



# Prediction of E-Commerce Shoppers' Purchasing Intention using Knn Algorithm

Ankush Verma<sup>1\*</sup>, Chetan Nagar<sup>2</sup>, Sharda Haryani<sup>3</sup> and Sumit Jain<sup>4</sup>

<sup>1,2,3,4</sup> Prestige Institute of Management & Research Indore, India

\*<sup>1</sup>ankush\_verma@pimrindore.ac.in

**Abstract.** An e-commerce web site is very much effective for visitors to buy on-line shopping achieving a high conversion rate which added a new explosion in the business sector. People tend to explore online for finding the items they need and buy in real time. For this companies use different machine learning algorithms to find the user behaviour and interest about the products. There are so many algorithms such as regression, Random Forest, Decision Tree, Knn, Naive Bayes, SVM, Logistic regression to predict whether a customer visiting the webpages of an e-commerce shop also show they will purchase or not. The analysing in real time predicts the history of customers shopping. In this paper we take a dataset of e-commerce purchasing and apply a knn algorithm to find out the purchasing intention of customers. We are also discussing the top companies using machine learning to generate revenue and analyze their data for better prediction.

**Keywords:** Knn, supervised learning, machine learning, natural language processing

## 1 Introduction

In today's world it is impossible to run successfully e-commerce business without Machine learning in such a huge competition. Natural language processing (NLP) is a technology that modern AI-based ecommerce search platforms utilise to detect the exact search intent of a visitor for sales prediction of the companies is the process of estimating the future growing sales using its own historical data. For better opportunity of any products these days most of the organisations are using machine learning tools and techniques to predict their sales. According to the Gartner and researcher said, up to 80% of customer interactions are managed by AI today and future. Many algorithms are created that can learn from given data set and explore different parameter of data to predict trends and outcomes of changes.

Most of the companies comply with and apply the supervised gadget studying approach. One of the techniques is class additionally part of information mining, category is a challenge to forecast category labels that's unknown earlier than to differentiate between one object to some other based totally on the attributes or capabilities [4]. In supervised learning, label data is a vital part that is used to classify

© The Author(s) 2025

S. Bhalerao et al. (eds.), *Proceedings of the International Conference on Recent Advancement and Modernization in Sustainable Intelligent Technologies & Applications (RAMSITA-2025)*, Advances in Intelligent Systems Research 192,

[https://doi.org/10.2991/978-94-6463-716-8\\_6](https://doi.org/10.2991/978-94-6463-716-8_6)

data into more distinct categories. The train-test split method is used to find how well machine learning algorithms perform and predict outcomes from data [1]. Peng and Yu use RBF neural community to expect product sales based totally on time series analysis and optimize the prediction version [8].

K-Nearest Neighbor (KNN) is a highly recognized classification technique and is considered to be among the top ten most essential and widely used algorithms in data analysis. KNN is known for its simplicity and clarity in implementation. It is primarily utilized for classification purposes. KNN works by finding the closest K number of neighbors based on a distance metric and using their class labels to determine the class of a new data point. The purpose of the K-Nearest Neighbor (KNN) algorithm is to categorize new objects based on their characteristics and training data. This is accomplished by using the Euclidean distance formula to determine the closest training data points to the new object [4]. KNN is a popular choice in situations where the data is noisy, easy to understand, and requires extensive machine learning. This algorithm is simple and straightforward, making it an attractive option for classification tasks [2].

## **2 Benefit of Using Machine Learning**

### **2.1 Product Recommendation**

People are curious about how effective the product, automatic product recommendations are widely used. Product recommendation is used in Netflix, amazon, Disney video, YouTube streaming platform is suggested to them via algorithms that analyses user interest and behaviour. Determining which products go well together is easy, in low lost and suitable for us. Artificial intelligence allows the suggestions which are pertinent and interesting, increasing the probability of altering. The preferences, choice of young and older customers will differ.

### **2.2 Search**

All major e-commerce stores have more than 800 million products in their inventory. Search numbers are increasingly important for providing accurate and relevant items for search algorithms is include with Machine learning, because there is no matter what is the quality of the product or the price of a product, if the item is not found it cannot generate sales. Here machine learning is able to pinpoint the particular trends and patterns needed to automatically determine from millions of products. There are estimated frequencies of specific search queries as well as the specific buyer profile

(e.g. previous product views, previous search queries, phrasing habits, age range). Search autocomplete is highly skilled to generate more revenue, raise more reformation up to 24%.

### **2.3 Misspelling**

Most of the people are not known the word of item so it will help great autocomplete. Now e-commerce giants are now combining. So, natural language processing and machine learning have to understand the kind of language and phrases for customers support. If the keyword are not satisfactory for those phrases, wrong spelling they correct it according to criteria and phrases resultant fast accessing.

