International Research Journal of Engineering and Technology (IRJET) e-ISSN: 2395-0056
Volume: 08 Issue: 04 | Apr 2021 www.irjet.net p-ISSN: 2395-0072

Product Research Analysis and Recommendations

S.E Viswapriya[1], Saisumanth Tallapragada[2], Sriakshata Cherukupalli[3]

[1]*Assistant Professor, Dept. Of CSE, SCSVMV (Deemed to be university), Kanchipuram, Tamil Nadu, India*
[2]*Student, Dept. Of CSE, SCSVMV (Deemed to be university), Kanchipuram, Tamil Nadu, India*
[3]*Student, Dept. Of CSE, SCSVMV (Deemed to be university), Kanchipuram, Tamil Nadu, India*

Abstract – *With the rapid growth in e-commerce industries, the recommendation system has gained its importance as well. Recommendation systems which are also known as information gathering systems use the existing related information about products or services to suggest the most relevant of them to users thus providing better customer satisfaction which is vital in e-commerce systems. The recommendation is generally based on user's preferences and interests based on purchases they made. The increase in the amount of information available online has resulted in an information overload problem making it very complex for users to get the useful information they require within time. They cannot give accurate searching results according to users' different backgrounds and needs. To overcome this situation, the concept of a personalized recommendation system was introduced. Traditional recommendation systems rely on ratings provided by users or just by using a single Machine Learning algorithm such as content-based or collaborative-based filtering. In this paper, we are trying to segment customers based on their product purchases first, and then use the purchase history from users in the same cluster to do user-based collaborative filtering, then train the model with various predictive algorithms for achieving better accuracy and recommend products.*

Key Words: Personalized recommendation, collaborative filtering, snowball stemming, predictive algorithms, clustering, content-based filtering, hybrid approach, ensemble learning.

1.INTRODUCTION

In today's generation, online retail shopping has taken over the world. The users are now able to choose from a wide variety of products or services. Along with the growth of the e-commerce industries the recommendation systems have also gained popularity. The usage of the Recommendation system is not only limited to e-commerce industries but also used in services such as books, music, videos recommendations as well.

Today almost every online retail site is using recommendation systems to enhance the user experience [1]. The main aim of the recommendation system is to assist the user to determine which product they want to buy depending upon their interests, preferences, and constraints. As a massive amount of data is added to the internet every day, with such a large data the user can access the required service or product they need, but at times the users might not be able to describe their needs properly thus resulting in a poor search result which is not relevant to the user's needs. Thus, it can be observed that the existing recommendation systems can process a large amount of data and recommend items, but the accuracy of the recommendation system has scope for development.

Earlier the recommendation systems were solely dependent on user ratings, but with the passage of time solely relying on user ratings resulted in less accurate recommendations. Nowadays the data filtering is carried out by content-based filtering recommendation algorithm or collaborative-based filtering recommendation algorithm. But the main problem associated with the existing recommendation systems is that these do not consider the differences between the interests of individual users. The item predicted to a particular user will not be identical to the interest of its neighbors thus the recommendation results in poor recommendations [2].

What more will a customer love, than having their personalized recommendations based on their past purchases. This is something every company wants to build rather than having a general recommendation on things like most sold- out products. This is where Machine Learning comes into the picture and makes the whole website look a lot fancier. The goal of this paper is to cluster users based on their purchases and use the clustered user's purchases alongside collaborative filtering and train the model with different predictive algorithms to build a personalized product recommendations system.

2.LITERATURE SURVEY

Previously researchers have worked on precision marketing for supply management [4]. Using the customer segmentation based on the type of products customers bought. They tried to create a model based on a single classification algorithm to cut down on which vendors' products are least sold. They have also tried to use this model for retailer categorization, to find more insights on business behavior, by using a data reduction technique to optimize the clustering. There is also research where they have implemented collaborative filtering and then from this, they implemented customer segmentation [5]. In this they have implemented collaborative first, products to be recommended to each specific customer are determined and second, customers are segmented following these

recommendations. This can be used for targeted promotional material. In most of the approaches we saw that a single ML model is used for a specific problem like clustering or classification. We know that every model has its advantages and combining those models can result in a better accuracy in the result. After coming across many such implementations where a single concept or algorithm is used, we have come up with an idea which aims to improve the accuracy. In our implementation we want to cluster users based on their purchases and use the clustered users purchases alongside collaborative filtering to build a personalized product recommendations system.

