Unfolding Customer Sentiment : A Machine Learning Approach to Sephora's Reviews

MSc Research Project
Data Analytics

Vaibhav Sonia
Student ID: 22136860

School of Computing
National College of Ireland

Supervisor: Taimur Hafeez

National College of Ireland
Project Submission Sheet
School of Computing

Student Name:	Vaibhav Sonia
Student ID:	22136860
Programme:	Data Analytics
Year:	2023
Module:	MSc Research Project
Supervisor:	Taimur Hafeez
Submission Due Date:	20/12/2018
Project Title:	Unfolding Customer Sentiment : A Machine Learning Approach to Sephora's Reviews
Word Count:	5135
Page Count:	21

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 required to use the Referencing Standard specified in the report template. To use other author's written or electronic work is illegal (plagiarism) and may result in disciplinary action.

Signature:	
Date:	14th December 2023

PLEASE READ THE FOLLOWING INSTRUCTIONS AND CHECKLIST:

Attach a completed copy of this sheet to each project (including multiple copies).	□
Attach a Moodle submission receipt of the online project submission, to each project (including multiple copies).	□
You must ensure that you retain a HARD COPY of the project, both for your own reference and in case a project is lost or mislaid. It is not sufficient to keep a copy on computer.	□

Assignments that are submitted to the Programme Coordinator office must be placed into the assignment box located outside the office.

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

Unfolding Customer Sentiment : A Machine Learning Approach to Sephora's Reviews

Vaibhav Sonia
22136860

Abstract

Customer reviews are core attributes of marketing segment and product objectives. Customers express their opinions through reviews either online or in stores. The reviews which are in textual form act as an interactive window between the consumer and business. Therefore, customer reviews are analysed and processed for computation in order to understand customer-enterprise relationship. These computational entities include deploying the available data and leveraging them to derive meaningful insights which acts as a catalyst as customer understanding which leads to business expansion and betterment. In this study, the dataset consists of review windows for the cosmetic retail chain Sephora. The dataset is fetched from Kaggle which is an open-source interactive platform. The models employed for this study are linear regression , logistic regression , decision tree , support vector classification and Naïve Bayes. The deployment of linear regression displays linear dependencies. On the other hand, logistic regression helps with identification of modelling binary entities which is a decision factor. Decision trees offers complex decision making into simpler comprehensive outcomes and support vector classification pioneers complex data structures and patterns. The use of Naïve Bayes gives a probabilistic frameworks for the study. The models contribute to the durable analysis for different parameters. The study aims to investigate the effect of these parameters and introduce the driving factors which unfolds patterns. The experiment is performed by using Python. The resultant outcome of this study not only reflects on the impact of vivid models in exploring the distinct dynamics amongst parameters but also it serves as a directional reference for Sephora in optimizing customer benefits thereby leading towards a successful information which caters business decisions. In other aspects, the study also demonstrates the critical role that customer reviews and algorithmic methodology plays in building the success course of a cosmetic retail chain like Sephora. The inferences drawn are showcased and they display the impact of models in understanding the relationship amongst the parameters.

1 Introduction

The pioneering growth of customer-centered business and marketing is evidently spearheading kinetics. A large amount of data is accessed and transformed as per requirements across almost every domain in the world. Upon the purchase of products, business enterprises seek customer reviews in order to understand their perceptions of products and services. Due to easy access to the internet and interactive interference, companies cater to their business goals and customer needs through customer reviews. It also assists with recommendations, advertisements, and other commercial variables. Consumer reviews are collected, stored, and analysed for marketing, business, and advertisement objectives. Real-time data can be fetched from sources and further studied. The customer reviews are not only computed for qualitative entities, but they're also leveraged for quantity inferences. Businesses adapt to advanced data handling and computing functions in order to maximize their revenue and design their business models in terms of production, distribution, and consumption. Thus, it is imperative for organizations to circulate and intend to administer customer reviews to yield fruitful outcomes and inferences. Upon successful procurement of products and services, businesses look for active customer reviews as vital sources of queries and insights that dive into understanding collective customer perception. The interactive environment provides a platform where enterprises can assess their business objectives while aligning with the expectations and needs of their consumers. The reviews serve as a vibrant feedback system that guides enterprises to refine their services and products, providing overall customer satisfaction by deploying appropriate engagement strategies.Jagdale et al. (2019)

In the sphere of data-driven decision-making, consumer reviews are computed for data handling and computing for analysis as per the business requirements. The circulatory attributes of enterprise establishments and succession depend on consumer needs and expectations on a large-scale basis. The relevance of recommendations is evident for the business ecosystem and it's functioning. It is driven by the wealth of customer feedback, which sheds light on the importance of fostering customer needs and queries through easy access of internet.Cambria et al. (2013) The availability of such platforms has unfolded consumer behavior, which allows individuals to engage and interact with the effortless acquisition of services as per their individual convenience. As a result, online reviews are assumed to play a vital role in the digital transaction circulatory period, acting as bundles of information and data that provide influence in the decision-making process for both consumers and enterprises. The availability of the free market enables the privilege of choices, and recommendations are a practical attribute for making an informed purchase. The approach goes beyond technical specifications to provide objective, grounded inferences on a consumer's intentions and interactions with a specific product. This qualitative attribute provides a degree of authenticity to the decision-making process, enabling consumers to make decisions aligned with the expectations and preferences. The adjoining of scaled sentiment scores with traditional rating entries approaches a new environment to the investigation. By minimizing the gap between subjective sentiment and objective quantitative or numerical entities, the study approaches an understanding of hidden patterns and correlations. The inclusion of sentimental scores and ratings allows for an exploration of user opinions but also displays merging of qualitative and quantitative data streams.*Determining the Link Between Consumer Sentiments and Automobile Sales Through Sentiment Analysis* (2023) Recommendations, being binary in nature, capture the apex of user encounters and act as powerful indicators of satisfaction or dissatisfaction. Unfolding the attributes that impact positive or negative recommendations has constituent capacity for enhancing consumer experiences across different sectors. The study addresses the wider concept of a recommendation system. The extraction of sentiment, quantitative ratings, and recommendations provides a multi-fold approach on user interaction with enterprises and the products and services they offer. By undertaking this approach, the study intends to advance the understanding of user preferences, behaviors, and the decision-making process. The study aims to extract fruitful insights that can drive information to decision-makers, who are the consumers for product consumption, and to production and service entities, who are the enterprises.

1.1 Motivation and Background

People rely on customer reviews to understand the common perception of the products or services. Such innovations empower consumers to browse and choose their options in a very short time frame. This saves time and money and cuts the paradox of options and choices. The accessibility provides a heavy acknowledgement to the consumer critic to choose their specifications as per their wants and needs. The internet expansion is evident in the increasing traffic in goods and services. Organizations run on the availability of the internet to interact with consumers and deal with their queries in every possible aspect. Sentiment analysis can quantify the positive and negative entities of reviews and provide a clear image of the initial perception of the products, goods, and services.Li and Wu (2010) Additionally, modelling of the quantified reviews, which are labelled as per the parameters, can be employed in order to predict commercial necessities. The objective of this study is to understand and study the impact of customer reviews, which is both sentimental and quantified for modelling purposes, and to find out the established relationship between them in order to observe the recommendations. Recommendations play a viral role for enterprises to ensure maximum revenue and excellent customer satisfaction. Recommendations are built on many parameters, and they are regulated across time in order for them to have the highest possible accuracy.

1.2 Research question

How do different predictive models integrated with ratings and quantified sentiment analysis perform in terms of accuracy in predicting recommendations?

1.3 Research objectives

In order to unfold the research question, five supervised learning methods were employed: linear regression, logistic regression, decision trees, support vector classification, and Naive Bayes. The algorithms were deployed and observed for evaluation. The accuracy was compared, and the resultant answer was showcased. The paper structure is as follows: Section 2 has a literature review, which is a collection of studies, and thereby the discussion. Methodology is discussed in Section 3. Evaluation consists of Section 4, which has observations and inferences. Section 5 directs towards the conclusion and future works.

2 Related Work

This segment offers a collective observation of studies that employed sentiment analysis and performed modelling in order to predict outcomes and evaluate whether the inferences were favourable or not. The study performed well, and the datasets used were leveraged to equip an appropriate classification of study papers. The advocated study is primarily distributed into conducting sentiment analysis and opting the most appropriate models which were employed in these research studies, which will be discussed below.

