Scientific Research Publishing

Theoretical Economics Letters, 2018, 8, 448-475
http://www.scirp.org/journal/tel
ISSN Online: 2162-2086
ISSN Print: 2162-2078

A Random Forest Approach for Predicting Online Buying Behavior of Indian Customers

Rohit Joshi[1*], Rohan Gupte[1], Palanisamy Saravanan[2]

[1]Indian Institute of Management, Shillong, India
[2]School of Commerce and Management, Central University of Tamil Nadu, Tamil Nadu, India
Email: *rj@iimshillong.ac.in

How to cite this paper: Joshi, R., Gupte, R. and Saravanan, P. (2018) A Random Forest Approach for Predicting Online Buying Behavior of Indian Customers. *Theoretical Economics Letters*, **8**, 448-475.
https://doi.org/10.4236/tel.2018.83032

Received: November 30, 2017
Accepted: February 10, 2018
Published: February 13, 2018

Abstract

Online retailing in India has shown remarkable growth in the recent years. Despite having a low internet penetration rate of 34.5%, India has the second largest number of internet users in the world after China. Given the growing importance of the online retail industry in India and its diverse set of sensitivities and region wise socio-psychological barriers, it is imperative for retailers to understand customer shopping preferences. In this paper, we attempt to understand various factors influencing the online buying behavior of Indian customers in different product categories, across geographic locations in India. Also, we developed and validated the Random Forest prediction model for each identified product category, to understand if the Indian online shopping market is ready for these product categories or the traditional channel is preferred over by customer. A questionnaire based survey is used to collected data from 124 Indian respondents from 18 states of India. The survey captured from both offline and online shopping environment to aggregate understanding of customers' shopping preferences. The high Sensitivity (above 85%) of the Random Forest model for Books and Electronics categories suggests inclination of purchase intension of customer towards online shopping. Retailers can use this model to predict the buying behavior of customers based on the location. However, for product categories like Movies, Sports equipment and Handbags, the high value of Specificity signifies the model prediction towards offline purchase intensions. So for these product categories retailers may like to focus more on customer services at retail stores.

Keywords

Online Purchase Intention, Prior Online Purchase Experience, Geographical Location, India, E-Commerce, Retailing, Random Forest Model

1. Introduction

With the advent of electronic medium of information, today's customers are excessively aware of their choices and what they demand is a high quality product at rock bottom prices. With constant connectivity, customers browse products online and then buy it from physical store or conversely, check the product physically before completing the online purchase ([1] [2]). The customers' action from product enquiry to ensuing purchase often involves multiple shopping channels ([3]). They maximize their shopping utility by comprehensively considering all possible alternatives across all possible channels.

In the context of developed economies, there is the significant number of research studies done on factors influencing the shopping pattern among the customers. The factors like, shopping preferences, prior experience of online shopping have shown significant effects on purchase orientation ([4] [5] [6] [7]). However, the relevance of these findings is required to be validated in Indian context.

As on June 30, 2017, India has significant low internet penetration rate of 34.4% percent as compared to UK (94.8%), Japan (94%), Germany (89.6%), USA (87.9%), Italy (86.7%), Thailand (83.5%), Russia (76.4%), Iran (70%) [8]. Despite having such a low internet penetration rate, India has the second largest number of internet users (462 millions) in the world after China. India's count of internet users has been increasing at a CAGR of 35 percent from 2007. From 100 million users in 2010, the number touched to 462 million users by June 2017 ([8]). This large internet base will have a direct impact in the Indian internet shopping.

The Indian e-commerce market has region-wise dissimilarity. Tier-I cities like Bangalore, Mumbai and Delhi have the higher share of online purchase than the tier-II cities such as Ahmedabad, Chandigarh, Bhubaneswar, Kozhikode ([9]). Also, South India buys online more than North India (Contribution to overall e-commerce sales is South India-41%, North India-32%, West India-21%, East, North East-6%). As the e-commerce penetration is yet to mature in India, disparity among the geographical locations has been the predominant influence on customer online purchase intention. Culturally, India has its own unique set of sensitivities and region wise socio-psychological barriers, which is equally pertinent in e-commerce business, especially in context to offline shopping ([10]).

Due to these diverse customer shopping preferences, Indian retailers face immense challenge in serving both online and off-line channels' efficiently ([11]). This challenge further aggravates when the product wise acceptance of these online and offline channels bring uncertainty. Therefore, in order to maximize the profit lines through the online and off line market, first and foremost, retailers need to have a clear understanding of the Indian online users' preferences, attitude and outlook within the framework of socio-economical, technical and environmental system. It is thus required to empirically investigate the factors impacting the buying behavior of Indian customers in different locational set-ups

across India. Also, it is required that the findings which are established in developed economies setup are to be validated in Indian context to identify the dissimilarities.

In this paper, our goal is to understand various factors influencing the online buying behavior of Indian customers in different product categories across different states of India. Also, we endeavored to identify consumer behavior influence on multi-channel retailer to understand if the Indian online shopping market is ready for these identified product categories or the traditional channel is preferred over by the customer. So accordingly geography wise retailers could decide upon to focus their efforts on online, offline or on both the channels.

Keeping above aspects on Indian e-commerce industry, we attempted to provide unique insights into the Indian online consumer's preferences, attitude and outlook by empirically deliberating upon the determinants of their online purchase intentions. We developed these preferences with Random Forest Model to bring insights on retailers' strategies for different product categories based on geographic location.

The remainder of this article is organized as follows. First, we provide a brief overview of the literature, deriving our conceptual framework and factors impacting the online behaviour of consumers. Then, we describe our study's methodology. Using Random Forest modelling, we test our data from online and offline consumers in the context of selected product categories. Finally, we discuss the results and their implications for theory and practice.

2. Literature Review and Theoretical Background

There is a growing requirement for theories, new knowledge of determinants and their inter-relationships to depict the behavior of Internet consumer. This understanding may help retailers to work on the matter of customer relations, marketing strategy and in turn the overall business strategy ([12] [13] [14]).

Several studies ([3] [4] [5] [6] [15] [16]) have focused on shopping orientation of customers that brings insights into understanding the emergence of internet retailing in the different countries. In the context of developed countries, these studies have stated that the shopping orientation of the customer has significant impact on purchase intention. Conversion of the preference into final purchase is the final consequence of a number of various factors in an online shopping context. Furthermore, for understanding the behavioural mind-set of online shopping customer, retailers are continually trying to explore the factors of customer online purchase intention, prior online purchase experience and trust on online shopping structure ([4] [7]) have significant influence e-commerce markets. [4] [15] specified the importance of the effect of demographic factors like age, gender, education, location and the like factors on adoption of customer online purchase. Also, ([17] [18]) mentioned about the slow transition was visible in the younger age group (21 - 35 years) which later scaled up with mushrooming of number of e-commence websites, where the responsiveness and in-

teractive web sites provide them more opportunities for customized products. However, the validation of these findings in Indian context requires an in-depth empirical research. Moreover, given the diversity of India, the validation is required on cultural, social economic, infrastructural differences as well as their impact on product wise preferences of the customers.

