Predicting online buying behaviour – a comparative study using three classifying methods

Sanjeev Prashar*, T. Sai Vijay and Chandan Parsad

Indian Institute of Management Raipur (IIMR),
Old Dhamtari Road, Sejbahar,
Raipur-492015, Chhattisgarh, India
Email: dr.sanjeev.prashar@gmail.com
Email: tsavijay@gmail.com
Email: chandanparsad@gmail.com
*Corresponding author

Abstract: Online retailing with its increasing foothold has made India one of the most anticipated destinations for both local and multinational retailers. The success of these online retailers will depend on their ability to attract more and more consumers to shop online. Therefore, it is pertinent to comprehend and understand consumers’ attitude and behaviour towards online shopping, besides predicting online buying potential. This empirical study investigates the accuracy of prediction of different classifiers used in determining online buying. We empirically compared the forecasting ability of neural network (NN), linear discriminant analysis (LDA) and k-nearest neighbour (kNN) in the context of consumers’ willingness to shop online. Statistical evidence has been provided that neural network significantly outperforms the other two models in terms of the predicting power. This study shall contribute to online retailers in reducing their vulnerability with respect to market demand and improve their preparedness to handle the market response.

Keywords: online buying; predicting buying behaviour; neural network; e-commerce; k-nearest neighbour; kNN; linear discriminant analysis; LDA.


Biographical notes: Sanjeev Prashar is a Professor at the Indian Institute of Management Raipur India in the area of Marketing Management. Prior to this, he worked with Institute of Management Technology (IMT), Ghaziabad, as a Professor in the area of Marketing Management. His areas of interest are impulse buying, online buying, mall selection, rural marketing, international marketing/exports management and marketing of services. He has published research papers/case studies in various journals like – Richard Ivey School of Business, Emerald Group Publishing Limited and Journal of Retailing and Consumer Services.

T. Sai Vijay is presently enrolled as a doctoral student in the area of Marketing at the Indian Institute of Management (IIM) Raipur. His research interest lies in: studying consumer buying behaviour for new and innovative products, application of game theory in consumer decision making, customer relationship marketing, and online consumer behaviour.
1 Introduction

Even though the future of e-commerce is highly volatile, many new companies have been entering to tap this mode of retail sales. It is estimated that online retailing was just 0.9% of the total retail sales in India, during the financial year 2014 to 2015 (Retail and Ecommerce, 2015). The industry is set to reach US$16 billion in 2015 from US$44.4 billion in 2010, of which 61% would be accounted by travel industry and 29% by e-tailing industry (PTI, 2015). The major factor that will fuel the growth of online retail sales in India is its small (19%) but quickly rising web population of 243 million in 2014 (PWC, 2015). Though the prospects for success of online retail stores are very high, still it will depend on ability of the online stores to attract substantial number of consumers who are interested and are willing to buy products through these stores (Gao and Bai, 2014). In the domain of e-commerce marketing, studies related to behavioural aspects is sparse and highly disjointed. Even though there is a continuous rise in the number of shoppers’ preferring to buy online products, in India, the growth of some online retail stores and the closure of some others, accentuates the necessity for examination of behavioural issues, along with concerns related to performance and quality. As a result, for an e-retailer, it is pertinent to understand and comprehend shoppers’ attitude and behaviour towards online shopping. Besides understanding the same, e-retailers must decipher the determinants of online shopping behaviour and develop the predictive models regarding shoppers’ willingness to buy. The importance of predictive models which would help online retailers in managing their forecasts accurately can be seen by the failure of Indian online retail giant, Flipkart’s, failure of managing the demand for products on ‘BigBillionDay’, which left many of the e-shoppers’ unhappy and dissatisfied (DNA, 2014).

Vast pool of empirical studies based on surveys and statistical analyses are available in the domain of online consumer buying. In their study, Liang and Huang (1998) elucidated consumers’ willingness to shop online and approval of this medium, by studying their perceptions of transaction-costs in context of shopping process, ambiguity therein and realisation of desired product. As per the study, consumers choose to shop in traditional retail stores, rather than on the internet-based retail stores. Also, it was found that consumers’ acceptance level of web stores varied with the product category they were interested in. Using the data from nearly three thousand respondents, Craig-Lees et al. (2013) attempted to delineate the elements that effected consumers’ commitment and revisit behaviour towards an online website. Browsing the web and previous experience with online stores emerged to be the predominant factors that could predict consumers’ intentions, regardless of the fact whether or not he/she had bought anything online. Later, to measure consumers’ online shopping experience, Hung et al. (2014) conducted a comprehensive study and identified the determinants of online satisfaction. Contentment with e-commerce was found to be positively associated with shoppers’
perceptions about convenience, information, websites content and design, offers and security of financial transactions of online stores.