Supervised learning acts as a foundation in the expanding landscape of machine learning. It represents an impactful framework where the algorithms are dormant processors of data points and active learners with the ability to make predictions and perform classifications. With respect to the core of supervised learning, anchors depend on the availability of labelled training data points, where each input entry is paired with a respective target. The characteristic aspect sets it apart from other strata of machine learning, enabling algorithmic entities to perceive patterns and establish relationships within data through a steering approach. The algorithm selection is based on the context and the extraction of inferences.Singh et al. (2016) Different supervised learning algorithms perform vividly under supervised learning, which makes it flexible as per requirements. Research studies often compare the machine learning models to evaluate the best fit as per requirements.Saifullah et al. (2021) The choice of algorithm is primarily associated with the type of problem. As per this study requirement, the most appropriate supervised learning models are discussed, along with the principles and applications they constitute. The landscape of supervised learning is diverse, with each algorithm deployed with its own set of principles and applications. As the principles and respective applications of supervised learning algorithms are observed, a fine understanding of their strengths, limitations, and real-time applications is acknowledged.

Sabapathi and Kaliyamurthie (2022) explains in a study that acknowledges the gaps in feature identification, selection, and further evaluation. The feature extraction process is crucial for retrieving meaningful insights. A study revealed that logistic regression outperformed VADER, which is valence aware dictionary, and sentiment reasoner by giving an accuracy of 0.79 by studying sentiment analysis in the context of airline tweets. The primary reason could be the classification factor of logistic regression Dhanalakshmi et al. (2023). The model performed well in the study due to it's course of classifying the sentiment of tweets, which was in textual form. Sentiment analysis was enhanced under this study to conduct modeling based on sentiment analysis, which showcased an accuracy of 0.78.Ajmain et al. (2022). The decision tree's performance unfolds after extracting attributes from textual data and modeling them with appropriate models.

The effectiveness of decision tree can be acknowledged by it's capacity to recognise patterns and aspects that aids to identifying sentiment classification and introducing decision boundaries. A study on understanding the sentiment based on smoking and the perception around it Mizan Khairul Anwar et al. (2022) observed that the decision tree model had an inference of 0.70 accuracy, which implies the ability of model to deal with polarity. The data set was classified, and then inferences were observed. The process of leveraging sentiment analysis is an automated mining of perceptions, opinions, and emotions from many sources of expression, like texts, reviews, social media, and database sources, through natural language processing. Kharde et al. (2016) The study enhances the availability of data in terms of many attributes represented in textual form, which are further processed for conducting sentiment analysis. The data in textual forms is extracted using natural language processing Astuti et al. (2022a). Naive Bayes classifies the sentiment as positive or negative, which displays the Naive Bayes modeling ability and computational efficiency.Astuti et al. (2022b) The outcome showcased adequate insight for classification, proving to be a good measure to conduct similar studies.

Naive Bayes is applicable for huge sets of data in textual form Troussas et al. (2013) with instances like social media and product review windows. Assigning a score to sentiment value is a good measure to study the impact of support vector classifier with n-gram folds. Naz et al. (2018) The study showcased classified sentiments equipped with support vector classifiers. Unigram with support vector classifiers displayed more accuracy compared to unigram without support vector classifiers. This is evident to the principle of support vector classifiers performing well under non-linear boundaries. In the study of European Conference on Machine learning , the discussion of support vector machine classifier tooled to train labelled data for studied involving sentiment and modeling was conducted.Joachims (1998). With the help of count vector, linear regression established a good relationship with sentiment analysis, such as rating systems, reviews, and inferences, which showcased a positive correlation between the indices.Sri Patibandla et al. (2023). A study conducted Li et al. (2020) proposed an approach for sentiment analysis of Chinese language knowledge using the sentiment information-based network model (SINM), which leverages LSTM as model components. Using a Chinese emotional dictionary, the texts were assigned as per their sentiment attribute.

3 Methodology

This section is a collection of the different methodologies followed that can be leveraged for the impact of text classification and modelling on the provided dataset. After careful examination of previous studies and papers, it was observed that different machine learning concepts such as logistic regression, support vector machines, Naïve Bayes, and decision trees were used. This study will also use them to understand and perform a detailed comparison. This study will also perform linear regression in order to understand the linearity, if any. This section is a systematic approach to unfolding the attributes of customer reviews and blending them with machine learning models to extract and compare detailed insights.

3.1 Data understanding

The dataset sourced for this study is from Kaggle,an interactive source for datasets to conduct.

Source: https://www.kaggle.com/datasets/nadyinky/sephora-products-and-skincare-reviews

The dataset consists of two sections. The first section is product information, with many columns displaying product details. The second section of the dataset consists of reviews in textual form, along with many other attributes. The required columns are listed below for better understanding.

reviews_1250-end.csv : This dataset contains entries from product reviews, which will be further leveraged to perform analysis and modeling.

product_info.csv : This dataset contains entries of product details

Below is distribution of columns:

- **rating:** The ratings provided by consumers for individual product on a scale of 1 to 5.
- **is_recommended:** Binary entities, with 1 being recommended and 0 being the contrary.
- **review_text:** Reviews left by consumers for products.
- **review_title:** The title of the reviews.
- **skin_type:** The skin type of each consumer who left a review.
- **hair_colour:** The hair colour of each consumer who left a review.
- **product_id:** The unique identifying component for each product.

3.2 Data processing

The complete data processing and analysis were performed using the Python programming language with a Jupyter notebook. The data was uploaded to a Jupyter notebook with the Python programming language.

- **Extracting Required Columns:** Identify and extract the columns from the dataset that are necessary for your analysis.
- **Checking and Removing Missing Values:** Perform a thorough check for missing values in the dataset and implement appropriate strategies, such as imputation or removal, to handle them.
- **Checking and Removing Null Values:** Examine the dataset for null values and apply suitable methods, like imputation or elimination, to deal with them.

The study requires the parameters of ratings, recommendations, review in textual format, name of product, and name of brand. The respective columns were called and listed.There were 3817 entries for recommended and 59 entries for reviews that were of no use for the study, and they were eliminated as part of the data preprocessing.The dropna.() function is utilized to eliminate unwanted entries.

A quick descriptive visualization was conducted to glimpse the data distribution in terms of a pie chart distribution and bar plot of the most preferred products and brands.

Figure 1 showcases the distribution of brands in the study.

3.3 Sentiment Analysis

Sentiment analysis is a collective study that optimizes computational processes to observe, examine, and unfold the sentiments and emotions underlying textual formats or interactions. It involves steps of mining textual data and leveraging it into meaningful insights and inferences.Medhat et al. (2014)

The steps are as follows

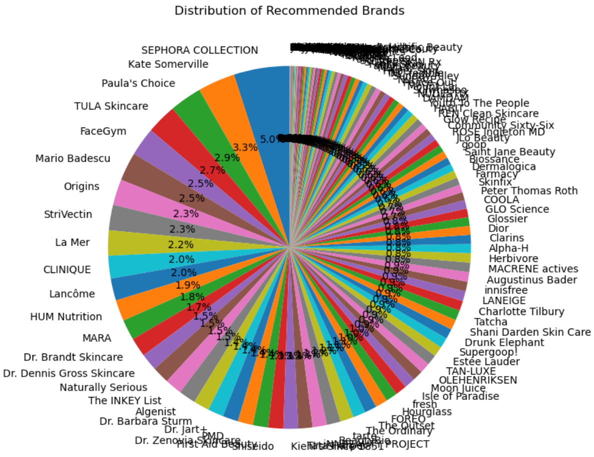


Figure 1: Distribution of brands

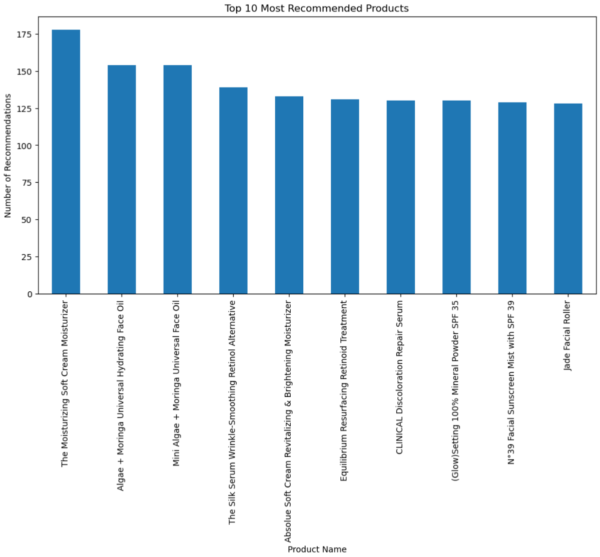


Figure 2: Most preferred products

- **Converting to lowercase:** Converting to lowercase for further processing.
- **Tokenizing:** Splitting the words into units which are termed as tokens.
- **Removing stop words:** Words such as 'a' and 'the' are removed from word queues to enhance performance.
- **Removing punctuations:** Punctuations are removed to have cleaner extraction.
- **Stemming:** The dimensions are reduced to have sharper extraction.
- **Lemmatization:** It identifies the emotional context of the query text.

