Data Analysis Using Machine Learning Algorithms

Women's Clothing E-Commerce Dataset

Final Project

Data Science in Business Course

Kiana Amani

Soheil PourMoradian

Yas Jilardi

Professor

Sajjad Heydari

4 Feb 2022

Abstract

Online shopping has made an underlying change in our lives and as the technology is advancing by leaps and bounds, online reviews aid the customers to choose the best item they are looking for. This have more shown up during the pandemic which most of our tasks are being undertaken through the internet, especially in the field of e-commerce. Thus, these messages, which people leave on the website when they purchase a product, can help others significantly. In this dataset (Women's Clothing E-Commerce) we presume that we are going to make the best decision based on the ratings and positive feedbacks to see whether the customers choose the products from different departments to buy or not (according to the reviews) and predict their attitude towards the items that have been mentioned. The machine learning methods used in this research are decision tree and k-nearest neighbors algorithm. All experiments were done in this project using python. We evaluate the model in terms of accuracy, precision, recall, F1-score and K-fold cross validation. These data suggest that the decision tree algorithm gives highest accuracy to classify the reviews, which is 95%.

Keywords: decision making, machine learning, classification, decision tree, K-nearest neighbors

1. Introduction

With the explosive growth of social media on the web (Liu and Zhang; 2012), large amounts of data and information are produced and shared across the social media every day (Ali et al.; 2019). While we face thousands of reviews, it is impossible to make decision to whether purchase an item or not. The process of decision analysis will help us to find the probability of an item being chosen by our customers based on the ratings and reviews.

1.1 Motivation and Background

Whenever you search for a piece of clothing online, as there is no imagination for the products and sense of trustworthy for the company, you find reading the messages and reviews from others below the product and gain information from distinctive opinions. Although sometimes they are confusing, ratings from 0 to 5 mostly summarize the feedbacks, especially when you lose your patience from reading irrelative and biased comments.

The aim of this project is to find the most reliable classification method of customer reviews based on online women clothing reviews by applying decision analysis, which can improve accuracy.

1.2 Research Question

Which machine learning algorithm can improve the accuracy of classifying decision about online women clothing reviews?

1.3 Research Objectives

To solve research question, four classification algorithms, which were Decision Tree Algorithm and K-Nearest Neighbors Algorithm, were selected to build the model. As a result, they were implemented and evaluated. Furthermore, we compared them with their accuracy and got the results.

The structure of the project is mentioned below:

In section 2, related literature review and previous study will be discussed. Methodology is presented in section 3. In section 4 shows how to implement the algorithms and methods. In section 5 evaluates the experimental results. Finally, we make conclusions and discuss future work.

2. Related Work

Nowadays it is the internet that play a key role in many aspects of human's life and shopping online is one of the most common relative habits. Sharing the experience on the network can be assumed as an inseparable part of online shopping. Decision analysis, which is known as one of the most initial and effective theories of all times, give accurate classifications and predictions.

Undoubtedly, sentiment analysis is more practical to use in this dataset but unfortunately it is impossible to utilize because of lack of knowledge. Hence, we focus on other theories that we mentioned before.

3. Methodology

Probability theory and statistics are the basis of data mining. Using models to represent simple, descriptive statistics makes it easier to help people understand what they are researching. Many procedure models such as KDD (Knowledge Discovery in Database), OSEMN (Obtain, Scrub, Explore, Model and Interpret) and CRISP-DM (Cross-industry Standard Process for Data Mining) have already been used in data mining. (Huber et al.; 2019) CRISP-DM is widely used in data mining as a standard process model.

The purpose of this research is to predict customer decision from reviews on women clothing e-commerce. Based on this aim, CRISP-DM is chosen to perform as a methodology for this research. There are six stages in CRISP-DM which are Business Understanding, Data Understanding, Data Preparation, Modeling, Evaluation and Deployment, as shown in Figure 1.