Urban et al. (1993) have posited that most of the empirical studies have adopted statistical techniques, which were generally known as preference regressions. All of these studies have assumed that consumers’ channel or medium evaluation is a linear combination of various factors and are linear compensatory in nature. Also, these models assume that underperformance or shortage in one element (e.g., touch and feel of the product) can be overcome by improving attributes of other elements of the channel (like lower price of the product). Because of their capability to replicate consumer choices and are assessed by various statistical techniques like ANOVA, discriminant analysis, logit and probit model, etc., these linear combinatory models have been extensively used by studies. However, the issue of reliability of these models has been questioned time and again. To overcome this, techniques from artificial intelligence and predictive modelling are being used as decision aids in a number of management disciplines (marketing, finance, manufacturing, etc.).

Despite of various predictive techniques, no study is available that has compared the predictive ability of such techniques in the domain of online buying. The objective of the present study is to compare the accuracy of three different binary classifying approaches – neural network (NN) analysis, linear discriminant analysis (LDA) and k-nearest neighbour (kNN) in predicting the consumers’ willingness to buy online. The findings of the study would enable online retailers and marketers to identify more accurate predictive technique in their endeavour to serve customers better with the least cost combinations of inventory carrying cost and opportunity cost.

Using the five indicators – accuracy, precision, negative predictive value, recall or sensitivity, specificity, the three predicting techniques were compared. The comparison was further tested through hypothesis testing, which confirmed the results. Area under receivers operating characteristic (ROC) curve was also used to further supplement the results. The paper has been organised as follow – opening section of the paper discusses the relevant theoretical background regarding the determinants of online shopping. Next, select predictive techniques used in this paper have been explained in brief, followed by description of the research methodology used. After the construction of hypotheses, the paper explains the results. The paper closes with discussion, conclusion and limitations.

2 Theoretical background

E-commerce had been defined as the medium of business that uses computer systems and internet technology to distribute goods and services. By using telecommunication networks, e-commerce helps in sharing business information by storing transactions and dealings that have been conducted (Zwass, 1996). Treese and Stewart (1998) defined e-commerce as the world-wide phenomenon that uses internet for ordering and selling of services and products. Shopping online is a complex phenomenon influenced by gamut of factors ranging from web atmospherics, availability of information, transaction convenience to augmented factors like image (Morganosky and Cude, 2000; Wolfinbarger and Gilly, 2003; Prashar et al., 2015). These influencing factors have been identified by different researchers in different markets.

According to Donthu and Garcia (1999), the inclination to shop online as compared to brick and mortar stores has been accredited to shoppers’ convenience orientation.
Convenience has been defined by Constantinides (2004) as simple and fast information browsing, shopping and completion of the transaction. As per Gehrt et al. (1996), convenience can be described in many ways depending on space or time. As per Morganosky and Cude (2000) this preference has become more significant because of presence of many constraints in visiting stores. This includes advantages of shopping from home (Eastlick and Feinberg, 1999; Jiang et al., 2013). In their study, Thomson and Laing (2003) identified three reasons of consumers opting for online shopping – drastic reduction in shopping time, freedom and flexibility to shop whenever they want, and the need of very little physical exertion for shopping. Consumers’ perceived usefulness of online stores and the perceived ease-of-use of these websites were found to have significant positive impact on repurchase intention at the same online stores (Aren et al., 2013; Gao and Bai, 2014). However, earlier, Kim and Stoel (2004) had reported that ease-of-use did not have any significant impact in determining customer satisfaction. Their study reported that content and transaction related qualities of a websites were more important for customer satisfaction, whereas design qualities were not significant.

Since online shopping provides consumers with benefits of easier comparison of products’ features and prices, the amount of time a consumer spends on the portals becomes an essential factor (Rowley, 1996). Frequent purchasers consider the websites design as the most important aspect of judging the quality of the products at the online stores. There exists a positive effect of websites design on customer satisfaction (Wolfinbarger and Gilly, 2003; Gao and Bai, 2014; Wu et al., 2014).

