce to grow in different aspects. Online customer reviews and ratings act for customers as a way of trust-building and to aid consumer decision-making. On the other end, the retailers can use those reviews to help drive sales, build trust between customers, enable problem-solving, and aid consumer decision-making.

Millions of review comments are being generated day by day. Therefore, it is difficult for retailers and product manufacturers to go through all these customer reviews of their products. Similarly, customers find it challenging to filter out the reviews when they are going to buy products. Thus,

it is crucial to analyze customer reviews, ratings, and features that affect reviews and ratings (E.g.: price) to identify valuable information from a large data set.

Machine learning and classification techniques have been heavily used in analyzing customer reviews while categorizing them into several groups or classes. For example, sentiment classification has been used in various fields to analyze movie reviews, travel destination reviews, and product reviews (Ye et al. 2009).

2.2 Related Work

This section presents several research studies conducted in analyzing user comments, ratings, and other associated factors in decision making. Due to the importance of online customer reviews, analysis of reviews and ratings has gained much attention recently.Before purchasing an online product or service, consumer predicts different types of perceived risks such as financial risk (loss of money), product risk (quality of the product as seen on the website), and non-delivery risk (Moshref et al. 2012). Amidst the risk, this study shows the possibility of searching for products and information 24 hours a day, and the availability of a wide selection of products are the main benefits. Hence the popularity of online shopping businesses increases every year (Ariff et al. 2013). Several other factors such as convenience, ease of use, low cost, time-saving, availability of various online products and brands compared to physical shops (Adnan, 2014) play a significant role in the popularity of e-commerce websites. According to Yoruk, Online shopping is the third most common use of the internet after web surfing and email use (Yoruk et al. 2011).

A survey (Hinckley, 2015) showed that 67.7% of consumers are effectively influenced by online reviews when making their purchase decisions. Also, another study says that looking and comparing text reviews can be frustrating for users as they feel submerged with information (Ganu, Elhada & Marian, 2009). Reviews on Amazon are provided not only for the product but also for the customer services. If users get clear bifurcation about product reviews and service reviews, it will be easier to decide on products (Aashutosh Bhatt et al.2015). Several techniques ranging from simple neural networks to more complex machine learning techniques have been used to classify customer reviews.

Using SVM and Naive Bayes classifiers, Pang, Lee, and Vaithyanathan (2002) tried to classify movie reviews into two classes, positive and negative. In terms of accuracy, all techniques showed

quite good results. In a recent survey that was conducted by Ye et al. (2009), three supervised machine learning algorithms, Naive Bayes, SVM, and N-gram model, have been used to analyze the online reviews gathered on different travel destinations in the world. In this research work, they found that well-trained machine learning algorithms perform very well for the classification of travel destination reviews in terms of accuracy. The authors J. Liu et al. used decision trees to classify high or low informative opinion phrases extracted from restaurant reviews (Liu, Seneff and Zue, 2012).

Many research works show that Naıve Bayes and SVM classifiers are the two most used approaches in classifying customer reviews (Joachims 1998; Pang et al. 2002; Ye et al. 2009**).** A survey study done by Karrupusamy, P. et al. (2020) classified the machine learning techniques used for analyzing the customer reviews into three categories; opinion mining/sentiment analysis for review helpfulness and rating prediction, Bias, spam and fake review detection, and collaborative filtering for recommendation system. According to that study, most of the existing approaches used supervised learning to classify customer reviews. Naïve Bayesian classifier is used heavily, and in some studies, performance issues were reported due to the large volume of data. In that regard, deep learning techniques can significantly impact the accuracy and performance of mining customer reviews.

The proposed study was conducted on the customer reviews for fashion products of the Amazon website. This dataset contains fashion products such as accessories, toys, etc., and hence the researchers provided not much attention. However, manufacturers of such products need to think of different ways to increase sales and motivate consumers to buy such products in the current context. Hence, the fashion products dataset of the Amazon website was selected in this study, and several machine learning techniques such as Naïve Bayesian classifier, Decision Tree, K nearest neighbor, and Linear SVC were used on the large dataset. The labeled data used in this study can accelerate many machine learning techniques to identify useful information from the dataset.

2.3 Natural Language Processing

Natural language processing is the most widely used area of computer science in which machine learning and computational linguistics are broadly used (Jain, Kulkarni and Shah, 2018). it is used in fields like machine translation speech recognition and text processing. The most of user generated text is in unstructured form. This huge amount of unstructured data has led to the creation of a collection of methods for computers to process content and understand text

(Veluchamy, Nguyen, L. Diop and Iqbal, 2021). This collection is known to natural language processing. In this study, we used Python libraries for processed and analyze textual data, such as Natural Language Toolkit (NLTK) which provides more than 50 collections of text and lexical resources and many necessary methods.

2.4 Machine Learning

Machine Learning is the process where machines can learn hidden patterns/trends from data and make predictions. Nowadays Machine Learning is widely used for various types of real time use cases. Classification and Regression are main tasks that are done in machine learning. Classification is a task where predictive models are trained to classify data into different classes. Regression is a task where models are built to predict continuous variables. Collecting data, prepare data, model selection, train machine model, evaluation and prediction are the steps involve for machine learning model development as shown in Figure 1.

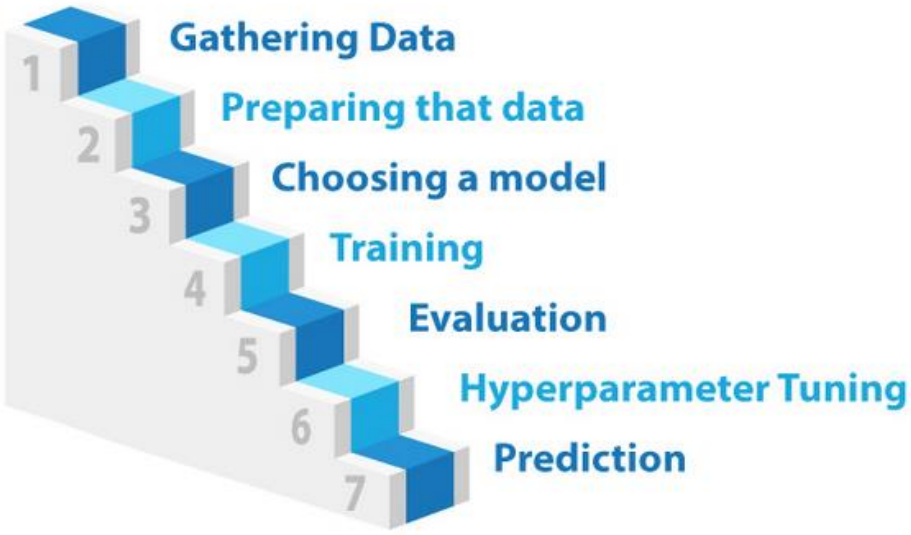

Figure 1: Steps for machine learning model development

This study is mainly focused on classification task. Because classification is the most widely used machine learning technique for classify reviews and ratings. This analysis mainly involves two steps. The first step is to learned the patterns from the training data, and the second step is to use

the learned pattern to identify new data (Jayashri Khairnar & Mayura Kinikar 2013). There are three main categories of machine learning.

a) Supervised Learning: In supervised learning training data should be labeled. If the classifier gets more labeled data, the accuracy of the result will be high. The aim of this method is that algorithm can correctly predict the output of new input data. If the system is provided with unlabeled data; it can either offer a false positive or a false, negative (Padraig Cunningham et al.2008).

b) Unsupervised Learning: In unsupervised learning, the model is trained with unlabeled data. This implies that common trends will be identified in the data to assess the performance without the right answers. Unsupervised learning challenges require clustering, one of the most significant approaches.

c) Semi-supervised Learning: In this type of learning, the dataset contains both labeled and unlabeled data. It has the benefit of both supervised and unsupervised learning. This is helpful because it is possible to gather data quickly, but labeling can take time and cost. The idea is to work on sparse datasets with the same accuracy of supervised learning techniques (Sebastian 2014).

Naïve Bayes, Linear SVC, Decision Tree and K-Nearest Neighbor are some of available classification algorithms used for this study.

2.4.1 Naïve Bayes

Naive Bayes is one of the powerful machine learning techniques. It is a simple classification algorithm. This classifier is based on Bayes theorem and relies on the assumption that the features are mutually independent. In spite of the fact that this assumption is not true, Naive Bayes classifiers have proved to perform surprisingly well (Rish 2001).

Bayes Theorem:

$$P(A|B) = \frac{P(B|A)P(A)}{P(B)}$$

We can find the probability of happening A,given that B has occurred using this Bayes theorem.Here A is the hypothesis and B is the evidence.The assumption made here is that presence of one paticula feature doesn't affect the other.Therefore this is the called Naïve.

2.4.2 Decision Tree

A Decision Tree is a supervised learning algorithm that can be used for both classification and regression problems. Most of the time, it is used for solving classification problems. It is a tree-structured machine learning classifier. Internal nodes of the tree represent the features of a dataset. Branches represent the decision rules and each leaf node represents the output. There are two nodes in a decision tree, which are the Decision Node and Leaf Node. Decision nodes are used to make any decision and have multiple branches. Leaf nodes are the output of those decisions. The decisions are performed on the basis of features of the dataset. It is a graphical representation for getting all the possible outcomes to a problem based on given conditions as shown in Figure 2. It is called a decision tree because it is similar to a tree, it starts with the root node, which expands on further branches and builds a tree-like structure.