3.4 Sentiment scores

For sentiment scoring, 'SentimentIntensityAnalyzer' which is a class provided by VADER, is employed. VADER - Valence-Aware Dictionary and Sentiment Reasoner is a tool that

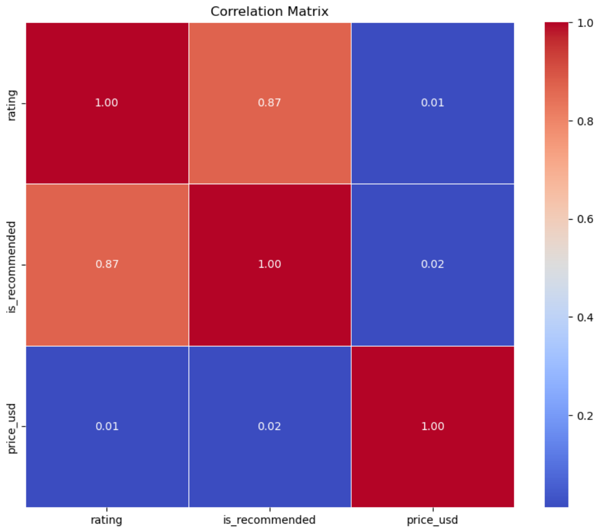


Figure 3: Correlation matrix

combines lexicon based on sentiment and grammatical entities to fetch the sentiment of a text.The gathered data was represented as sentiment scores which was performed under VADER. It is a sentiment analysis tool designed to perform sentiment breakdown. It is used in natural language processing entities.Stanley et al. (2023)VADER assigns sentiment scores to words based on their emotional context. The VADER outstands in performance with the lexicon and the inclusion of phrases and emoticons. The tool has the capability to distribute awareness of sentiment intensity. Considering the polarity of sentiment intensity, it elaborates on the scope of sentiment decoding Krishna et al. (2023). The most extensive instance of leveraging VADER is extraction from social media to understand common perceptions, customer feedback through review windows, and other domains. The limitations VADER faces is of not being able to adapt with processing sarcasm or ironyStanley et al. (2023).The reviews are assigned a sentiment score and displayed to understand the sentiment in quantified contexts. The tweaked sentiment values are easy to understand and can be used for further computation.

3.5 Scaling

The sentiment score had negative entries, which would make it hard to compute and draw inferences. Therefore, the entries underwent scaling, which is the process of normalizing the range of features in the entries. The ‘cleaned$_{r}eview_{d}f'dataframehadcompoundscoresthatwerescale$

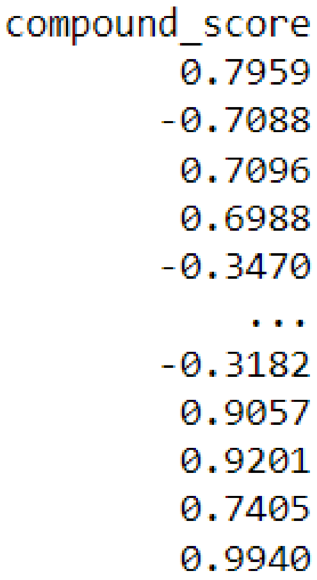
$$((compound_{s}core + 1)/2) * 9 + 1. \tag{1}$$

After successful scaling of the revised text scores, the scaled score exists within the range of 1–10, all being positive entities.

Below is the representation of scaled score:

3.6 Data Modeling

To establish modeling, a depth analysis of previous works was studied. These studies delivered extensive knowledge and understanding of the machine learning models, along



```
compound_score
        0.7959
       -0.7088
        0.7096
        0.6988
       -0.3470
           ...
       -0.3182
        0.9057
        0.9201
        0.7405
        0.9940
```

Figure 4: Compound score

with appropriate inferences to support the research query.After a careful glossary of studies, the most appropriate models chosen for this study are linear regression, logistic regression, decision trees, support vector classification, and Naive Bayes. Along with these, VADER was used to perform detailed sentiment analysis, as discussed above.The model accuracy are then compared to undertsand the best model performance and the factors that drive its accuracy.

Linear Regression

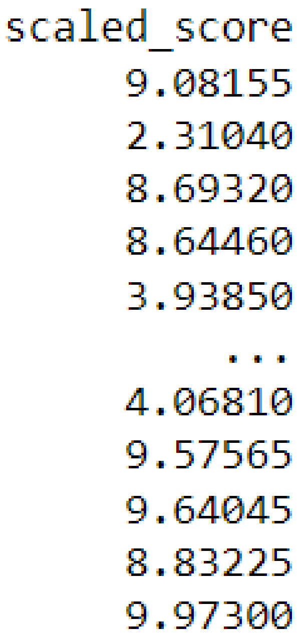
Linear regression is performed to explore underlying linear relationships within customer reviews, enabling a detailed examination of specified factors that may impact business dynamics.

Logistic Regression

Logistic regression is considered for its ability to model binary entities, providing crucial binary outcomes in terms of recommendations and customer reviews. It can unfold factors impacting customer reviews, offering insights into whether certain attributes create positive or negative impacts.

Decision Tree

Decision trees are employed to minimize the complexity of the decision-making process. By offering an unfolding of complex patterns, decision trees provide a transparent understanding of the attributes contributing to the study.



```
scaled_score
     9.08155
     2.31040
     8.69320
     8.64460
     3.93850
         ...
     4.06810
     9.57565
     9.64045
     8.83225
     9.97300
```

Figure 5: Scaled sentiment score

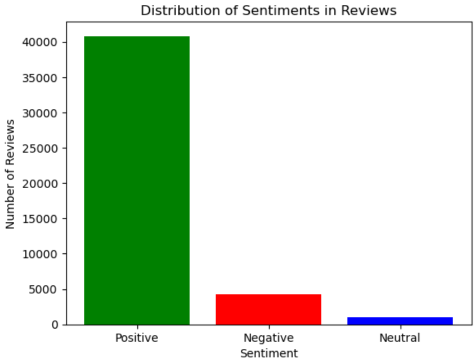


Figure 6: Sentiment distribution

Support Vector Classification

Support vector classification is applicable to the analysis of complex data structures and patterns. The model excels at exploring relationships that can provide insights.

3.7 Model Evaluation

The dataset used to train the models is further split into train and test sets with an 80% distribution for training and 20% for testing. The performance of each model is evaluated using accuracy, precision, recall, and F1 score values.

The evaluation metrics are defined as:

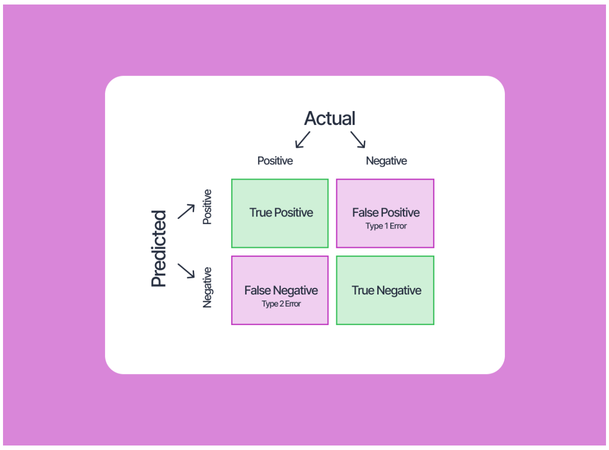


Figure 7: Evaluation matrix

- **Accuracy:** It measures the correctness of the model. It is defined as the total number of correct predictions made by the model, divided by the total predictions.
- **Precision:** Precision is the percentage of positive predictions from the total predicted positive instances. It calculates the accuracy for the minority class.
- **Recall:** Recall measures the model's capability to detect positive specimens. It provides an indication of missed positive predictions.
- **F1 score:** F1 score is the harmonic mean of precision and recall. The contribution of precision and recall makes it a balanced score, and the model's performance can be considered better with a higher F1 score.