In addition to convenience and websites design, competitive price was found to significantly influence shoppers in favour of online buying, across all demographic variables. Shergill and Chen (2004) posited that shoppers believed that online medium provides goods/services at very competitive price. On-time and proper delivery of goods are found to have a significant positive impact on customer satisfaction (Xia et al., 2008; Muhmin, 2011). The provision of warranty was found irrelevant and did not have any influence on the buyers’ satisfaction or trust (Sonia San and Carmen, 2009). Transaction capability of the web portals and their receptiveness were found to have substantial influence on consumer satisfaction with respect to online shopping (Kim et al., 2004; Liu et al., 2008). As per Kim and Stoel (2004), the ability of the websites to respond quickly had strong positive significant impact on buyers’ satisfaction. However, Xia et al. (2008) did not find any impact of rapid response time on customers’ satisfaction. Besides these two, the influence of payment mode on customer choice of an online web portal was established by various studies (Kim and Stoel, 2004; Xia et al., 2008; Muhmin, 2011).

The quality of services being provided by online web stores is a strong determinant in customers’ satisfaction and loyalty towards an e-commerce websites (Lin and Sun, 2009; Sonia San and Carmen, 2009). The buyers’ intention of purchasing goods from an online store is determined by the perceived value he/she gets from that portal (Ching-Wen and Hsi-Peng, 2007). A study by Kim and Stoel (2004) confirmed the positive impact of adequacy of information available on the websites. Online shopping customers who felt that the information provided by these websites is adequate for their task were found to be satisfied. Another study reported that buyers were likely to be committed to online stores when there would be satisfaction from the information provided and other relational benefits (Park and Kim, 2003). Even the customers’ loyalty towards online stores and their buying behaviour were found to be strongly influenced by information provided by these stores.
In the meanwhile, a study conducted by Chang and Fang (2013) opined that consumers were highly sceptical about the online security of their transactions as well as about their personal information. However, for people having considerable amount of online shopping experience, this concern was lesser than the new shoppers. Also, better educated consumers are more likely to spend time online, and so are less concerned regarding internet privacy and transactional systems security (Hui and Wan, 2007; Muhmin, 2011). To gain trust of the online shoppers, websites use compatible terms and conditions, which again is another significant factor influencing online buying. The simplified process and easy-return policy were found to be important in developing trust in an online buying environment (Tan, 1999; Wang et al., 2004). Research had identified that clear communication regarding return-policy and compensation to customers, in such cases, does have an impact on customers’ online buying behaviour.

One of the most significant issues for online retailers/managers is their ability to predict conversion rate of a shopper on online stores. According to Sismeiro and Bucklin (2004), “predicting and understanding online-buying behaviour is of utmost importance for e-commerce website managers”. This was investigated for the first time by Moe and Fader (2001). The results of their study on predicting consumer online behaviour were much better than the earlier linear compensatory models. Other studies too have been focusing on improving the conversion rate by understanding and scrutinising the shopping drivers. A study by Padmanabhan et al. (2001) forecasted the likelihood that the incomplete shopping trip during a visit leads to buying and whether that e-shopper would make a purchase in any future visit. Sismeiro and Bucklin (2004) have shown that browsing behaviour and experiences are predictive of online buying. For online retailers, it is highly crucial to tackle the issue of low conversion rate of website visitors. For this, they must be able to comprehend in details the antecedents that influence shoppers’ decision to buy online or not (Bucklin et al., 2002). According to the authors, most of the times people visit websites either to search for information and/or to purchase a product from the online site, which is not very clear for the web stores owners. The authors ascertain that the online retailers must attempt to comprehend individual consumer’s motive of shopping online, which will help the retailers/marketers to outline the best predictions for online shopping. Similarly, Rajagopal (2011) observed that most of the studies with respect to modelling consumer behaviour assume that consumers’ choices and preferences do not change over the period of time. The study provides framework that enables prediction of intangible elements associated with consumer decision making.