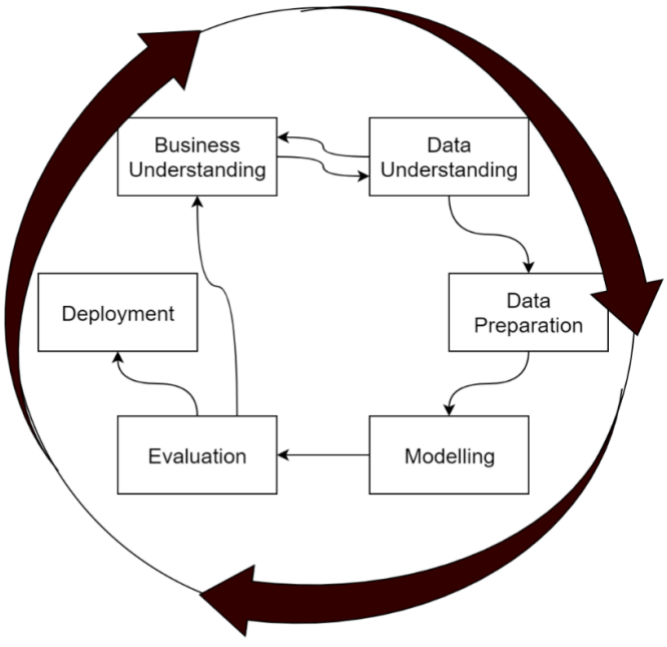


Figure 1: CRISP-DM

4. Implementation

4.1 Data Collection

This dataset can be obtained from Kaggle. It is a csv file which includes customer reviews for Women's clothing.

4.2 EDA (Exploratory Data Analysis)

This is an unavoidable step in data analysis in order to deeply understand the dataset and different aspects and insights which we can interpret from interactive graphs and tables. In this project, we have obtained useful information from Pandas, Seaborn and NumPy libraries in Python, as well as Google Data Studio and Tableau which can be observed in the figures 3,4,5,6,7,8.

Totally, most reviews were written by customers in their 39's. 75% of reviews received 4 or 5-star ratings and 82% reviews recommended the product therefore we can assume that the recommended IND variable correlates with the rating review. We have changed all reviews above 3 stars to positive reviews and reviews below 3 stars as negative reviews. After eliminating the 3-star reviews 88% of the reviews were classified as positive reviews.

4.3 Data Preparation

4.3.1 Pre-Processing

At this step we are going to transform the raw dataset into information to gain useful pieces of knowledge from our machine learning algorithms.

Categorical Data

First of all, we look for our categorical data to handle and drop (due to our restrictions for lack of knowledge in text mining). There is an unexpected feature named “Unnamed” in our dataset which we removed and furthermore the features “Title” and “Review Text” had been omitted from our data. We decided to keep “Division Name”, “Department Name” and “Class Name” columns and encode them to see whether they become useful or not subsequently.

Missing Values

Due to this fact that missing values can directly affect our model, it is crucial to handle them, even choose to replace them with 0 or just delete them. Fortunately, we did not have any missing observations to delete.

Outliers and Duplicated Data

Although it is not necessary to handle and delete outlier and duplicated data for decision tree algorithm, we had to drop all of them for KNN algorithm as it is a distance-based method.

4.3.2 Feature Selection

As it can be observed in the heatmap below (figure2), there is no eye-catching correlation between different features but it can be interpreted that Rating has a significant impact on our model.

As a result, we considered the Min Max Scaler method to initially scale our dataset and afterwards, chose Chi2 method for feature selection step. Finally, Rating was selected to be taken into the algorithm and Recommended IND as our target feature.

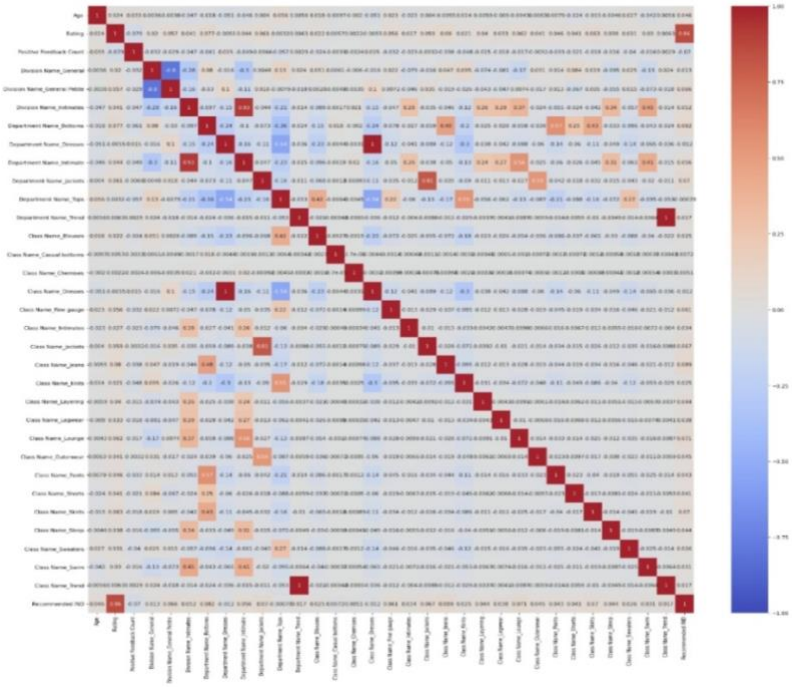


Figure2: Feature Selection Heatmap

4.4 Models

The new data set was chosen 70% as train data and the other 30% as test data. Sklearn model selection import train test split was applied for splitting data set.