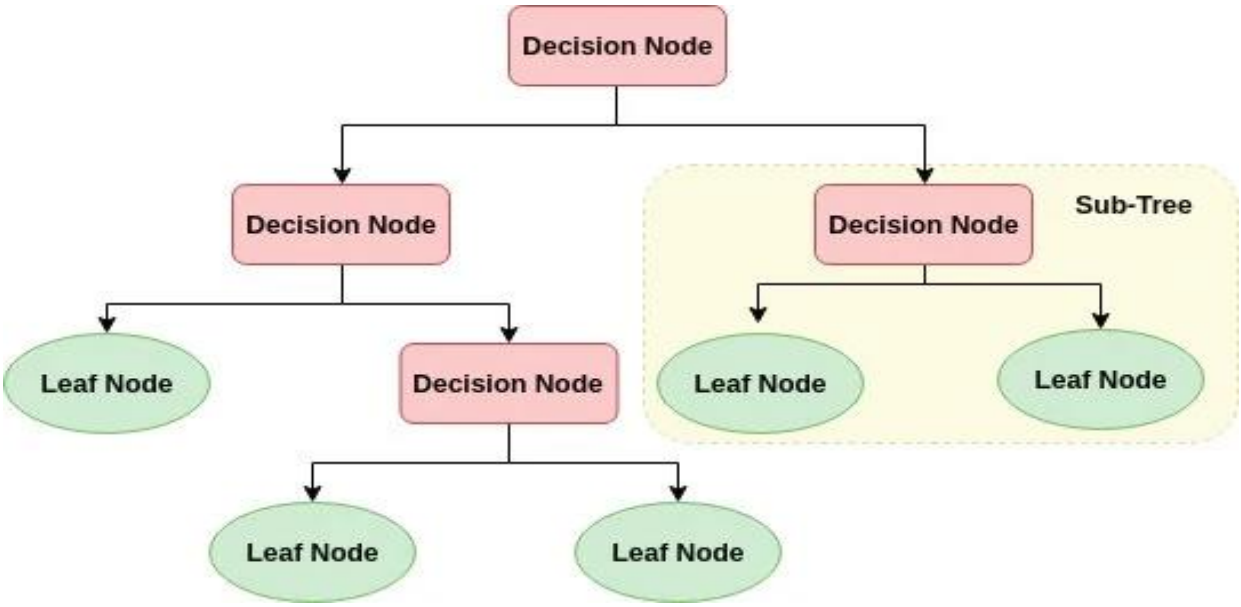

Figure 2: Decision Tree

2.4.3 Linear SVC (Support Vector Classifier)

Support Vector Classifier is a supervised learning method that can be used for classifying sentiments (Cristianini & ShaweTaylor 2000). The purpose of the Linear SVC is to fit to the data you provide and returning the best fit hyperplane that divides the data as shown in Figure 3. After getting the hyperplane, then can feed some features to the classifier to identify predicted class.

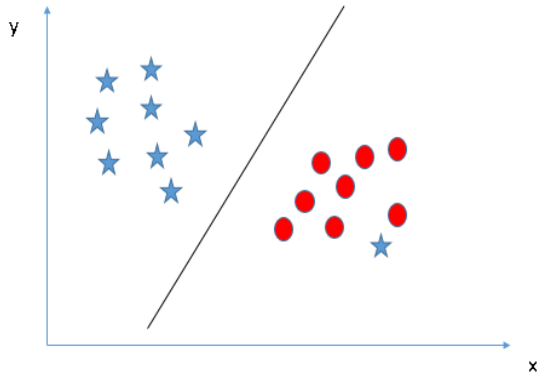

Figure 3: Hyperplane that divided the data

2.4.4 K- Nearest Neighbor (KNN)

The KNN algorithm is used to solve the problems related to classification and regression. It is based on supervise learning technique. When new data points come in, the KNN algorithm will try to predict that to the nearest of the boundary line. So that, larger k value indicates smother curves of separation resulting in less complex models (Figure 4). It's very essential to have the correct k-value when analyzing the dataset to prevent overfitting and under fitting of the dataset. There are two properties that define K-NN well.

- **Non-parametric algorithm** - That means it does not make any assumption on underlying data. This algorithm is also known as
- **Lazy learning algorithm** − KNN does not have a specialized training phase and uses all the data for training while classification.

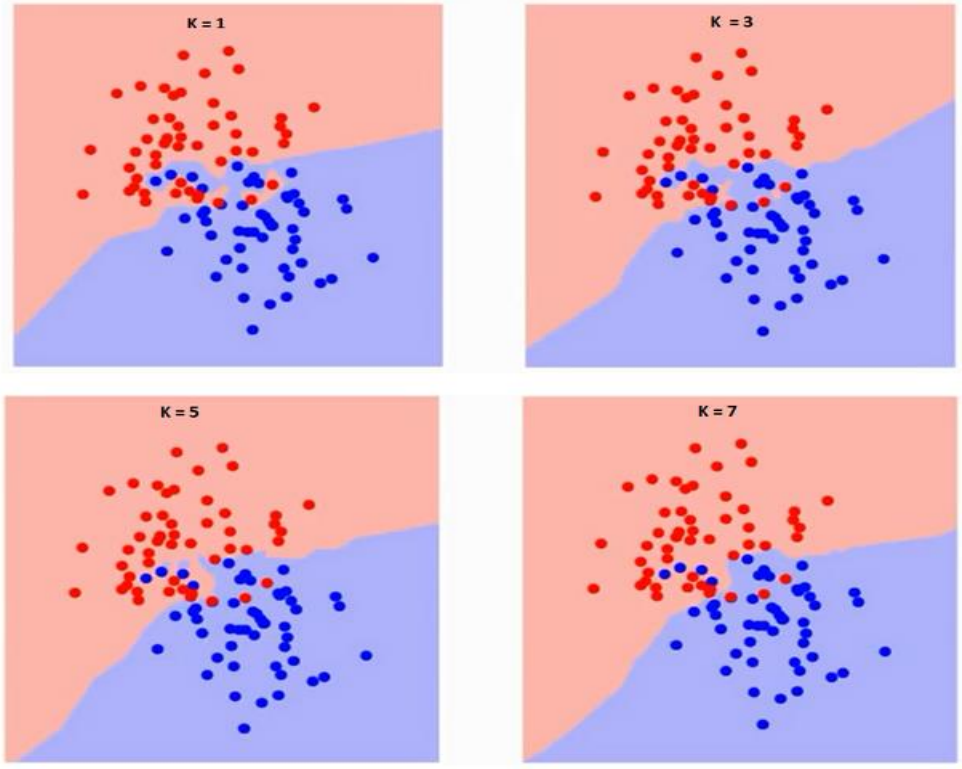

Figure 4: Separating the two classes with different values of K

2.5 Statistical approach to analysis of features

There are few statistical analysis methods also plan to use for this study to achieve above mentioned objectives.

2.5.1 Correlation Analysis

This analysis is used to identify the association between variables. The correlation coefficient (Pearson correlation), denoted by r, ranges between -1 and +1. It quantifies the direction and strength of the linear association between the two variables. The correlation between two variables can be positive or negative (Figure 5). The sign of the correlation coefficient shows the direction of the association between variables. The magnitude of the correlation coefficient indicates the strength of the association.

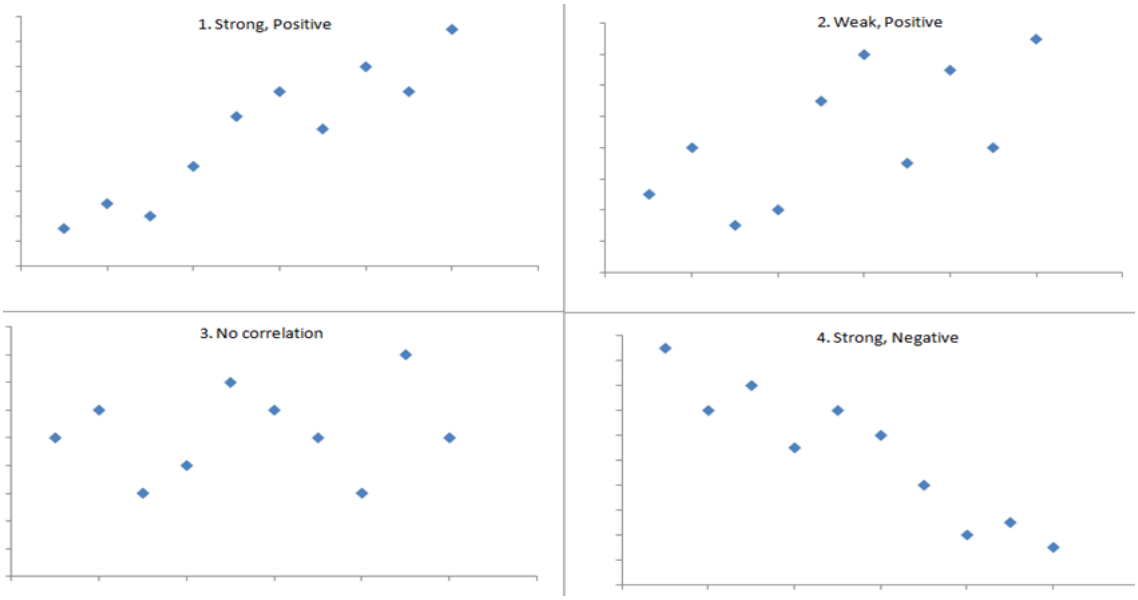

Figure 5: Correlation coefficient

2.5.2 Regression Analysis

Regression analysis is a set of statistical methods used for the estimation of relationships between a dependent variable and one or more independent variables. It can be used to assess the strength of the relationship between variables. Also modeling the future relationship between them. There are several types of regression analysis available as categorize in Figure 6.

Figure 6: Types of Regression Analysis

Ordinary least squares (OLS) regression is the one of the statistical methods of analysis. It estimates the connections between one or more independent variables and a dependent variable. The method estimates the relationship between variables by minimizing the sum of the squares in the distinction between the observed and predicted values of the dependent variable configured as a straight line. OLS regression discuss in the context of a bivariate model. There is only one independent variable (X) predicting a dependent variable (Y).

CHAPTER 3

METHODOLOGY

This chapter discusses the methodology adopted in this study and process carried out to achieve the goals of this study. More well defined steps require to develop a classification model such as data gathering, data preprocessing, feature selection, training and finally testing the classification algorithm.

3.1 Environment

The model was developed using the Python programming language. Python was preferred over other programming languages considering that it is more suitable for data analytics and machine learning tasks. The dependencies used with python are Sklearn, Matplotlib, and Nltk. Sklearn machine learning algorithms to implement the models in this study. As for the data preprocessing and handling, the Numpy, Pandas and Nltk libraries were used. Nltk was used to clean the review text used for the study. Lastly, for the data visualization, the matplotlib and seaborn Python libraries were used. Anaconda Spyder used as coding environment for this project. Also python is used for analyzing correlations and OLS regression.

3.2 Data Gathering

In order to get required dataset to achieve this study, fashion product dataset from Amazon were used. This is a subset of a bigger dataset that was created by extracting data from amazon and contains more than 10,000 records with sixteen attributes as discussed above in the introduction chapter.