3 Predictive modelling techniques used

Predictive models basically work on the premise of defining rules for forecasting the values of one or more elements in a dataset (outputs) from the values of other elements in the dataset (inputs). Many predictive models have been developed over the years. These modelling methods use the procedures that have been developed by various methodological researches in the areas of statistics, pattern recognition and machine learning. In this study, we have used three predictive modelling techniques – NN, LDA and kNN to analyse online shoppers’ data. In this section, we present basic theory of three techniques that have been used as predictive models.
3.1 Linear discriminant analysis

One of the most commonly used techniques for classification purpose, LDA was developed by Fisher. A classification method, discriminant analysis is used for developing models that explains a crucial data class and/or to generate predictions for use on future datasets. In the conditions where the primary populations are multivariate normal and different groups have equal covariance structures, this method is considered ideal. In case, multivariate populations having unequal covariance structures, quadratic discriminant analysis is used. The resultant combination or linear classifier helps in dimensionality reduction before undertaking later classification. It explains a categorical variable by the values of continuous independent variables. LDA has been found to be very robust to deviations from them and works well with many ‘well-behaved’ datasets. As is the case with regression analysis, here also it is a common practice to remove all the insignificant variables from the study. Cox and Prasad (1988) used LDA to understand and examine the features of acquisitions of banks which have failed after some period of time. Kumar and Ravi (2007) used it for reviewing bankruptcy prediction techniques.

3.2 Nearest neighbours

A non-parametric statistical technique used for classification and regression purposes is kNN algorithm. Here, the purpose for which it is used – classification or regression analysis, determines the output of the analysis. Considered as a lazy learning, kNN algorithm brings together closest k records of the training sample data that have the maximum resemblance to the test, instead of building a model or function. Then, a majority vote is performed among the selected k records to identify the class label, which is allocated to the query record.

For kNN classification, the neighbours are taken from a set of objects for which the class is known. If the data consists of many dimensions, it is suggested that techniques like PCA and LDA be applied before using k-NN. In their study related to credit scoring, Henley and Hand (1996) had used kNN.

3.3 Neural network

The technique, NN was developed as generalised outcomes of mathematical models of human cognition through biological neurons. It is considered as an information processing system that has certain performance characteristics in common with human neural biology. NN are a family of methods that use a number of simple processors which are linked together to ‘learn’ the relationships between sets of variables. Just as the individual neurons in the brain provide ‘intelligent learning’ through their constantly evolving network of interconnections and reconnections, artificial NN functions by constantly adjusting the values of the interconnections between neural units. Besides being used as models for investigating the way that brain works, this technique has also been exploited for its mathematical properties as signal processors and nonlinear statistical models.

To recognise and categorise the credit worthiness of consumers’ (both urban and rural), Lee and Jung (1999) used the predictive capability of logistic regression model and NN models. The outcome of the study was varied; in case of rural consumers, NN
model gave better predicted values, whereas for urban consumers’ logistic regression performed superior. In another study, Tam and Kiang (1992) used discriminant analysis model and NN models to scrutinise the reasons behind failure of banks. The results of the study suggested that NN model had higher accuracy in predicting failure than discriminant analysis model. Similarly, West et al. (1997) showed that NN model had better predictive power as compared to discriminant analysis and logistic regression in the context of consumers’ brand choice decision.

4 Research methodology

The study was carried out using multi-method research design. Cross-sectional design has been used with the assumption that the sample is representation of the entire population. The first phase involved comprehensive investigation of existing literature online buying for identifying varied dimensions related to decision making in online shopping. A list comprising 28 antecedent factors was prepared and shared with expert panel comprising of academicians and industry professionals. This seven member expert panel was asked to examine the significance and influence of these factors on online buying. Finally, a structured questionnaire with 25 statements was constructed. Each statement measured the level of agreement about the significance of specific variable effecting consumers’ online buying decision. A five-point Likert’s scale was used to record the responses, where ‘1’ and ‘5’ represented a least and highest agreement with the statement respectively. Respondents’ willingness to buy online was measured using the statement ‘Will prefer buy online’, for which the responses were – ‘yes’ and ‘no.’ This responses to this statement formed the dependent variable for the study. Besides these, the respondents had share information regarding demographic variables like age, gender, income class, educational qualification and occupation.

This was followed by descriptive design phase that involved collecting data from the desired sample. Population for the present study comprised of people above 18 years in India who have personal access to internet and use the same at least three hours a week. Since the objective of the study is to identify the factors that prompt online shopping and understand the respondents’ preference to buy using web portals, sample unit included people who had experience of shopping online. Convenience sampling technique was used for data collection and the questionnaire was distributed to 340 respondents in the emerging Indian cities of Raipur, Lucknow, Dehradun and Ranchi. These four state capital cities are smaller than major metropolitan cities. Sampling extent was defined as anyone who had the experience of minimum three online transactions.

