ISSN 2321 3361 © 2021 IJESC

IJESC

Research Article Volume 11 Issue No.08

Predicting Contact Lens Purchase Preferences by Consumers using Machine Learning

Bandna Kumari
Reva Academy for Corporate Excellence, Bengaluru, India

Abstract:
Study on consumer engagement and their online purchase behavior is an important concept in today's word. Consumer behavior information is playing an important role in gaining market share for companies especially during COVID 19 pandemic. Online consumer behavior is an emerging area and is being studied at length by lots of researchers. The latest statistics shows that there is a huge shift in consumer behavior for e-purchase from traditional shopping method. It is very evident because you see most of the retailers are also offering online interface to consumers. Online shopping help people to explore and purchase contact lenses online without any intermediaries along with many other facilities such as product return, cash on product delivery etc. This research paper presents interesting insights about consumer preference on contact lenses and social media influence in making those purchases. The goal of this study is to know the consumer behavior towards buying contact lenses and recommend organization to prepare their marketing plan towards addressing the outcome. Researcher will focus on the classification of consumer preference by using different classification algorithms like KNN, Random Forest, Naive Bayes, Decision Tree and Logistic Regression.

Keywords: Consumer Behavior, consumer purchase preference, contact lenses, e-commerce, purchase preferences, classification, Machine learning, KNN, Logistic Regression, Random Forest.

I. INTRODUCTION

The global market for contact lenses is evaluated at USD 12.79 million and is expected to grow at a CAGR of 5.7%. India being a millennial country this market is continuously growing too and is further expected to grow at the CAGR of 7.05% from 2020 to 2025[1].

Indian contact lens market is segmented by category, by type, and by distribution channel. The factor contributing to contact lens market expansion is the organization's reach to consumers through different mediums. In the era of the internet, there is a fundamental change in every human's life[2].

People, including families and friends, switching to online shopping much faster especially during the pandemic. However, there are a set of people who still prefer to make offline purchases[3].

Contact lenses as a product are no different in the market space, consumers are buying contact lenses online. Most of the population in India have access to the internet, and in some way or otherthey are exposed to social media.

For consumers who make purchase offline, lots of factors such as trust credibility and security come their way while deciding on the mode of purchase[3]. Still, the online shopping trend has been continuously picking up the pace by providing not only products but the best of services that are flexible and unique[4].

We will understand the pattern of purchases done by the consumers on the contact lenses and build a model that classifies the consumer's future purchase preference. The fig.1 shows the image of contact lens.

Figure.1.Contact lens Image

Image source: vision direct

The researchers aim to study the purchase pattern of the contact lenses via online/offline sources and to predict the future purchase preferences of the contact lenses.

II. RESEARCH METHODOLOGY

The research methodology adopted in this paper is doctrinal & non-doctrinal. It is based on online resources and a consumer survey. Accumulation of data, questionnaires, and surveys were the best-suited mode for the collection of data. Classification techniques/algorithms (KNN, Naive Bayes, Logistic Regression Random Forest) are used for statistical modeling.

A. Data Collection:

Data Collection is the most vital part of our study. We obtained the primary data through surveys and used it for the analysis. We have collected data from a total of 1000 consumers. Respondents shared their views about their purchase preferences of the contact lenses and purchase frequency through online/offline platforms. Asking multiple questions resulted in a dataset that is used for analysis and deriving conclusions. The dataset had 1000 records and 12 variables.

III. DATA OUTLOOK

The starting point for analyzing the dataset is the data cleaning step. Performed initial and basic clean-up and some feature engineering. The missing values were imputed with the average for age and mostly repeated value for response. Age variable was log transformed to remove any skewness. Non numeric variables were encoded to convert into numeric. One-Hot Encoder technique was used for converting the text variables into 0 and 1. Variables that were highly correlated and were not useful for further analysis were removed. The data is split using an 80/20 ratio for train and test set respectively.

A. Survey Respondent's Profile & Demographics Details

At the beginning of the survey respondents were briefed about the survey and its relevance. The respondents were asked their age in the survey questionnaire. Most of the respondents, male and female were of the age between 18-35 years and were exposed to internet and social media. About 62.3% of the respondents were female's and 37.7% of the respondents were males. Female's average age is 24years and male's average age is 29 years. Fig. 2 shows the male and female ratio in the collected dataset.

