

Customer Behavior Analysis: Identifying risky customers based on their purchased product on e-commerce.

Pratik Thorat¹, Dhanamma Jagli²

¹PG Student, Vivekanand Education Society's Institute of Technology, Dept. of MCA, Mumbai, India

²Assistant Professor, Vivekanand Education Society's Institute of Technology, Dept. of MCA, Mumbai, India

Abstract - Today's world many people are using e-commerce to buy their day to day products. E-commerce plays a vital role in purchasing products as the e-commerce industry grows many problems related to this industry start growing. The main problem the e-commerce industry is facing is to identify risky customers based on their purchase of products. Many criminals are today using e-commerce sites to buy antisocial or harmful or unusual or explosive/chemicals products to attempt crime. Many criminals and terror organizations are using e-commerce sites to buy chemicals, acids, electronics components. This research paper focuses on identifying Risky customers based on purchased products using machine learning methods.

Key Words: Machine Learning, Data Mining, KNN Algorithm,

1. INTRODUCTION

E-commerce sites selling many types of risky/harmful products. People could use this product to attempt any type of crime like robbery, terror attack, homicide etc. In Pulwama attack, Kashmiri 17-year-old teen used Amazon to buy chemical ingredients for making improvised explosive devices (IED).[1] Many internet shopping platforms, Alibaba, amazon is selling chemicals without understanding the background information of customers.[2] For e-commerce sites, it is necessary to understand the background information of customers and stop unusual buying of products. There are a lot of methods and techniques available to analyze Customer behavior but most of them are only effective in the domain of Products Marketing, Customer Rating, and Market basket analysis but very few methods and techniques that are effective in the security domain. Customers who visit e-commerce sites leave some important information and also e-commerce sites have some important information about customers that customers provided to them while Registration process. We can use machine learning methods and data mining techniques to identify customer behavior. In this research paper, we use data mining and KNN (K Nearest Neighbor) classification method to identify risky customers.

Approach to Customer Behavior Analysis:

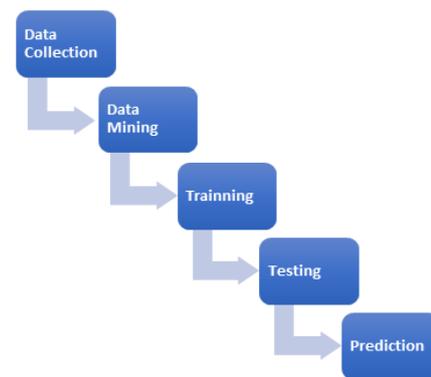


Fig 1- Process Flow

1.1 Literature Review

This paper is a study of customer behavior based on products purchased by them on e-commerce sites. Many Research papers are published in the customer behavior prediction domain. The past literature was taken from the published studies of similar work "K. Maheswari and P. P. A. Priya, Predicting customer behavior in online shopping using SVM classifier"[3]. This paper uses the SVM classification algorithm to classify customers according to their buying behavior. Classification is done by considering how the customer spends their valuable time, day in buying decisions. Most frequent items bought and quantity of buy. "Kareena and N. Kapoor, A Review on Consumer Behavior Prediction using Data Mining Techniques,"[4] this paper predicts customers behavior prediction using data mining and SVM Algorithm for this it uses data either past purchase history or customer review. "A Customer Classification Prediction Model Based on Machine Learning Techniques".[5] this paper predicts which types of customers are responding to offers and discount sales using KNN classification. Most of the customer Behavior analysis papers are related to Marketing, product rating, product reviews domain but a very limited search is carried out in customer behavior analysis based on purchased products to predict customer risky or not. Motivation to write this research paper is Pulwama terrorist attack case where

terrorist purchase chemicals and electronic equipment on amazon "Pulwama terror attack: Srinagar boy bought chemicals for making a bomb from Amazon, arrested by NIA"

2. Background Study

Data mining is a process that involves searching, collecting, filtering, and analyzing the data. Data mining is a set of procedure and function to determine following from huge dataset [6]

- Previously Unknown
- Legal information
- New information
- Constructive model
- Clear models

In business and organizations data mining techniques used to perform the following operation.

