**Abstract.**

Recommendation System is the knowledge discovery techniques and use of statistics to deliver users with personalized content and service. It is used to solve the interaction with the customers which are targeted to provide product recommendation issue. Hence, Advertisements are one of the medium to recommend to the customers and learner from their previous experiences. We found this a challenging task to build a recommendation system by Ad analysis. Customers these days are dependent on recommendations whether it is for products to purchase, news on recent launches, restaurants to visit or services to avail. One other challenge is to Recommend the Ads based on the User types. Hence the root to provide the necessary recommendation is Ads is the Recommender system and also it is important to identify the provided Advertisements are Legitimate. Recommender systems solve this problem of searching through large volume of dynamically generated information to provide users with personalized contents and services. In this project, we attempt to understand different kind of algorithms for Analyzing the Ads systems and compare their performance on E-commerce dataset which is picked from the known e-commerce site. The reason to pick up the dataset from sites is straight simple that there are bucketlist of different aspects of our analysis. We have used Supervised Learning Algorithms like Gaussian Naïve and also implemented Natural Language Processing using Bag of Words, TF-IDF and Hashing for Sentiment Analysis, and Stacked Ensemble model for our recommender system.

# INTRODUCTION

# Social media are computer-mediated tools that allow people and organizations to create, share, or exchange information, career interests, ideas, pictures/videos in virtual communities and networks. It facilitates the development of online social networks by connecting a user’s profile with other individuals and/or groups. Since people post unstructured content in social media, it is relatively difficult to extract product related information from such content. Using a users’ profile it is easy to gather thinking patterns such as interests, likes, hobbies, thoughts and so on. However, social media content has a huge potential in delivering more personalized advertisements. Most of the commercial advertisements and web applications are not based on users’ actual preferences. They do not have social media related advertisements classification system, self-updating or a continuously updating user character profile more personalization and preference based advertisements pushing mechanism. Therefore users, advertisers, and corporations are not often harnessing the full potential of social media based advertising. In this paper, we present an efficient way to provide personalized advertisements to the right people at the right time as well as advertisements pushing mechanism based on user preferences. We used semantic analysis and ontology mapping, advertisements classification mechanism to achieve this target in our AdSeeker advertisement engine.

# Objectives

# The over all objective of this thesis project is to investigate the relation between sentiment extracted from social media and the performance of generating Ads. The goal is to analysis the User reviews on a product and specific Ads related to the product and draw a relationship to analysis Ads and recommend.

# This results in the following more specific objectives:

# • Develop techniques for sentiment analysis of data from Social medias

# • Use machine learning and train a model to predict the performance of a site

# • Develop a prototype tool for the proposed method.

# Module design

Traditionally, people used to buy product from the stores. But now -a- days people prefer online shopping. Massive adoption of internet/web as an e- commerce platform led to change in the way business interact with their users. Recommendation System is the method to provide customers with suggestions about the product they could buy.

In this project, we have taken the dataset from data. world and we tried to provide recommendations to our users for several products having similar taste as of our other customer. Our dataset has relevant columns like UserID, ProductID, ProductDescription, Rating, DoReccomended, DidPurchase columns to do predictive analysis. We applied various machine learning and deep learning algorithms based on regression, clustering classification and found Collaborative filtering ALS Method to be most suitable algorithms for Recommendation Systems.

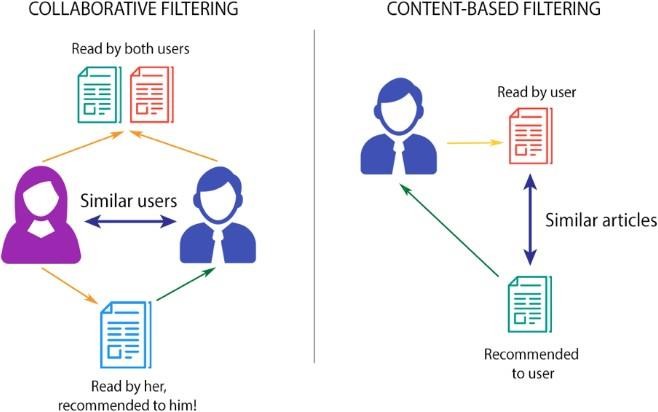
The recommendation system is produced basically by two types of approaches:

* 1. **Collaborative Filtering** (Item based, User based)

Collaborative Filtering is the most common technique used when it comes to building intelligent recommender systems that can learn to give better recommendations as more information about users is collected.