After discarding incomplete and unsuitable questionnaires, 203 questionnaires were taken for further analysis. Classifying techniques – NN; LDA and kNN, were used to develop predictability model. Statistical software SPSS 21 was used to generate required outcomes. The data was partitioned into two subsets – the first subset, training set, contained 65% of the cases; and the second subset, test set, contained the remaining 67 cases. Data points in the training set are excluded from the test (validation) set. The training set was used for training/developing the model(s) and the outcomes were validated using the test set. Of the dataset with 203 responses, 65% (136) was used as training set, while remaining 35% (67) comprised the test set.
Since the model is assumed to be generalisable, its suitability was checked using split-sample validation. The derived models were assessed against test/validation set. For this, the actual value and the predicted value of each observation were classified as either 0 or 1. In case the estimated probability is more than 0.5, the predicted value of an observation is rounded off to 1 and if the estimated probability is below 0.5, then it is rounded to 0.

The performance of the algorithm undertaken was presented in a specific table layout – classification matrix/contingency table. Here, columns present the instances in a predicted class and rows represent the actual class. The combinations are shown in Table 1.

Table 1  Classification matrix

<table>
<thead>
<tr>
<th>Observed</th>
<th>Predicted</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Will not buy online</td>
<td>True negative (TN)</td>
<td>False positive (FP)</td>
</tr>
<tr>
<td>Will buy online</td>
<td>False negative (FN)</td>
<td>True positive (TP)</td>
</tr>
</tbody>
</table>

Using the values, the core measure – accuracy of predictions (predictability) of three models was calculated. It is the fraction of observations with correctly predicted classifications and is calculated by dividing (TP + TN) by (TP + FP + FN + TN). Besides this, other predictive indicators – precision or positive predictive value, negative predictive value, recall or sensitivity and specificity, were also calculated.

4.1 Hypothesis

Further, to observe statistical significance and importance of results, statistical tests are carried. This is aimed to comparatively assess the accurateness of predictions among the three predictive models – NN, kNN and LDA.

Following hypotheses are proposed:

H$_{01}$  There is no difference in prediction accuracy of NN and k-NN methods.

H$_{02}$  There is no difference in prediction accuracy of NN and LDA methods.

H$_{03}$  There is no difference in prediction accuracy of LDA and k-NN methods.

5  Results and analysis

The demographic profile of sample respondents is present in Table 2.

5.1 Classification tests: predictive accuracy results

To evaluate the forecasting ability of these techniques, empirical evaluation based on test set (out-sample) was undertaken and classification tables were drawn. The classification table for NN model is present in Table 3.
Table 2  Profile of respondents

<table>
<thead>
<tr>
<th>Demographic characteristic</th>
<th>Demographic sub-group</th>
<th>Count</th>
<th>Percentage</th>
</tr>
</thead>
<tbody>
<tr>
<td>Gender</td>
<td>Male</td>
<td>136</td>
<td>67</td>
</tr>
<tr>
<td></td>
<td>Female</td>
<td>67</td>
<td>33</td>
</tr>
<tr>
<td>Age group (years)</td>
<td>18–24</td>
<td>55</td>
<td>27.1</td>
</tr>
<tr>
<td></td>
<td>24–30</td>
<td>91</td>
<td>44.8</td>
</tr>
<tr>
<td></td>
<td>30–36</td>
<td>38</td>
<td>18.7</td>
</tr>
<tr>
<td></td>
<td>36–42</td>
<td>17</td>
<td>8.4</td>
</tr>
<tr>
<td></td>
<td>42 and above</td>
<td>2</td>
<td>1.0</td>
</tr>
<tr>
<td>Marital status</td>
<td>Unmarried</td>
<td>102</td>
<td>50.2</td>
</tr>
<tr>
<td></td>
<td>Married</td>
<td>101</td>
<td>49.8</td>
</tr>
<tr>
<td>Qualification</td>
<td>Schooling</td>
<td>6</td>
<td>3.0</td>
</tr>
<tr>
<td></td>
<td>Graduation</td>
<td>57</td>
<td>28.1</td>
</tr>
<tr>
<td></td>
<td>Post-graduation and above</td>
<td>119</td>
<td>58.6</td>
</tr>
<tr>
<td></td>
<td>Professional qualification</td>
<td>21</td>
<td>10.3</td>
</tr>
<tr>
<td>Occupation</td>
<td>Student</td>
<td>74</td>
<td>36.5</td>
</tr>
<tr>
<td></td>
<td>Government service/private sector</td>
<td>102</td>
<td>50.2</td>
</tr>
<tr>
<td></td>
<td>Professional (self-employed)</td>
<td>7</td>
<td>3.4</td>
</tr>
<tr>
<td></td>
<td>Business</td>
<td>13</td>
<td>6.4</td>
</tr>
<tr>
<td></td>
<td>House wife/retired</td>
<td>3</td>
<td>1.5</td>
</tr>
<tr>
<td></td>
<td>Others</td>
<td>4</td>
<td>2.0</td>
</tr>
<tr>
<td>Monthly family income (INR)</td>
<td>Less than 10,000</td>
<td>18</td>
<td>8.9</td>
</tr>
<tr>
<td></td>
<td>10,000–25,000 INR</td>
<td>12</td>
<td>5.9</td>
</tr>
<tr>
<td></td>
<td>25,001–40,000 INR</td>
<td>46</td>
<td>22.7</td>
</tr>
<tr>
<td></td>
<td>More than 40,000</td>
<td>127</td>
<td>62.6</td>
</tr>
</tbody>
</table>