3.3 Exploratory Data Analysis (EDA)

In order to derive some meaningful information, data preprocessing and handling missing values, this study conducted EDA on Amazon fashion products dataset. Below Figure 7 describes about basic statistical concepts like mean, median, mode, percentiles of available quantitative data on data set.

	price(£)	number_of_reviews	avg_review_rating
count	10025.000000	10011.000000	10011.000000
mean	19.054426	9.113275	4.695405
std	168.011411	33.681415	0.425863
min	0.000000	1.000000	1.000000
25%	2.650000	1.000000	4.500000
50%	8.700000	2.000000	5.000000
75%	18.120000	6.000000	5.000000
max	16268.000000	1399.000000	5.000000

Figure 7: Data Description

3.3.1 Data Preprocessing & Feature Selection

For preparing the desired data, a simple code was written in python to remove the useless features. The features like number of answered questions, customers who bought this item also bought, items customers buy after viewing this item and customers questions and answers were removed and not considered for this study due to high number of null values (see Figure 8). Also description, information and sellers were not considered since it gives similar information and not supported by the development ide. The null values of all other features handled using techniques available in python and excel.

uniq_id	0
product_name	0
manufacturer	7
price	1435
number_available_in_stock	2500
number_of_reviews	18
number_of_answered_questions	765
average_review_rating	18
amazon_category_and_sub_category	690
customers_who_bought_this_item_also_bought	1062
description	651
product_information	58
product_description	651
items_customers_buy_after_viewing_this_item	3065
customer_questions_and_answers	9086
customer_reviews	21
sellers	3082

Figure 8: Null values based on features

3.3.2 Data Labeling

The rating score that is given by the reviewer includes a number of stars on scales of 1 to 5. This data set contains the average rating for a particular product. Therefore, reviews that were rated more than 4 stars were considered as a Good product and those with less than three stars or equal were not recommended or considered as a bad product. Reviews that were rated between 3 to 4 or equals to 4 considered as a moderate product. These kinds of review scores usually contain many mixed reviews and are hard to be labeled into a recommended or not recommended category (Figure 9).

Multi Class Classification:

- Good Product – Satisfied by the product
- Moderate Product – Neutral product
- Bad Product- Unsatisfied with the product

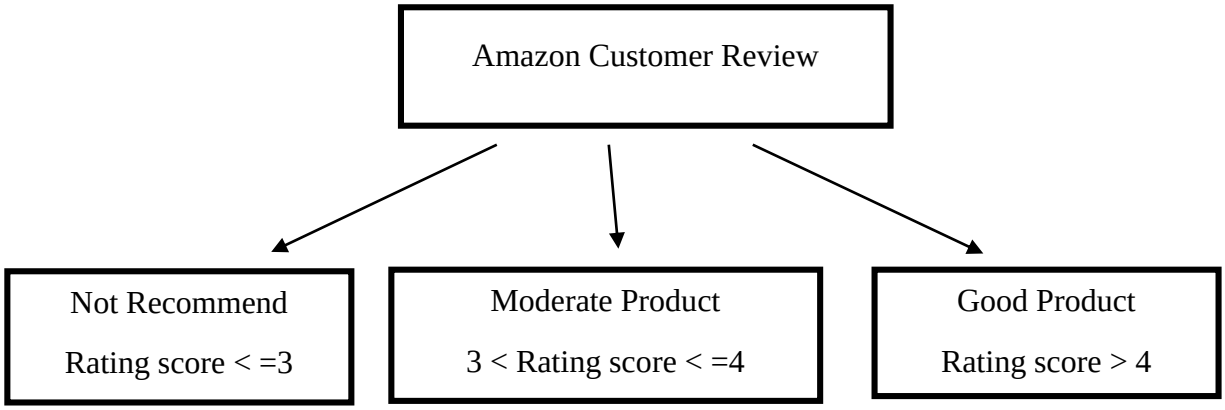

Figure 9: Class variations

3.3.3 Text Pre-processing using NLP methods

Review comment text is the most unstructured form of all available data in the dataset. Because various types of noise are present in it. Therefore, it is hard to analyze without preprocessing. Following preprocessing was applied for review comment text (See Figure 10).

- Tokenize text and removing punctuations - In here, we use a simple program to split the sentences into distinct words by splitting them at whitespaces. Removing punctuation

means, this is for words that may have several accepted forms, or words with punctuation in them. (Ex: don't, 3D, 3-D, period., period?)

- Removing words that containing numbers
- Removing stop words – The words with frequent occurrence in the document are called as stop words. It consists of conjunctions, prepositions, articles, and frequently occurring words like 'an', 'the', 'a' etc. Stop words are those words that has a little or no meaning in the text.so removing them from a sentence is good. Removing stop words from review helps to improve performance.
- Removing empty tokens
- Removing words with only one letter
- Lemmatize text - The process of lemmatization is to remove word affixes to get to a base/simplest form of the word.

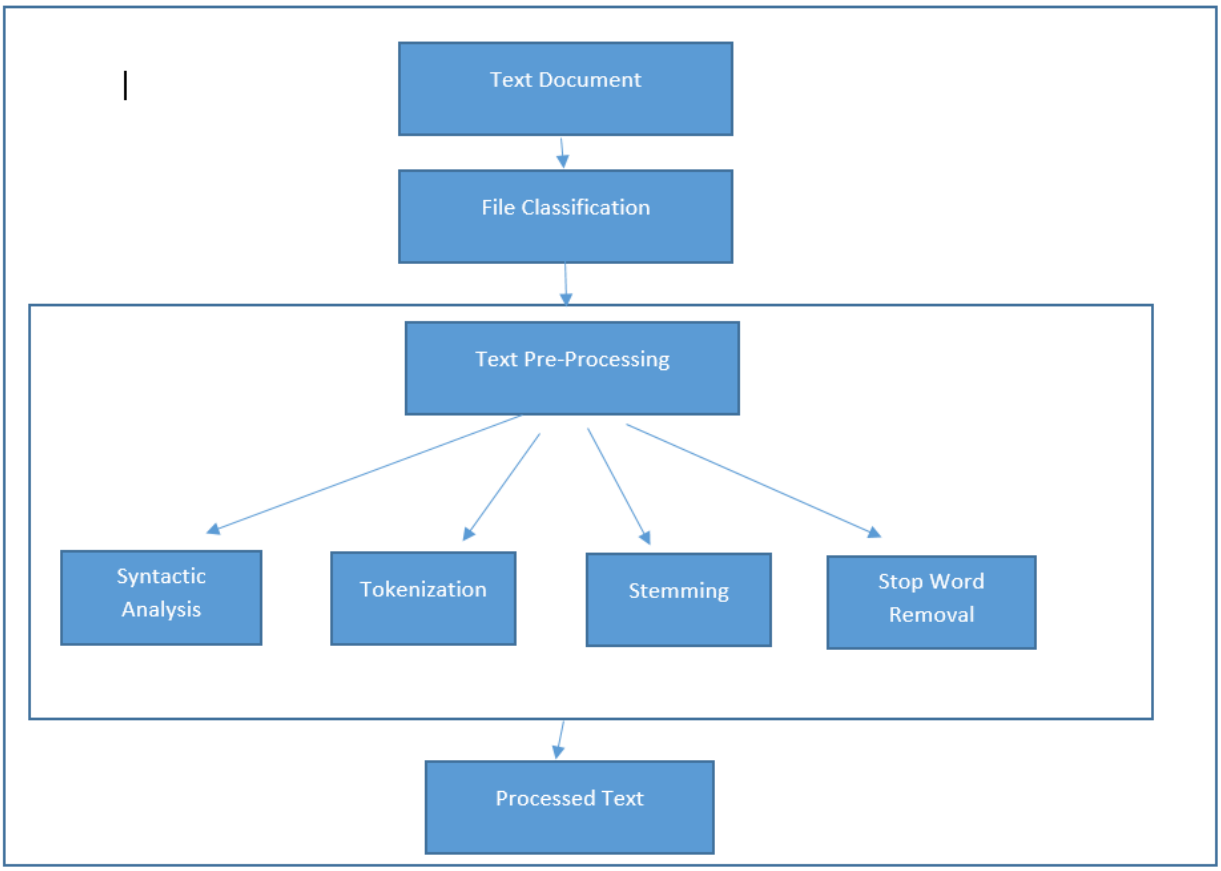

Figure 10: Text Processing

3.4 Building and Evaluate Model

The purpose of this model is recommend products based on customer reviews and ratings which means predicting a class of the products. Therefore, this study has selected four classification algorithms according to the objective. Then divided preprocessed dataset into testing and training datasets and applied classification algorithms. Model is trained with the training dataset. Finally evaluated model against test dataset. Few evaluation metrics were used to evaluate the performance of model as discussed below since this dataset is imbalanced.

Confusion Matrix: A table showing relation between correct predictions and types of incorrect predictions. Confusion Matrix allows to measure Recall, Precision, Accuracy are the metrics to measure the performance of the model.

Table 2: Confusion Matrix

	Predicted Positive	**Predictive Negative**
Actual Positive	TP	FN
Actual Negative	FP	FN

There are four categories of predictions can be encountered for a binary classification as explained in Table 2.

- **True Positive (TP):** predictions predicted as positive
- **True Negative (TN):** predictions predicted as negative
- **False Positive (FP):** predictions wrongly predicted as positive
- **False Negative (FN):** predictions wrongly predicted as negative (In here predictions should be predicted as positive but were predicted as negative)

Precision: This shows the exactness. The number of true positives divided by all positive predictions. Low precision indicates a high number of false positives.

$$\text{precision} = \frac{TP}{TP + FP}$$

Recall: This shows the completeness**.** The number of true positives divided by the number of positive values in the test data.

$$\text{Recall} = \frac{TP}{TP + FN}$$

F1 Measure: This shows the weighted average of precision and recall.

$$\text{FMeasure} = \frac{2 * \text{Precision} * \text{Recall}}{\text{Precision} + \text{Recall}}$$

Accuracy: This is one of the most common performance evaluation parameter. It is calculated as the ratio of number of correctly predicted review comments to the number of total number of review comments.

$$\text{Accuracy} = \frac{TP + TN}{TP + FP + TN + FN}$$

CHAPTER 4

PROJECT DESIGN & APPROACH

Labeled amazon fashion product data set has been used to conduct this study which consist of more than 10000 records. First performed exploratory data analysis on data set and then preprocessed. From the preprocessed dataset, potential features are selected. Then applied several classification algorithms to retrieve the results. Step-wise presentation of used approach is shown in the Figure 11.

4.1 Steps Followed for Classification and Data Analysis

Step 1: Amazon fashion product data set has been used to conduct this study which consist of more than 10000 records and 16 attributes.

Step 2: Performed exploratory data analysis in order to identify features that containing null values and useful information. Then removed irrelevant features and preprocessed remaining data. Customer review comments are cleaned with the help of Natural Language Processing techniques. Below steps are applied during preprocessing reviews:

- Convert text in to lower case and tokenize text and remove punctuations
- Remove words that contain numbers
- Remove stop words
- Remove empty tokens
- Pos tag text and lemmatize text

Step 3: After cleaning the dataset in step 2, features are used for statistical analyzing and model building. Performed Pearson correlation analysis and OLS regression analysis on selected features to identify relationship between variables and how it affects to each other. Then identified most popular products, product categories and manufacturers based on rating analysis.

Step 4: Preprocessed data is divided in to training dataset and testing dataset. The 75% of data is used as training data and 25% of data is used as testing data. Then different classification algorithms are applied on training data and those models are evaluated against testing data. The algorithms are mentioned in below used to perform classification:

- Naïve Bayes
- Decision Tree
- Linear SVC
- K-Nearest Neighbor – trained model with different k values.

For each model, values of accuracy, precision, recall and F-1 score as performance evaluation metrics are found out and detailed output is generated. The confusion matrix is also generated. Finally, these obtained results are compared to identifying better algorithm to do this classification.