Most websites like Amazon, YouTube, and Netflix use collaborative filtering as a part of their sophisticated recommendation systems. You can use this technique to build recommenders that give suggestions to a user on the basis of the likes and dislikes of similar users.

# Content- based Filtering.

In collaborative filtering we find how many of users or items in the data are like other user and using correlation and cosine similarity find item for others. While the Content- based recommendation works on the rating/data provided explicitly or implicitly by the user. BBR provides recommendation which are highly personalized by matching user interest and description.

Machine learning algorithms are commonly used to do meaningful analysis and allow computer to learn human behavior or nature and improve the performance of new task on old analysis. Additional benefits of the Advertisement analysis are:

1. **Convert Visitor to Buyer:** help consumers find items that best fit the customers interests and inclinations and many times these lead to unplanned purchases driven by the buyer just because of the suggestion made.
2. **Increase in Cross-sell:** Giving Recommendation for products helps improve in cross selling by suggesting more products or services to customers. If the suggestions are perfect, the average order size increases.
3. **Create Value- added relationship:** In a competitive world where everything is one click away, building customer-loyalty becomes an essential aspect of business. Each time a customer visits a website, the system “learns” more about that customer’s preferences and interests and is increasingly able to operationalize this information to e.g. personalize what is offered. By providing each customer with an increasingly relevant experience, a corresponding improvement in the likelihood of that customer returning is achieved.

**HARDWARE REQUIREMENTS**

* System : Pentium IV 2.4 GHz.
* Ram : 4 GB
* Any desktop / Laptop system with above configuration or higher level

**SOFTWARE REQUIREMENTS**

* Operating system : Windows 8 and above.
* Coding Language : Python 2.7 and above.
* Scripting tool : Jupyter notebook
* Libraries : Pandas, scipy, numpy, matplotlib, etc.

Module Explanation

**PROPOSED ARCHITECTURE**

Feature selection

Predictions

Evaluation

testing data

raw data

Processed data

training data

Predictor model

Machine learning models

Selected data

Review Dataset

Preprocessing

# Dataset Explanation

The dataset which we are using is from the different eccomerce websites with 100k rows and 25 columns. We have an E- commerce dataset with details of Product like EAN number, Product Id, Name, Category, Date Added, Date Updated, Rating, Product URL, Product Manufacturer etc. Dataset also contains user information like User Id, Username, User Review for product, User Rating for product, User City, User Purchase and User Recommendation for product. User recommendation columns consist of True/False value to show if the user recommends a product or not. Rating consist of value between 1 to 5. The E- commerce dataset has approximately 72K rows and 25 columns after filtering and removing rows which are cannot be used by performing Exploratory Data Analysis and Data Cleaning.

In Algorithms used, we split dataset into two partitions: Test and train dataset by sampling in the ratios 20% and 80% respectively.

We aim to achieve the following for our system:

1. **Suggest:** Provide related items for the users from relevant and irrelevant collection of items.
2. **Predict:** Given a data of items purchased by Customers, we are trying to predict items for the user which can be useful for them based on user’s past purchase history and location of the user.
3. **Forecast:** Demand forecasting can also be made of items according to the item sold in a country/continent.

# Data Visualization

After having a quick look, we understand that the dataset consists columns of different datatypes such as object, float, integer. Few rows/columns are empty and have no suitable values (NaN) which needs to be processed to get more accurate results.

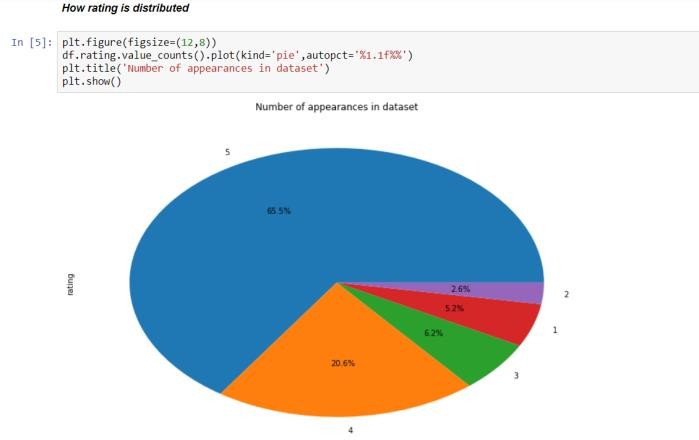


Fig. 2. Distribution of Rating

Figure 2 shows the range for giving product rating

i.e. from 1 to 5 with 1 being the least rating and 5 being the highest rating.