Clustering is the process of grouping a set of physical or abstract items into classes of similar items where the groups are either meaningful or useful, or both. In product recommendation, the dataset is categorical in nature hence K-means algorithm shall not be befitting the dataset nature, since it is based on the Euclidean distance between two quantities hence there is a chance of clustering absolutely non-relevant data into one cluster since any type data can be assigned with any arbitrary distance. Many of the implementations have used the K-means algorithm. [6] Many existing implementations have used single concepts and their accuracy has a scope for improvisation.

3.EXISTING SYSTEMS

Many existing systems have used a single approach to recommend products to users. Main drawback of the existing systems is their prediction accuracy. Combining several models will certainly improve the accuracy as every model holds a unique ability using which advantage of each model can contribute to higher accuracy. The technologies or methodologies used by the existing papers can be broadly divided into usage of Content Filtering, usage of Collaborative Filtering and recommendations based on user reviews.

4.PROPOSED SYSTEM

In most of the existing approaches, we saw that a single Machine Learning model is used for a specific problem like clustering or classification. Every model has its advantages and combining those models can result in better accuracy in the result. So, we want to perform an analysis of each classification model and try combining various models to get the best possible integrated classification model that gives high accuracy of customer clustering.

As part of our project, we are trying to segment customers based on their product purchases first, and then use the purchase history from users in the same cluster to do user-based collaborative filtering and training the model with different types of predictive algorithms to recommend products. This would help in giving our recommendations precisely from the segment the user belongs to.

Thus, by building this model, an online shopping website would not only be able to retain its older customers but also use this as a Key Business Strategy to get new customers.

5.ACTUAL WORK

To test our proposed recommendation approach, we have used the dataset from UCI machine learning repository [3]. This data set contains all the transactions occurring between the year 2010 December to 2011 December for a UK-based online retail. The company mainly sells all-occasion gifts. Many customers of the company are wholesalers, thus making it mostly Business to Business marketing.

There are a total of 5,41,909 rows with 8 columns which are invoice number, stock code, product description, invoice date, customer id, country, unit price, and quantity. The data set initially had roughly 25% of null values. There were 4,06,829 data after the removal of null values. The data also had 5225 duplicate entries. There were 3684 unique products, 22190 unique transactions, and 4372 unique customers. The data also had 16% of canceled orders.

A. ATTRIBUTE INFORMATION

- InvoiceNo: Invoice number, A 6-digit number uniquely assigned to every transaction. If this code starts with alphabet 'c', it indicates a cancellation.
- StockCode: Product code, A 5-digit number uniquely assigned to every product.
- Description: Name of the product
- Quantity: Number, the quantity of each product per transaction.
- Country: Name of the country where the customer places an order from.
- CustomerID: Customer number, A 5-digit number uniquely assigned to every customer.
- InvoiceDate: Invoice Date and time, Number, the day and time when every transaction is generated.
- UnitPrice: Unit price, Number, Price per unit.

B. OBSERVATIONS

- In the dataset, there are few transactions not associated with any customer id, which can be deemed as not useful rows and can be discarded.
- There are also few duplicate records in the

given dataset.

- There are a total of 37 countries from which the orders have been placed and the majority of the orders are from the UK (Where it is actually basedat).
- There are a total of 4,372 unique customers that bought 3,684 products in 22,190 transactions.
- We also observe that some of the products only vary by color or shape.

C. METHODOLOGY

Flowcharts, diagrams, pseudocode, or formulas are added whenever appropriate. As mentioned above there are 5,41,909 rows and 8 features available. Following are the steps of evaluation:

1) Initially we have removed all the null entries in the data set.

2) We realized that there are few duplicates, precisely 5225, which we removed from the data set.

3) We have then concentrated on cancelled orders, there are the ones with their unique id that has "C" appended to it.

4) We tried to remove all the rows with C in the id. Later, we realized that the orders with C have different parent orders, which is still considered as the order was successful, although they were cancelled with different id. So, we tried to map all the cancelled orders with the parent orders and removed all of them. After removing mapped entries, there are still ids with "C" in them, which we removed, assuming they were placed before this data set is created. We store this in retail_data_cleaned dataframe.