- Clustering
- Classification
- Prediction and
- Rule induction
- Pre-processing or cleaning
- Association rule mining

2.2 Classification

There are many classification methods available some of them are followings [7]

ID3 Algorithm

- C4.5 Algorithm
- K Nearest Neighbors Algorithm
- Naïve Bayes Algorithm
- SVM Algorithm
- Linear classifiers
- Logistic Regression
- Decision trees
- Random forests.

K Nearest Neighbors Algorithm is one of the simplest algorithms used in Machine Learning for regression and

classification problems. KNN algorithms use data and classify new data points based on similarity measures (e.g. distance function). Classification is done by a majority vote to its neighbors. The data is assigned to the class which has the nearest neighbors. As you increase the number of nearest neighbors, the value of k, accuracy might increase.[8]

3. Research Methodology

Data collection is the first process in every Research project. There are many ways to collect data. For Data collection, Google Form was Created to Conduct an online survey and shared among various participants and collect their responses. Later the responses of the form were converted into an excel sheet to be used as a source of data. The Objective of Survey form is to understand the Real-time Behaviour of Customer o e-commerce and their preference for products to buy on e-commerce sites. Many types of question were asked through google form to collect this data. Participants asked to buy any 5 types of product which they wish to buy. Various types of products were added into form and asked participants to select a product they wish to buy. Products belonged to different domains like IT, Clothing, Chemical, house-appliances, stationary, Electronics etc.

Data Preparation:

Data collected through google form is in textual format. Each row contains various products that fall under various domains like appliances, stationery, Electronics, IT, Clothing, chemicals etc. For KNN classification algorithm data must be in numerical format. Using python LabelEncoder library data is transformed into Numerical Format [Fig 3.1]. Later data is shuffle using the python shuffle library. The objective of the data shuffle is to improve training as well as testing accuracy. [Fig 3.2]

15	4	8	2	5	3	1
23	7	3	8	0	10	1
16	8	3	6	1	1	1
24	1	5	6	1	3	1
3	2	5	1	0	9	1
22	2	7	7	6	2	1
60	2	5	7	0	9	1
127	0	8	3	8	8	0
61	7	3	5	6	10	1
45	7	5	0	0	2	1
46	4	3	1	0	9	1
83	8	5	6	0	3	1
77	7	4	5	7	0	1
76	4	9	2	7	2	1
36	2	3	4	0	3	1
11	8	5	6	5	10	1
...
120	9	2	9	8	8	0
64	2	8	1	2	2	1
40	4	5	0	0	3	1
52	1	5	9	0	9	1
18	7	6	9	1	1	1
113	9	1	9	8	4	0
..

[Fig 3.1]

9	7	3	7	4	9	1
10	3	5	6	5	3	1
11	8	5	6	5	10	1
12	2	3	5	0	10	1
13	8	5	0	6	9	1
14	8	5	5	0	3	1
15	4	8	2	5	3	1
16	8	3	6	1	1	1
17	8	5	8	0	1	1
18	7	6	9	1	1	1
19	4	5	6	1	2	1
20	8	4	6	0	9	1
21	6	5	1	5	1	1
22	2	7	7	6	2	1
23	7	3	8	0	10	1
24	1	5	6	1	3	1
25	5	5	1	0	2	1
26	2	6	7	0	3	1
27	6	7	4	0	7	1
28	3	3	1	0	3	1
29	2	1	1	0	10	1
...
116	9	0	9	2	8	0
117	9	1	5	8	6	0
118	1	7	3	8	5	0

[Fig 3.2]

Dataset Format:

Each Rows contain 6 columns, column 1 to column 5 contain different products and column 6 contain Class

Product1	Product 2	Product 3	Product 4	Product 5	Class
----------	-----------	-----------	-----------	-----------	-------

For the Successful implementation of the KNN Algorithm, we required a training Dataset and the testing dataset for this single dataset is split into 70:30 ration to form a training dataset and testing dataset.