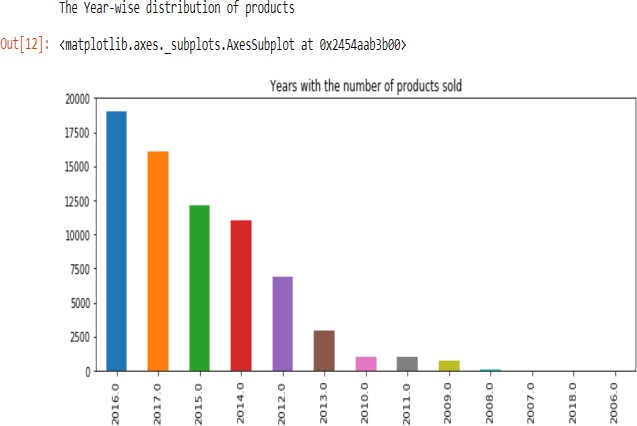


Fig. 3.Year-wise distribution of products

Figure 3 shows year-wise distribution of products from 2006-2016.

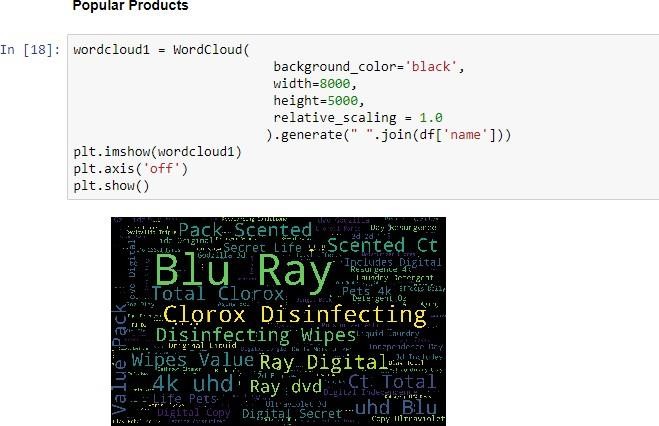


Fig. 4. Word Cloud

In addition, we also analyzed the words which are most repeated in the Ads and also corresponding comments by the user for products. Using this we can build a analysis Model.

# Data Processing

We did filtering data and defining which columns will be used for recommendation algorithms. In most of the cases, we have not used the rows which have any kind of reviews or purchases and no Ads content while in some cases, NaNs have been replaced with suitable values. The columns which are irrelevant have been dropped so that they do not affect predictions.

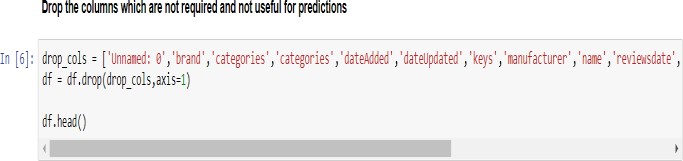


Fig. 5. Drop columns not useful for predictions

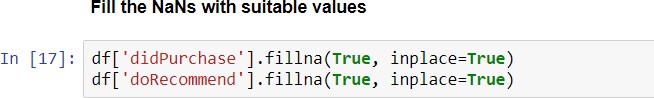


Fig. 6. Fill Nans with suitable values

# Feature Selection

In this system, we select the most relative feature by generating the score of each feature and selecting the most relevant one. We have columns like Rating, Reviews, Legitimate (True/False), didPurchase (True/False) which are very much effective in our Adanalysis. Since We analyze user Comments along with the Ad content and make the ML model learning together of this to make it more efficient.

# MACHINE LEARNING ALGORITHMS

# Gaussian Naïve Bayes

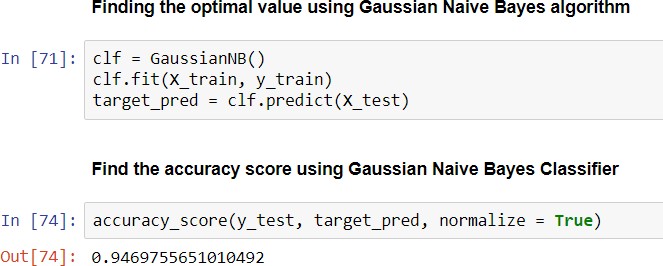
Naive Bayes algorithm based on Bayes’ theorem with the assumption of independence between every pair of features. This classifier work well in many real-world situations such as document classification

Fig. 9. Accuracy using Gaussian Naïve Bayes

and spam filtering. It requires a small amount of training data to estimate the necessary parameters and are extremely fast compared to more sophisticated methods. However, it is known to be a bad estimator.

