Amazon Product Recommendation System Using Pearson Correlation

Dhairya Umrania 1814063
Tirth Thaker 1814061
Onkar Sanap 1814035

IT Dept., KJSCE
Mumbai, India

I. INTRODUCTION

In this world of information age, it is imperative that data has become the biggest asset in the world. In an ever-growing world where people are influenced by the most whimsical of things, it is crucial for any industry to understand what their customers want. That is how it used to work in the older times as well, it is the basic rule of demand and supply, if a product has more demand, then increase its supply. Basic business and marketing principles work on this concept, you display and sell what your customers want. But how can businesses know what their customers want? That is exactly what we'll be discussing in this paper. Modern day has blessed us with tremendous algorithms which can predict preferences of any user, we see recommendations everywhere on every site, be it video streaming sites like Netflix, YouTube, or retail sites like Amazon, Flipkart. These recommendations are not randomly generated, they are generated using well thought out algorithms such as KNN classification, collaborative filtering, content-based filtering, etc. Almost all of these algorithms take note of the data that a user is consuming and then recommend similar products which have similar data to the users. Therefore, all of these algorithms need a similarity measure, something that can compare two products and tell us definitely how similar they are or how similar they are to the users' preferences. Similarity measures could be anything from testing different numeric measures to testing different textual measures, anything that can definitely say if two things are similar or not. Once we have ascertained a similarity measure which correctly depicts the nature of similarities of our products, we feed that parameter to these previously mentioned algorithms and they then filter all the products based on our similarity measure and display the products which are most likely to be nearest to the product we want.

In this paper we'll be discussing how we can use Amazon ratings as a measure to test user preferences and based on them we'll use KNN classification to ascertain other products the user might like. The similarity measure to be used for KNN classification will be Pearson Correlation. Pearson Correlation is also known as centered cosine similarity. In cosine similarity, the cosine of the angle between two vectors is calculated and the answer is known as the cosine similarity between the two. The issue with this is that for values which don't exist, cosine similarity treats them as 0, when comparing with its corresponding similarity value, this can lead to significant deviations for the similarity measure. Instead, what we do in Centered Cosine Similarity is that, we treat the missing values as the mean of all the values. This is a fairer way of finding similarity and gives us a much better similarity measure. KNN will use this similarity measure as its parameter for finding the most similar neighbor of our desired value.

Formula:

$$r = \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum (x_i - \bar{x})^2 \sum (y_i - \bar{y})^2}}$$

r = correlation coefficient
x_i = values of the x-variable in a sample
$\bar{x}$ = mean of the values of the x-variable
y_i = values of the y-variable in a sample
$\bar{y}$ = mean of the values of the y-variable

II. APPLICATIONS OF SIMILARITY MEASURE

- **Noise Reduction**

Noise reduction is the technology which aims to reduce background noise from a speech so that words can be picked up better and more clearly. Normally in sound signals a filter is added to the signal which makes the sound cleaner and suppresses unwanted noise. Many different technologies have been used to implement this filter and have achieved a certain degree of success, but research is ongoing constantly to make it even better. Recently, the Pearson correlation coefficient was implemented for the filter. The coefficient can be used for identifying correlation and similarity between noise and speech. In noise reduction what is to be done is, the correlated speech which is what is desired is to be maximized and the uncorrelated part which is the noise, is to be minimized. Pearson Correlation helps us in doing that to achieve a clear, smooth and a high-quality sound.

- **Air Pollution Prediction**

Every year, high levels of pollution in the air destroy an estimated seven million people around the world. High levels of toxins in the air can cause a variety of illnesses, and pollution has resulted in an unprecedented number of deaths in many areas. Pollutants have serious consequences for people, particularly the elderly. As a result, effective methods for predicting pollutant levels are needed, which will aid in reducing the negative consequences of high pollutant levels in the air. KNN algorithm can be used to analyze the pollution data and make future predictions. Choosing the best similarity for this is also a significant problem. After using various similarity measures, Pearson Correlation Coefficient values were much better and reduced the root mean square error significantly.
Correlation analysis using Pearson Correlation Coefficient is done for our data set to analyze how species are correlated with each other. This analysis helped in knowing the importance of the predictors in the prediction of target variables. In addition, the weights to the attributes in Attribute Weighted k-NN are given based on the correlation scores.
This way using Pearson Correlation, we can find out predicted concentration of pollutants in air.