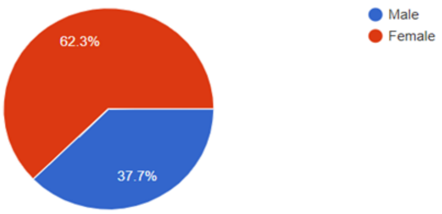

Figure.2. Male and Female Ratio

B.Buying behavior:

In the buying behavior, understanding whether they are buying contact lens from an optical store or an online merchant was attempted.The variables collected during the survey are age, gender, vision correction needed, preferred choice for vision correction, wearing spectacles or contact lenses, regular contact lens wearer or occasional, purchasing online or offline, why the preference, the social media impact on their choices, purchase frequency.

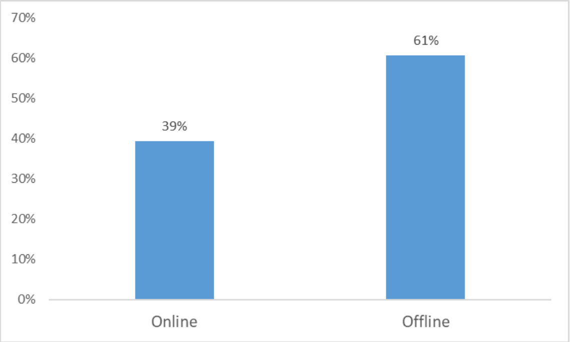

Figure. 3. Purchase Pattern online vs offline

Figure 3 shows that 39% of consumers prefer buying contact lenses online. Out of total consumers buying contact lenses online, 71% female and 29% male are making the purchase online.The percentage of consumers buying contact lenses online is much lower in the normal scenario, at around 21%, revealed the data. COVID 19 pandemic has brought the shift towards online purchases. The aim of the study is to see if this behavior shift toward online purchases will continue and becomes a way of life. Data shows around 61% of the customers buy contact lenses offline. Most of the respondents informed that due to Covid 19 pandemic they prefer to visit the stores less and prefer to use spectacles instead of purchasing the contact lenses online. Whereas during normal times, the respondents prefer to visit the stores regularly for the purchase of the contact lenses for their day to day life use.

C. Preferred Online Channels

Consumers buying contact lens online mentioned during the survey that Lenskart® is the first choice to buy contact lenses. The respondents get information about the latest trend on contact lenses beauty and fashion segment from Facebook®, Instagram®, and YouTube®. Fig 4 shows the preferred channels for contact lens online purchase

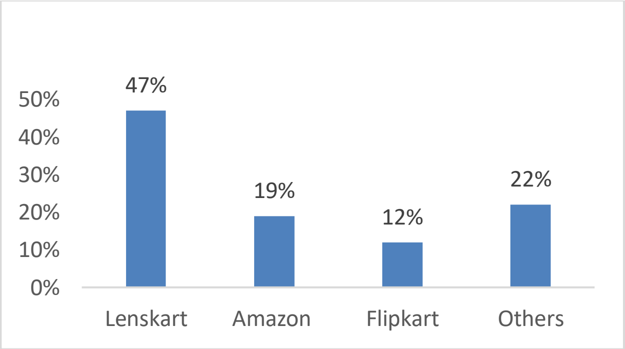

Figure.4. Preferred Online Channels

IV. MACHINE LEARNING APPROACH

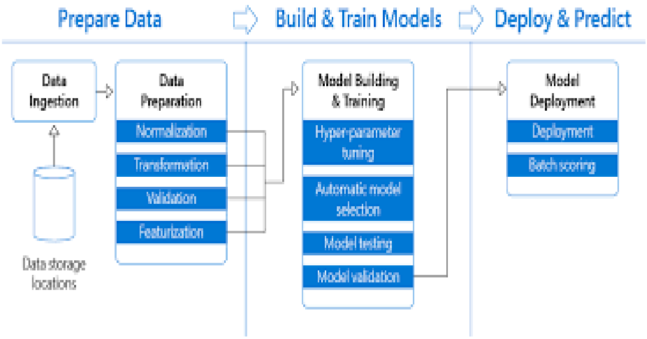

Figure.5. Stages of the Machine learning lifecycle (Image Source: Medium)

Figure 5 illustrates the working of a machine learning algorithm. The first step in our machine learning approach is to build a classification model that would be used to classify consumer purchase preferences in the future. This is a binary classification problem[5]. The researchers have tried multiple classification algorithms e.g. k-nearest Neighbors, Decision Tree, Random Forest, Logistic Regression, and Naive Bayes to classify target variables that are to understand thefuture purchase preference for the contact lenses, (Yes/No). The data was divided into training and test datasets in the ratio 80% and 20% respectively. Modeling is done on the training dataset and the test dataset is used for making predictions.