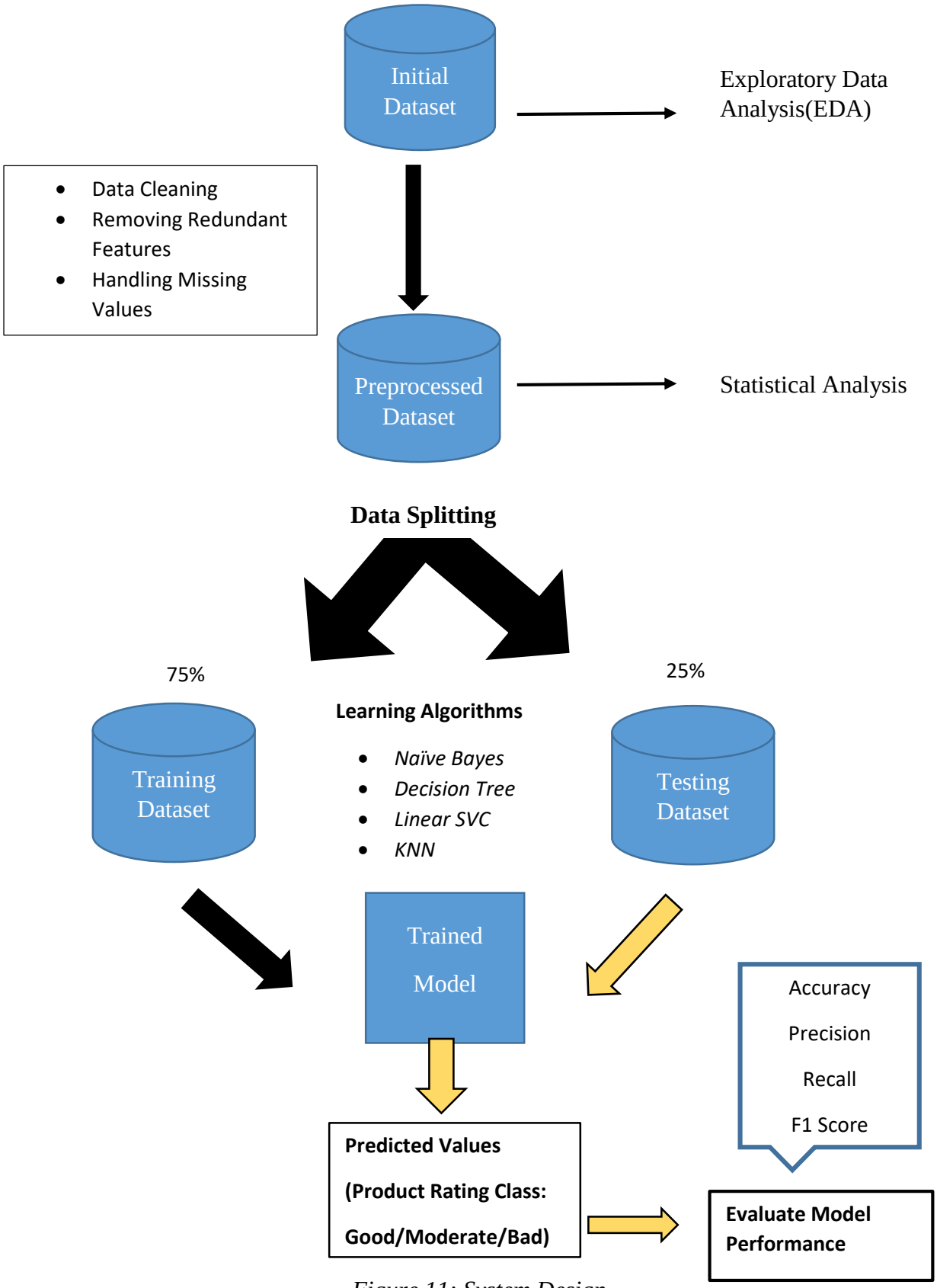

Figure 11: System Design

CHAPTER 5

IMPLEMENTATION

This chapter contains the approach and implementation part of this project describing every step taken to achieve them through code snippets.

5.1 Course of Action

a) Importing required libraries:

In this step, all relevant libraries are imported from nltk, sklearn, Matplotlib etc.

```
import pandas as pd
import numpy as np
# Visualizations
import matplotlib.pyplot as plt
import seaborn as sns
import matplotlib.colors as colors

from sklearn.impute import SimpleImputer
from sklearn.neighbors import KNeighborsClassifier
from sklearn.feature_selection import SelectKBest, chi2, SelectPercentile, f_classif
from sklearn.model_selection import train_test_split
import nltk
from tqdm import tqdm
tqdm.pandas(desc="progress-bar")
nltk.download('stopwords')
from sklearn.feature_extraction.text import TfidfVectorizer
from sklearn.metrics import classification_report, confusion_matrix, accuracy_score
from sklearn.pipeline import Pipeline
from time import time
import string
from nltk import pos_tag
from nltk.corpus import stopwords
from nltk.tokenize import WhitespaceTokenizer
from nltk.stem import WordNetLemmatizer
# return the wordnet object value corresponding to the POS tag
from nltk.corpus import wordnet
# Text preprocessing and occurance counting
from sklearn.feature_extraction.text import CountVectorizer
from sklearn.feature_extraction.text import TfidfTransformer
from sklearn import metrics, model_selection
```

b) Importing dataset and extracting relevant features:

Loading original dataset and extract only relevant features for this study.

```
mydataset=pd.read_csv('amazon_co-ecommerce_processed.csv', engine='python')
df = pd.DataFrame(mydataset, columns= ['uniq_id','product_name','manufacturer',
                                       'price(£)','number_available_in_stock','number_of_reviews',
                                       'avg_review_rating','amazon_category_and_sub_category'
                                       'customer_reviews','rating_class'])
```

c) Preprocessing and clean customer review comments:

Here we called clean_text function to clean customer review text that execute several transformation techniques. Called get_word_net(pos_tag) function with in the clean_text method to assign a tag to every word to define if it relates to a noun, a verb etc.

```
################################
###Text preprocessing
################################

def get_wordnet_pos(pos_tag):
    if pos_tag.startswith('J'):
        return wordnet.ADJ
    elif pos_tag.startswith('V'):
        return wordnet.VERB
    elif pos_tag.startswith('N'):
        return wordnet.NOUN
    elif pos_tag.startswith('R'):
        return wordnet.ADV
    else:
        return wordnet.NOUN

def clean_text(text):
    # lower text
    text = str(text).lower()
    # tokenize text and remove puncutation
    text = [word.strip(string.punctuation) for word in text.split(" ")]
    # remove words that contain numbers
    text = [word for word in text if not any(c.isdigit() for c in word)]
    # remove stop words
    stop = stopwords.words('english')
    text = [x for x in text if x not in stop]
    # remove empty tokens
    text = [t for t in text if len(t) > 0]
    # pos tag text
    pos_tags = pos_tag(text)
    # lemmatize text
    text = [WordNetLemmatizer().lemmatize(t[0], get_wordnet_pos(t[1])) for t in pos_tags]
    # remove words with only one letter
    text = [t for t in text if len(t) > 1]
    # join all
    text = " ".join(text)
    return(text)

# clean text data
df["review_clean"] = df["customer_reviews"].apply(lambda x: clean_text(x))
```

d) Feature Engineering:

Label the customer reviews into categories/classes based on their ratings. Then splitting the dataset into train and test set and print the size of train and test data. The training data is 75%, and testing data is 25 %.

```
############################
###Review classification
############################

df['rating_class'] = df['avg_review_rating'].apply(lambda x : 'Not_Recommended' if x <= 3 else ('Moderate_Product' if x <= 4 else 'Good_Product'))

X = df['review_clean']
y = df['rating_class']

# Splitting Dataset into train and test set
X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.25, random_state=42)
# Print shape of train and test data
print ('Train Set Shape\t\t:{}\nTest Set Shape\t\t:{}'.format(X_train.shape, X_test.shape))
```

e) Apply vectorization and frequency/inverse document frequency:

This will create feature vectors and count the number of words in each document and reduce the weightage of more common words like (the, is, an etc.) which occurs in all document

```
count_vect = CountVectorizer()
X_train_counts = count_vect.fit_transform(X_train)
X_train_counts.shape


tfidf_transformer = TfidfTransformer(use_idf=False)
X_train_tfidf = tfidf_transformer.fit_transform(X_train_counts)
X_train_tfidf.shape
```

f) Build and fit the model:

This code explains about building a pipeline from extracted data and then test model against the testing data. Below code has applied to Naive Bayes, Linear SVC, Decision Tree and K-Nearest Neighbor classifiers on the dataset for predict the class of the product. Also applied KNN algorithm with different k values. Time taken for training data and prediction also displayed.

```
print('Training data...')
t = time()

count_vect = CountVectorizer()
X_train_counts = count_vect.fit_transform(X_train)
X_train_counts.shape


tfidf_transformer = TfidfTransformer(use_idf=False)
X_train_tfidf = tfidf_transformer.fit_transform(X_train_counts)
X_train_tfidf.shape

print('Using K-Nearest Neighbor...')
pipeline = Pipeline([
            ("vect", CountVectorizer()),
            ("tfidf", TfidfTransformer()),
            ('clf_nominalNB', KNeighborsClassifier(n_neighbors=7))])

model = pipeline.fit(X_train, y_train)

print('Training data completed!')
print('Training time: ', round(time()-t, 3), 's\n')

print('Predicting Test data...')
t = time()

y_pred = model.predict(X_test)

print('Prediction completed!')
print('Prediction time: ', round(time()-t, 3), 's\n')
```

g) Evaluate model performance:

Prediction of test data is completed and Confusion Matrix of prediction is displayed. Also calculated accuracy, precision, recall, f1 score and support are displayed in the classification report.

```
############################
### Confusion Matrix #########
############################

cm=confusion_matrix(y_test, y_pred)
df_cm=pd.DataFrame(cm, index= ['Good_Product','Moderate_Product','Not_Recommended'],columns=['Good_Product','Moderate_Product','Not_Recommended'])
print(df_cm)
print()


########################################
# Compute and print the classification report##
########################################
print('Evaluated Result...')
t = time()

print(classification_report(y_test, y_pred))
```

CHAPTER 6

EVALUATION AND RESULTS

This chapter presents the findings and the evaluation of the research and include results obtained and critical evaluation of the research work.