### **2.4 Dynamic Pricing**

Airlines domain were among the first in the world to benchmark the pricing concept and adjust prices automatically. It can simply mean raising fares when demand on a particular route is high and lowering them when demand is low. But there are lots of other parameters that can be used to evaluate optimum prices such as competitor's prices, time of day when the price is increase or decrease, repository stock or season of sale. Adapting this concept e-commerce companies are also going to operate dynamic pricing it will help to generate more revenue. These are some fundamental guidelines and tactics used by online retailers nowadays to set dynamic pricing changes often [7, 6].

- i. Determining each product's optimum price.
- ii. Foreseeing the best price that will be offered to each consumer.

### **2.5 Customer Support**

If you want to solve various all type of issues with manpower, it will generate a large cost and time consuming and not at all efficient, because customers generate millions of query in a day, a lot of issues will require human assistance and some of the companies for dealing with such queries that could be solved by redirecting the customer to an FAQ page. Customers always complaint about long waiting times for his solution, difficult to explain and reevaluate their problem numerous times, unqualified suggestions or unknown of items summary.

Robots that can answer phone calls can help automate this process with the use of system learning. In addition to phone conversations, machine learning also adds additional support channels, such as email automation, email labelling (e.g., questions, complaints, requests), or support through AI chatbots. The chatbot will require lot of time to learn, to get to know the products and services as well as the customers profile and their way of communication [4].

### **2.6 Fraud Detection**

Fraud detection is a big concern for ecommerce businesses having a lot of transaction in a day. For this machine learning is used to detect them early stage and prevent solution to the customer. Machine learning models are trained to determine all the patterns associated with fraud activities like isolated high order values or customers setting orders from new IP addresses etc and many others. A few massive e-trade platforms like ebay, Alibaba & Amazon constructed their fraud detection structures that use machine learning algorithms to get over fraud in real time.

### **3 Top Companies using Machine Learning in Ecommerce**

#### **3.1 Amazon**

Invention of Alexa product is one of Amazon visible artificial assistant applications. Amazon use Machine learning algorithms to predict targeted marketing plan, allowing the company what merchandise customers will maximum in all likelihood interest to buy and also to offer them custom designed tips based on their involved searches. Automation perform a very important function in the evolution technique to warehouse operations. The company has reported that about 100,000 robots managed operations at its warehouse facilities throughout the world. This will increase delivery effectiveness and cut shipping costs for e-commerce businesses. In order to track when products are removed from and returned to the warehouse, the company argues that deep learning when used in conjunction with sensor, computer vision, and technology can be beneficial.

#### **3.2 JD.com**

JD.com joined partnership with Siasun Automation and Robot. This is a method of using automated technologies, including robots, to enhance all types of data mining operations. This effectively speeds up the process of categorising, sorting, and delivering products; which result in cost savings and more sales.

According to the company, there are over 250 centres in China that are in charge of managing and supervising warehouses, including automated logistics. The corporation wants to deploy more AI in order to use fewer resources over a ten-year period and boost its profit margin.

#### **3.3 Alibaba**

The leading e-commerce company, Alibaba, claims that AI algorithms help automate customer care tasks like search and suggestion as well as product delivery. Alibaba's software monitors client interactions and areas of interest in order to make a variety of product recommendations.

To plan the most environmentally friendly transit routes, AI is utilised. According to Alibaba, clever logistics have led to a 10% reduction in vehicle use and a 30% reduction in travel distances.

### **3.4 eBay**

In October 2016, the eBay Shopbot, a chatbot that can be accessible through Facebook Messenger, underwent its initial testing. The bot functions as an AI assistant to easily guide users to interesting products using natural language. Customers can use photographs from their smart phones that are relevant to a specific product to communicate with the bot via text, speech, or images.

On eBay, we observed and used machine learning approaches for the tasks of item categorization and responsibility, price prediction, and object-to-product matching.

## **4 KNN**

KNN is a machine mastering algorithm that used for each regression and classification type. It is a supervised machine learning which gaining knowledge of set of rules which have labelled data set to study a characteristic of records that produces the precise output although given unlabeled facts set. while k-Nearest associates examines the labels of quantity of information points surrounding a target facts factor, to be able to make a prediction approximately the class that the facts point falls into it. K-Nearest neighbors (KNN) is a totally easy and effective set of rules for this purposes it's one of the maximum famous system gaining knowledge of algorithms [5].

The KNN concept determines the nearest k neighbours based on the distance between the new pattern and the training data of the sample. Based on the class to which the neighbour belongs, the new sample is then classified, If they are all in the same class and fit, the new pattern will also fall into that class; if not, each publish-elite category is rated, and the new pattern category is then chosen in accordance with the rules of desirable deeds.