5) We created a new column called "TotalPrice" in the data frame, that represents the multiplication of total quantity ordered and each item price.

6) After this, we have created keywords from list distinct product descriptions, using Snowball stemmer. Filtered the keywords to remove any small words or colors.

7) Create One hot encoding matrix for all product descriptions. We populate the matrix 1 if the product description has the keyword otherwise, we populate zero. We also include columns to indicate the price range of the product.

8) We use this matrix as a feature set for product clustering using KModes.

9) We choose the number of clusters 'k' based on the silhouette score.We obtained the maximum silhouette score for k=5.

10) We populate a categ_product column in the retail_data_cleaned indicating product category.

11) Aggregate the data to obtain below details for each customer and store it in dataframe selected_customers: a. The amount spent in each product cluster b. minimum spent in a single transaction, c. maximum spent in a single transaction, d. number of transactions e. average spent across all transactions.

12) We cluster the users using KMeans ++ into 11 categories (k selected based on user distribution across clusters and silhouette score).

13) We populate the cluster column in selected_customers dataframe which is our target variable for training.

14) Using this dataset to train different models to predict customer clusters. Below are the different models used.

a. SVC.
b. Logistic Regression.
c. KNN.
d. Decision Trees.
e. Random Forest.
f. Ada Boost.
g. Gradient Boosting.

15) For the above trained models, we used GridSearchCV with K Folds=5 to obtain the best parameters for the above models.

16) We have used ensemble voting classifiers to obtain higher performance. For ensembling we used Random forests, Gradient boosting, and KNN trained above step having the best parameter combination.

17) Using the ensemble model, we predicted customer category for the unseen customer and calculated accuracy by comparing the prediction output of the ensemble model with the cluster data obtained while we did KMeans++.

18) Next, we will do Product Recommendation using collaborative filtering by following the below steps:

a. Firstly, we need to restructure the data to form a matrix representing products bought by the customer populated with quantity if bought or else 0.
b. Next, we will predict the customer category for a customer from the test set based on his spending pattern we discovered in the previous steps.
c. Now compute the cosine similarity for all the users in that cluster.

19) At last, we recommend products bought by the most similar customers.

D. EVALUATION STRATERGY

- We will evaluate the accuracy of each of the below classification models individually followed by an ensemble of classification models for the cross-validation dataset and choose the model with higher accuracy. Accuracy over here signifies the number of correctly classified customers.
- For train and test data we split the clean dataset

we obtained after data cleaning using sklearn's train_test_split and chose 80% of the data as train and 20% as test, randomly.

- We chose to train our model on:
  1. SVC.
  2. Logistic Regression.
  3. KNN.
  4. Decision Tree.
  5. Random Forest.
  6. Ada Boost.
  7. Gradient Boosting.
- For the above trained models, we used GridSearchCV with K Folds=5 to obtain the best parameters for the above models and train data will be the same.
- We have used ensemble voting classifiers to obtain higher performance. For ensembling we used Random forests, Gradient boosting, and KNN trained above step having the best parameter combination.
- On the test data, using the ensemble model we predicted the customer category for the unseen customer and calculated accuracy by comparing the prediction output of the ensemble model with the cluster data obtained while we did KMeans++.

6.RESULTS AND DISCUSSIONS

The data after cleaning i.e., after removal of null entries, duplicate entries, and cancelled orders. We choose the number of clusters to be formed using the silhouette scores. From the below table 1 and Fig 1. we observe that the maximum silhouette score was obtained for clusters k=5.

Table -1: Silhouette scores

N clusters	Average Silhouette Scores
n_clusters = 3	0.188826593018323 16
n_clusters = 4	0.247419132075104 44
n_clusters = 5	0.282573124213848
n_clusters = 6	0.27058539224994
n_clusters = 7	0.275373373429765 76
n_clusters = 8	0.267325506814204 03
n_clusters = 9	0.274336549895199 7