6.1 Model Training and Evaluation

The main goal of this study is to analyze customer ratings/reviews and identify proper classification to recommend products. In order to achieve this goal, analyzed preprocessed review text and predict whether that product is recommended, Moderate or Worst/Not recommended product. And also determined which machine learning algorithm performs better in the task of text classification. In order to do that, the trained model on training data and tested model on test data using three classifiers. This was accomplished by using the Amazon fashion products as data set. The data set was divided into 75% for the training dataset,25% for the testing dataset. First model is trained with 75 % of training data. After training is completed, the model is tested against the remaining 25 % of data. However, take note that the frequency distributions for classes in the recommendation are imbalanced (See Figures 12 and 13). There are more recommended classes than moderate and not recommended. Accuracy is not the best metric to use when evaluating imbalanced datasets as it can be misleading. So this study used several metrics to evaluate the performance of the model like precision, recall, f1 score and confusion matrix. Also Figure 14 and 15 show Word cloud generated using positive word list and negative word list in the dataset.

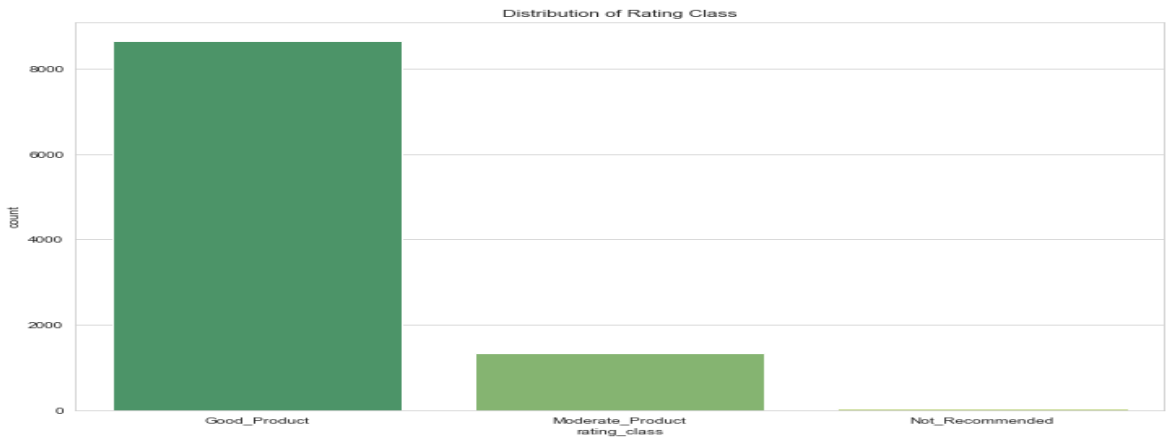

Figure 12: Distribution of Rating Classes

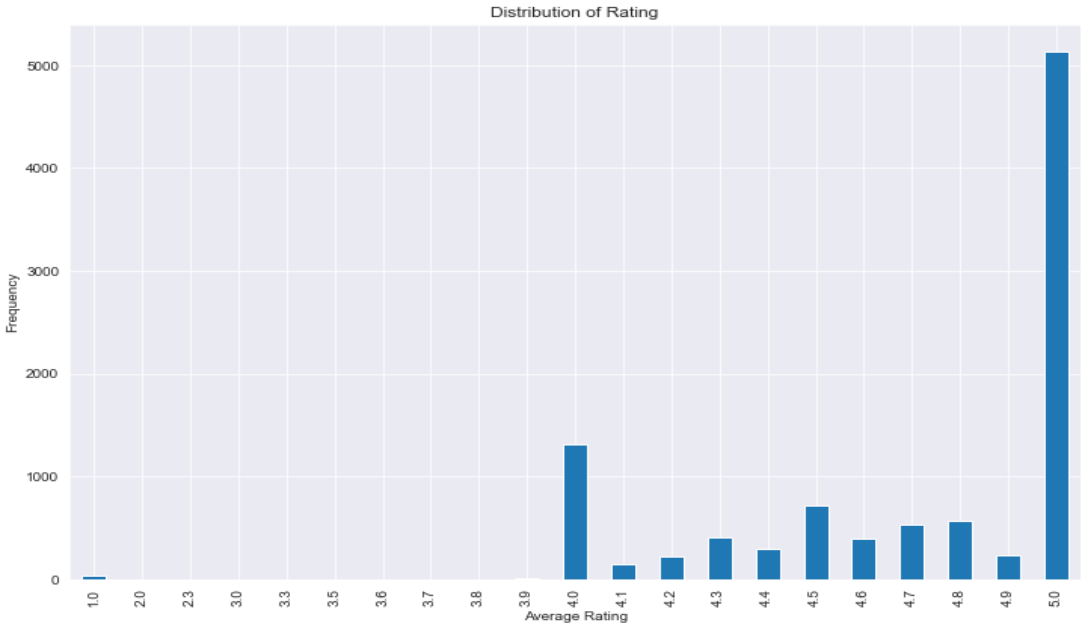

Figure 13: Distribution of Ratings

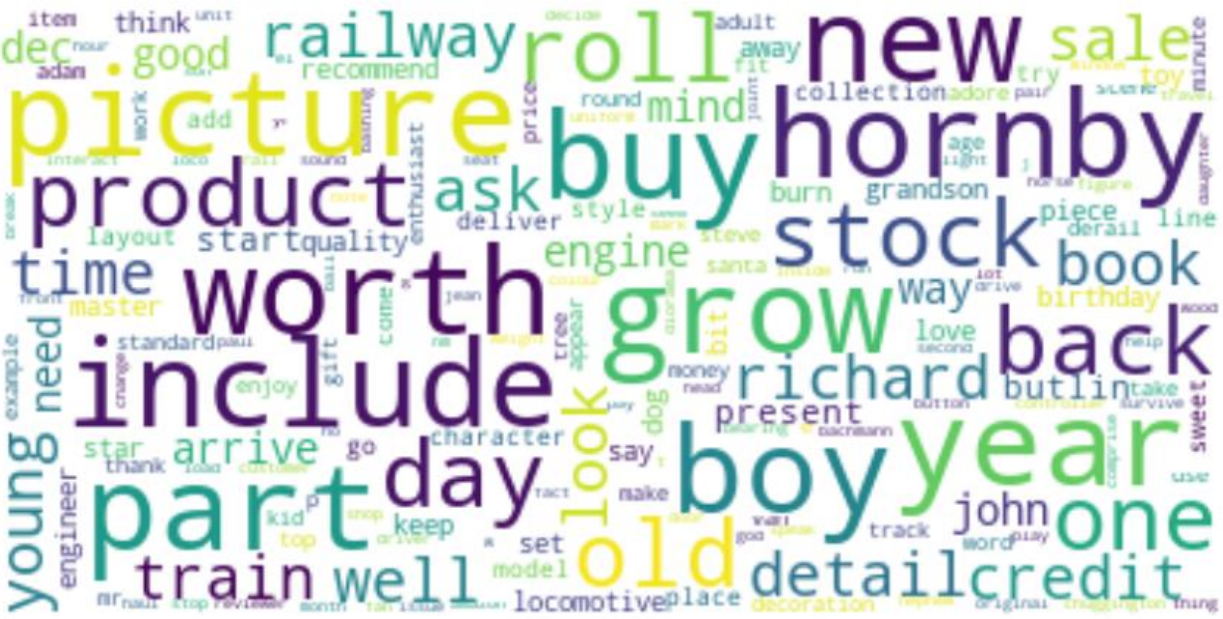

Figure 14: Word cloud generated using positive word list

Figure 15: Word cloud generated using negative word list

6.2 Experimental Results

In this section presents experimental results from all four different classifiers(Figure,16,17,18,19) and results obtained using statistical techniques on Amazon fashion product dataset. The confusion matrix that classifies the product into Good, Moderate and Not Recommended are also generated with detailed outputs.

a) Identify proper classification to recommend products

Output of Decision Tree Classifier

```
Train Set Shape         :(7518,)
Test Set Shape          :(2507,)
Training data...
Using Decision Tree Classifier...
Training data completed!
Training time:  4.56 s

Predicting Test data...
Prediction completed!
Prediction time:  0.224 s

                 Good_Product  Moderate_Product  Not_Recommended
Good_Product             1975               187                0
Moderate_Product          219               119                0
Not_Recommended             1                 0                6

Evaluated results...
                 precision    recall  f1-score   support

    Good_Product       0.90      0.91      0.91      2162
Moderate_Product       0.39      0.35      0.37       338
 Not_Recommended       1.00      0.86      0.92         7

        accuracy                           0.84      2507
       macro avg       0.76      0.71      0.73      2507
    weighted avg       0.83      0.84      0.83      2507
```

Figure 16: Screenshot of the Decision Tree classifier output

Output of Naïve Bayes Classifier

```
Train Set Shape         :(7518,)
Test Set Shape          :(2507,)
Training data...
Using Naive Bayes Classifier...
Training data completed!
Training time:  0.804 s

Predicting Test data...
Prediction completed!
Prediction time:  0.259 s

                 Good_Product  Moderate_Product  Not_Recommended
Good_Product             2162                 0                0
Moderate_Product          338                 0                0
Not_Recommended             7                 0                0

Evaluated results...
                 precision    recall  f1-score   support

    Good_Product       0.86      1.00      0.93      2162
Moderate_Product       0.00      0.00      0.00       338
 Not_Recommended       0.00      0.00      0.00         7

        accuracy                           0.86      2507
       macro avg       0.29      0.33      0.31      2507
    weighted avg       0.74      0.86      0.80      2507
```

Figure 17: Screenshot of the Naïve Bayes classifier output

Output of Linear SVC Classifier

```
Train Set Shape         :(7518,)
Test Set Shape          :(2507,)
Training data...
Using Linear SVC Classifier...
Training data completed!
Training time:  0.925 s

Predicting Test data...
Prediction completed!
Prediction time:  0.217 s

                 Good_Product  Moderate_Product  Not_Recommended
Good_Product             2118                44                0
Moderate_Product          245                93                0
Not_Recommended             1                 0                6

Evaluated results...
                 precision    recall  f1-score   support

    Good_Product       0.90      0.98      0.94      2162
Moderate_Product       0.68      0.28      0.39       338
 Not_Recommended       1.00      0.86      0.92         7

        accuracy                           0.88      2507
       macro avg       0.86      0.70      0.75      2507
    weighted avg       0.87      0.88      0.86      2507
```

Figure 18: Screenshot of the Linear SVC classifier output

Output of K-Nearest Neighbor Classifier

```
Train Set Shape         :(7518,)
Test Set Shape          :(2507,)
Training data...
Using K-Nearest Neighbor...
Training data completed!
Training time:  1.53 s

Predicting Test data...
Prediction completed!
Prediction time:  1.642 s

                 Good_Product  Moderate_Product  Not_Recommended
Good_Product             2153                 8                1
Moderate_Product          312                24                2
Not_Recommended             1                 0                6

Evaluated Result...
                 precision    recall  f1-score   support

    Good_Product       0.87      1.00      0.93      2162
Moderate_Product       0.75      0.07      0.13       338
 Not_Recommended       0.67      0.86      0.75         7

        accuracy                           0.87      2507
       macro avg       0.76      0.64      0.60      2507
    weighted avg       0.86      0.87      0.82      2507
```

Figure 19: Screenshot of the KNN classifier output

Here in Table 3, we can see that all classifiers performed the best. In terms of accuracies and time taken for prediction, Linear SVC tends to do better than other three algorithms although there is not much difference. The KNN algorithms took the highest time for prediction. Also this study tried the KNN (k-nearest algorithm) algorithm for different k values. Then the accuracy is increased for higher k values (see Table 4).