If we glimpse into the Indian e-commerce sector, we can find that with rapid increase in the number of internet users and robust investment in the sector, India is expected to become the world's fastest growing e-commerce market. It is expected that Indian retail market will grow at a Compound Annual Growth Rate (CAGR) of 12 per cent from the US$ 641 billion in 2016 to US$ 1.6 trillion by 2026 ([19]). Indian retail market is divided into "Organised Retail Market" which is valued at $60 billion which is only 9 per cent of the total sector and Unorganised Retail Market constitutes the rest 91 per cent of the sector ([9]). According to a study conducted by Federation of Indian Chambers of Commerce and Industry (FICCI) and Indian Institute of Foreign Trade (IIFT), total potential of Business to. Indian e-commerce sales are expected to reach US$ 120 billion, by 2020 from US$ 30 billion in 2016 ([9] [19]). Further, India's e-commerce market is expected to reach US$ 220 billion in terms of gross merchandise value and 530 million shoppers by 2025, led by faster speeds on reliable telecom networks, faster adoption of online services and better variety as well as convenience ([9] [19]).

There are obvious reasons for such phenomenal growth in Indian e-commerce market like-raising per capita income of the middle class, busy schedules of working professionals, convenience of online shopping, changes in the supporting ecosystem, 3G, 4G services, launch of schemes like digital India by the Government of India, electronic medium encouragement from banks, public services, railways ([9]). Also, an inadequately developed distribution network in terms of supply chain has also contributed towards increasing online users in smaller cities to go online for shopping ([10]). However, there are few concerns also when it comes to the usage of e-commerce platform like, low average broadband speed and flat average internet speed, online payment landscape marred by low penetration of credit and debit cards, high failure rate of online payment transactions and most importantly the security issues. Probably that is the reason, though a majority of customers search for information on product categories online but relatively a smaller percentage of them actually buy online (Vazquez *et al.*, 2009).

Select studies in Indian context have focused on shopping orientation of customers that brings insights into understanding the emergence of internet retailing. [20] mentioned in their studies that socio-psychological factors and infrastructure are influential factors, while the perceived risk and gender level behaviour of male and female also impact the online buying behaviour. [21] and [22] found that Indian students' intention to purchase online is influenced by utilitarian value, attitude toward online shopping, availability of information, and

hedonic values. Their study was confined to college going students. A trend analysis on the online shopping in India by [23], points out that there is a growing awareness of getting more information through websites. There is an increasing trend of using Internet for booking tickets, buying books and music but the scene has not transformed dramatically in case of India. Kiran *et al.* (2008) mentioned that In India there is a growing awareness of getting more information through websites, however, when it comes to buying online, the scene has not transformed dramatically in case of India. [24] studied the mediatory influence of online trust. **Table 1** presents the factors considered for the study.

[25] suggested the retailer can optimize both profits of internet and traditional channel, without changing any change in traditional channel price, but that will result in less profit to the internet channel. In the similar line, [26] also reiterated that that the manufacturer can keep the wholesale price same as before and change retail price to optimize profits. The strategy to keep the retail price and wholesale price only when the e-tail channel is relatively hard to reach or not very convenient. Moreover, [27] presented that traditional retailers have the highest prices, followed by multichannel retailers, and pure play e-retailers.

Although a lot of research has gone into the models for direct retail channel and a distribution channel, a little has been mentioned about determining the demand from the customer end for a direct e-retail channel. To the best of our knowledge, there is no significant scholarly research that focuses on the retailers strategies based on the interplay of factors influencing Indian consumers' online purchase intentions. According to [28], the online buying behavior of products differs from location to location. Thus, location is another factor which requires attention and was adequately addressed in the earlier research. This research has increased the scope of demographic range *i.e.* this research is applicable on a much wider range of people than the current literature considered in Indian context.

Underpinning Theories It is important to note that for the purpose of this study to empirically test construct and relationships in Indian Context, rather than inclusion of conceptual approaches. The observational field study that used survey research method to collect information over a period of six months

Table 1. Identified factors having causal impact on consumer buying behavior.

	Demographics	Behavioral	Experience
Factors	Age	Information search	Years of internet use
	Occupation	Hedonic shopping motive	Cash on delivery preference
	Gender	Online purchase intention	Years of e-commerce use
	Marital status	Utilitarian shopping motive	Past purchase count
	Income	Attitude	Trust propensity
	Geography		Hours of daily internet use
		Source. ([4] [5] [6] [9] [15] [16] [18] [20] [21] [22] [23] [24])	

during November 2016-April 2017, was situational grounded on contextual idiosyncrasies. By conducting an in-depth investigation of the relationship among the factors, we intended to elaborate the theory of reasoned action (TRA) under the framework of consumer behavior. This provides a relatively simple basis for identifying where and how to target consumers' behavioral change attempts. The TRA family theories include the Theory of Planned Behaviour (TPB) ([29]), the Technology Acceptance Model (TAM) and the Unified Theory of Acceptance and Use of Technology (UTAUT) ([30] [31]).

As the consumer behavior literature suggest, customers are more likely to buy from a store that has a positive image on considerations like price or customer service. Over many years, this is an approach that has been demonstrated for traditional stores and shopping centers ([32] [33]). Intentions are the consequence of attitudes which has an influence of social aspects ([34]). Since online shopping is emerging over time, still in India online purchases are perceived as riskier than online ones and an online shopper therefore relies profoundly on experience qualities acquired through prior purchase ([35]). As subjective norm and attitude and cannot be the exclusive factors of behavior where an individual's control over the behavior is incomplete, the TPB implies to improve on the TRA by adding "perceived behavioral control" defined as the ease or difficulty that the person perceives of performing the behavior. The resource-based view (RBV) of the firm has been widely used in retail management research ([36] [37]). This theory is suitable for investigating the effects of people, technology, and information resources across service delivery systems ([38]).

The ontological position of this research is constructionism whereby we accept that the need and desire to make profits by retailers is built on the twin logic of enhanced customer service and customer satisfaction as the dominant outcome. Additionally, Equity Theory suggests that customers who perceive delivery of service quality of a retailer in conjunction with product quality are likely to attribute greater influence to the relationship with that retailer. The epistemological standing of this research is the critical realism in a sense that we recognize the need to identify the customer requirement and develop retailer strategies that provide optimum customer satisfaction. We sought situational groundedness using the conceptual framework with elaborated theoretical abstraction ([39]). This approach allowed us to explain and predict online consumer behavior in India and decision-making process of consumers, in general, and to examine the online behavior from the retailer perspective.

We have explored the determinants of customer online purchase intention for understanding the online shopping preferences, attitude and outlook. Therefore, in this paper, we attempt to examine impact of different product categories on customer purchase intention in the Indian context. Also, the study includes a proxy factor for location that is Geography or the state that the customers belong to. Thus, the study will be applicable to people from a larger part of India. This study includes a total of 18 factors their impact on eight product categories.

Thus, objectives of our study are:

1) To empirically test the impact of various factors influencing the online buying behavior of Indian customers in different product categories across geographical locations in India; and

2) To develop a Random Forest prediction model for each identified product category to categorize the customers preferences between online or offline purchase, and thus, retailers may modify their strategies accordingly.

3. Research Methodology

After careful examination of e-commerce websites namely—Flipkart, Amazon, Snapdeal and Quikr, a limited number of product categories of retail sector are chosen as the context of this study. These product categories are characterized by product attributes for which a continuum of information (from very less to too much) can be gathered prior to purchase. Also, these product categories are ranged from low to high involvement items, and consumers are typically engaged in a problem-solving task of no complexity to moderate or high complexity.

Following are the categories identified for the study:

1) Books;

2) Movies, Games & Music;

3) Electronics;

4) Home & Kitchen;

5) Sports & Fitness;

6) Jewelry;

7) Handbag and Luggage;

8) Cars & Bike.