Decision Tree

Decision Tree models allow us to create classifications in order to predict events based on decision rules. For this project, the max depth of the decision was 1. The accuracy of Decision Tree was 95%.

As it can be interpreted from the confusion matrix, 5619 reviews show that customers recommended the product and we predict correctly that they did.

119 reviews show that customers recommended the product but our prediction was incorrect.

384 reviews show that customers did not recommend the product but our prediction was wrong.

5467 reviews show that customers did not recommend the product, as we predicted.

K-Nearest Neighbor

KNN captures the idea of similarity (sometimes called distance, proximity, or closeness) with some mathematics we might have learned in our childhood—

calculating the distance between points on a graph and it can be used for both classification and regression models and it is easy and simple to implement.

In this project, we chose to implement this algorithm with Manhattan metric and the accuracy was 91%.

Also it can be interpreted from the confusion matrix, 4627 reviews show that customers recommended the product and we predict correctly that they did.

514 reviews show that customers recommended the product but our prediction was incorrect.

316 reviews show that customers did not recommend the product but our prediction was wrong.

4643 reviews show that customers did not recommend the product, as we predicted.

5. Conclusion

This research used two machine learning algorithms: Decision Tree and KNN algorithms to classify customer ratings. We focused more on online women clothing reviews on feature of rating. Moreover, we compared our results with previous research indicated that Naive Bayes was the preferred classifier but this project was modeled with decision tree algorithm which had higher accuracy.