Table 3: Different Classifiers and Accuracies

Algorithm	Accuracy	Time Taken for Prediction
Decision Tree	84 %	0.224 Seconds
Naïve Bayes	86 %	0.259 Seconds
K-Nearest Neighbor(k=7)	87 %	1.642 Seconds
Linear SVC	88 %	0.217 Seconds

Table 4: Different k values and Accuracies

K value	Accuracy	Time Taken for Prediction
1	82 %	1.55 Seconds
3	86 %	1.507 Seconds
5	86 %	1.547 Seconds
7	87 %	1.642 Seconds

b) Identifying most popular top 10 products

Figure 20 shows the products name of top 10 rated fashion products from the e-commerce. The products like Zoo Animal Hand Sock Glove Finger Puppets Sack Plush Toy Cow, Bananagrams Game, Tommy Pop-Up Pirates belong to the most reviewed popular products.

	product_name	Total_reviews
4485	Zoo Animal Hand Sock Glove Finger Puppets Sack Plush Toy Cow	1545
9307	Bananagrams Game	1399
8864	TOMY Pop-Up Pirate	1040
1251	Viskey 600 Loom Rubber Bands Bracelet Making Clips Tools Children Games,Hot Pink	802
9330	Orchard Toys Shopping List	690
1927	Temporary Tattoos (5 sheets) - Nitefall(TM) Wounds	649
1582	Welecom(TM) 100Pcs 10mm 12mm 15mm 18mm 20mm 22mm Mixed Size Googly Eyes	600
7285	Original Stomp Rocket	585
6568	Crayola Supertips Washable - Pack of 12	561
1365	Loom Bandz - Rainbow Colours - White 600 Count	518

Figure 20: Most popular top 10 products

c) Identifying most popular top 10 products categories

Figure 21 shows the top 10 rated products categories from the e-commerce. Die-Cast & Toy Vehicles, Figures & Playsets and Arts & Crafts are most popular selling product categories.

	amazon_category_and_sub_category	ratings_sum_categories
795	Die-Cast & Toy Vehicles > Toy Vehicles & Accessories > Scaled Models > Vehicles	4207.1
749	Figures & Playsets > Science Fiction & Fantasy	2271
1192	Arts & Crafts > Children's Craft Kits > Bead Art & Jewellery-Making	1730.7
242	Characters & Brands > Disney > Toys	1618.8
2347	Hobbies > Trading Cards & Accessories > Packs & Sets	1503.7
399	Games > Dice & Dice Games	1393.7
2789	Party Supplies > Decorations > Balloons	1384.9
1577	Party Supplies > Banners, Stickers & Confetti > Banners	1312.9
4552	Puppets & Puppet Theatres > Hand Puppets	1142.5
5146	Games > Card Games	1133.4

Figure 21: Most popular top 10 product categories

d) Identifying most popular products manufacturers

Figure 22 and 23 show the top 10 manufacturers of popular fashion products from the e-commerce. The products manufactured by LEGO, Disney and Oxford Diecast are the most popular products among customers.

	manufacturer	ratings_sum_manufacturer
613	LEGO	812.7
242	Disney	783.1
803	Oxford Diecast	747.7
603	Playmobil	674.8
236	Star Wars	569.5
86	Mattel	532.2
206	The Puppet Company	517.4
390	Hasbro	505.2
8593	MyTinyWorld	457.9
54	Corgi	431.2

Figure 22: Most popular product manufacturers

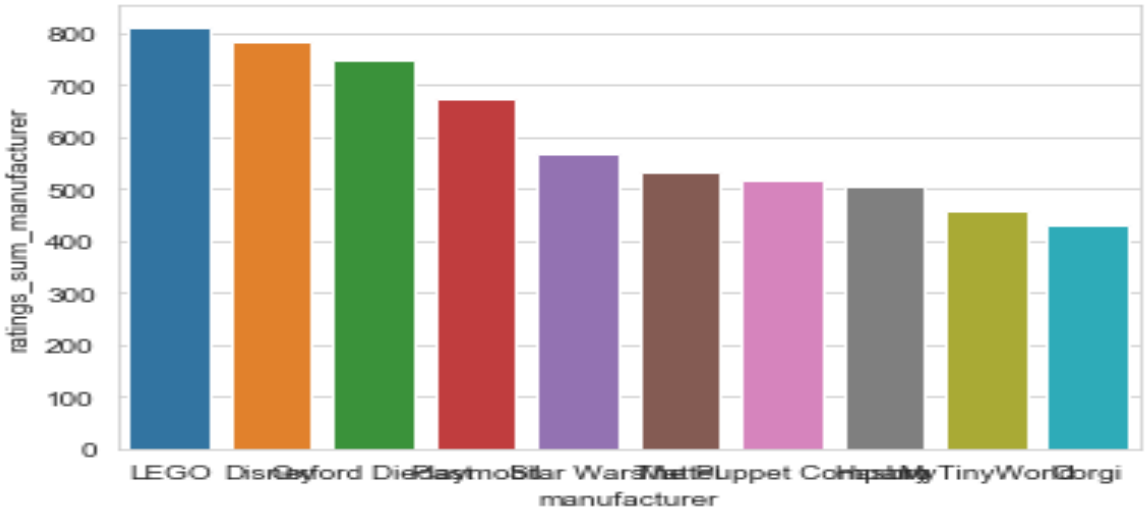

Figure 23: Bar Chart -Most popular top 10 product categories

e) Identifying relationship between product prices, number of reviews and ratings

OLS Regression Results

Dep. Variable:	price	R-squared:	0.000
Model:	OLS	Adj. R-squared:	-0.000
Method:	Least Squares	F-statistic:	0.1817
Date:	Fri, 03 Sep 2021	Prob (F-statistic):	0.834
Time:	20:59:27	Log-Likelihood:	-65508.
No. Observations:	10011	AIC:	1.310e+05
Df Residuals:	10008	BIC:	1.310e+05
Df Model:	2		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
Intercept	26.1195	18.728	1.395	0.163	-10.591	62.829
avg_review_rating	-1.4549	3.962	-0.367	0.714	-9.222	6.312
number_of_reviews	-0.0255	0.050	-0.509	0.611	-0.124	0.073

Omnibus:	38594.893	Durbin-Watson:	1.995
Prob(Omnibus):	0.000	Jarque-Bera (JB):	31717158923.837
Skew:	90.570	Prob(JB):	0.00
Kurtosis:	8721.069	Cond. No.	398.

Figure 24: OLS Regression Results

Based OLS regression results shown in Figure 24, we can say that number of reviews and average review rating do not have much impact on the price. In here we have used price as dependent variable and review rating as explanatory variables.

Assume below hypothesis:

H_0 – Variables has no significant influence on price

H_1 - Variables has impact on price

The higher p values indicate that we cannot reject the null hypothesis that price has no effect on average review rating and number of reviews. R^2 is the coefficient of determination that tells us that how much percentage variation dependent variable can be explained by independent variables. Here, the zero R-squared shows that the model explains none of the variability of the response data around its mean. The partial regression plots in Figure 25 help to interpret regression analysis

results more intuitively. The trend indicates that the predictor variables (average review rating, number of reviews) do not provide significant information about the response (price).

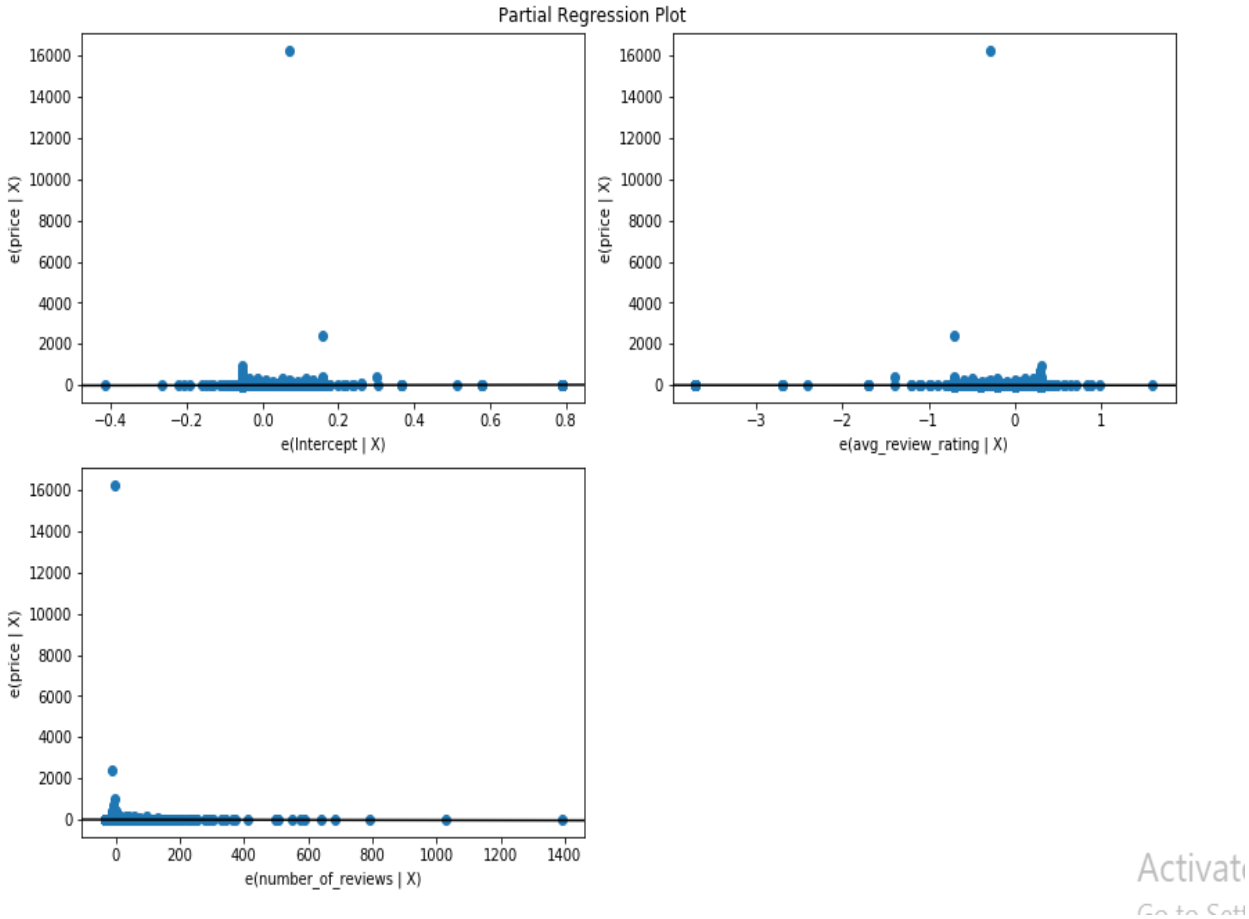

Figure 25: Partial Regression Plots

Also as per the Pearson correlation matrix shown in Figure 26, we can quickly see that there are no strong relationships between price, average review rating and number of reviews. The coefficient close to 1 means that there is a very strong positive correlation between the two variables. But here all 3 variables show weak negative correlations. Also here average review rating shows considerably strong negative relationship with number of reviews.