A questionnaire based survey was chosen as a data collection instrument. The questionnaire was prepared by considering the aforementioned literature. The first part of the questionnaire had questions on demographic details like gender, age group, education level, income Geography. The second part focused on aspects like, information search, purchase intensions, attitude, and shopping motive, experience with ecommerce, device and mode of payment. Each of these question was presented as a five-point Likert scaled-response question with 1 being "strongly disagree" to 5 "strongly agree". The questionnaire is pre-tested by five PhD students and two professors with marketing and domain expertise to get their view on understandability, relevance and order of questions. Furthermore, the pilot test was conducted with thirty MBA students. Based on the suggestions, a few changes were done in the questionnaire. After incorporating these changes, the questionnaire was reviewed by the same two professors for the comments and finally, 8 product category, 18 variable were selected.

Respondents were screened for their age and their level of purchase experience with selected product categories. We invited the respondent proved to be eligible for the study (that is, he/she must be over 18 years, and should have

purchased at least once among the product categories product online or offline in the last six months). The questionnaire administered to the respondent to cover all strata of consumers. The online survey as well as paper survey method was chosen to cover more geographic area in terms of reaching out to the target sample, time, and cost. The final sample consisted of data from 124 consumers from 18 states. Table 2 shows the characteristics of respondents.

Random Forest Method (RFM) is used to establish the causal relationship among the factors determining the online purchasing behavior. [40] proposed the RFM that is a collaborative method that fits many classifications of trees re-sampled by the bootstrap method and then combines the predictions from all the trees. To achieve good prediction ability RFM uses variable importance to find the smallest set of predictor variables ([41] [42] [43]). RFM has gained some attention in past decade. We preferred RFM over logistic regression as unlike the logistic regression, tree-based methods do not assume a pre-specified relationship between the response and predictors. A tree-based method generates primarily the classification tree on the predictor variables constructed by recursively partitioning the data into successively more homogeneous subsets with respect to the variables of interest ([44]). Unlike logistic regression, where a statistical model that was likely to have generated that data is specified by the researcher prior to estimation, no "model" in the conventional sense is generated by Random Forests. Also, logistic regression analyses demonstrate the importance of each predictor to be able to explain the outcome variable. The odds ratios which is an important statistic in logistic regression does not provide information about relative priorities or importance among the predictive variables ([45]).

Random Forest Method is used where each tree is built based on recursive partitioning, and the prediction is made on the average of an ensemble of trees rather than of a single tree. Random forest trees generate predominantly the classification tree on the predictor variables, created by recursively partitioning

Table 2. Characteristics of the respondents.

Measure	Items	Frequency	Percentage	Measure	Items	Frequency	Percentage
Gender	Male	73	58.87	Occupation	Household	32	25.81
	Female	51	41.13		Professional	65	52.42
					Student	27	21.77
Age	18 - 24	43	34.68				
	25 - 34	51	41.13	Income	less than 1 L*	18	14.52
	35 - 44	23	18.55		1 L – 4 L	24	19.35
	Above 44	7	5.65		4 L – 8 L	45	36.29
					8 L – 10 L	23	18.55
Marital status	Married	47	37.90		Above 10 L	14	11.29
	Single	77	62.10	1 L* = 0.1 Million Rupees (1 USD = Rupees 65)			

the data into sequentially more homogeneous subsets pertaining to the variables of interest (44). Further, the most discriminative variable is chosen to partition the dataset into subsets, and partitioning is continually repetitive until the nodes are homogeneous. The output is a tree diagram where with the splitting rules determines the branches with series of terminal nodes containing the response frequency. The Gini criterion is used to denote the decrease in the node impurity function. The Gini index is one of the most commonly used tree-building criteria to measure node impurity for categorical target values. The Gini index measures purity of data, which equals 0 for a pure node. The Gini index can be obtained by where P_j is a relative frequency of class j in a node.

$$\text{Gini index} = 1 - \sum_{j=1}^{r} P_j^2$$

The data received was cleaned and imputations were carried out to fill the missing data. The data was divided into the eight datasets, each containing one dependent variable corresponding to the 8 research questions. The eight random forest models were used to determine the impact of factors on predicting the online purchasing behavior for all the eight product categories. R software is used to perform all the operations on data including imputations, analysis and Figure creation. ROCR[1] curve was seen to determine the cut-off probability for the prediction. Lift curve was seen to determine the benefit the model gives us when we are targeting customers. The random forest models could also predict the probability of a customer with the specific characteristics purchasing the product from the specific category online.

4. Analysis and Finding

4.1. Data Treatment and Imputations

Although most of the survey was filled correctly, there were some questions which either the respondents did not want to fill or they did not remember the required details, e.g. one of the question was "When was your first online purchase made". Around 7% of the respondents did not remember even the approximate year for their first online purchase. Some of the respondents did not want to reveal their annual income, thus some kind of imputation was required before the development of the prediction model. Imputations were carried out to fill the missing data with data points. "Multiple Imputations by chained equations" were used to fill in the missing data. There were multiple types of data. The following treatment was used for different level of measurement:

1) Continuous numerical variables (like age): Predictive mean mapping (PMM);

2) Binary variables (Like COD option): Logistic Regression Imputations;

[1]ROCR is a flexible tool for creating cutoff-parameterized 2D performance curves by freely combining two from over 25 performance measures. Curves from different cross-validation or bootstrapping runs can be averaged by different methods, and standard deviations, standard errors or box plots can be used to visualize the variability across the runs.

3) Variables with multiple options (Like State): Bayesian polytomous regression.

We ensure that the missing data is random in nature and the missing values are not forming a pattern with other variables. We saw that no discernable connection was present among the missing values. We plotted "First online purchase year" variable with the variable "age". We saw that for lower ages, the "First online purchase year" variable is missing. The peak comes at around an age of 25 - 27. However, no logical explanation could be made regarding this behavior. Also, since those missing values were less in number, we proceeded with the imputation using MICE (Multiple Imputations using Chained Equations). After imputations, the plausibility of the imputations was checked by going over the imputations and making sure that they were within the range of the other data in the variables. Table 2 presents the characteristics of the respondents.

As required in Random Forest Model, the data was converted either in numerical or binary format. The character variables like, "States" were converted into the dummy variables. For instance, if the respondent provides an input as "Maharashtra" for the variable State, the binary values 1 is assigned against "Maharastra" and for all other responses for State are assigned with 0. Random forest algorithm is used to split the data set into a Training dataset (70% of data) and Validation dataset (the remaining 30 % of data). The eight datasets were created with eight dependent variables to run the model. The model is built on training datasets and tested on validation datasets.

4.2. Model Development

4.2.1. Category 1: Books

1) *Random Forest Model* (*RFM*)

As discussed above the data is split into Training dataset and Validation dataset and the Random Forest model is built on the training dataset with independent variable as "Books". The model is built using maximization of the decrease in Gini Score. After every iterative step the preference of the variables was decided based on the Gini Score. Figure 1 shows the plot for Random Forest model in the category "Books" with the mean decrease in Gini factor. We can see that for the category "Books", Occupation is the most important factor for determining if a customer will purchase books online, followed by the geographical location (the State) and the Annual Income of the customer. The Random Forest model is built on 86 data points and Table 3 shows the confusion matrix for the training dataset.