4 Design Specification

The design specification for the study is shown in figure [8]. The data for the study was sourced from kaggle. For this study, many packages in the Python programming language were leveraged as per the configuration of data frame and inferences needed. From Kaggle the data source, the data was scrapped and stored as a csv file and jupyter notebook is used to conduct the study. The dataset had many columns which were not significant hence they were eliminated. The data was represented as a dataframe and then converted to corpus for text analysis methods including elimination of punctuation , turning the words into lowercase , stemming , lemmatization and removing other unwanted entities.

The gathered data was represented as sentiment scores which was performed under VADER. The sentiment score had negative entries which would make it hard to computing and conducting inferences. Therefore, the entries underwent scaling which is the process of normalizing the range of features in entries. The data frame was converted from compound scores to scaled within the range of 1-10. The scaled score was ready to compute and drawn to conduct further analysis.

The data was modeled using different python programming languages after installation and import.

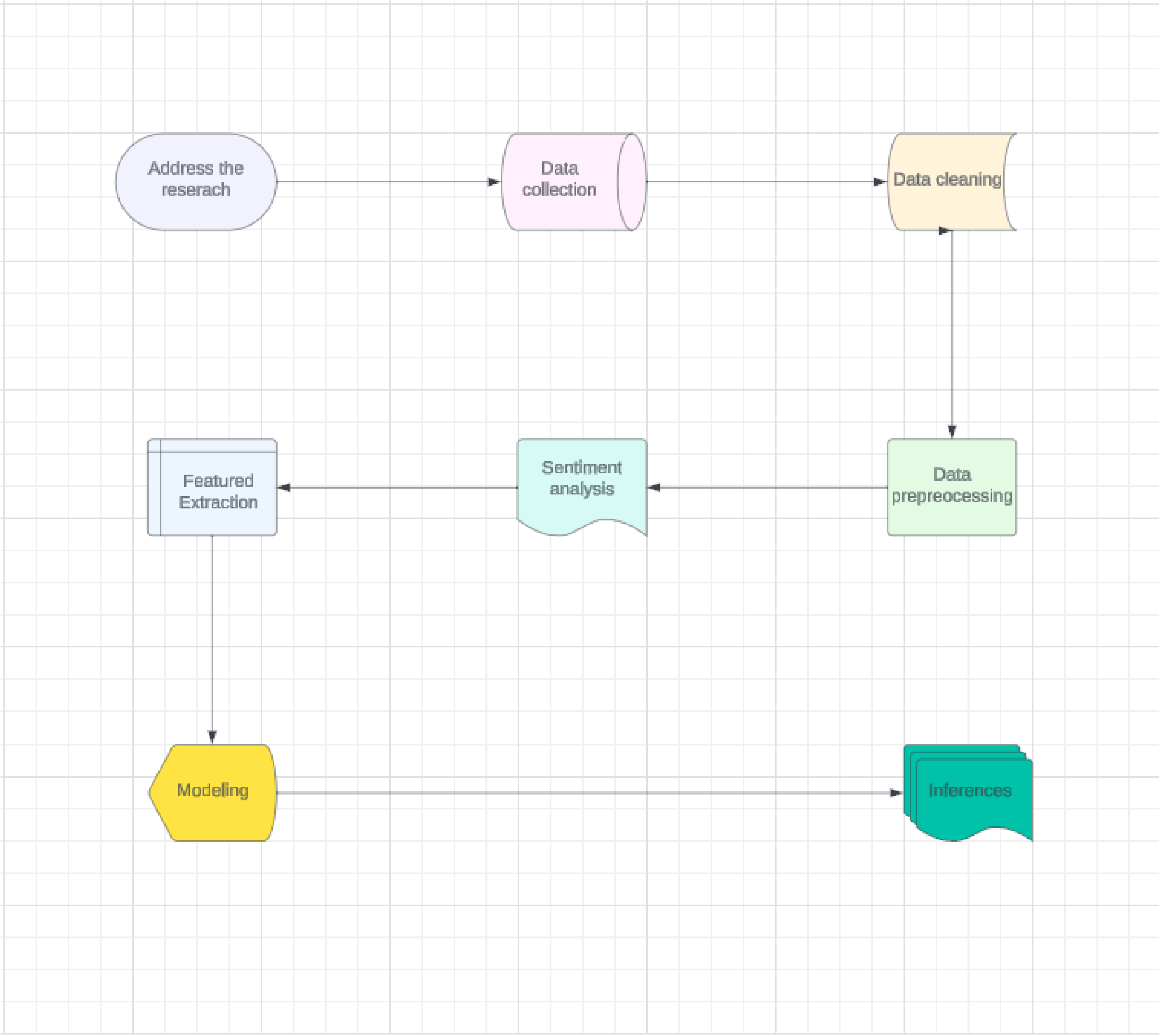


Figure 8: Design specification

5 Implementation

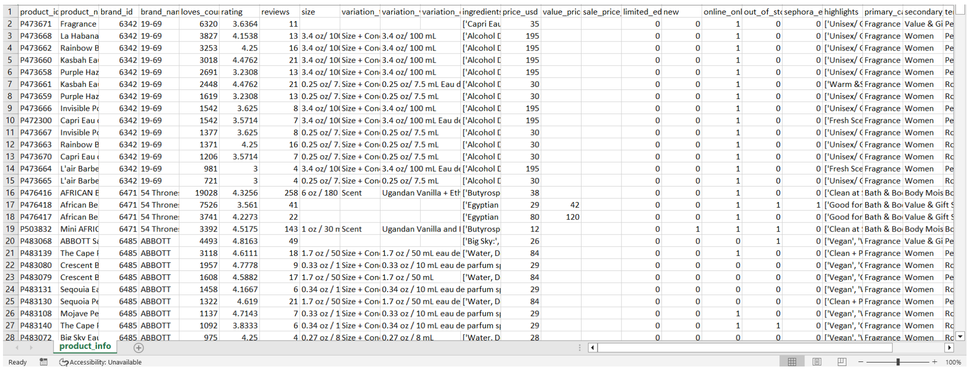
5.1 Data collection and pre processing

5.1.1 Data collection:

The data in this study can be fetched from interactive source Kaggle. As displayed in figure. There are about two csv files which represents product information and customer reviews respectfully. In this study , there were multiple entries of customer review and the csv file of review window named '1250-end' is used because it had adequate number of entries for the study. Figure 9 and 10 represent the csv files.

5.1.2 Removing unexpected entries:

Figure 11 provides column entries as raw data which had unexpected entries and columns. The unexpected entries and columns were eliminated.