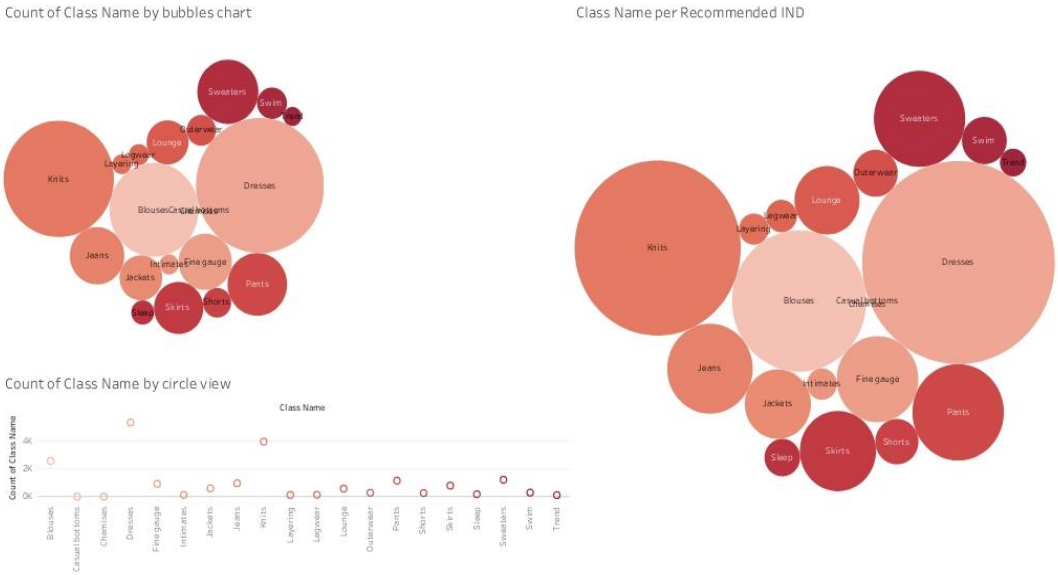


Figure3

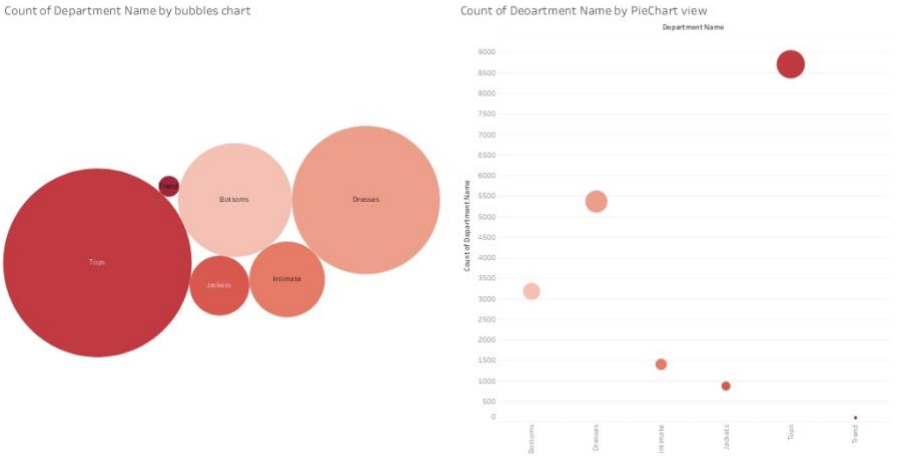


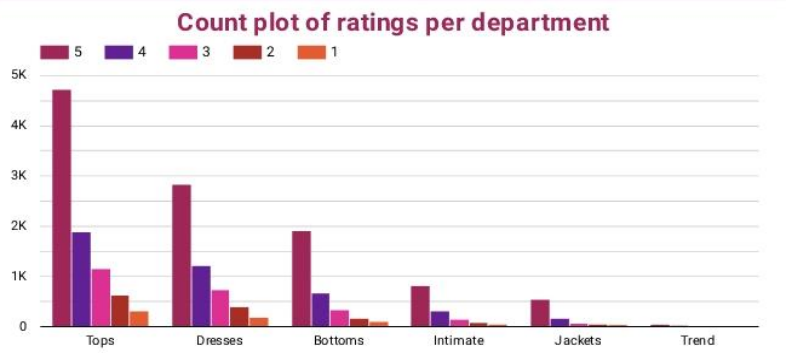
Figure4

	Rating	Departme...	Departme...
1.	5	Bottoms	1,903
2.	5	Tops	4,719
3.	5	Dresses	2,832
4.	5	Intimate	819

1 - 30 / 30

	Rating	Class Na...	Class Name
1.	1	Lounge	19
2.	1	Blouses	97
3.	1	Legwear	9
4.	1	Skirts	27
5.	1	Layering	2