Based on correlation analysis and OLS analysis, we can say that price has no significant impact on ratings and number of reviews. So customers mainly rely on reviews when buying products.

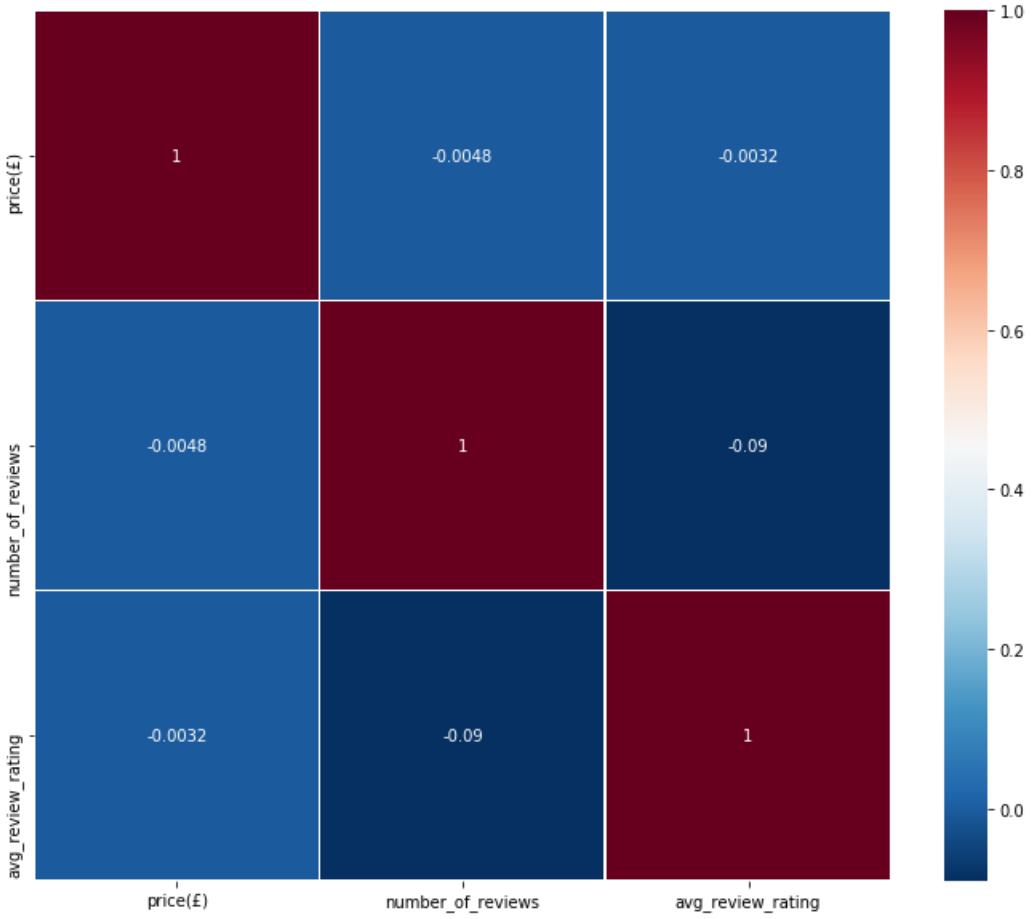

Figure 26: Pearson correlation

CHAPTER 7

CONCLUSION AND FUTURE WORK

An evolutionary shift to eCommerce markets has increased the dependency of customers and sellers on online reviews and ratings to a vast extent. This dependency helps for building trust, confidence and influencing the customer buying decision. Nowadays, due to the enormous growth of technology lot of customers use eCommerce sites. Therefore, thousands of reviews and ratings are generated daily for a particular product. Going through each customer review is very time-consuming and a complex task. So, there is a need to handle this large volume of reviews,analyze to derive valuable data for recommending products for customers. People are always more eager to purchase products that others have already recommended. Customers rely on product reviews to make up their minds to make better purchase decisions. Therefore, it is vital to analyze ratings and reviews to provide better options for consumers to make actionable decisions. On the other end, it can also help businesses extend sales and improve the product by understanding customer needs while building trust to reduce the business failure risks. Hence, this study allows sellers and marketers to make better business decisions and increase their revenue while assisting customers in choosing the right product.

The main aim of this study was to analyze customer ratings and reviews and analyze the possibility of recommending the products to make actionable decisions. Thereby, classification approaches were carried out toidentify the most popular products, product categories, and manufacturers among the fashion products in Amazon. Further studies were carried out to determine the relationship between product prices, customer reviews, and ratings, which directly impact customers' or marketers' decisions.

In that regard, a model was designed to classify the product into three main categories, namely good, moderate, or Not recommended. Sentiment analysis was applied to the product rating and reviews and trained using machine learning algorithms. The study used four different machine learning algorithms, Naive Bayes, Linear SVC, Decision Tree, and K nearest neighbor, and applied on customer reviews and ratings of the Amazon fashion products. The results from the study showed that in terms of accuracyand prediction time, the Linear SVC approach achieves better

results than other approaches. The performance of the K-Nearest Neighbor algorithm classifier was further improved when the values of k were increased.

This study also applied the OLS statistical model and Pearson correlations on price, rating, and numberof reviews. Based on the results, the average rating and number of reviews do not directly affect the product price. Hence irrespective of the product price, customers can mainly rely on ratings and reviews when purchasing products from eCommerce sites. This study also identified the most popular products, product categories, and manufacturers based on product ratings. The positioning of the words in a text is not considered in the bag of words model. Hence, this could negatively affect the semantic of a review, which is a limitation of this model. For example, although the overall review is negative, a machine would classify the review as positive due to the number of positive words it contains.

This dataset is mainly based on fashion products like toys and accessories, and many have not looked at these types of products. The findings of this study will help eCommerce sellers identify customer preferences for these kinds of products and improve the consumer experience. Besides being beneficial to consumers and sellers, these predictions can also be of great use for manufacturers. Overall, based on the reviews, they can quickly identify products that have been poorly rated. Therefore, they can improve the quality of their products as per the customers' requirements.

As per future work, the accuracy of the proposed model can be further improved by fine-tuning the classifier. Conversely, a Grid Search with LinearSVC classifier pipeline can be performed by choosing the best parameters from a grid of possible values. In addition to that, similar classification algorithms such as maximum entropy classifier, Stochastic gradient classifier, and XGBoost can be applied with different datasets

APPENDICES

Source Codes:

Model development and evaluation (Full Source Code):

```
import pandas as pd
import numpy as np
# Visualizations
import matplotlib.pyplot as plt
import seaborn as sns
import matplotlib.colors as colors

from sklearn.impute import SimpleImputer
from sklearn.neighbors import KNeighborsClassifier
from sklearn.feature_selection import SelectKBest, chi2, SelectPercentile, f_classif
from sklearn.model_selection import train_test_split
import nltk
from tqdm import tqdm
tqdm.pandas(desc="progress-bar")
nltk.download('stopwords')
from sklearn.feature_extraction.text import TfidfVectorizer
from sklearn.metrics import classification_report, confusion_matrix, accuracy_score
from sklearn.pipeline import Pipeline
from time import time
import string
from nltk import pos_tag
from nltk.corpus import stopwords
from nltk.tokenize import WhitespaceTokenizer
from nltk.stem import WordNetLemmatizer
# return the wordnet object value corresponding to the POS tag
from nltk.corpus import wordnet
# Text preprocessing and occurance counting
from sklearn.feature_extraction.text import CountVectorizer
from sklearn.feature_extraction.text import TfidfTransformer
from sklearn import metrics, model_selection

mydataset=pd.read_csv('amazon_co-ecommerce_processed.csv', engine='python')
df = pd.DataFrame(mydataset, columns= ['uniq_id','product_name','manufacturer',
                                       'price(£)','number_available_in_stock','number_of_reviews',
                                       'avg_review_rating','amazon_category_and_sub_category',
                                       'customer_reviews','rating_class'])
```

```
##############################
###Text preprocessing
##################################

def get_wordnet_pos(pos_tag):
    if pos_tag.startswith('J'):
        return wordnet.ADJ
    elif pos_tag.startswith('V'):
        return wordnet.VERB
    elif pos_tag.startswith('N'):
        return wordnet.NOUN
    elif pos_tag.startswith('R'):
        return wordnet.ADV
    else:
        return wordnet.NOUN

def clean_text(text):
    # lower text
    text = str(text).lower()
    # tokenize text and remove puncutation
    text = [word.strip(string.punctuation) for word in text.split(" ")]
    # remove words that contain numbers
    text = [word for word in text if not any(c.isdigit() for c in word)]
    # remove stop words
    stop = stopwords.words('english')
    text = [x for x in text if x not in stop]
    # remove empty tokens
    text = [t for t in text if len(t) > 0]
    # pos tag text
    pos_tags = pos_tag(text)
    # lemmatize text
    text = [WordNetLemmatizer().lemmatize(t[0], get_wordnet_pos(t[1])) for t in pos_tags]
    # remove words with only one letter
    text = [t for t in text if len(t) > 1]
    # join all
    text = " ".join(text)
    return(text)
```

```
df["review_clean"] = df["customer_reviews"].apply(lambda x: clean_text(x))


##############################
###Review classification
##############################

df['rating_class'] = df['avg_review_rating'].apply(lambda x : 'Not_Recommended' if x <= 3 else ('Moderate_Product' if x <= 4 else 'Good_Product'))

X = df['review_clean']
y = df['rating_class']

# Splitting Dataset into train and test set
X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.25, random_state=42)
# Print shape of train and test data
print ('Train Set Shape\t\t:{}\nTest Set Shape\t\t:{}'.format(X_train.shape, X_test.shape))

print('Training data...')
t = time()

count_vect = CountVectorizer()
X_train_counts = count_vect.fit_transform(X_train)
X_train_counts.shape


tfidf_transformer = TfidfTransformer(use_idf=False)
X_train_tfidf = tfidf_transformer.fit_transform(X_train_counts)
X_train_tfidf.shape

print('Using K-Nearest Neighbor...')
pipeline = Pipeline([
            ("vect", CountVectorizer()),
            ("tfidf", TfidfTransformer()),
            ('clf_nominalKNN', KNeighborsClassifier(n_neighbors=7))])

model = pipeline.fit(X_train, y_train)

print('Training data completed!')
print('Training time: ', round(time()-t, 3), 's\n')
```

```
print('Training data completed!')
print('Training time: ', round(time()-t, 3), 's\n')

print('Predicting Test data...')
t = time()

y_pred = model.predict(X_test)

print('Prediction completed!')
print('Prediction time: ', round(time()-t, 3), 's\n')

##############################
### Confusion Matrix ##########
##############################

cm=confusion_matrix(y_test, y_pred)
df_cm=pd.DataFrame(cm, index= ['Good_Product','Moderate_Product','Not_Recommended'],columns=['Good_Product','Moderate_Product','Not_Recommended'])
print(df_cm)
print()


##########################################
# Compute and print the classification report##
##########################################
print('Evaluated Result...')
t = time()

print(classification_report(y_test, y_pred))
```