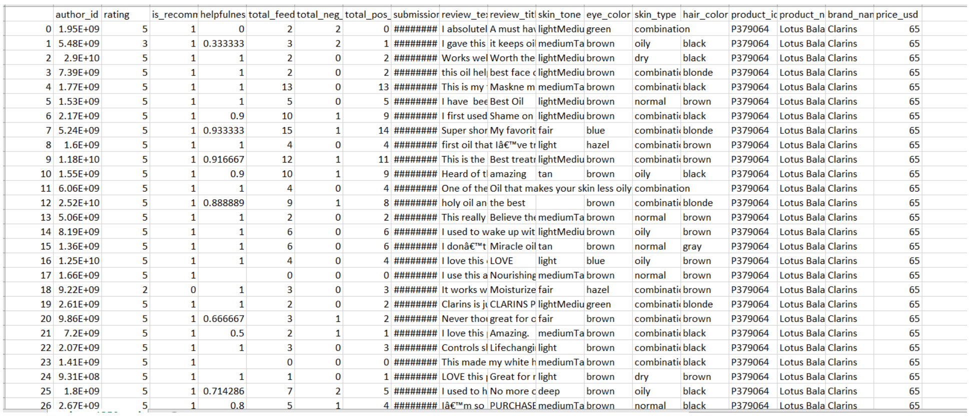


1	product_ic	product_n	brand_id	brand_nan	loves_cou	rating	reviews	size	variation_t	variation_v	variation_d	ingredients	price_usd	value_pric	sale_price_	limited_ed	new	online_onl	out_of_st	sephora_e	highlights	primary_ca	secondary_	te
2	P473671	Fragrance	6342	19-69	6320	3.6364	11					['Capri Eau	35			0	0	1	0	0	['Unisex/ C	Fragrance	Value & Gi	Pe
3	P473668	La Habana	6342	19-69	3827	4.1538	13	3.4 oz/ 10(	Size + Con	3.4 oz/ 100 mL		['Alcohol D	195			0	0	1	0	0	['Unisex/ C	Fragrance	Women	Pe
4	P473662	Rainbow B	6342	19-69	3253	4.25	16	3.4 oz/ 10(	Size + Con	3.4 oz/ 100 mL		['Alcohol D	195			0	0	1	0	0	['Unisex/ C	Fragrance	Women	Pe
5	P473660	Kasbah Ea	6342	19-69	3018	4.4762	21	3.4 oz/ 10(	Size + Con	3.4 oz/ 100 mL		['Alcohol D	195			0	0	1	0	0	['Unisex/ C	Fragrance	Women	Pe
6	P473658	Purple Haz	6342	19-69	2691	3.2308	13	3.4 oz/ 10(	Size + Con	3.4 oz/ 100 mL		['Alcohol D	195			0	0	1	0	0	['Unisex/ C	Fragrance	Women	Pe
7	P473661	Kasbah Ea	6342	19-69	2448	4.4762	21	0.25 oz/ 7.	Size + Con	0.25 oz/ 7.5 mL Eau d		['Alcohol D	30			0	0	1	0	0	['Warm &S	Fragrance	Women	Ro
8	P473659	Purple Haz	6342	19-69	1619	3.2308	13	0.25 oz/ 7.	Size + Con	0.25 oz/ 7.5 mL		['Alcohol D	30			0	0	1	0	0	['Unisex/ C	Fragrance	Women	Ro
9	P473666	Invisible Po	6342	19-69	1542	3.625	8	3.4 oz/ 10(	Size + Con	3.4 oz/ 100 mL		['Alcohol D	195			0	0	1	0	0	['Unisex/ C	Fragrance	Women	Pe
10	P472300	Capri Eau (	6342	19-69	1542	3.5714	7	3.4 oz/ 10(	Size + Con	3.4 oz/ 100 mL Eau de		['Alcohol D	195			0	0	1	0	0	['Fresh Sce	Fragrance	Women	Pe
11	P473667	Invisible Po	6342	19-69	1377	3.625	8	0.25 oz/ 7.	Size + Con	0.25 oz/ 7.5 mL		['Alcohol D	30			0	0	1	0	0	['Unisex/ C	Fragrance	Women	Ro
12	P473663	Rainbow B	6342	19-69	1371	4.25	16	0.25 oz/ 7.	Size + Con	0.25 oz/ 7.5 mL		['Alcohol D	30			0	0	1	0	0	['Unisex/ C	Fragrance	Women	Ro
13	P473670	Capri Eau (	6342	19-69	1206	3.5714	7	0.25 oz/ 7.	Size + Con	0.25 oz/ 7.5 mL		['Alcohol D	30			0	0	1	0	0	['Unisex/ C	Fragrance	Women	Ro
14	P473664	L'air Barbe	6342	19-69	981	3	4	3.4 oz/ 10(	Size + Con	3.4 oz/ 100 mL Eau de		['Alcohol D	195			0	0	1	0	0	['Fresh Sce	Fragrance	Women	Pe
15	P473665	L'air Barbe	6342	19-69	721	3	4	0.25 oz/ 7.	Size + Con	0.25 oz/ 7.5 mL		['Alcohol D	30			0	0	1	0	0	['Unisex/ C	Fragrance	Women	Ro
16	P476416	AFRICAN B	6471	54 Thrones	19028	4.3256	258	6 oz / 180	Scent	Ugandan Vanilla + Eth		['Butyrosp	38			0	0	1	0	0	['Clean at S	Bath & Bo	Body Mois	Bo
17	P476418	African Be	6471	54 Thrones	7526	3.561	41					['Egyptian	29	42		0	0	1	1	1	['Good for	Bath & Bo	Value & Gift S	
18	P476417	African Be	6471	54 Thrones	3741	4.2273	22					['Egyptian	80	120		0	0	1	0	0	['Good for	Bath & Bo	Value & Gift S	
19	P503832	Mini AFRIC	6471	54 Thrones	3392	4.5175	143	1 oz / 30 m	Scent	Ugandan Vanilla and I		['Butyrosp	12			0	1	1	1	0	['Clean at S	Bath & Bo	Body Mois	Bo
20	P483068	ABBOTT Sa	6485	ABBOTT	4493	4.8163	49					['Big Sky:',	26			0	0	0	1	0	['Vegan', 'V	Fragrance	Value & Gi	Pe
21	P483139	The Cape F	6485	ABBOTT	3118	4.6111	18	1.7 oz / 50	Size + Con	1.7 oz / 50 mL eau de		['Water, D	84			0	0	1	0	0	['Clean + P	Fragrance	Women	Pe
22	P483080	Crescent B	6485	ABBOTT	1957	4.7778	9	0.33 oz / 1	Size + Con	0.33 oz / 10 mL eau de parfum sp			29			0	0	0	0	0	['Vegan', 'C	Fragrance	Women	Ro
23	P483079	Crescent B	6485	ABBOTT	1608	4.5882	17	1.7 oz / 50	Size + Con	1.7 oz / 50 mL		['Water, D	84			0	0	0	0	0	['Vegan', 'C	Fragrance	Women	Pe
24	P483131	Seqouia Ea	6485	ABBOTT	1458	4.1667	6	0.34 oz / 1	Size + Con	0.34 oz / 10 mL eau de parfum sp			29			0	0	0	0	0	['Vegan', 'V	Fragrance	Women	Ro
25	P483130	Sequoia Pe	6485	ABBOTT	1322	4.619	21	1.7 oz / 50	Size + Con	1.7 oz / 50 mL eau de		['Water, D	84			0	0	0	0	0	['Clean + P	Fragrance	Women	Pe
26	P483108	Mojave Pe	6485	ABBOTT	1137	4.7143	7	0.33 oz / 1	Size + Con	0.33 oz / 10 mL eau de parfum sp			29			0	0	1	0	0	['Vegan', 'V	Fragrance	Women	Ro
27	P483140	The Cape F	6485	ABBOTT	1092	3.8333	6	0.34 oz / 1	Size + Con	0.34 oz / 10 mL eau de parfum sp			29			0	0	1	1	0	['Vegan', 'C	Fragrance	Women	Ro
28	P483072	Big Sky Eau	6485	ABBOTT	975	4.25	4	0.27 oz / 8	Size + Con	0.27 oz / 8 mL		['Water, D	28			0	0	0	0	0	['Vegan', 'V	Fragrance	Women	Ro