Table 3 shows the predicted values generated by the Random Forest model and is compared with the actual values present in the training data set. The confusion matrix for training dataset depicts 24 actual positive values (online buying preference) that the model has predicted accurately. Also there are 9 actual negative values (off line buying preference) that model has predicted accurately. So

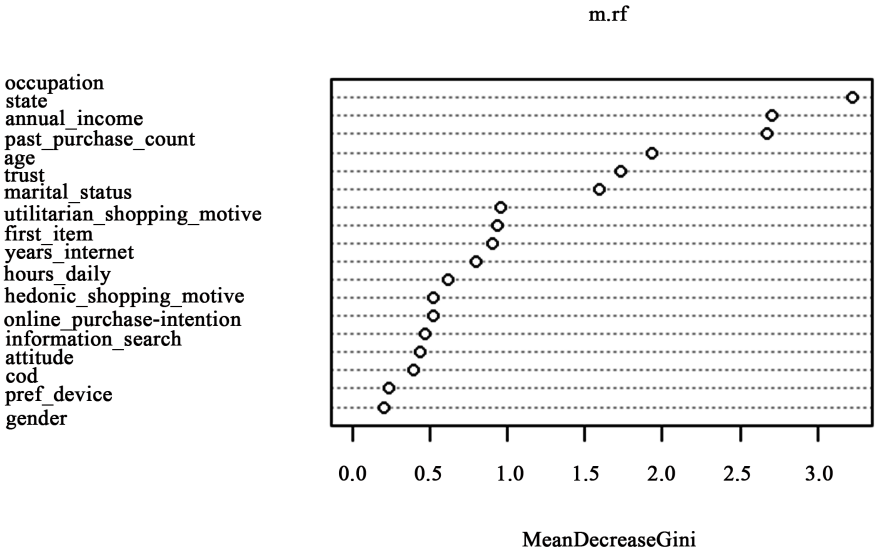

Figure 1. Books random forest plot.

Table 3. Confusion matrix for the training dataset versus validation dataset for RFM.

		Actual data	
		No	Yes
Model prediction	No	9	2
	Yes	11	64
		Training data set	

		Actual data	
		No	Yes
Model validation	No	5	3
	Yes	8	22
		Validation data set (Cut-off 0.5)	

out of 86 data points 73 are predicted correctly by the Random Forest model. The accuracy rate of the model is around 85%. However, if we talk about the error of the prediction model, we can see that 13 data point (11 Negative values that model predicted positive and 2 positive values that model predicted negative) are predicted inaccurately Thus, this model is fairly performing well if we wish to focus on people who are willing to go for the online retail channel but not good otherwise.

Now we need to determine the performance of the predicted model using the Receiver Operator Characteristic Curve (ROC Curve). **Figure 2** shows the ROC curve for the variable "Books".

The ROC curve has three axes. It plots the curve between True Positive Rate and False Positive Rate and the third axis represents the Area under Curve (AUC). True positive rate is the percent of positive values that the model is able

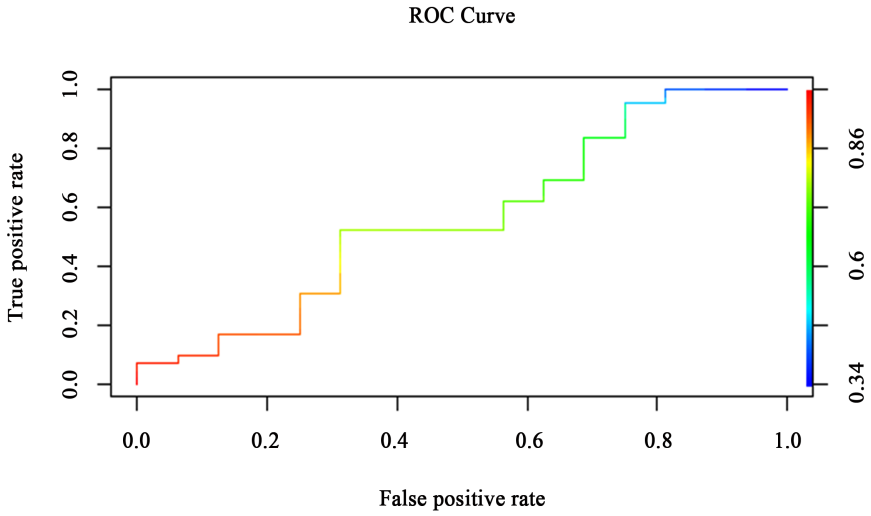

Figure 2. Books ROC curve plot.

to predict accurately of the total positive values actually present in the dataset (*i.e.* total number of "Yes" that the model has predicted accurately divided by the total number of actual "Yes" in the training dataset. Likewise, the False positive rate is the percent of positive values that the model predicts falsely of the total negative values (*i.e.* total number of "No" that the model has predicted falsely divided by the total number of actual "no" present in the dataset). An ideal model has maximum true positive rate and minimum false positive rate. The third axis depicts the cumulative probability which is used for deciding the cut-off probability and is represented by shade of colors. By default, the model considers 0.5 as the cut-off probability. So, according to the model prediction, if the probability of buying online by a customer is greater than 0.5, the customer is classified as "purchasing online", if it is less than 0.5, the customer is classified as "not purchasing online". With the help of ROC curve we decide on the changing the cut-off probability. The probability at which the true positive rate is maximum and true negative rate is minimum is to be taken as the cut-off probability.

In the "Books" category, we see that there is no clear point where we can keep a cut-off probability which will demarcate a high true positive rate and a low false positive rate. The ROC curve suggests a cut-off probability of 0.5 as fairly satisfying. The cut-off probability 0.5 will come near the center of the plot where the true positive rate is more than the false positive rate. When we run the algorithm on the validation dataset, we get the result provided in **Table 4**. This is very similar to our training dataset result. We see a good performance when it comes to predicting the people who are willing to buy from the online channel, but we the result is not favorable in predicting the people who are not willing to shift. **Table 4** provides the performance of the model.

4.2.2. Product Category 2: Movies, Music & Games

2) *Random Forest Model for Movies, Music & Games*

Table 4. Model performance–random forest for books.

False Positive Rate	61.54%
False Negative Rate	11.68%
Sensitivity (True Positive Rate)	88.32%
Specificity (True Negative Rate)	38.46%
Error	15.11%

Now in the same sequence, the Random forest model is built for the training dataset with independent variable category namely "Movies, Music & Games". The model was built by maximizing the decrease in Gini Score after every iterative step *i.e.* the preference of the variables was decided based on Gini Score. Based on the mean decreasing Gini score, **Figure 3** shows the plot for Random Forest model for the product category "Movies, Music & Games".

We see that for "Movies, Music & Games" product category, the location (State from where the respondent belong to) is the most important factor for determining if a customer is willing towards the online purchase of "Movies, Music & Games", which is followed by Annual income and the Year of e-commerce usage. **Table 5** shows the model prediction for the training dataset.

Table 5 compares the predicted values from the Random Forest model with the actual values from the training dataset. Here, we see that the model accurately predicted the respondents who are not willing to buy from online channel. However, the model has not performed well in capturing the online behavior of the respondents. The accuracy rate of the prediction model is around 65%. Thus, this model is moderately good if we only want to capture customers who are willing to go for the offline channel purchase. This result is exactly opposite to the one we observed in the product category "Books". To determine the performance of the prediction model we use ROC curve. The ROC Curve is shown below in **Figure 4**.

In this ROC curve, we see that from a false positive rate of 0.1 to 0.4, there is no change in true positive rate, thus, we have to avoid that area. We can see the faint blue/green line where the true positive rate is particularly high as compared to false positive rate. The exact cut-off probability is found by experimentation. At the default cut-off probability of 0.5, the true positive rate is specifically low, thus, we try increasing the cut-off probability to 0.38 instead of 0.5. The result is checked on the validation data. The value of 0.38 is decided after continuous experimentation with different values of the cut-off probability. When we run the algorithm on the validation dataset, after changing the cut-off probability to 0.62, we get the result shown in **Table 5**. Based on running the Random Forest algorithm on the validation dataset, the performance of the model is determined as shown in **Table 6**.