A. Classification Approach

The goal of the classification problem is to identify the class/category for the new data based on the past/historical learnings by an algorithm. The dataset was split into train and test datasets, 80%, 20% respectively to build classification models. Data cleaning, removing the nuances from the data is the key to building any machine learning model.

B. K-Nearest-Neighborhood Classifier

K-nearest Neighbors (KNN) is a simple yet accurate data classification technique. This classifier works on analogy, a comparison between test tuples and similar training tuples. In this algorithm, the data structure is not assumed[6].

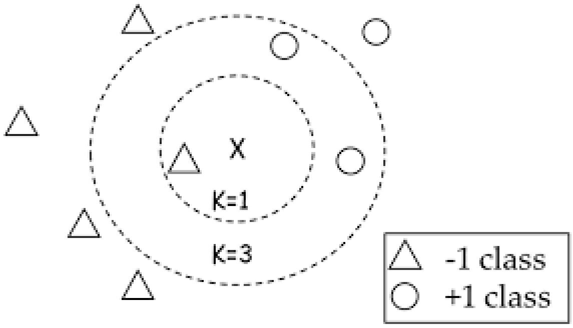

Figure.6. Illustration of KNN Algorithm
Image Source: research gate

Fig. 6 shows how a KNN algorithm works. The method used for continuous variables is Euclidian Distance. The formula for Euclidian Distance between two points is defined as:

$$d(p,q) = \sqrt{}(\sum\_(i=1)^n 〖(q\_i - p\_i)〗^2)$$

p, q = Two points in Euclidean space

p_i, q_i *= Euclidean vectors starting from the origin in space (initial points)*

n = n-space

The KNN algorithm has a disadvantage that is the value of ‘k’, it is complex to determine the value of ‘k’ while applying the KNN. ‘k’ represents the number of nearest neighbors. We considered the value of ‘k’ as 10 in our model after parameter tunic. The formula used to decide the value of ‘k’ is √n.

C. Random Forest

An important feature of Breiman’s algorithm is the variable importance calculation[7]. This algorithm will make a tree and will sort attributes based on data values just like a conventional tree. Each tree works on a random vector. The tree has branches and nodes where each node represents the variable group for classification. Each tree will have similar distribution for all the trees in the forest. The advantage of random forest is, that it can be used for both classification and regression problems. Prediction in the random forest is made by aggregating the predictions of the ensemble[8]. It is very easy to measure the relative importance of the features in this algorithm using the sklearn library. Usually, the problem of overfitting in machine learning is faced, however, it is not the case with Random Forest.

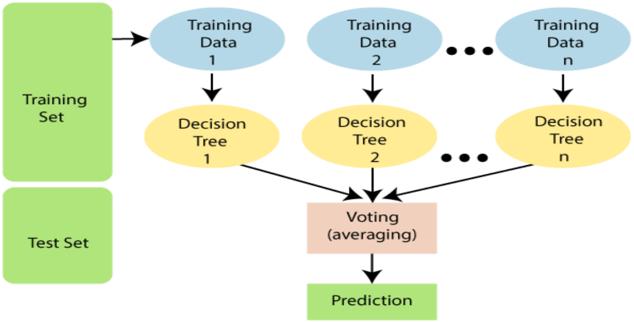

Figure.7. Random Forest Model
Image Source: javat point

This model is similar to the decision tree but harder to interpret. This model was implemented using the Random Forest Classifier package in python. We build the model on random data and considered 75% of the total data. During the model training process, each decision tree produces some predictions. In this technique, the nodes are expanded until all leaves are pure. When a new data point is added, based on the high prediction, it is ranked according to the result of the prediction. Multiple trees were built to arrive at the most accurate number.