# Natural Language Processing (NLP):

Natural Language Processing, or NLP for short, is broadly defined as the automatic manipulation of natural language, like speech and text, by software. NLP is a collective term referring to automatic computational processing of human languages. This includes both algorithms that take human-produced text as input, and algorithms that produce natural looking text as outputs.

We have analyzed the Review column of our dataset to understand customer reviews for each product. We have done basic pre-processing for the text such as removal of special characters (@,#), digits, punctuations, stopwords etc.

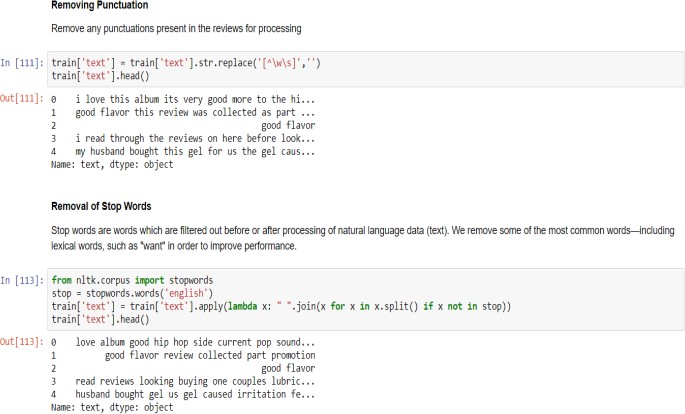


Fig. 17. Removing punctuations and stopwords

# Bag of Words

In this model, a text (such as a sentence or a document) is represented as the bag (multiset) of its words, disregarding grammar and even word order but keeping multiplicity. In the given dataset we considered the column “text” which contains the reviews given by a user for a product to achieve this concept. Using regular expression, we search a pattern and clean the data by removing the punctuation marks and special characters. By importing stop words from natural language toolkit, we later try to segregate the important and frequent used terms from the familiar words of English dictionary. We used Count Vectorizer to count the number of times a word occurs in a corpus. Hence, we display a list of words with their counts which are important and frequent in a corpus.

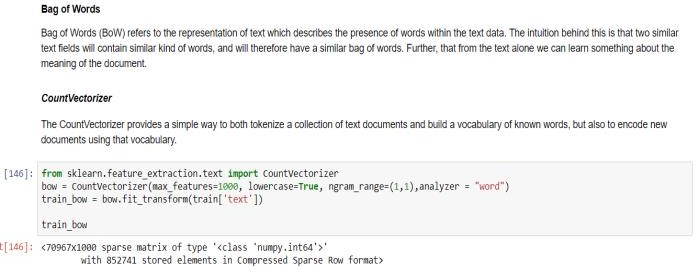


Fig. 18. Bag of Words using CountVectorizer

# TD-IDF

TF-IDF stands for "Term Frequency, Inverse Document Frequency." It's a way to score the importance of words (or "terms") in a document based on how frequently they appear across multiple documents or corpus. If a word appears frequently in a document, it's important. Give the word a high score. But if a word appears in many documents, it's not a unique identifier. Give the word a low score. Therefore, common words like "the" and "for," which appear in many documents, will be scaled down. Words that appear frequently in a single document will be scaled up.

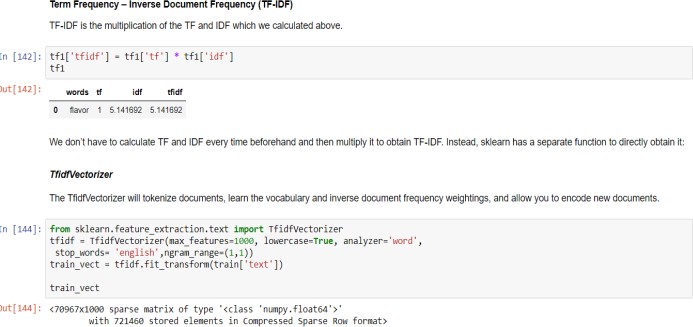


Fig. 19. TF-IDF using TfidfVectorizer

# Hashing

Hashing Vectorizer applies a hashing function to term frequency

counts in each document or corpus. Hash functions are an efficient way of mapping terms to features; it doesn’t necessarily need to be applied only to term frequencies but that’s how Hashing Vectorizer is employed here. Depending on the use case for the word vectors, it may be possible to reduce the length of the hash feature vector (and thus complexity) significantly with acceptable loss to accuracy/effectiveness (due to increased collision). If the hashing matrix is wider than the dictionary, it will mean that many of the column entries in the hashing matrix will be empty, and not just because a given document doesn't contain a specific term but because they're empty across the whole matrix. If it is not, it might send multiple terms to the same feature hash - this is the 'collision'.