3.2 Machine Learning Algorithm:

KNN classification algorithm falls under the type of supervised learning. In Supervised Learning, A model is prepared through a training process in which it is required to make predictions and is corrected when those predictions are wrong. The training process continues until the model achieves a desired level of accuracy on the training data. K nearest neighbors is a simple algorithm that stores all available cases and classifies new cases by a majority vote of its k neighbors. The case is assigned to the class is most common amongst its K nearest neighbors measured by a distance function

Following Steps are performed for prediction:

Step 1) import all python libraries required for ML and Data mining.

Step 2) dataset is transformed into a numerical dataset.

Step 3) dataset shuffle and split into a training dataset and testing dataset

Step 4) The KNN Classification model is defined. Set value of K=3.

```
Out[611]: KNeighborsClassifier(algorithm='auto', leaf_size=30, metric='minkowski',
metric_params=None, n_jobs=None, n_neighbors=3, p=2,
weights='uniform')
```

Step 5) Training is performed on training dataset and accuracy is checked.

Step 6) Testing is Performed on Testing Dataset and accuracy is checked.

3.3 EXPERIMENTAL RESULTS.

The customers are classified based on their purchased products. Data mining methods are applied on the dataset and later KNN Classification algorithm is applied on the training dataset. KNN classifier are build using the dataset. When customers visit e-commerce sites for buying any product and if customers purchase any product which is

highly risky products or group of high risky product that are threatening to society and mankind then that customer classified as high risky customer and if a customer buys any usual products or group of the usual product then it is classified as Low risky products.

The Customer is classified as High risky or Low Risky based on their purchased Products

Risk Factor	Types of Purchased Product
High	Chemicals, Knife, Dagger, Unusual books and other Unusual products
Low	Usual Product, Home Appliances and other usual Products

Chart -3: Risk Classification

Model trains using training data and tested on testing data accuracy of training and testing data are as follows Fig [3.3]:



Fig: 3.3 - Accuracy Test

4. CONCLUSION

Customer Behaviour Analysis is done by data mining and using machine learning KNN classification algorithm from the experimental result, the customer who buys any unusual product or group/combination of unusual products is classified as High Risky customer and customer who buys usual products are classified as Low Risky Customer.

REFERENCES

[1] "Pulwama attack: Chemicals for making bomb bought online from Amazon - The Week." <https://www.theweek.in/news/india/2020/03/07/Pulwama-attack-Chemicals-for-making-bomb-bought-online-from-Amazon.html> (accessed Jun. 19, 2020).

[2]"ChemicalWatch." <https://chemicalwatch.com/67611/feature-e-commerce-and-the-challenges-of-enforcing-a-chemicals-policy> (accessed Jun. 19, 2020).

- [3] K. Maheswari and P. P. A. Priya, "Predicting customer behavior in online shopping using SVM classifier," Proc. 2017 IEEE Int. Conf. Intell. Tech. Control. Optim. Signal Process. INCOS 2017, vol. 2018-Febru, pp. 1–5, 2018, doi: 10.1109/ITCOSP.2017.8303085.
- [4] Kareena and N. Kapoor, "A review on consumer behavior prediction using data mining techniques," Proc. 2019 6th Int. Conf. Comput. Sustain. Glob. Dev. INDIACom 2019, pp. 1089–1093, 2019.
- [5] T. K. Das, "A customer classification prediction model based on machine learning techniques," Proc. 2015 Int. Conf. Appl. Theor. Comput. Commun. Technol. iCATccT 2015, pp. 321–326, 2016, doi: 10.1109/ICATCCT.2015.7456903.
- [6] "Data Mining and Its Importance." <https://www.loginworks.com/blogs/217-data-mining-and-its-importance/> (accessed Jun. 19, 2020).
- [7] "Introduction to Classification Algorithms - DZone AI." <https://dzone.com/articles/introduction-to-classification-algorithms> (accessed Jun. 19, 2020).
- [8] "Commonly Used Machine Learning Algorithms | Data Science." <https://www.analyticsvidhya.com/blog/2017/09/common-machine-learning-algorithms/> (accessed Jun. 19, 2020).