Table 3  NN: classification table

<table>
<thead>
<tr>
<th>Observed</th>
<th>Predicted</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Will not buy online</td>
</tr>
<tr>
<td>Will not buy online</td>
<td>27</td>
</tr>
<tr>
<td>Will buy online</td>
<td>2</td>
</tr>
</tbody>
</table>

Similarly, Tables 4 and 5 provide classification tables associated with LDA and kNN, respectively.

Table 4  LDA: classification table

<table>
<thead>
<tr>
<th>Observed</th>
<th>Predicted</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Will not buy online</td>
</tr>
<tr>
<td>Will not buy online</td>
<td>21</td>
</tr>
<tr>
<td>Will buy online</td>
<td>4</td>
</tr>
</tbody>
</table>
Predicting online buying behaviour

Table 5  kNN analysis: classification table

<table>
<thead>
<tr>
<th>Observed</th>
<th>Will not buy online</th>
<th>Will buy online</th>
</tr>
</thead>
<tbody>
<tr>
<td>Will not buy online</td>
<td>17</td>
<td>13</td>
</tr>
<tr>
<td>Will buy online</td>
<td>1</td>
<td>36</td>
</tr>
</tbody>
</table>

On the basis of classification tables, performance of three predictive models was compared on five criteria:

a  overall percentage of accurate classifications
b  percentage of ‘yes’ accurately classified
c  percentage of ‘no’ accurately classified
d  sensitivity (or recall)
e  specificity.

The same has been present in Table 6.

Table 6  Performance indices of the three models

<table>
<thead>
<tr>
<th>Indicators</th>
<th>NN (%)</th>
<th>LDA (%)</th>
<th>k-NN (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Accuracy</td>
<td>92.5</td>
<td>80.6</td>
<td>79.1</td>
</tr>
<tr>
<td>Precision</td>
<td>92.1</td>
<td>78.7</td>
<td>73.5</td>
</tr>
<tr>
<td>Negative predictive value</td>
<td>93.1</td>
<td>84</td>
<td>94.4</td>
</tr>
<tr>
<td>Recall or sensitivity</td>
<td>94.6</td>
<td>89.2</td>
<td>97.3</td>
</tr>
<tr>
<td>Specificity</td>
<td>90</td>
<td>70</td>
<td>56.7</td>
</tr>
</tbody>
</table>

The term accuracy related to classification matrix is the fraction of the total number of predictions that were correct. It tells the researcher how often is the model correct? For the present study, the three methods – NN, LDA and kNN gave the accuracy of 92.5%, 80.6% and 79.1%, respectively. On the basis of predictive accuracy, we note that NN model, with accuracy of 92.5%, performed much better in predicting correctly than the other two models. Thus, more than nine predictions made out of ten, on the basis of the model, are correct. This implies that NN model correctly classifies 92.5% of web-visitors as either online buyers or non-buyers.

NN model also outperformed the other two models in two other criteria that we tested – precision and specificity. Precision tells us the percentage of predicted positive cases that were correct. For the NN model, the precision value came to be 0.921. This indicates that when the neural model predicts yes, 92% of the time it is correct. Specificity, on the other hand, represents the situation in which a case turns out to be a non-online buyer (‘no’) and the model had correctly predicted. It deals with the portion of dataset that was found actually negative of all the negative outcomes. This measures the goodness of a model at avoiding false alarms. NN model had the specificity value of 0.90. This implies that the model was able to identify 90% of the negative cases that were actually negative.