Fig. 20. Hashing using HashingVectorizer

# Sentiment Analysis

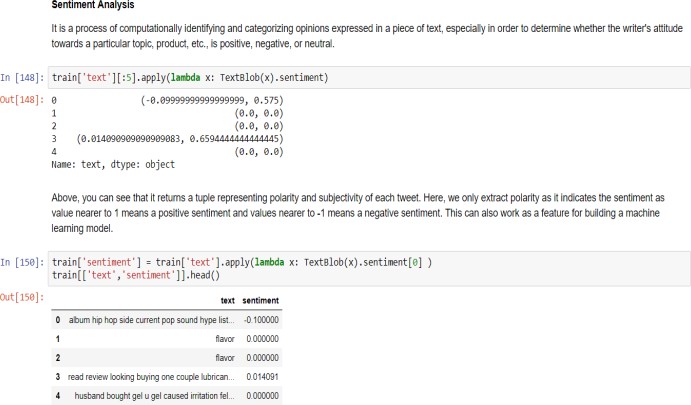
Sentiment is like a combination of tone of voice, word choice, and writing style all rolled into one. Natural language with labels about positivity or negativity (or any other spectrum we want to gauge), we can develop agents that can learn to understand the sentiments underlying new messages. Using NLP, we can understand what people like and dislike about products by crawling the thousands of reviews already posted on the website.

Fig. 21. Sentiment Analysis for the text.

# Stacked Ensemble Model

Ensemble modeling is the process of running two or more related but different analytical models and then synthesizing the results into a single score or spread in order to improve the accuracy of predictive analytics and data mining applications. Stacking is a way of combining multiple models, that introduces the concept of a meta learner. Applying stacked models to real-world big data problems can produce greater prediction accuracy and robustness than do individual models. The model stacking approach is powerful and compelling enough to alter your initial data mining mindset from finding the single best model to finding a collection of good complementary models. Of course, this method does involve additional cost both because you need to train a large number of models and because you need to use cross validation to avoid overfitting.

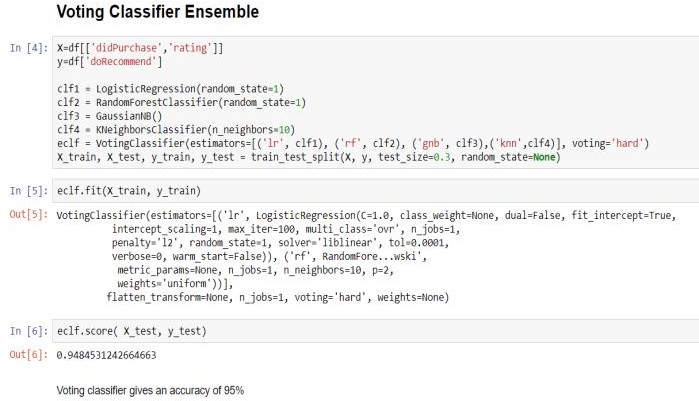


Fig. 28. Accuracy using Stacked Ensemble for Voting Classifier

Even after performing Cross validation, the accuracy did not increase. Instead it decreased the accuracy by approximately by 1.5%. But this gives us a clearer picture of the actual accuracy of the model.

RESULTS

Keeping the independent and dependent variables same across various algorithms, we found acceptable results of various algorithms used for our E- commerce dataset.

**Dependent variable:** doRecommend **Independent variable:** ProductId, didPurchase, Username, Rating.

Since we take into consideration various independent variables and find the correlation between them, our aim to predict the likelihood for the Advertisement being legitimate to the user and percentage of the correctness.

|  |  |
| --- | --- |
| Gaussian Naïve Bayes | 94.69% |
| Natural Language Processing for Sentiment Analysis | Bag of Words TF-IDF  Hashing, and sentiment analysis |
| Stacked Ensemble model | Voting Classifier  ~94% |

# CONCLUSIONS

By building this Analysis system, we are able to find products which are similar and can be recommended to a set of users who have similar buying patterns through a set of ads generating based on the current analysis. After using various algorithms, we concluded that, Gaussian Naïve Bayes, NLP technique are most suitable for accurately predictions. NLP using Bag of Words is very much suitable for our dataset and for performing the required sentiment analysis serves as a way to let us know products demand and the kind of Ad based on them.