4.2.3. Product Category 3: Electronics Items

3) *Random Forest Model*

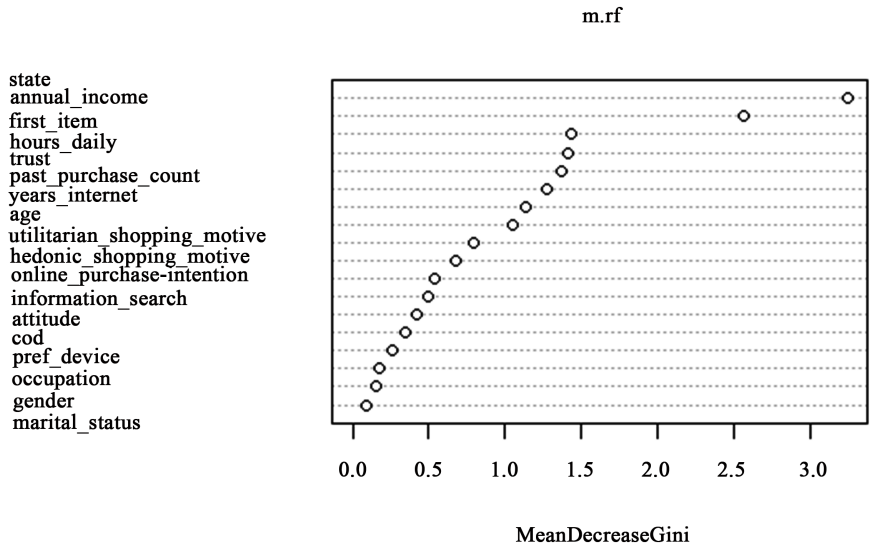

Figure 3. Movies, music & games random forest plot.

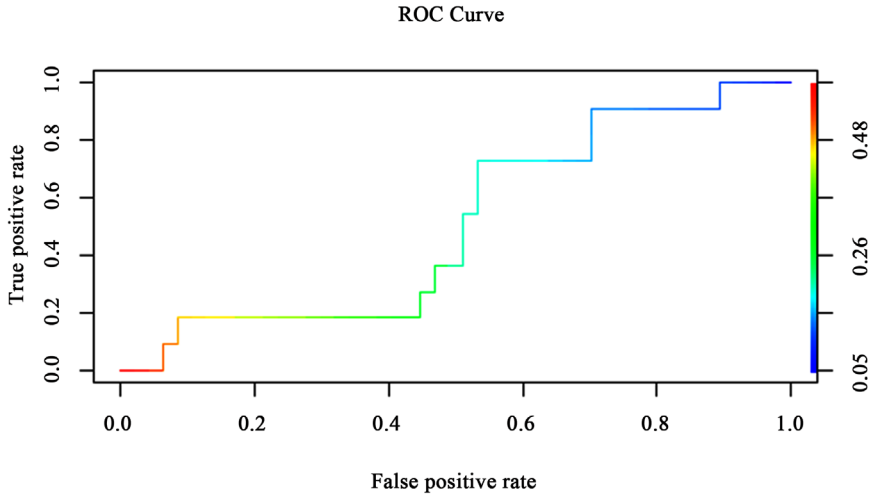

Figure 4. Movies, music & games roc curve plot.

Next, the Random Forest model is built for the training data set with independent variable as "Electronics". The model was developed by maximizing the decrease in Gini Score after every iterative step. Figure 5 is the plot for Random Forest model in electronics:

According to the mean decrease in Gini score of the prediction model, the most significant factors affecting purchase behavior are, past purchase count, Location (the State respondent belonged to) and Years of internet usage. Also, the annual income has significant impact on the purchase behavior. Following is the confusion matrix (Table 7) for the product category.

The training dataset based Random Forest model depicts that the respondents who are willing to buy from online channel. The accuracy rate of the prediction model is around 87%. Figure 6 shows the ROC Curve to determine the model performance.

Table 5. Confusion matrix for movie, music & games.

		Actual data	
Model prediction		No	Yes
	No	24	9
	Yes	3	1
		Training data set	

		Actual data	
Model validation		No	Yes
	No	54	22
	Yes	8	2
		Validated data set (Cut-off to 0.62)	

Table 6. Model performance-random forest for movies, music & games.

False positive rate	6.38%
False negative rate	90.91%
Sensitivity (True positive rate)	9.09%
Specificity (True negative rate)	93.62%
Error	34.88%

Table 7. Confusion matrix for electronics item.

		Actual data	
Model prediction		No	Yes
	No	4	9
	Yes	2	71
		Training data set	

		Actual data	
Model validation		No	Yes
	No	2	4
	Yes	1	31
		Validated data set	

In this ROC curve, we can observe that at cut-off probability of 0.5, the true positive rate and false positive rate are significantly high. Ideally, a model like this doesn't really form unless there is some specific issue with it. Here we see that for a cut-off probability of 0.5, the false positive rate is specifically high, thus, we try increasing the cut-off probability to 0.65 instead of 0.5. It is corresponding to the faint green line. The result is tested on the validation data. When

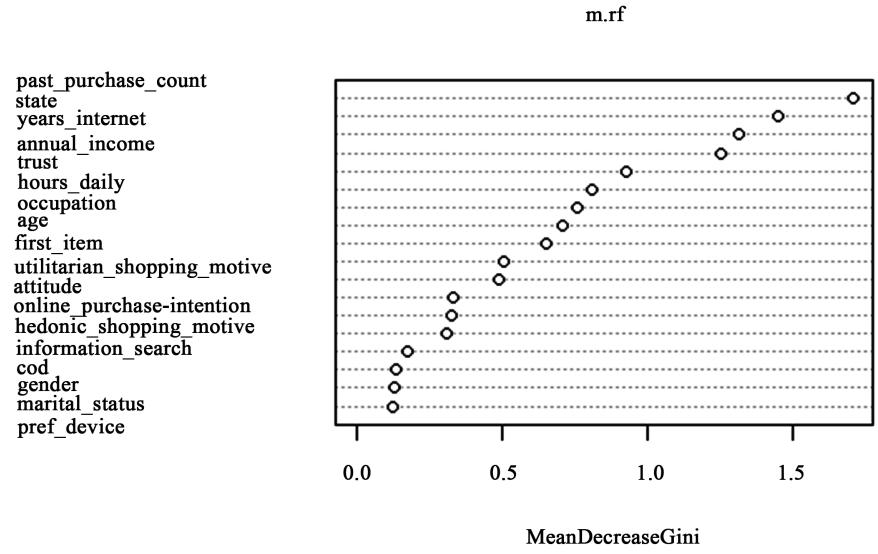

Figure 5. Electronics random forest plot.