Here also, we see that NNs model is found to be more qualified than the other two models – LDA and kNN.
However, with respect to negative predictive value and sensitivity (or recall), kNN reflects better results. While negative predictive value represents the fraction of predictive negatives that is correct, sensitivity reflects the proportion of positive instances that were identified correctly by the model under consideration. This is basically to check how good a model is at identifying the positives. The kNN model gave a very high recall value of 0.973. This implies that the model correctly identified 97% of the positive cases. Though Table 6 shows the different predictive abilities of the three tested models, we are more interested in ascertaining whether the visible differences are really significant enough. Unless these differences are significant, it will not matter if we make use of one model or other for our forecasting purposes. To test this assertion, we had framed three hypothesis as discussed in the previous section.

5.2 Hypotheses testing

Since the data used for prediction in all the models is same, paired t-test on prediction accuracy is carried out to test the three hypotheses. The results of paired t-tests are shown in Table 7.

<table>
<thead>
<tr>
<th>Tests</th>
<th>df</th>
<th>t-Stat</th>
<th>p-value</th>
<th>Conclusion</th>
</tr>
</thead>
<tbody>
<tr>
<td>H01: NN vs. k-NN</td>
<td>202</td>
<td>–2.384</td>
<td>0.018</td>
<td>$\mu_{NN} &gt; \mu_{kNN}$</td>
</tr>
<tr>
<td>H02: NN vs. LDA</td>
<td>202</td>
<td>–6.721</td>
<td>0.000</td>
<td>$\mu_{NN} &gt; \mu_{LDA}$</td>
</tr>
<tr>
<td>H03: LDA vs. kNN</td>
<td>202</td>
<td>–5.047</td>
<td>0.000</td>
<td>$\mu_{LDA} &gt; \mu_{kNN}$</td>
</tr>
</tbody>
</table>

Since $P < 0.05$ in all the three cases, H01, H02 and H03 are rejected. This evidence indicates that the prediction accuracy of NN model is significantly different from the prediction accuracy of nearest neighbour model (reject H03). Similarly, in case of the second hypothesis, it can be inferred that there is significant difference in prediction accuracy between NN model and LDA model (reject H02). Same holds true for third hypothesis. Accordingly, the prediction accuracy of LDA model is significantly different from the prediction accuracy of nearest neighbour model (reject H03).

5.3 ROC or ROC curve

Using data from the validation set, ROC curves were drawn for the present work. These curves exhibit the performance of a binary classifier system as its discrimination threshold is varied. The true positive rate (sensitivity) is plotted in y-axis against the false positive rate (1-specificity) in x-axis for various decision cut-off points. Area under the curve (AUC) for an ROC curve describes the effectiveness of a prediction model/system by defining the model’s capability to correctly anticipate the happening or non-happening of pre-defined events (Brenning, 2005). It is known that the closer AUC is to 1, the better is the prediction accuracy. It shows perfect predictive accuracy, when it equals 1 (Lee and Dan, 2005). ROC analysis facilitates in selecting possibly optimal models and rejecting suboptimal ones. After obtaining prediction-rate results, areas under the curves (AUC) for three models, which represent the area under the prediction rates, were calculated. The results of the ROC curve test for three models – NN analysis, LDA and kNNs are illustrated in Figure 1.
From the figure, we note that ROC curve for NN is nearest to the upper left corner. Thus, it is inferred that NN models are the most appropriate predictive methods for the present case. Even though k-NN has better negative predictive value and recall level, its ROC curve is well below the other two curves of NN and LDA. From this, we can conclude that NN is a superior model as compared to LDA and k-NN classifiers.

Figure 1  ROC curve (see online version for colours)

A composite measure on the relative position of the ROC curve, the AUC of three classifiers used in the study is reported in Table 8.

Here again, as expected, the AUC is highest for NN classifier at 0.975, confirming NN model’s better predictive ability.

Table 8  Area under the curve

<table>
<thead>
<tr>
<th>Test result variables</th>
<th>Area under curve</th>
<th>Sig</th>
<th>Accuracy (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Neural network</td>
<td>0.975</td>
<td>0.000</td>
<td>92.5</td>
</tr>
<tr>
<td>Linear discriminant analysis</td>
<td>0.882</td>
<td>0.000</td>
<td>80.6</td>
</tr>
<tr>
<td>k-nearest neighbour</td>
<td>0.777</td>
<td>0.000</td>
<td>79.1</td>
</tr>
</tbody>
</table>

Note: Diagonal segments are produced by ties.
6 Discussion and conclusions

In this research, we performed a three-way comparison of prediction accuracy involving NN, LDA and kNN models. These three models had binary dependent variable and independent variables were all continuous. There have been many comparative studies on classifiers conducted in the past, but very few of these have used three models as used in this study. Also, very few of these have been on developing predictive models for online shopping behaviour.