D. Decision Tree

A decision tree is a supervised learning technique best used for classification problems. This is a hierarchical design that works on the divide and conquer approach[7]. It is a flowchart in which an internal node represents a “test” on an attribute, each branch represents the outcome of the test and each leaf node represents a class label[9]. Tree-based methods provide high accuracy, stability, and enable easy analysis of the predictive models. The below graph illustrates the prediction process by a decision tree algorithm for a binary class**.**

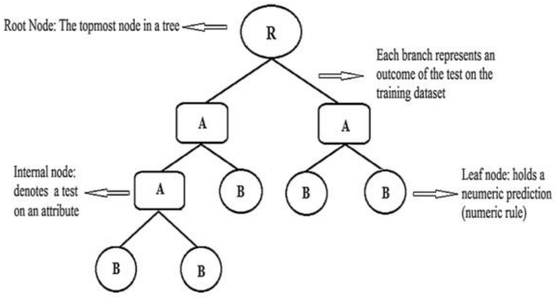
Root Node: The topmost node in a tree
R
Each branch represents an outcome of the test on the training dataset
A
A
Internal node: denotes a test on an attribute
A
B
B
B
Leaf node: holds a neumeric prediction (numeric rule)
B
B

Figure.8. Decision Tree Structure
Image Source: towardsai

The decision tree is used for both regression and classification problems. The decision tree is expensive to compute and often shows the problem of overfitting. To avoid overfitting, the splitting needs to be stopped beyond a certain number of nodes.

E. Logistic Regression

Logistic Regression (LR) is one of the most widely used techniques for classification by many statisticians[10]. It is pretty similar to linear regression. The challenge with linear regression is that it assumes all the variables are independent of each other, which may not be true for most of the datasets. Method to study the Logistic function also called sigmoid function. It is an S-shaped curve that can take any real-valued number and map it into a value between 0 and 1. If an input is greater than 0.5 then the corresponding value is considered from class 1.

$$1/((1+e^{\wedge}(-value)))$$

The logistic regression equation is

$$y = e^{\wedge}((b\_o + b\_1\,x))/(1+e^{\wedge}((b\_o + b\_1\,x))\,)$$

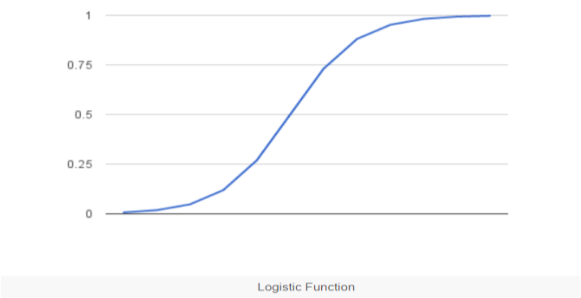

Figure.9. Logistic Function
Image Source: machine learning mastery

F. Naive Bayes Classifier

Naive Bayes classifier works on the Bayesian Probability theorem[11]. It is considered a semi-supervised classification method because it can be used both for clustering and classification. While we apply the Naive Bayes algorithm, it requires both an input and the target variable. Naive Bayes is a way of calculating conditional probability

$$P(c|x) = (P(x|c) \times P(c))/(P(x))$$

$$P(c|x) = P(x\_1\,|c) \times P(x\_2\,|c) \times \cdots \times P(x\_n\,|c) \times P(c)$$

- P(c|x) is the posterior probability of the target
- P(c) is the prior probability of the target
- P(x|c) is the likelihood which is the probability target
- P(x) is the prior probability of predictor.

Naive Bayes classifier is a probabilistic classifier, it is a collection of classification algorithms. In Naive Bayes, every variable is classified as independent of each other. The equation which is used is defined as:

$$P(A|B) = (P(B|A) \times P(A))/(P(B))$$

V. MODEL VALIDATIONS

We have used some criteria for validating the outcome of the statistic models in this paper.

- ROC curve: this curve plots two parameters, a) True Positive Rate

- True positive rate (TPR) is defined as

$$TPR = TP/(TP+FN)$$

- False-positive rate (FPR) is defined as

$$FPR = FP/(FP+TN)$$

We have also evaluated the model's accuracy by the most effective machine learning metrics.