Implementing Pearson correlation and OLS regression analysis:

```
import pandas as pd
import numpy as np
# Visualizations
import matplotlib.pyplot as plt
from matplotlib.colors import ListedColormap
import seaborn as sns
import matplotlib.colors as colors
from tabulate import tabulate
import statsmodels.api as sm
from statsmodels.formula.api import ols
from statsmodels.compat import lzip

mydataset=pd.read_csv('amazon_co-ecommerce_processed.csv', engine='python')
df = pd.DataFrame(mydataset, columns= ['uniq_id','product_name','manufacturer','price(£)',
                                       'number_available_in_stock','number_of_reviews','avg_review_rating',
                                       'amazon_category_and_sub_category','customer_reviews','rating_class'])

df_corr = pd.DataFrame(df, columns= ['price(£)','number_of_reviews','avg_review_rating'])

########################################
## Pearson Correlation
########################################


pearsoncorr = df_corr.corr(method='pearson')
sns.heatmap(pearsoncorr,
            xticklabels=pearsoncorr.columns,
            yticklabels=pearsoncorr.columns,
            cmap='RdBu_r',
            annot=True,
            linewidth=0.5)
```

```
########################################
## OLS Regression Analysis
########################################

df_corr = df_corr.rename(columns={'price(£)': 'price'})
df_corr['price(£)'] = df_corr['price'].astype(float)
df_corr['avg_review_rating'] = df_corr['avg_review_rating'].astype(float)
df_corr['number_of_reviews'] = df_corr['number_of_reviews'].astype(float)

result_model = ols(" price ~ avg_review_rating+number_of_reviews",data=df_corr).fit()
print(result_model.summary())

fig = plt.figure(figsize=(12,8))
fig = sm.graphics.plot_partregress_grid(result_model, fig=fig)
```

Implementing word cloud:

```
from nltk.tokenize import RegexpTokenizer
from nltk import FreqDist
import seaborn as sns
from sklearn.manifold import TSNE
from wordcloud import WordCloud
import matplotlib.pyplot as plt


good_words = df[(df['avg_review_rating'] > 4)]

def RegExpTokenizer(Sent):
    tokenizer = RegexpTokenizer(r'\w+')
    return tokenizer.tokenize(Sent)

ListWordsgood = []
for m in good_words['review_clean']:
    n = RegExpTokenizer(str(m))
    ListWordsgood.append(n)
#print(ListWordsgood)



def Bag_Of_Words(ListWordsgood):
    all_words1good = []
    for m in ListWordsgood:
        for w in m:
            all_words1good.append(w.lower())
    all_words2 = FreqDist(all_words1good)
    return all_words2

plt.figure(figsize = (8,6))

all_words5good = Bag_Of_Words(ListWordsgood)
count = []
Words  = []
for w in all_words5good.most_common(10):
    count.append(w[1])
    Words.append(w[0])
sns.set_style("darkgrid")
print(sns.barplot(Words,count))


all_words5 = Bag_Of_Words(ListWordsgood)
ax = plt.figure(figsize=(15,10))
# Generate a word cloud image
wordcloud = WordCloud(background_color='white',max_font_size=40).generate(' '.join(all_words5.keys()))

# Display the generated image:

plt.imshow(wordcloud, interpolation='bilinear')
plt.axis("off")
print(len(all_words5good))
```

Source code for finding most popular manufactures:

```
import pandas as pd
import numpy as np
import matplotlib.pyplot as plt
from matplotlib.colors import ListedColormap
import seaborn as sns
import matplotlib.colors as colors
from tabulate import tabulate

mydataset=pd.read_csv('amazon_co-ecommerce_processed.csv', engine='python')
df = pd.DataFrame(mydataset, columns= ['uniq_id','product_name','manufacturer','price(£)','number_available_in_stock',
                                       'number_of_reviews','avg_review_rating','amazon_category_and_sub_category',
                                      'customer_reviews','rating_class'])
```

```
#####################################
##  Top 10 manufacturers
#####################################
top_manufacturers = pd.DataFrame(df.groupby(['manufacturer'])['avg_review_rating'].sum()).rename(columns = {'avg_review_rating': 'ratings_sum_manufacturer'})
top10_manufacturer = top_manufacturers.sort_values('ratings_sum_manufacturer', ascending = False).head(10)
top10_popular_manufacturer=top10_manufacturer.merge(df,left_index = True, right_on = 'manufacturer').drop_duplicates(
    ['manufacturer'])[['manufacturer','ratings_sum_manufacturer']]
print(tabulate(top10_popular_manufacturer, headers = 'keys', tablefmt = 'fancy_grid'))
```

REFERENCES

Aashutosh Bhatt, Ankit Patel, Harsh Chheda, Kiran Gawande, 2015. Amazon Review Classification and Sentiment Analysis. international Journal of Computer Science and Information Technologies, Vol. 6 (6) 2015, 5107-5110.

Adnan, H. (2014). An Analysis of the Factors Affecting Online Purchasing Behavior of Pakistani Consumers. International Journal of Marketing Studies. Vol. 6(5), p.133-148.

Ariff, M.S.M., Yan, N.S., Zakuan, N., Bahari, A.Z., Jusoh, A. (2013). Web-based Factors Affecting Online Purchasing Behavior. IOP Conf. Series: Materials Science and Engineering. Vol. 46, pp. 1-10.

Bell, C., 2021. *E-Commerce Models - Business to Consumer - B2B | B2C | C2B | C2C | B2G | Chris Bell*. [online] Chrisbell.com. Available at: <https://www.chrisbell.com/SNHU/IT-647-website-construction-and-management/ecommerce-models-business-to-consumer-B2C.php>.

Bo Pang, Lillian Lee, and Shivakumar Vaithyanathan. Thumbs up? sentiment classification using machine learning techniques. In Proceedings of the ACL-02 conference on Empirical methods in natural language processingVolume 10, pages 79–86. Association for Computational Linguistics, 2002.

D. Yörük, S. Dündar, L. Moga and M. Neculita, "Drivers and Attitudes towards Online Shopping: Comparison of Turkey with Romania", Communications of the IBIMA, pp. 1-13, 2011. Available: 10.5171/2011.575361.

Forbes, 2021. Amazon, Already The Nation's Top Fashion Retailer, Is Positioned To Grab Even More Market Share. [online] Available at: <https://www.forbes.com/sites/pamdanziger/2020/01/28/amazon-is-readying-major-disruption-for-the-fashion-industry/?sh=5f4415ef67f3> [Accessed 13 September 2021].

Ganu, G., Elhadad, N. & Marian, A. (2009). Beyond the Stars: Improving Ratings Predictions using Review Text Content. WebDB, 1-6.

Hinckley, D. (2015). New Study: Data Reveals 67% of Consumers are influenced by Online Reviews. MOZ. Retrieved from https://moz.com/blog/new-data-reveals-67-ofconsumersare-influenced-by-online-reviews

Inc.com. 2021. 84 *Percent of People Trust Online Reviews As Much As Friends. Here's How to Manage What They See| Inc.com*. [online] Available at: <https://www.inc.com/craig-bloem/84-percent-of-people-trust-online-reviews-as-much-.html>

Irina Rish. An empirical study of the naive bayes classifier. In IJCAI 2001 workshop on empirical methods in artificial intelligence, volume 3, pages 41–46. IBM, 2001.

Jain, A., Kulkarni, G. and Shah, V., 2018. Natural Language Processing. *International Journal of Computer Sciences and Engineering*, 6(1), pp.161-167.

Jayashri Khairnar and Mayura Kinikar. Machine learning algorithms for opinion mining and sentiment classification. International Journal of Scientific and Research Publications, 3(6):1–6, 2013.

Liu, J., Seneff, S. and Zue, V., 2012. Harvesting and Summarizing User-Generated Content for Advanced Speech-Based HCI. *IEEE Journal of Selected Topics in Signal Processing*, 6(8), pp.982-992.

Mitra, Abhijit (2013), "e-commerce in India- a review", International journal of marketing, financial services & management research, vol.2, no. 2, pp. 126-132

Moshrefjavadi, M., Rezaie Dolatabadi, H., Nourbakhsh, M., Poursaeedi, A. and Asadollahi, A., 2012. An Analysis of Factors Affecting on Online Shopping Behavior of Consumers. International Journal of Marketing Studies, 4(5).

Nello Cristianini and John Shawe-Taylor. An introduction to support vector machines and other kernel-based learning methods. Cambridge university press, 2000.

Padraig Cunningham, Matthieu Cord, and Sarah Jane Delany "Supervised learning. In Machine learning techniques for multimedia," pages 21–49. Springer, 2008.

Poornima, P. and Chithra, S., 2019. Optimization of Sensing Time in Cognitive Radio Networks Based on Localization Algorithm. *Sustainable Communication Networks and Application*, pp.38-48.

Qiang Ye, Ziqiong Zhang, and Rob Law. Sentiment classification of online reviews to travel destinations by supervised machine learning approaches. Expert systems with applications, 36(3):6527–6535, 2009.

Veluchamy, A., Nguyen, H., L. Diop, M. and Iqbal, R., 2021. *Comparative Study of Sentiment Analysis with Product Reviews Using Machine Learning and Lexicon-Based Approaches*.

Sebastian Raschka. Naive bayes and text classification i-introduction and theory. arXiv preprint arXiv:1410.5329, 2014.

Sinha, S. and Sandhya, M., 2021. *Analysis of Consumer Reviews by Machine Learning Techniques.*

Search Engine Journal. 2021. *15 Online Review Stats Every Marketer Should Know*. [online] Available at: <https://www.searchenginejournal.com/online-review-statistics/329701/#close> [Accessed 14 September 2021].

Thorsten Joachims. Text categorization with support vector machines: Learning with many relevant features. In European conference on machine learning, pages 137–142. Springer, 1998.