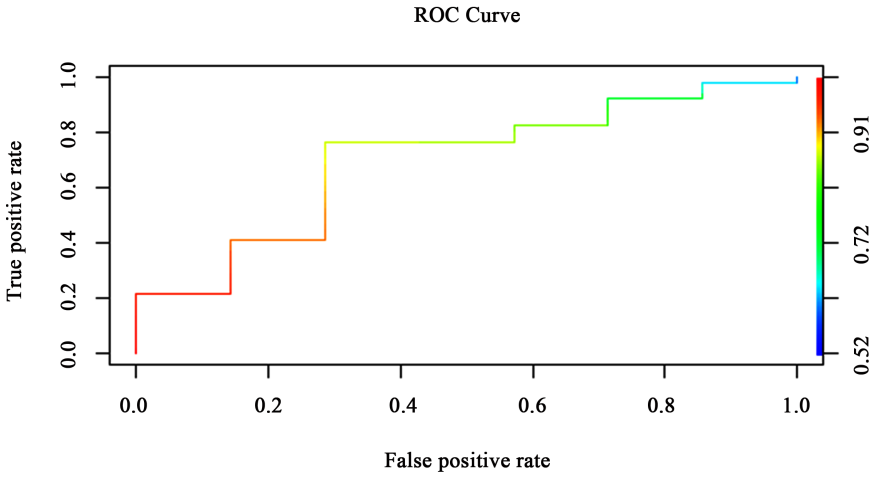

Figure 6. Electronics ROC curve plot.

we run the algorithm on the validation dataset, we get the following result shown in Table 7. This result is very similar to the result for our training data which further validates the model. After changing the cut-off probability to 0.68, no significant change was observed. Thus, we see excellent performance when it comes to predicting the people who are willing to buy from the online channel, but we do not favorable while predicting the people who are not willing to shift. Table 8 gives the performance of the model.

4.2.4. Product Category 4: Home Appliances

4) *Random Forest Model*

Random forest model was run for the dataset with independent variable, Home Appliances. The model was built by maximizing the decrease in Gini Score after every iterative step *i.e.* the preference of the variables was decided based on Gini Score. Figure 7 is the plot for Random Forest model in Home

Table 8. Model performance-random forest for electronics.

False positive rate	33.36%
False negative rate	9.43%
Sensitivity (True positive rate)	90.57%
Specificity (True negative rate)	66.67%
Error	1.16%

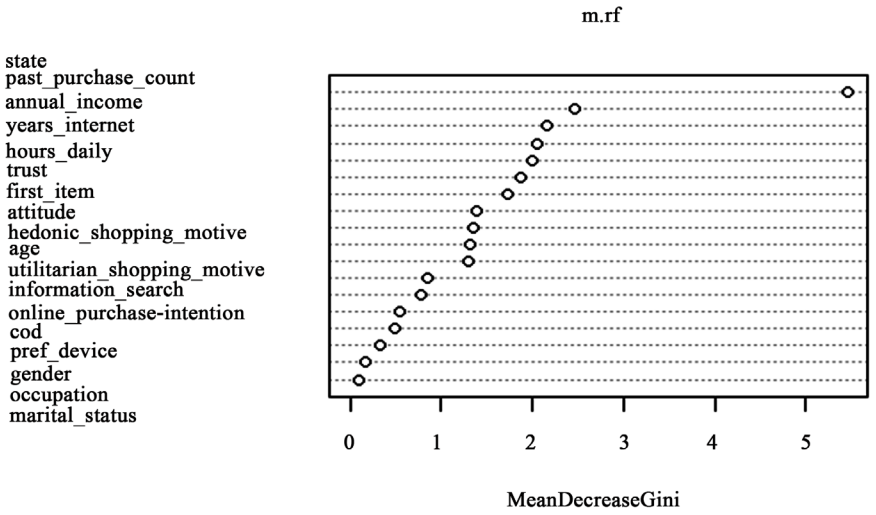

Figure 7. Home appliances random forest plot.

Appliances.

The plot suggest that the factor Location (the State respondent belonged to) has the significant impact on the buying behavior in this category. **Table 9** presents confusion matrix derived from the prediction model.

The prediction model suggests that we have predicted accurately around 78 % of respondents who are willing to buy from online channel. The accuracy rate of the prediction model is around 70%. The ROC Curve shows below in **Figure 8**.

Although ROC curve for home appliances is almost linear in nature. Here we see that at a cut-off probability of 0.5, the false positive rate is high. When we tried increasing the cut-off probability from 0.5 to 0.6, the light green line near the middle of the plot showed no significant difference in the true positive rate and the true negative rate. The result is further checked on the validation data. When we run the algorithm on the validation dataset, we get the following result: This result is very similar to the result for our training data which further validates the model. Thus, we can see excellent performance when it comes to predicting the people who are willing to buy from the online channel, but the model is not favorable while predicting the people who are not willing to shift. **Table 10** gives the performance of the model.

4.2.5. Product Category 5: Sports & Fitness

5) *Random Forest Model*:

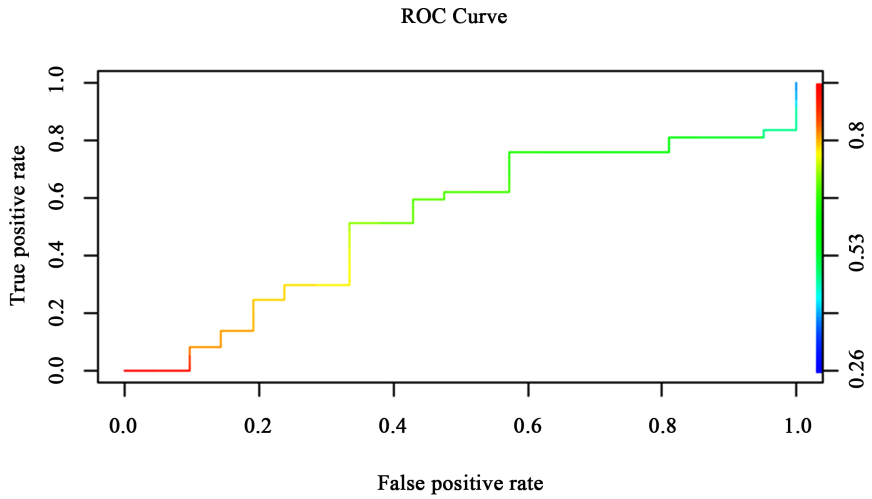

Figure 8. Home Appliances ROC Curve Plot.

Table 9. Confusion matrix for home appliances.

		Actual data	
Model prediction		No	Yes
	No	8	15
	Yes	10	53
	Validated data set		

		Actual data	
Model validation		No	Yes
	No	3	6
	Yes	7	22
	(Validated data set after cutoff at 0.6)		

Table 10. Model performance–random forest for home appliances.

False Positive Rate	76.19%
False Negative Rate	18.92%
Sensitivity (True Positive Rate)	81.08%
Specificity (True Negative Rate)	23.81%
Error	29.06%

Now for the product category 'Sports & Fitness", the Random Forest model was run for the training dataset. The model was built by maximizing the decrease in Gini Score after every iterative step *i.e.* the preference of the variables was decided based on Gini Score. **Figure 9** is the plot for Random Forest model in Sports category.

The top three factors affecting purchase behavior according to the model are

the location (State), Years of e-commerce usage and Annual income. Following is the confusion matrix, as shown in Table 11.

Here, we compare the predicted values with the actual values for the data. Here, we see that we have predicted good number of the people who are willing to buy from online channel properly, but we fail to capture those who are not willing to buy from online channel. The accuracy rate is around 76.74%. Figure 10 depicts the ROC curve for Sports.

If we look at the cut-off probability of 0.5, we go to the top right of the Figure, which is not really giving us any insights. Here we see that for a cut-off probability of 0.5, the false positive rate is specifically high, thus, we try increasing the cut-off probability to 0.7 instead of 0.5, which is near the center-right of the Figure, such that the sensitivity and false positive rate are optimized to the extent possible. The result will be checked on the validation data. When we run the algorithm on the validation dataset, we get the following result.