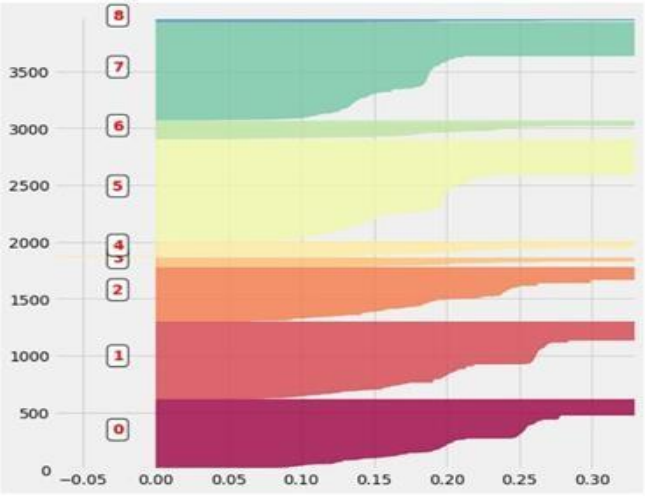

Fig -1: Product clustering graphical representation.

With this score we develop product categories clusters below Fig 2. depicts the same. We also added minimum, maximum and mean to analyze the overall spending pattern of the user in each category.

	CustomerID	InvoiceNo	TotalPrice	cat_0	cat_1	cat_2	cat_3	cat_4	InvoiceDate
1	12347	537626	711.79	124.44	23.40	187.2	293.35	83.40	2010-12-07 14:57:00.000001024
2	12347	542237	475.39	38.25	84.34	130.5	169.20	53.10	2011-01-26 14:29:59.999999744
3	12347	549222	636.25	38.25	81.00	330.9	115.00	71.10	2011-04-07 10:42:59.999999232
4	12347	556201	382.52	19.90	41.40	74.4	168.76	78.06	2011-06-09 13:01:00.000000256

Fig -2: Product categories based on k=5.

In Fig 3. We cluster the users using KMeans ++ into 11 categories (k selected based on user distribution across clusters and silhouette score). We use the updated dataset to train different models to predict customer clusters. We used the following models (SVC, Logistic Regression, KNN, Decision Tree, Random Forest, Ada Boost, Gradient Boosting) with GridSearchCV with K Folds=5 to obtain the best parameters for the models to train the dataset.

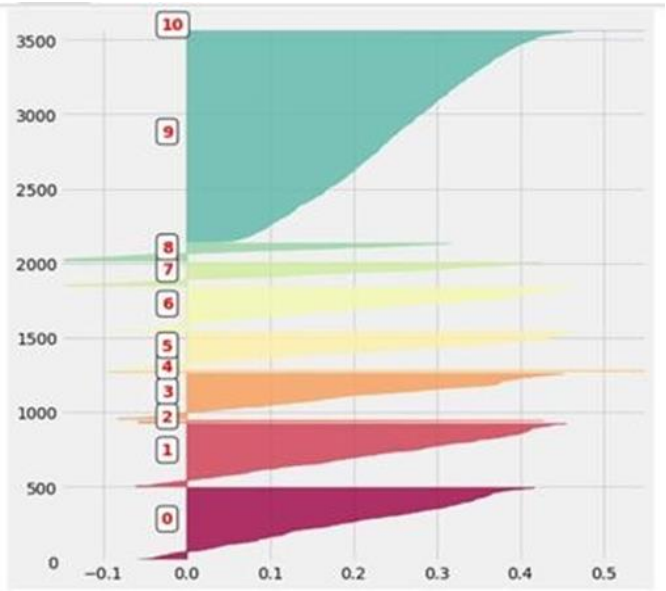

Fig -3: Customer based cluster classification.