- **Collaborative Filtering**

Preferences are something we perceive as unique to us own but we'll be surprised if we actually went to see how common our preferences really are. Many people share similar preferences and many items share similar traits. Therefore, if a certain person with particular preferences likes an item, there is a high chance that another person with similar tastes will too like that item. We can also go the other way, if an item with certain traits is liked by someone, that same person might like another item which has similar traits. This is exactly what collaborative filtering is. Collaborative filtering is used for recommendation systems and predictive systems. We can give out recommendations to users based on their preferences based on User-User based filtering which is what the first example I gave in the above passage indicates and item-item based filtering which is the latter. This is used in many leading websites, such as YouTube, Netflix and Amazon, which is the topic of our research paper. To test the similarity, collaborative filtering uses Pearson Correlation, which is acts as the similarity measure and the main parameter for collaborative filtering.

III. METHODOLOGY

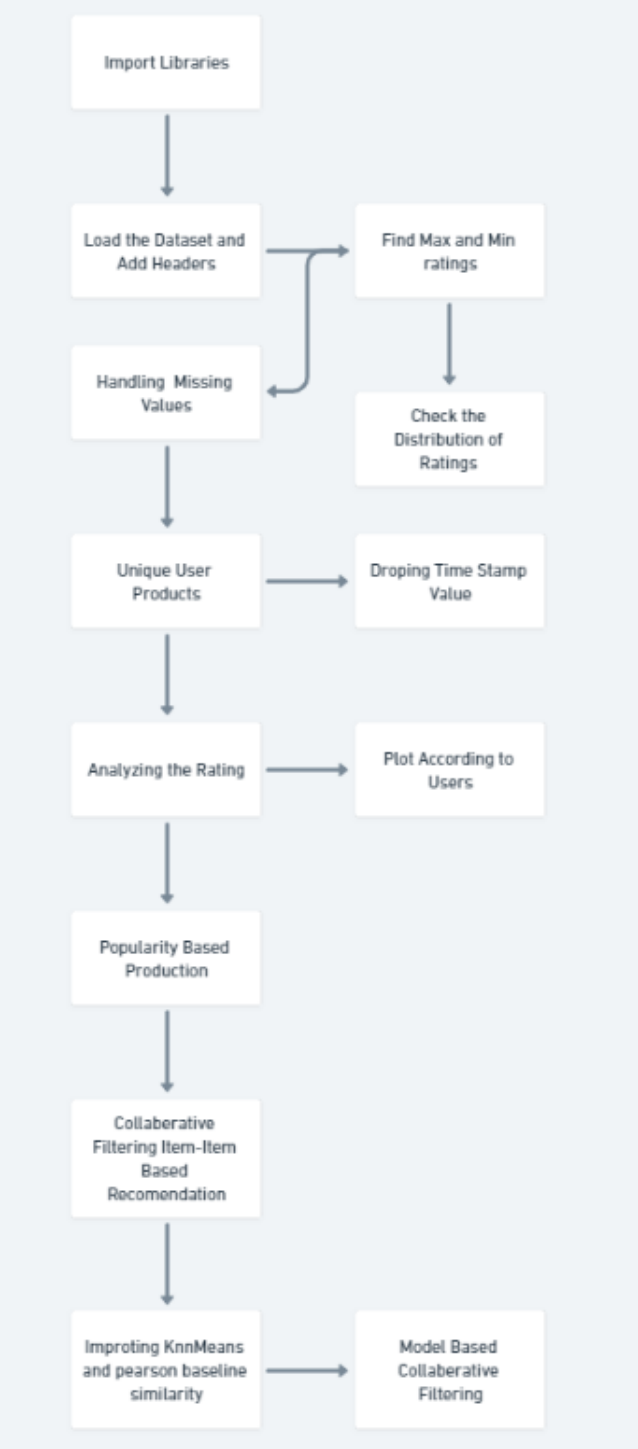

Fig.2: Flowchart of application

ALGORITHM:

1)Start Importing Libraries
2)Load the dataset and Add headers to print shape, head, dtypes of dataset
3)Finding max and min ratings of the datasets
4)Checking the distributions of Ratings
5)After Finding the Max and Min ratings, Finding the Missing Values in dataset
6)Finding Any unique User Products in dataset
7)Dropping the Timestamp Values
8)Analyze the Ratings and Plot According to Users
9)Finding Popularity Based Productions
10)Applying Collaborative Filtering item-item based Recommendation
11)Importing KNN means and Pearson baseline similarity
12)Finally Applying Collaborative Model Based Filtering

IV. IMPLEMENTATION

Numpy Pandas: Used for importing Dataset.