product_info

Ready Accessibility: Unavailable 100%

Figure 9: Product info csv



	author_id	rating	is_recomm	helpfulnes	total_feed	total_neg_	total_pos_	submissior	review_tex	review_titl	skin_tone	eye_color	skin_type	hair_color	product_ic	product_n	brand_nan	price_usd
0	1.95E+09	5	1	0	2	2	0	########	I absolutel	A must hav	lightMediu	green	combination		P379064	Lotus Bala	Clarins	65
1	5.48E+09	3	1	0.333333	3	2	1	########	I gave this	it keeps oil	mediumTa	brown	oily	black	P379064	Lotus Bala	Clarins	65
2	2.9E+10	5	1	1	2	0	2	########	Works wel	Worth the	lightMediu	brown	dry	black	P379064	Lotus Bala	Clarins	65
3	7.39E+09	5	1	1	2	0	2	########	this oil hel	best face c	lightMediu	brown	combinati	blonde	P379064	Lotus Bala	Clarins	65
4	1.77E+09	5	1	1	13	0	13	########	This is my	Maskne m	mediumTa	brown	combinati	black	P379064	Lotus Bala	Clarins	65
5	1.53E+09	5	1	1	5	0	5	########	I have bee	Best Oil	lightMediu	brown	normal	brown	P379064	Lotus Bala	Clarins	65
6	2.17E+09	5	1	0.9	10	1	9	########	I first used	Shame on	lightMediu	brown	combinati	black	P379064	Lotus Bala	Clarins	65
7	5.24E+09	5	1	0.933333	15	1	14	########	Super shor	My favorit	fair	blue	combinati	blonde	P379064	Lotus Bala	Clarins	65
8	1.6E+09	5	1	1	4	0	4	########	first oil that Iâ€™ve tr		light	hazel	combinati	brown	P379064	Lotus Bala	Clarins	65
9	1.18E+10	5	1	0.916667	12	1	11	########	This is the	Best treatr	lightMediu	brown	combinati	brown	P379064	Lotus Bala	Clarins	65
10	1.55E+09	5	1	0.9	10	1	9	########	Heard of tl	amazing	tan	brown	oily	black	P379064	Lotus Bala	Clarins	65
11	6.06E+09	5	1	1	4	0	4	########	One of the	Oil that makes your skin less oily			combination		P379064	Lotus Bala	Clarins	65
12	2.52E+10	5	1	0.888889	9	1	8	########	holy oil an	the best		brown	combinati	blonde	P379064	Lotus Bala	Clarins	65
13	5.06E+09	5	1	1	2	0	2	########	This really	Believe the	mediumTa	brown	normal	brown	P379064	Lotus Bala	Clarins	65
14	8.19E+09	5	1	1	6	0	6	########	I used to wake up wit		lightMediu	brown	oily	brown	P379064	Lotus Bala	Clarins	65
15	1.36E+09	5	1	1	6	0	6	########	I donâ€™t	Miracle oil	tan	brown	normal	gray	P379064	Lotus Bala	Clarins	65
16	1.25E+10	5	1	1	4	0	4	########	I love this	LOVE	light	blue	oily	brown	P379064	Lotus Bala	Clarins	65
17	1.66E+09	5	1		0	0	0	########	I use this a	Nourishing	mediumTa	brown	normal	brown	P379064	Lotus Bala	Clarins	65
18	9.22E+09	2	0	1	3	0	3	########	It works w	Moisturize	fair	hazel	combinati	brown	P379064	Lotus Bala	Clarins	65
19	2.61E+09	5	1	1	2	0	2	########	Clarins is ju	CLARINS P	lightMediu	green	combinati	blonde	P379064	Lotus Bala	Clarins	65
20	9.86E+09	5	1	0.666667	3	1	2	########	Never thou	great for c	fair	brown	combinati	brown	P379064	Lotus Bala	Clarins	65
21	7.2E+09	5	1	0.5	2	1	1	########	I love this	Amazing.	mediumTa	brown	combinati	black	P379064	Lotus Bala	Clarins	65
22	2.07E+09	5	1	1	3	0	3	########	Controls sl	Lifechangi	light	brown	combinati	black	P379064	Lotus Bala	Clarins	65
23	1.41E+09	5	1		0	0	0	########	This made my white h		mediumTa	brown	combinati	black	P379064	Lotus Bala	Clarins	65
24	9.31E+08	5	1	1	1	0	1	########	LOVE this	Great for r	light	brown	dry	brown	P379064	Lotus Bala	Clarins	65
25	1.8E+09	5	1	0.714286	7	2	5	########	I used to h	No more c	deep	brown	oily	black	P379064	Lotus Bala	Clarins	65
26	2.67E+09	5	1	0.8	5	1	4	########	Iâ€™m so	PURCHASE	mediumTa	brown	normal	black	P379064	Lotus Bala	Clarins	65

Figure 10: csv file containing reviews

5.1.3 Missing values:

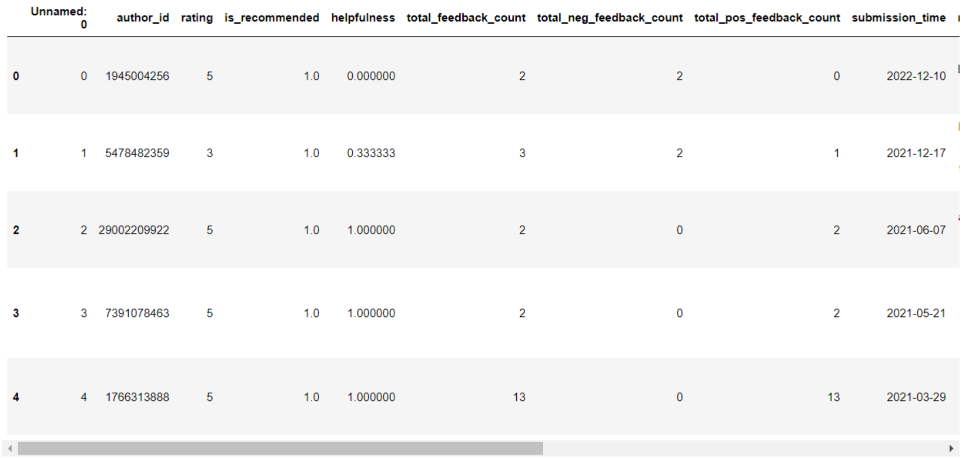
Missing values create an impact with the model performance. Missing values namely NaN can be replaced by 0 or eliminated. In this study , the missing values were eliminated . There are options such as replacing with mean value. However, the decision to eliminate them was taken. Figure 12 illustrates the unexpected values.

5.1.4 Data Exploration:

To understand the nature of text embedded in the reviews, and to have an initial perception of the reviews, a WordCloud is generated. From figure 13 it can be observed that the most frequent words were 'skin','product','use','face','love','serum','using'.

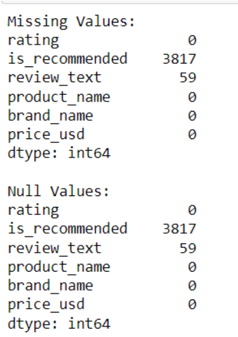
5.1.5 Feature Selection:

The dataset has many features, however not all features are vital for the study. Therefore , feature selection can help with understanding the most important features and amalgamate other required features. Figure [3] is matrix showing relationships between variables. The figure observation explains a good relationship between recommended and ratings. It seems there is no significant relationship between price and ratings



	Unnamed: 0	author_id	rating	is_recommended	helpfulness	total_feedback_count	total_neg_feedback_count	total_pos_feedback_count	submission_time
0	0	1945004256	5	1.0	0.000000	2	2	0	2022-12-10
1	1	5478482359	3	1.0	0.333333	3	2	1	2021-12-17
2	2	29002209922	5	1.0	1.000000	2	0	2	2021-06-07
3	3	7391078463	5	1.0	1.000000	2	0	2	2021-05-21
4	4	1766313888	5	1.0	1.000000	13	0	13	2021-03-29

Figure 11: Overview of raw data in jupyter notebook



```
Missing Values:
rating              0
is_recommended   3817
review_text        59
product_name        0
brand_name          0
price_usd           0
dtype: int64

Null Values:
rating              0
is_recommended   3817
review_text        59
product_name        0
brand_name          0
price_usd           0
dtype: int64
```

Figure 12: Displaying the missing and null values

5.2 Model:

After careful study of literature and understanding the dynamics of study, modelling is performed through selected features . The features used in for modelling will be rating, customer reviews through texts and binary entity of recommendations. The new dataframe computable for modelling and analysis has 80% train data and 20% test data. Sklearn.model_selection imports train_test_split and deployed for splitting the data entries. The models are evaluated using metrics like accuracy, confusion matrix and other metrics.

5.2.1 Linear Regression :

Linear regression model is used to model the relationship between variables namely a dependent variable and one or multiple independent variables. The model assumes a linear relationship between assigned input and output. In this study ,linear regression model performed and inferences of Mean Square Error was 0.00359. This inference leaves room for more inscrutability. Figure 14 shows the resultant outcome of linear regression model.