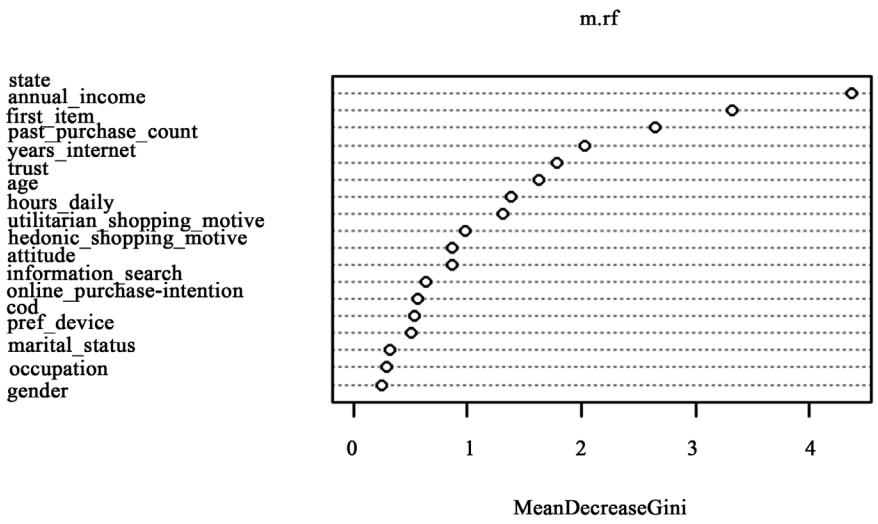

Figure 9. Sports equipment random forest plot.

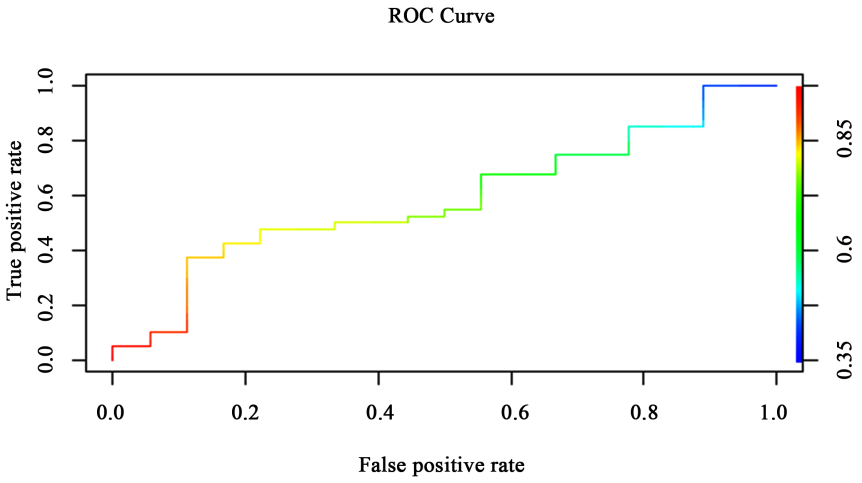

Figure 10. Sports equipment ROC curve plot.

This result is very similar to the result for our training data which further validates the model, after changing the cut-off probability to 0.6. Thus, the new result is not satisfactory, but at least better than the one we got with cut-off probability of 0.5. Table 12 gives the performance of the model:

4.2.6. Product Category 6: Handbags

6) *Random Forest Model for Handbags*

Now for the category Handbags, the model was built by maximizing the decrease in Gini Score after every iterative step *i.e.* the preference of the variables was decided based on Gini Score. Figure 11 is the plot for Random Forest model in Handbags:

The top three factors affecting purchase behavior according to the model are Location (State), Past purchase count and Annual income. Table 13 shows the confusion matrix.

Here we see that our positive prediction is very good, but the negative prediction is not extremely good. However, it is still better than targeting random members from the audience. Figure 12 presents the ROCR Curve.

Here we see that for a cut-off probability of 0.5, the false positive rate is specifically high, thus, we try increasing the cut-off probability to 0.38 instead of 0.5. The result was checked on the validation data. When we run the algorithm on the validation dataset, we get the result shown in Table 13. This result is very similar to the result for our training data which further validates the model. After changing the cut-off probability to 0.38, we got a better prediction model than our previous prediction. Thus, this model performed fairly good to predict

Table 11. Confusion matrix for sports and fitness equipment.

	No	Yes
No	12	13
Yes	7	54
	Training data set	

	No	Yes
No	7	6
Yes	3	22
	(Validated data set after cutoff at 0.6)	

Table 12. Model performance-random forest for sports & fitness.

False positive rate	29.97%
False negative rate	21.43%
Sensitivity (True positive rate)	78.57%
Specificity (True negative rate)	70.03%
Error	15.52%

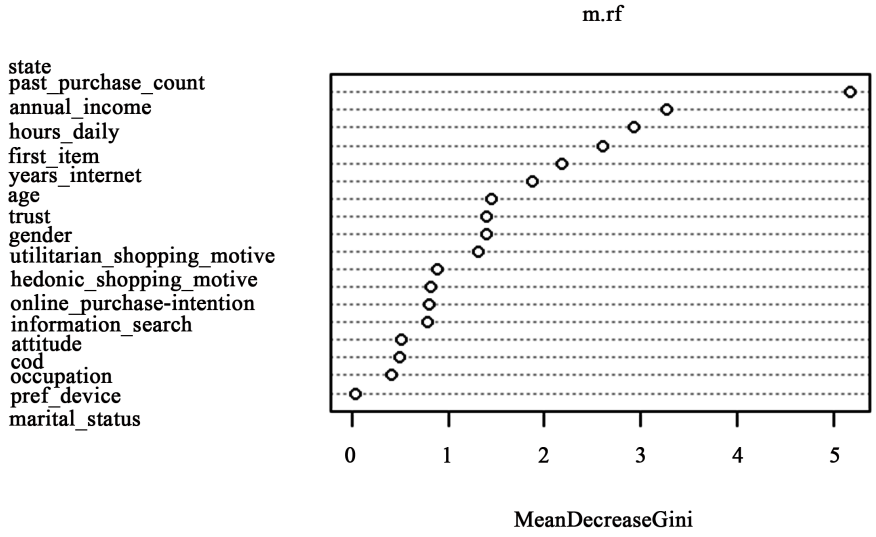

Figure 11. Handbag random forest plot.

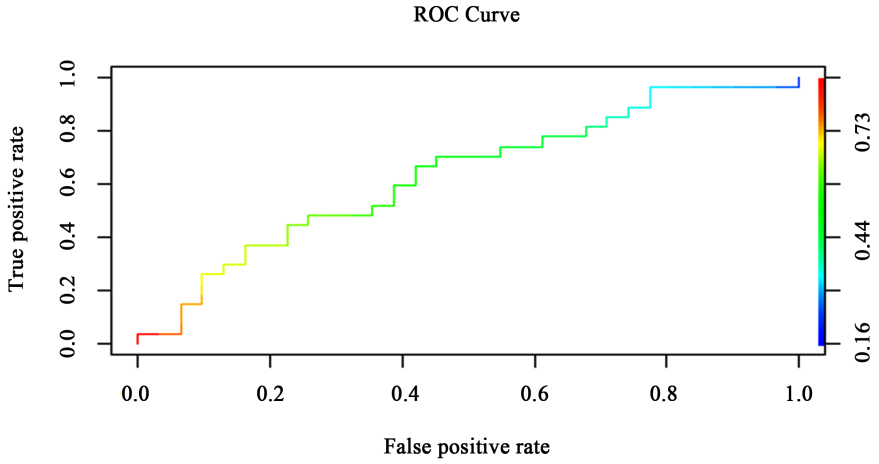

Figure 12. Handbag ROC Curve Plot.

Table 13. Confusion matrix for the training dataset versus validated dataset for RFM.