Matplotlib: Used for plotting graphs

Sklearn: Used for statistical analysis

Surprise Package:
Surprise is a Python scikit for building and analysing recommender systems that deal with explicit rating data.
We have us this in step 10 and step 11 that is surprise.prediction_algorithms.knns.KNNWithMeans
To perform KNNWithMeans(k=5, sim_options={'name': 'pearson_baseline', 'user_based': False})using pearson basline cosine similarity

from sklearn.decomposition import TruncatedSVD
This transformer performs linear dimensionality reduction by means of truncated singular value decomposition (SVD). Contrary to PCA, this estimator does not center the data before computing the singular value decomposition. This means it can work with sparse matrices efficiently.

Correlation matrix:
A correlation matrix is a table showing the value of the correlation coefficient (Correlation coefficients

are used in statistics to measure how strong a relationship is between two variables.) between sets of variables.
In our Case it shows it gives the ratings like weather it is positive or negative according to product id and there reviews.

Recommend Product:
After performing above steps of correlation matrix and then finding correlation product is, we found the recommended product if customer has already bought, we remove it otherwise we keep it

V. RESULTS & DISCUSSIONS

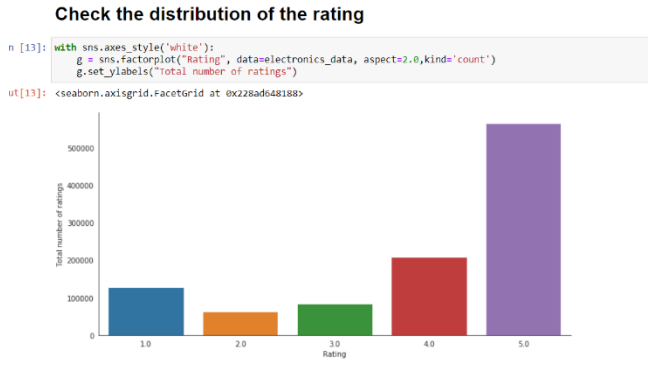

Fig.3: Rating frequency distribution

In the above graph we can see the count of total number of ratings. On the X-axis are the ratings ranging from 1 to 5. On the Y-axis we have the total number of ratings for each rating. As we can see, most of the users have given rating of 5.

The lowest frequency is of 2.

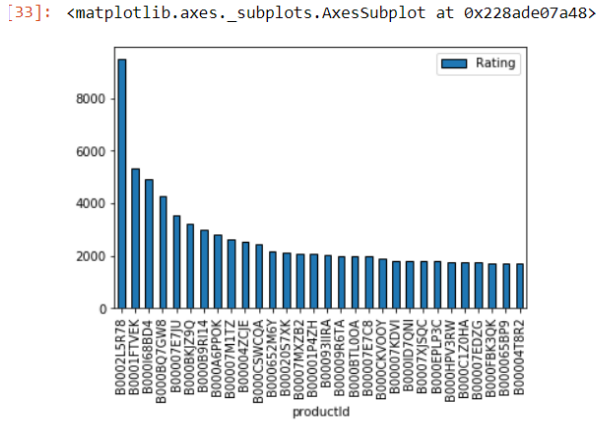

Fig. 4: Most Popular products

Collaborative filtering is commonly used for recommender systems. These techniques aim to fill in the missing entries of a user-item association matrix. We are going to use collaborative filtering (CF) approach. CF is based on the idea that the best recommendations come from people who have similar tastes. In other words, it uses historical item ratings of like-minded people to predict how someone would rate an item. Collaborative filtering has two sub-categories that are generally called memory based and model-based approaches.
Here the figure gives the products in descending order of the number of ratings given to each product. On the X axis there are products with their respective product ID. On the Y axis is the number of ratings.