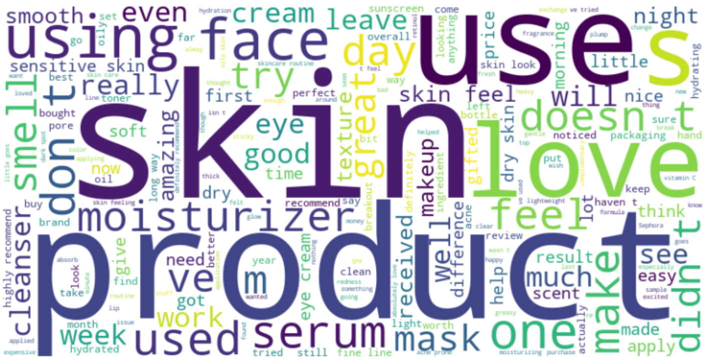
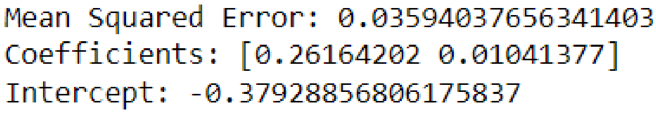


Figure 13: WordCloud

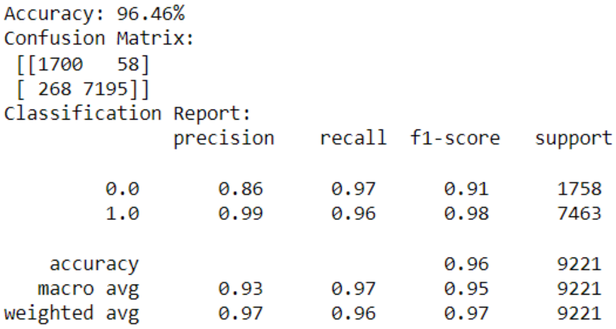


```
Mean Squared Error: 0.03594037656341403
Coefficients: [0.26164202 0.01041377]
Intercept: -0.37928856806175837
```

Figure 14: Linear regression result

5.3 Logistic Regression :

Logistic regression model is an instance of a classification model that provided in the context of conditional probability distribution. For the study , sklearn package is leveraged to carry logistic regression. The accuracy of logistic regression was 96.46 %. Figure 15 illustrates the execution and result.

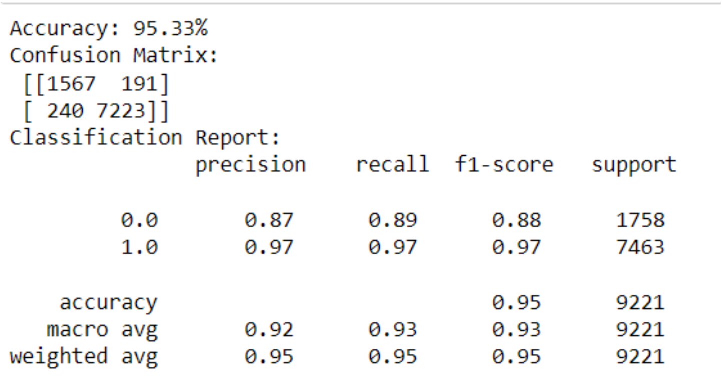


```
Accuracy: 96.46%
Confusion Matrix:
 [[1700   58]
 [ 268 7195]]
Classification Report:
              precision    recall  f1-score   support

         0.0       0.86      0.97      0.91      1758
         1.0       0.99      0.96      0.98      7463

    accuracy                           0.96      9221
   macro avg       0.93      0.97      0.95      9221
weighted avg       0.97      0.96      0.97      9221
```

Figure 15: logistic regression result

5.4 Decision tree:

Decision tree modelling predicts the target variable with structure being represented as a set of decision rules which directs towards classification. In this study the decision tree model performed an accuracy of 95.3%. Figure 16 gives the illustration.

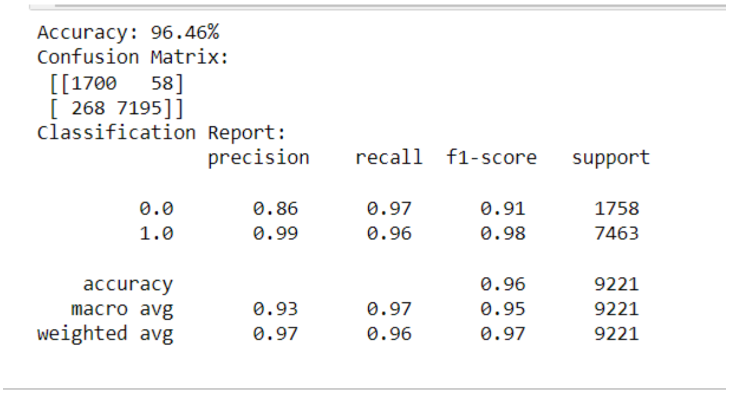


```
Accuracy: 95.33%
Confusion Matrix:
 [[1567  191]
 [ 240 7223]]
Classification Report:
              precision    recall  f1-score   support

         0.0       0.87      0.89      0.88      1758
         1.0       0.97      0.97      0.97      7463

    accuracy                           0.95      9221
   macro avg       0.92      0.93      0.93      9221
weighted avg       0.95      0.95      0.95      9221
```

Figure 16: Decision tree result

5.5 Support Vector Classification:

Support vector classification is designed to carry binary classification entities which aims to look for a hyperplane that segregates datapoints into different classes. Support vector classifications performs well to deal with complex and high dimensional data. In this study the accuracy for support vector classification is 96.46%. Figure 17 represents the execution and result.



```
Accuracy: 96.46%
Confusion Matrix:
 [[1700   58]
 [ 268 7195]]
Classification Report:
              precision    recall  f1-score   support

         0.0       0.86      0.97      0.91      1758
         1.0       0.99      0.96      0.98      7463

    accuracy                           0.96      9221
   macro avg       0.93      0.97      0.95      9221
weighted avg       0.97      0.96      0.97      9221
```

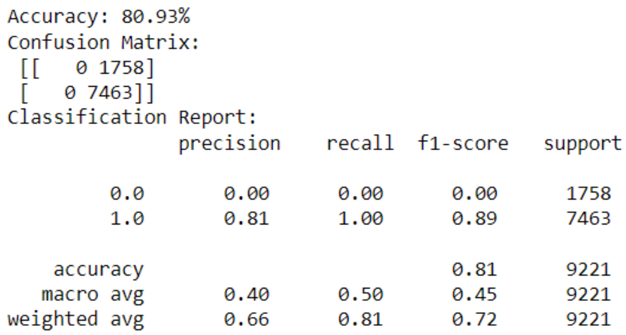
Figure 17: Support vector classification result

5.6 Naive Bayes:

Naive Bayes is based on Bayes theorem. It is a probabilistic classification algorithm. Naive Bayes algorithm is preferred to read and explore text classification tasks which involves identifying sentiment. Naive Bayes performs well with huge data and has good computational abilities. In this study Naïve Bayes performed with an accuracy of 80.93%. The illustration can be observed in figure 18.

6 Evaluation

The performance of algorithm was observed and the models performed will be evaluated by accuracy, precision , recall and F1- score.



```
Accuracy: 80.93%
Confusion Matrix:
 [[   0 1758]
 [   0 7463]]
Classification Report:
              precision    recall  f1-score   support

         0.0       0.00      0.00      0.00      1758
         1.0       0.81      1.00      0.89      7463

    accuracy                           0.81      9221
   macro avg       0.40      0.50      0.45      9221
weighted avg       0.66      0.81      0.72      9221
```

Figure 18: Naive Bayes result

Algorithm	Accuracy (%)	Precision	Recall	F1-score
Logistic Regression	96.46	0.99	0.96	0.98
Decision Tree	95.33	0.95	0.95	0.95
Support Vector Classification	96.46	0.97	0.96	0.97
Naive Bayes	80.93	0.66	0.81	0.72

Table 1: Performance Metrics of Classification Algorithms

6.1 Experiment /Linear regression

Linear regression was an experimental approach however it did not perform to meet any of the expectations. It has Mean square error of 0.035940.

6.2 Experiment / Logistic Regression

From table 1 it is evident that logistic regression perform with satisfactory accuracy of 96.46% with precision and recall being 0.99 and 0.96 respectively.

6.3 Experiment / Decision Tree

Decision tree experiment showcased an accuracy of 95.33% and the precision and recall values were 0.97 and 0.97 respectively

6.4 Experiment / Case Study Support Vector Classification

: The accuracy for this experiment was 96.46% with precision and recall being 0.99 and 0.96 respectively

6.5 Experiment / Naive Bayes

: The experiment observed an accuracy of 80.93% with precision and recall being 0.81 and 1 respectively.

6.6 Discussion

The experiments were conducted involving vivid machine learning entities linear regression , logistic regression , decision tree , support vector classification and Naive Bayes.

The models inferenced gave opportunity of discussion. These models are applicable for many real-world problems and the evaluation of these models based on evaluation metrics are optimized for solutions and computations.