1 - 92 / 92



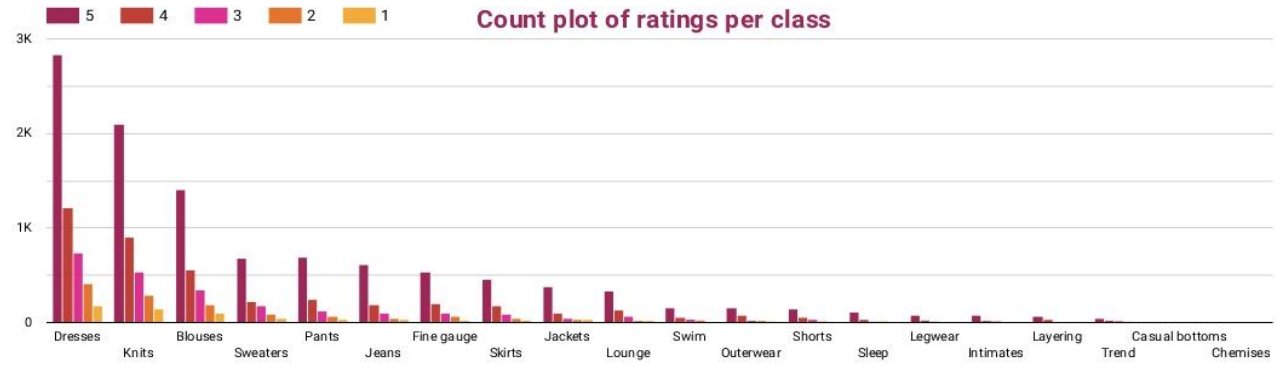


Figure5

	Rating	Rating
1.	Negetive	2,407
2.	Posetive	18,194

1 - 2 / 2

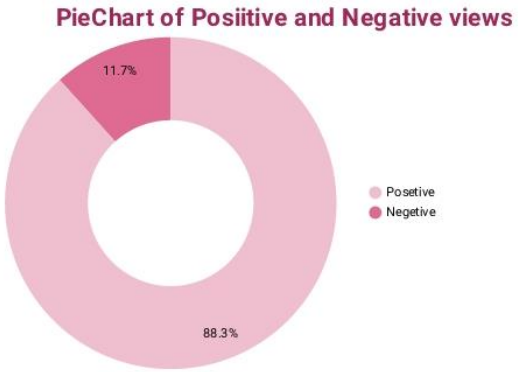


Figure6

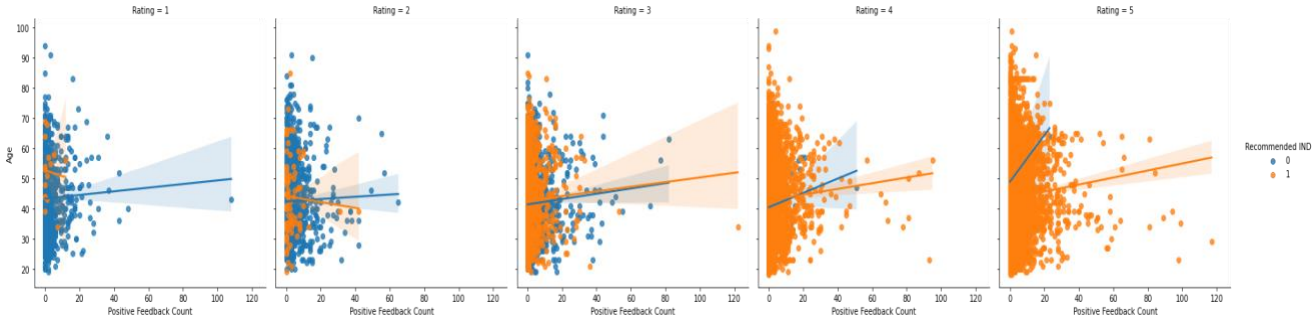


Figure7

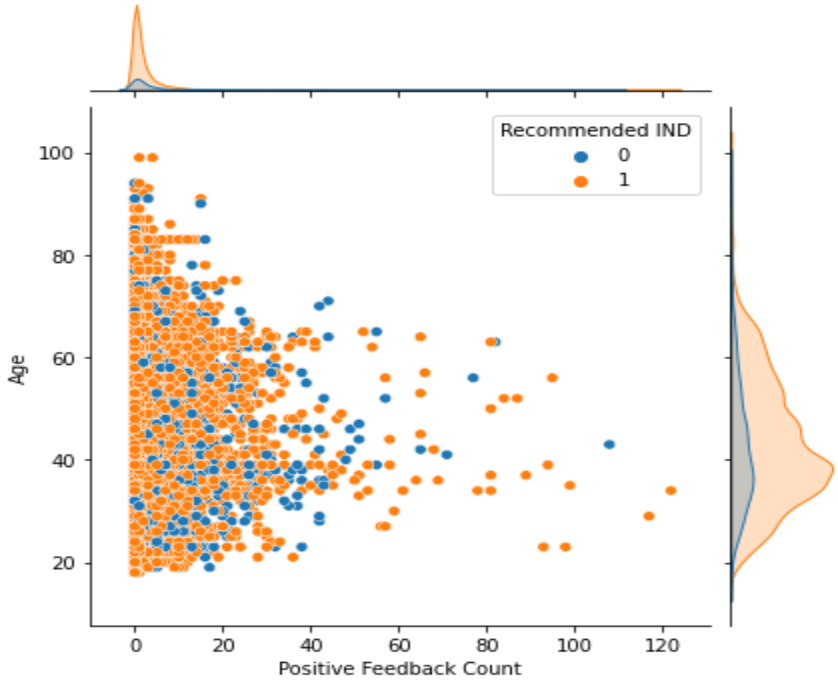


Figure8

6. References

Alrehili, A. and Albalawi, K. (2019). Sentiment analysis of customer reviews using ensemble method, 2019 International Conference on Computer and Information Sciences (ICCIS).

Huber, S., Wiemer, H., Schneider, D. and Ihlenfeldt, S. (2019). Dmme: Data mining methodology for engineering applications – a holistic extension to the crisp-dmmodel, 12th CIRP Conference on Intelligent Computation in Manufacturing Engineering 79: 403–408.

Liu, B. and Zhang, L. (2012). A survey of opinion mining and sentiment analysis, Springer Science+Business Media, LLC 2012 pp. 415–463.

https://pro.arcgis.com/en/pro-app/latest/help/editing/merge-features-into-one-feature.htm

https://pro.arcgis.com/en/pro-app/latest/tool-reference/conversion/feature-class-to-feature-class.htm

https://medium.com/@abcgconsultinggroup/womens-e-commerce-review-dataset-data-analysis-fab4bccf99b4

https://github.com/ya-stack/Women-s-Ecommerce-Clothing-Reviews

https://github.com/NadimKawwa/WomeneCommerce

https://github.com/EbinAbraham/Womens-Clothing-E-Commerce-Reviews-NLhttps://github.com/EbinAbraham/Womens-Clothing-E-Commerce-Reviews-NLP