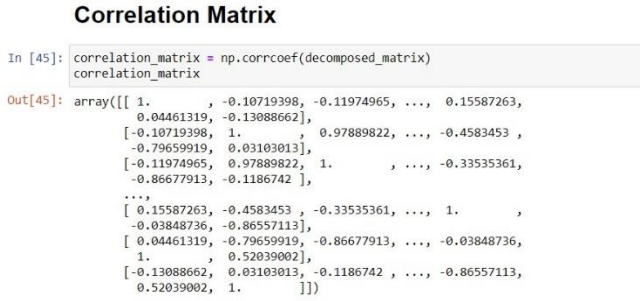

Fig. 5: Correlation Matrix

After calculating the correlation with all other products, we make a matrix of it. This matrix shows all the Correlation coefficients we got from Pearson Correlation.

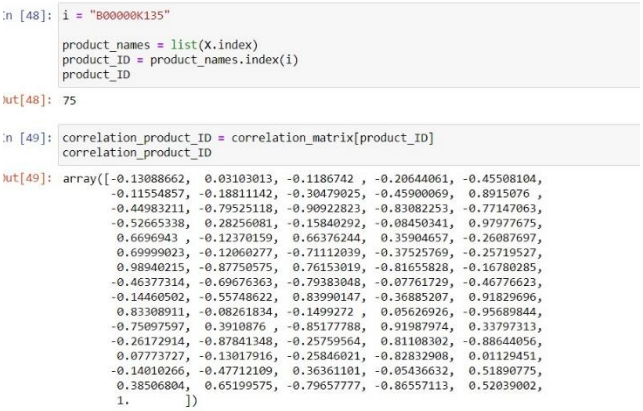

Fig. 6: Full correlation list w.r.t one product

We then choose a product ID for which we want to see the recommendations, so first we display all the coefficients with respect to all other products, which is displayed on the image above.
We will then Select a particular threshold and we will only display the products whose similarity index is higher than given threshold.

Fig. 6: List of recommended Products

We have taken our threshold as 0.65 and created a list of all the products that cross that threshold. Hence we get a list of the most similar products and that has been displayed in the image above.

VI. CONCLUSION

The Pearson Correlation is an effective tool that can be used to detect the similarity between ratings which is crucial to determine which users are similar and which are dissimilar. KNN classification is also a great method to implement collaborative filtering so that we can get the nearest neighbours of our desired product, and we have listed every product whose similarity is greater than the index of 0.65 because we believe that is the right amount for filtering the products which will be most similar to our product because if we increase the similarity index, it is possible we will get very few to no matches at all, and that is not something we want, and if we decrease the index too much, we'll get many products who in reality won't actually be that similar. So an index of 0.65 satisfies both the needs and provides us with a list which has enough similar products to display.

VII. REFERENCES

1. EXAMPLES OF OPTIMAL NOISE REDUCTION FILTERS DERIVED FROM THE SQUARED PEARSON CORRELATION COEFFICIENT Jiaolong Yu1, Jacob Benesty2, Gongping Huang1, and Jingdong Chen1

2. Image Processing Using Pearson's Correlation Coefficient: Applications on Autonomous Robotics A. Miranda Neto, A. Correa Victorino, I. Fantoni, D. E. Zampieri, J. V. Ferreira and D. A. Lima

3. Pearson Correlation Coefficient Based Attribute Weighted k-NN for Air Pollution Prediction Ankita Jain Department of Computer Science and Engineering National Institute of Technology, Delhi New Delhi, India Email: jainankita4567@gmail.com Rajya Lakshmi Lella Department of Computer Science and Engineering National Institute of Technology, Delhi New Delhi, India Email: rajya99@gmail.com

Originality report

COURSE NAME
EDA IA2

STUDENT NAME
DHAIRYA UMRANIA

FILE NAME
DHAIRYA UMRANIA - EDA IA2

REPORT CREATED
Apr 30, 2021

Summary

Flagged passages	5
Cited/quoted passages	2

Web matches

scikit-learn.org	2
apache.org	1
surpriselib.com	1
coursehero.com	1
towardsdatascience.com	1
statisticshowto.com	1

1 of 7 passages
Student passage FLAGGED

...is also known as centered cosine similarity. In cosine **si**
vectors is calculated **and the** answer is known as the cos

Top web match

The **similarity** measurement is a measure of **the cosine**
A **and** B. Suppose **the** angle **between the two** vectors wa

Understanding Cosine Similarity And Its Application | by R
... https://towardsdatascience.com/understanding-cosine-

2 of 7 passages
Student passage FLAGGED

Surprise is a Python scikit for building and analysing **r**
rating data.