### **4.1 Mathematical Model of Knn Algorithm**

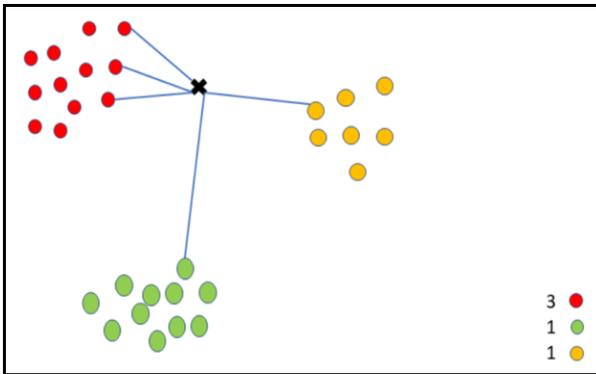
The basis for utilising the closest neighbour method (Fig. 1) to forecast values is only the assumption that the values of the known items can be predicted similarly. Finding the mesh points in the multidimensional space  $R_n$  that are closest to the unknown patterns and determining the size of the unknown pattern based on the categories of k

points is the basic premise of the nearest neighbour rule set [3]. According to the method, each time corresponds to points in a non-dimensional area. According to the standard Euclidean distance, the closest neighbour of an instance is described. if possible, let  $x$ 's eigenvector be:

where  $x_r$  represents the feature for the  $r$ th attribute of example  $x$ . the space between two times is  $x_i$  and  $x_j$  is defined as  $d(x_i, x_j)$ , wherein [10]:

$$d(x_i, x_j) = \sum_{r=1}^n (x_{ir} - x_{jr})^2$$

The discrete object classification function in closest neighbour learning is given by  $f: R^n \rightarrow V$ , where  $V$  is a finite collection of various sets of categories ( $v_1, v_2, \dots, v_s$ ). The quantity and level of dispersion in each type of sample determines the nearest neighbour  $k$  value, and various  $k$  values can be used for a variety of purposes [10].



**Fig.1.** The K Nearest Neighbor Model

## 5 Implementation of Knn

We apply a real dataset to the implementation of the knn algorithm using following steps [9].

### Step 1: Load the dataset

Three independent variables here are “Category” , “Sub Categaory” and “Gender”

The dependent variable is “Amount”

```
import numpy as np
```

```
import matplotlib.pyplot as plt
```

```
import pandas as pd
```

```
import sklearn
```

**Step 2:** After importing, divide a dataset into independent and dependent variables.

```
import pandas as pd
dataset = pd.read_csv('Test_Details.csv')
X = dataset.iloc[:, :-1].values
y = dataset.iloc[:, 3].values
```

We must encode our dataset using LabelEncoder because it contains character variables.

```
import LabelEncode from the sklearn.preprocessing module
labelEncoder_gender = LabelEncoder()
X[:,0] = labelEncoder_gender.fit_transform(X[:,0])
X = np.vstack(X[:,:]).astype(np.float)
```

A train and test set was created by splitting the data. Since we are using 0.20 as the test size, our training sample comprise of 1500 count and test set consist of 80 counts.

```
X_train, X_test, Y_train, Y_test = train test split(X, Y, test size = 0.20, random state = 0).
```

Next, Feature scaling is done on the training and test set.

```
import StandardScaler
sc = StandardScaler()
X_train = sc.fit_transform(X_train)
X_test = sc.transform(X_test)
```

Create and train the K Nearest Neighbor model with the training set (Fig. 2, Fig.3, Fig. 4)

```
import KNeighborsClassifier
knn = KNeighborsClassifier(n_neighbors=4)
knn.fit(X_train, y_train)
print(knn.score(X_test, y_test))
```

Using different k values to implement knn model

```
In [23]: from sklearn.neighbors import KNeighborsClassifier
         knn = KNeighborsClassifier(n_neighbors=4)
         knn.fit(X_train, y_train)
         print(knn.score(X_test, y_test))
```

0.71

**Fig.2.** The KNN Classification Model using k=4

```
In [10]: from sklearn.neighbors import KNeighborsClassifier
knn = KNeighborsClassifier(n_neighbors=7)
knn.fit(X_train, y_train)
print(knn.score(X_test, y_test))
```

0.7066666666666667

Fig.3. When kNN classification model using k=7

```
In [22]: from sklearn.neighbors import KNeighborsClassifier
knn = KNeighborsClassifier(n_neighbors=10)
knn.fit(X_train, y_train)
print(knn.score(X_test, y_test))
```

0.6833333333333333

Fig. 4. When kNN classification model using k=10

```
import pandas as pd
import numpy as np
import matplotlib.pyplot as plt
import seaborn as sns
%matplotlib inline
dataset = pd.read_csv('D:\Test_Details.csv')

plt.figure(figsize=(12, 6))
plt.title('Purchase by Gender')
sns.countplot(dataset ['Product line'], hue = dataset.Gender)
```