In the present study, NN and kNN models produced superior prediction accuracy than linear discriminant model. On three prominent indicators – accuracy, precision, specificity, NN had better predictive ability. k-NN, on the other hand, was clearly ahead of NN on the indicator recall. Further, the superiority of NN was also established by the statistical technique (t-test) and area under the ROC curve.

Though NNs tend to provide best predictive accuracy, their disadvantage lies in the fact that they do not have any probability associated with it as is the case with other multivariate techniques (like LDA and logistic regression). Brown et al. (1993) have posited that NNs do better than statistical models on multimodal classification problems where datasets are large with few attributes. NN and nearest neighbour analysis were found to have better predictive ability than LDA on issues where the data is nonlinear (Curram and Mingers, 1994). Further supported by the hypothesis testing, NN model outperformed both LDA and kNN in its prediction accuracy.

Considering the exponential growth of e-commerce in India, this research contributes to the literature in diverse ways. To the best of our knowledge, it is first time that NN and other predictive techniques have been used and compared in the domain of online consumer buying behaviour within the environment of the Indian e-commerce industry. The predictor variables developed in this study can be used by online retailers as input variables for predicting online buyer behaviour. An examination of each of the predictor variable will show its relevance in contributing to online buying and this would help online marketers in understanding consumers’ decision whether to buy online or not. The major e-commerce giants may reduce the risk of failure because of understocking of products and ill-preparedness in terms of technological capabilities. There have been instances like Flipkart’s ‘BigBillionDay’ which failed miserably owing to lack of proper market forecasts (DNA, 2014).

The other major contributions of this study include – the use of three predictive modelling methods for predicting consumers’ willingness/preference to shop at online retail stores and comparison thereof. Further, this study compares the predictability in consumers’ willingness to shop online. According to Allenby et al. (2002), quality of a model depends on the accuracy of the prediction. For online retailers and managers, a good predictive model plays a significant role in analysing online consumers’ behaviour and will help them in forecasting sales. It is obvious from the study results that NN model provides better prediction compared to LDA and k-NN models when the predictor variables are binary and the dependent variable continuous. For application standpoint, either one of NN and nearest neighbour models may be used for prediction and would provide better predictability over LDA. However, more studies and different scenarios/conditions need to be explored in order to establish a clear distinction of performance between these models.

Though the NN models have great predictive advantage over the other statistical models, NNs have been criticised for having few drawbacks. The identification of
Predicting online buying behaviour

different factors such as number of hidden layers, number of nodes in the hidden layer, etc. related with NNs, is not direct and determining the ideal structure of NNs is a very intense process. Lack of interpretability of the weights between various nodes obtained at the time of model building process is another drawback of this method. In this aspect, statistical techniques are much better as they are easily interpretable and the values of the associated coefficients could be easily examined and due to the assumptions of these models, inferences can also be drawn regarding the significance of certain variables in prediction or classification problems. The study was limited to Indian market and so generalisation of the results are difficult. Future studies can minimise this by testing the predictability of methods using cross-country comparison. The study had employed cross-sectional design which could not have captured the dynamic aspects in predicting online buying. Future studies can incorporate longitudinal design to gain more insights and get more accurate results. This paper did not study the impact of demographic variables on the predictability of methods. Further research may study the impact of the same on predictability of these three methods.

References


Predicting online buying behaviour


Appendix

Variables influencing online buying

- Easy access to web portals
- Graphics used by the portals
- Design and layout of the websites
- Effectiveness of search process
- Faster navigability
- Ease of use
- Image of web portals
- Portfolio of products and services available
- Extent of information available on websites
- Quality of information available on websites
- Competitive pricing
- Less transaction time
- Terms and condition of sales
- Different payment options available
- Security of payment during transaction
- Delivery of quality products
- Conducive delivery time
- Privacy of information
- Discounts and promotional offers available
- Online and offline interaction
- Personalised features
- Previous experience
- Media reviews
- Return and exchange policy
- Availability of physical display