A. Accuracy: It is the most commonly used measure for classification model performance. It is defined as:

$$Accuracy = (TN+TP)/(TP+FP+TN+FN)$$

B.Precision: It tells how precise the model is predicting positives. The formula is:

$$Precision = TP/(TP+FP)$$

C. Recall: It tells the percentage of true positives the model can identify. It is identified as:

$$Recall = TP/(TP+FP)$$

The above parameters are used to determine the model accuracy.

B. Model Evaluation:

Classification Algorithm	Precision	Recall	f1-score	Support
KNN	0.87	0.69	0.72	13
Decision Tree	0.66	0.62	0.63	13
Naive Bayes	0.74	0.77	0.75	13
Logistic Regression	0.51	0.46	0.49	13
Random Forest	0.66	0.62	0.63	13

The best output is provided by the Naive Bayes model with an accuracy of 77%. We can improve the accuracy even further by adjusting and fine-tuning the parameters. There is future scope of improving the model accuracy by using deep learning models.

VI. RECOMMENDATIONS

Results of the survey are quite evident and show the change in consumer's purchasing preferences in recent times and their choices to purchase contact lenses through the online channel. The study results shows that at present it is very important for organization to engage with consumers through various online platforms. The classification model's outcome suggests that in future majority of consumers are going to purchase the contact lenses through online sources. Organization must make digital marketing strategy to touch base consumers and spend marketing budgets to get consumer's attention at several points. People even in tier III cities are now moving towards online shopping so it is a good opportunity to reach out and engage with consumers through online forums. Online sampling for contact lenses or consultation by ophthalmologist could be some of the areas to invest in for better return on investment for future. In addition to this, a product complaint handling service or user interface can be an added advantage. However, it is also important to be mindful of the fact that the still lot many consumers are buying offline and the retail presence cannot be ignored. The study tells that how combined services, online and offline are the most efficient ways to reach to the consumer and increase your brand reach.

Limitation of the Study: The limitation of the study was the geographical coverage, the survey was done in Himachal Pradesh, District Hamirpur in north India state, so the results cannot be generalized. It would help the organization to take up informative business decisions in the state only. The future scope of the study is to conduct a consumer survey for all of India and make the prediction for a larger organizational benefit. Also collecting more variables would be more beneficial.

VII. REFERENCES:

[1]. BUSINESS WIRE, "India's Contact Lens Market," Bus. Wire, 2019, [Online]. Available: https:/ /www. businesswire. com/ news/home/20190830005342/en/Indias-Contact-Lenses-Market-to-2025---ResearchAndMarkets.com.

[2]. M. A. Rahman, M. A. Islam, B. H. Esha, N. Sultana, and S. Chakravorty, "on Dhaka City, Bangladesh," 2018.

[3]. M. A. Rahman, M. A. Islam, B. H. Esha, N. Sultana, and S. Chakravorty, "Consumer buying behavior towards online shopping: An empirical study on Dhaka city, Bangladesh," Cogent Bus. Manag., vol. 5, no. 1, p. 1514940, 2018.

[4]. S. Aravinth, "A Study on Customer Preference towards," vol. 2, no. 1, pp. 44–51, 2012.

[5]. A. E. Mohamed, Comparative Study of Four Supervised Machine Learning Techniques for Classification, vol. 7, no. 2. 2017.

[6]. T. M. Cover and P. E. Hart, "Nearest Neighbor Pattern Classification," pp. 1–12, 2018.

[7]. J. Ali, R. Khan, N. Ahmad, and I. Maqsood, "Random Forests and Decision Trees," vol. 9, no. 5, pp. 272–278, 2012.

[8]. Y. L. Pavlov, "Random forests," Random For., pp. 1–122, 2019, doi: 10.1201/9780429469275-8.

[9]. K. Shim and R. Pastogi, "PUBLIC: A Decision Tree Classifier that Integrates Building and Pruning," Data Min. Knowl. Discov., vol. 768, pp. 1–28, 2000.

[10]. J. Lever, M. Krzywinski, and N. Altman, "Points of Significance: Logistic regression," Nat. Methods, vol. 13, no. 7, pp. 541–542, 2016, doi: 10.1038/nmeth.3904.

[11]. D. Berrar, "Bayes� theorem and naive Bayes classifier," Encycl. Bioinforma. Comput. Biol. ABC Bioinformatics; Elsevier Sci. Publ. Amsterdam, Netherlands, pp. 403–412, 2018.