We observed the **Classifier Accuracies:**

- SVC: 80.21%
- Logistic Regression: 89.22%
- KNN: 79.83%
- Decision Tree: 98%
- Random Forest: 99%
- Ada Boost: 53.05%
- Gradient Boosting: 99.07%
- Ensemble Learning: 95.38%

```
from sklearn.metrics.pairwise import cosine_similarity
test_data.apply(lambda x:predictClusterProducts(x),axis=1)
-----------------------------------------
12967
0.24500793879906752
{'similarity', 'HOT WATER BOTTLE TEA AND SYMPATHY', 'HOT WATER BOTTLE KEEP CALM', 'CHICK GREY HOT WATER BOTTLE', 'RED WOOLLY
HOTTIE WHITE HEART.'}
-----------------------------------------
13369
0.28870137031989895
{'similarity', 'HEART IVORY TRELLIS LARGE', 'LONDON BUS COFFEE MUG', 'QUEENS GUARD COFFEE MUG'}
-----------------------------------------
13565
0.28912858437570166
{'similarity', 'SILVER HANGING T-LIGHT HOLDER', 'TRADITIONAL KNITTING NANCY', 'ENAMEL FLOWER JUG CREAM', 'LARGE WHITE HEART O
F WICKER', 'LARGE DECO JEWELLERY STAND', 'SMALL WHITE HEART OF WICKER', 'HEART OF WICKER SMALL', 'LOVE BUILDING BLOCK WORD',
'DRAWER KNOB CRACKLE GLAZE PINK'}
-----------------------------------------
17886
0.27429384112278
{'DOLLY GIRL CHILDRENS BOWL', 'PICNIC BOXES SET OF 3 RETROSPOT ', 'SET OF 3 REGENCY CAKE TINS', 'DOLLY GIRL CHILDRENS CUP',
```

Fig -4: Final Output

The Fig 4. represents the results, for each customer we are showing his/her customer id, similarity score and products recommended to the user.

6. CONCLUSIONS

We clustered customers based on their spending patterns by aggregating the available data. We achieved a silhouette score of 0.2 for product clustering. We have trained used grid search for parameter selection K-Fold cross validation for training various classifiers. We were able to predict the customer categories with up to 96% accuracy using ensemble Voting C classifier. We have achieved this by using the products the customer bought, amount spent in each category of products and spending range. We have clustered the products using nltk using silhouette score to get proper clustering of products and uniform distribution of products. We finally predict products for new customers using cluster prediction and recommend products using collaborative filtering. For collaborative filtering we are using a user-item matrix and populated it with the quantity bought by a user in each cell resulting in a sparse matrix as there are many unique products. We are receiving varying numbers of recommendation products for each customer due to low similarity scores between the new customer and existing customers in the clusters in certain cases. We could improve this if the data included rating of product given by the user which can be used in the user-item matrix as collaborative filtering works better when we have ratings of products.

We could further improve the clustering and recommendation if we are provided with more data as the dataset, we have used for training aggregation has only approximately 4000 users.

With the rapid development of information technology, more and more kinds of information content gradually have fused together and formed a massive data warehouse. Every moment more data are being generated, most of which are useless. However, there are much useful data in it, including preferences and so on. To help general user to get more personalized service, recommend system come into being. The system has been used in E-commerce field, and it's also good to help enterprises to improve competitiveness. However, for future work, we are looking forward to investigating other methods such as clustering algorithms with the principal component analysis. Additionally, we are aiming at conducting other experiments with big datasets to confirm the effectiveness of our method.

REFERENCES

[1] Rahul Kumar Chaurasiya, Utkarsh Sahu," Improving Performance of Product Recommendations Using User Reviews" 3rd International Conference and Workshops on Recent Advances and Innovations in Engineering, 22-25 November 2018 (IEEE Conference Record # 43534).

[2] Yang Xiao qing, "An intelligent E-commerce recommendation algorithm based on collaborative

filtering technology" 2014 7th International Conference on Intelligent Computation Technology and Automation.

[3] Data set link- http://archive.ics.uci.edu/ml/machine-learning- databases/00352.

[4] Zhen You, Yain-Whar Si, Defu Zhang, XiangXiang Zeng, Stephen C.H. Leung c, Tao Li, " A decision-making framework for precision marketing", Expert Systems with Applications, 42 (2015).

[5] SnowballStemmer: https://kite.com/python/docs/nltk.SnowballStemmer

[6] SongJie GONG and HongWu YE "Joining User Clustering and Item Based Collaborative Filtering in Personalized Recommendation Services .", 2009 International Conference on Industrial and Information Systems.

BIOGRAPHIES

I am S.E Viswapriya working as an assistant professor in Computer Science Engineering department in SCSVMV (Deemed to be University), Kanchipuram.

I am Saisumanth Tallapragada currently a final year student in B.E- CSE from SCSVMV (deemed to be university), Kanchipuram.

I am Sriakshata Cherukupalli currently a final year student in B.E- CSE from SCSVMV (deemed to be university), Kanchipuram.