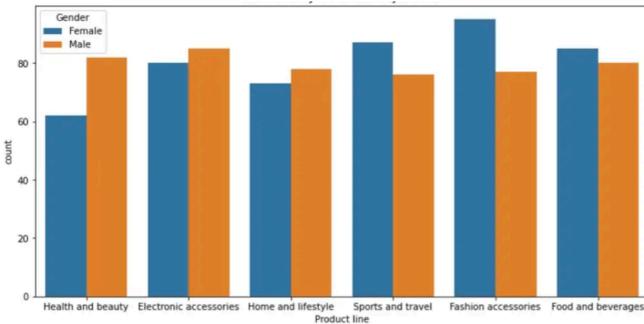
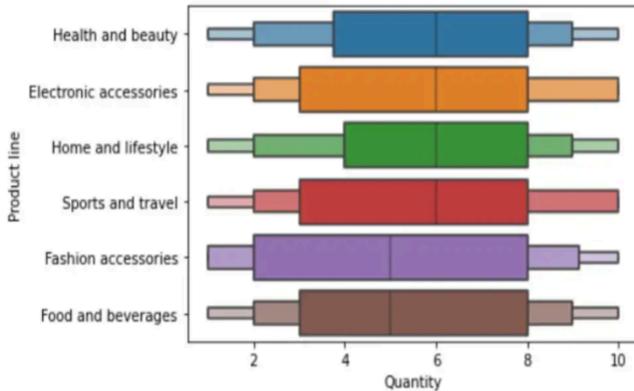


Fig.5. Bar Graph of male and female purchasing product criteria

sns.boxenplot(y = 'Product line', x = 'Quantity', data= dataset)



**Fig.6.** Box plot of products with the quantity

Bar Graph and Box Plot are shown in Fig. 5, Fig. 6).

## 6 Conclusion

K-NN increases the predictive model's freedom to match data, which can be used to enhance forecast accuracy. Despite being straightforward, K-NN has a surprisingly high predicted accuracy under the correct conditions. To anticipate a new point, K-NN uses the  $k$  parameter from 1 to  $i$ th value to look at points nearby and aggregate the labels of those points. A model that is low will react overfit and be excessively flexible. The converse is also true if a model is overly rigid if the value of  $k$  is too large. The genuine decision boundary appears to be much more closely approximated, accurate and close to prediction between  $k = 9$  to  $k = 15$  respectively. If the value of  $k$  is again increase and close to 20 then it is not accurate and find a new classification. In our paper we found that purchasing prediction of female is more in fashion and accessories and for male product are electronic accessories.

## References

1. Agrawal R.: K-Nearest Neighbor for Uncertain Data. International Journal of Computer Applications, 105 (11), 13-16,0975-8887, 2014.
2. Hamid P., Hoseinali A. and Behrouz M.: A Modification on K-Nearest Neighbor Classifier. Global Journal of Computer Science and Technology, 2010, 37-41.

3. Taunk, Kashvi, Sanjukta De, Srishti Verma, and Aleena Swetapadma. "A brief review of nearest neighbor algorithm for learning and classification." In 2019 international conference on intelligent computing and control systems (ICCS), pp. 1255-1260. IEEE, 2019.
4. Larose D.T.: *Discovering Knowledge in Data An Introduction to Data Mining*. Wiley Interscience, 90-106.
5. Liu Z.G., Pan Q., & Dezert J: Research and Implementation of Machine Learning Classifier Based on KNN. IOP Publishing, 74 (1), 2019, 119-132.
6. Motwani,B. & Haryani, S.: Prediction of Customer Buying Intention due to Digital Marketing: Application of TAM Model. *Jour of Adv Research in Dynamical & Control Systems*, 10 (14), 2019, 1873-1879.
7. Okunola O. & Onyekwelu B.: Predicting Consumer Behaviour in Digital Market: A Machine Learning Approach. *International Journal of Innovative Research in Science, Engineering and Technology*, 2019, 8(8).
8. Peng H. & Yu S.: Sales forecast of vending machines based on time series analysis. *Computer Science*, 2015.
9. Verma A., Nagar: Predicting House Price in India Using Linear Regression Machine Learning Algorithms. *International conference of Intelligent Engineering and Management*, 2022.
10. Wang L.: Research and Implementation of Machine Learning Classifier Based on KNN. *IOP Conference Series: Materials Science and Engineering*, 2019.

**Open Access** This chapter is licensed under the terms of the Creative Commons Attribution-NonCommercial 4.0 International License (<http://creativecommons.org/licenses/by-nc/4.0/>), which permits any noncommercial use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license and indicate if changes were made.

The images or other third party material in this chapter are included in the chapter's Creative Commons license, unless indicated otherwise in a credit line to the material. If material is not included in the chapter's Creative Commons license and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder.