Linear regression was an experimental approach to the study. The model was deployed to examine if there is any significant linear relationship between variables. It is noted that linear regression model did not meet the expectations . The mean square error MSE of 0.003590 suggests non-significant deviation from expected values. Logistic regression encouraged with promising outcome. With accuracy of 96.46% , it showcased an impressive performance. The precision and recall values of 0.99 and 0.96 respectively support the performance with proper identification of positive entities. Decision trees expressed accuracy of 95.33% indicating good predictions. Precision and recall scores of 0.97 respectively illustrated well balance and effective understanding and interpretability. Support vector classification known to handle non linear decision boundary and ability to adapt complex data and relationships performed well with 96.46% accuracy with precision and recall values of 0.99 and 0.96 respectively. The Naive Bayes model performed with lower accuracy comparatively with others of 80.93%. The precision and accuracy scores were 0.81 and 1 respectively.

The tailored requirements of the study was to determine the model with highest accuracy and best fit for the study. Logistic regression and support vector classification showcased high accuracy with balanced precision and recall scores. Decision trees on the other hand were fruitful to unfold interpretability. Naive Bayes could have performed better with more optimization.

7 Conclusion and Future Work

The study conducted analysis of customer reviews with aligning factors such as ratings and recommendations. The study employed five machine learning algorithms: linear regression, logistic regression, decision tree, support vector classification, and Naive Bayes, to unfold customer reviews and explore the capabilities of different algorithms. The study focused on reviews for the cosmetic brand Sephora using customer review constituents such as ratings, reviews in textual fold, and recommendations, which are binary entities. The accuracy of each model was compared, and it was observed that logistic regression and support vector classification performed proficiently with accuracy of 96.46%. The fruitful inference from logistic regression can be imputed to its ability to handle binary outcomes with high impact, while decision trees with accuracy of 95.33% are attributed to handling complex relationships with the data. The high accuracy obtained by logistic regression suggests a relationship between ratings and sentiment scores when used to predict recommendations. It concludes that the fluctuation of the sentiment score towards positive or negative leads to a corresponding likelihood of positive or negative recommendations. Decision trees were successful in observing non-linear relationships, which showcased decision boundaries. The ambiguity suggests handling complex data interactions between sentiment and ratings, which expand the course of strong performance. Naive Bayes had accuracy of 80.93% . The high precision scores of logistic regression , decision tree, support vector classification were 0.99 , 0.95, 0.97 respectively. This indicates the model ability to interpret returning of relevant entities by these algorithms. Logistic regression and support vector classification returned recall score of 0.96 which elaborates the ability of models to return relevant values. This indicates the deployment of logistic regression , decision tree and support vector classification to execute recommendation prediction. Naive Bayes despite having the ability to perform well with complex high dimensional data , performed acceptable. One of the reason it could be is the frequency distribution of entries. Overall, the models preferred for the study performed proficiently and it can be concluded that these models are good fit for recommendation systems and automation.

The future works can be focused on different aspects like external data integration such as economic or financial parameters, geographical location coordinates, weather, and more factors to explore demographic attributes and integrate for a better understanding of the relationships. Furthermore, more algorithms would suggest an additional opinion on the accuracy of the models based on the context of the study. Advanced or fine feature engineering could enhance the predictive ratio of the models; this consists of sentiment analysis on more textual aspects such as colour of eye, skin type, hair colour, and such entities, which can provide more hyper-parameters to conduct more impacting analysis on. Overall, the study has business as well as machine learning implications while business implications being helping with revenue models and machine learning implications helping as a guiding tool for advanced computing.

8 Acknowledgement

My heartfelt gratitude extends to my project supervisor, Taimur Hafeez, for his invaluable guidance and support throughout the entire project. I am also deeply appreciative of the entire faculty of Research In Computing at the National College of Ireland and extend my thanks to my friends for their unwavering support in aiding my understanding of the project and its documentation.

References

Ajmain, M. R., Khatun, M. F., Bandan, S. S., Rejuan, A. R., Ria, N. J. and Noori, S. R. H. (2022). Enhancing sentiment analysis using machine learning predictive models to analyze social media reviews on junk food, *2022 13th International Conference on Computing Communication and Networking Technologies (ICCCNT)*, pp. 1–7.

Astuti, Y., Wahyuni, S. N., Maulina, D. and Sidiq, F. M. (2022a). The data leakage sentiment analysis using naive bayes algorithm based on machine learning approach, *2022 5th International Seminar on Research of Information Technology and Intelligent Systems (ISRITI)*, IEEE, pp. 215–220.

Astuti, Y., Wahyuni, S. N., Maulina, D. and Sidiq, F. M. (2022b). The data leakage sentiment analysis using naive bayes algorithm based on machine learning approach, *2022 5th International Seminar on Research of Information Technology and Intelligent Systems (ISRITI)*, IEEE, pp. 215–220.

Cambria, E., Schuller, B., Xia, Y. and Havasi, C. (2013). New avenues in opinion mining and sentiment analysis, *IEEE Intelligent systems* **28**(2): 15–21.

Determining the Link Between Consumer Sentiments and Automobile Sales Through Sentiment Analysis (2023). Master's thesis, Dublin, National College of Ireland. Submitted.

Dhanalakshmi, P., Kumar, G. A., Satwik, B. S., Sreeranga, K., Sai, A. T. and Jashwanth, G. (2023). Sentiment analysis using vader and logistic regression techniques, *2023 International Conference on Intelligent Systems for Communication, IoT and Security (ICISCoIS)*, pp. 139–144.

Jagdale, R. S., Shirsat, V. S. and Deshmukh, S. N. (2019). Sentiment analysis on product reviews using machine learning techniques, *Cognitive Informatics and Soft Computing: Proceeding of CISC 2017*, Springer, pp. 639–647.

Joachims, T. (1998). Text categorization with support vector machines: Learning with many relevant features, *European conference on machine learning*, Springer, pp. 137–142.

Kharde, V., Sonawane, P. et al. (2016). Sentiment analysis of twitter data: a survey of techniques, *arXiv preprint arXiv:1601.06971* .

Krishna, Y. L. S., Paramesh, P., Kumar, Y. T. and Gopi, A. (2023). Sentiment analysis of product reviews by using naive bayes and vader models, *2023 5th International Conference on Smart Systems and Inventive Technology (ICSSIT)*, pp. 1–6.

Li, G., Zheng, Q., Zhang, L., Guo, S. and Niu, L. (2020). Sentiment infomation based model for chinese text sentiment analysis, *2020 IEEE 3rd International Conference on Automation, Electronics and Electrical Engineering (AUTEEE)*, pp. 366–371.

Li, N. and Wu, D. D. (2010). Using text mining and sentiment analysis for online forums hotspot detection and forecast, *Decision support systems* **48**(2): 354–368.

Medhat, W., Hassan, A. and Korashy, H. (2014). Sentiment analysis algorithms and applications: A survey, *Ain Shams engineering journal* **5**(4): 1093–1113.

Mizan Khairul Anwar, M. K., Yusoff, M. and Kassim, M. (2022). Decision tree and naïve bayes for sentiment analysis in smoking perception, *2022 IEEE 12th Symposium on Computer Applications Industrial Electronics (ISCAIE)*, pp. 294–299.

Naz, S., Sharan, A. and Malik, N. (2018). Sentiment classification on twitter data using support vector machine, *2018 IEEE/WIC/ACM International Conference on Web Intelligence (WI)*, pp. 676–679.

Sabapathi, P. R. and Kaliyamurthie, K. (2022). Analysis of customer review and predicting future release of the product using machine learning concepts, *2022 International Conference on Communication, Computing and Internet of Things (IC3IoT)*, pp. 1–5.

Saifullah, S., Fauziah, Y. and Aribowo, A. S. (2021). Comparison of machine learning for sentiment analysis in detecting anxiety based on social media data, *arXiv preprint arXiv:2101.06353* .

Singh, A., Thakur, N. and Sharma, A. (2016). A review of supervised machine learning algorithms, *2016 3rd International Conference on Computing for Sustainable Global Development (INDIACom)*, pp. 1310–1315.

Sri Patibandla, J., Nilofuer, S., Sidhartha, M. and Sai, R. (2023). Product prediction using sentiment analysis and linear regression, *2023 International Conference on Inventive Computation Technologies (ICICT)*, pp. 546–552.

Stanley, M., Kr, A. and G, D. (2023). Sentiment analysis of covid vaccine tweet with vader, and implementation of different machine learning models, *2023 14th International Conference on Computing Communication and Networking Technologies (ICCNT)*, pp. 1–6.

Troussas, C., Virvou, M., Espinosa, K. J., Llaguno, K. and Caro, J. (2013). Sentiment analysis of facebook statuses using naive bayes classifier for language learning, *IISA 2013*, IEEE, pp. 1–6.