	No	Yes
No	23	22
Yes	11	30

Training data set

	No	Yes
No	11	9
Yes	4	14

(Validated data set after cutoff at 0.6)

people who are willing to purchase online, but for the ones who are not willing, it does not perform that good. Table 14 gives the performance of the model:

Table 14. Model performance-random forest for handbag & accessories.

False positive rate	26.67%
False negative rate	39.14%
Sensitivity (True positive rate)	60.86%
Specificity (True negative rate)	73.34%
Error	29.66%

4.2.7. Other Product Categories: Jewelry and Cars/Bikes

For the product category viz., Jewelry and Cars/bikes the result only suggested that India is not ready for these two categories to go online.

5. Managerial Implication

This research establishes the product category wise impact of shopping motives, location, prior online purchase experience and attitude on the Indian customer purchase intention. The study has implications to online retailers, off line retailers, brand and marketing managers, vendors and customers in India. Online and Offline retailers specifically will be able to develop effective strategies to attract customers based on their preferences.

Many researchers (citations) have studied the buying behavior of Indian consumers and developed models to predict online buying behavior of Indian population. However, it may be observed that most of the studies are generic in nature where the general online purchase behavior was considered for the study but not for specific categories. In this paper we attempt to develop models that can predict the buying behavior of Indian customer for different product categories. We used Random Forest Models (RFM) to develop the causal relationship between the identified factors and their varied implication on purchase behaviors of Indian customers. In this study, RFM are developed for eight different categories so that the prediction results could be more specific in nature and can be applied to the explicit category instead of using a generic model for all products. Also, it may be observed that some of the researches done in Indian context have very limited demographics variance (mostly done on students in a limited geography). We tried to develop robust models with professional variability that included students, households and professionals. Moreover, the geographical location which is an important factor is inadequately addressed in literature especially in Indian context. Through our model we have established how important the location is for determining online customer purchase behavior. The causal impact of different factors on customer buying behavior in different product categories varies and have diverse implication on retailers' strategy. Table 15 shows the top three significant factor in respective product categories. We can see that Location is a very important factor that has significant impact on product categories.

Table 16 depicts the overall performance of the model for predicting buying

behavior of customers for different product categories. The model performance is divided into three scales namely, Good, Moderate and Bad, as per the level of performance. The detailed understanding of the performance can be seen from Table 17.

Table 17 compares the models for all the six categories based on six metrics: True Positive Rate, True Negative Rate, False Positive Rate, False Negative Rate and Error Rate. As discussed in earlier sections, the confusion matrix shows the predicted values generated by the Random Forest model and is compared with the actual values present in the training data set. The online buying preferences the model predicted correctly represents True positive rate (TPR). However the offline buying preference that model has predicted accurately signifies the True negative rate (TNR). Both TPR and FPR contributed to the overall performance of the Random Forest model. TPR is stated in terms of Sensitivity and TNR is indicated as Specificity of the Model.

Sensitivity and specificity are the important characteristics of the Random

Table 15. Most significant factors to predict online purchase behavior.

Importance	Books	Movies	Electronics	Home appliances	Sports equipment	Handbags
1	Occupation	Location	# Items purchased	Location	Location	Location
2	Location	Income	Location	No. of items purchased	Income	# Items purchased
3	Income	Time using e-commerce	Time using internet	Income	Time using e-commerce	Income

Table 16. Model performance.

	Books	Movies	Electronics	Home appliances	Sports equipment	Handbags	Jewelry	Cars/Bikes
Predicting people purchasing online	Good	Bad	Good	Good	Good	Good	Market not ready	Market not ready
Predicting people purchasing offline	Good	Good	Bad	Bad	Moderate	Moderate		

Table 17. Detailed comparison of the model performance.

	Books	Movies	Electronics	Home appliances	Sports	Handbag
False positive rate	61.54%	6.38%	33.36%	76.19%	29.97%	26.67%
False negative rate	11.68%	90.91%	9.43%	18.92%	21.43%	39.14%
Sensitivity (true positive rate)	88.32%	9.09%	90.57%	81.08%	78.57%	60.86%
Specificity (true negative rate)	38.46%	93.62%	66.67%	23.81%	70.03%	73.34%
Error	15.11%	34.88%	1.16%	29.06%	15.52%	29.66%

Forest model developed for different product categories. Sensitivity of the model represents the percent of customers inclined towards online purchase that the model is able to predict accurately. Likewise, the Specificity is the percent of offline purchase behavior of customers that the model predicts correctly. As we can see from Table 16 the Sensitivity of the Random forest model for Books and Electronics categories is above 85%. Which can be considers as a good performance for online purchase intension of customer. Retailers can use this model to predict the buying behavior of customers based on the location. However, we can also see that for product categories like Movies, Sports equipment and Handbag categories the high value of Specificity signifies the model prediction for off-line purchase intensions. So for these retailers may like to focus more on customer service and satisfaction in physical store. The categories like Jewelry and Cars/Bikes will absolutely not be received by the respondents for online purchase in the coming future. If a firm intends to target an audience for their product, it is acceptable if they send their advertisements to more people than required, however, if they were to send it to less people and potentially lose some customers, they may incur a huge opportunity cost. This model tries to do the same, it predicts almost everyone who is willing to buy online, but for the ones who are not, the error rate is a more.

A retailer may have a question: what is the chance that a person with an online purchase behaviour prediction truly has the online purchase intension? The model can run through the given values of the factors and the retailer may calculate the probability of the customer buying behaviour. Thus, this research is beneficial for retailer to determine if one should go for online channel or offline channel or both. Further, the model helps in identifying which target segment the retailers should focused on. At macro level, the top management could also get insights on which geographies are the best for their online and offline channels.

Furthermore, the models can be created and stored in the database. To find if the target audience will be attracted with the product offering of a particular company, one can enter the characteristic of the target audience, and the model will predict the likelihood of that audience actively purchasing the product. For instance, Leather Handbags Limited, a hypothetical company is planning to start an e-commerce application in Maharashtra, India. Their target audience is a college going, unmarried student whose family has an annual income more than 10 Lac rupees (10 Lac = 1 Million; 1 Rs = 65 USD). The Random Forest model prediction suggests that in Maharashtra, 40% of the people prefer to buy handbags, 50% of students in Maharashtra prefer to purchase handbags and 100% of the people in the defined annual income in those 50% of the students prefer purchasing handbags. Thus, Maharashtra would be a good market to sell handbags for that particular handbag manufacturer. As an alternate to that analysis and data collection, one can enter the details in the model which will immediately predict that handbags can be sold to a particular customer along with the

likelihood of her purchasing the product through online application.

6. Conclusion and Future Research Suggestions

Today consumers are keen to maximize their shopping utility by comprehensively considering all possible channels. With this diverse customer shopping preferences, retailers face immense challenge in serving both online and off-line channels efficiently. This problem further intensifies with the diverse social and economic conditions in India. In this paper, we attempted to understand various factors influencing the online buying behavior of Indian customers in different product categories across the different States in India. Also, using Random Forest model for each product category, we tried to establish consumer behavior influence on multi-channel retailer to understand if the Indian online shopping market is ready for these identified product categories or the traditional channel is preferred over by the customer.

In two ways, the paper contributes to the theoretical domain. First it establishes the influence of socio-psychological and economic factors upon the buying behavior of consumer. Second, it provides a thorough interplay of the factors and their impact on online and offline buying behaviors in emerging economies like India. Indian situation may echo to the situation in most of the Asian developing nations and similar solutions may apply there also ([46]). Mostly, in all these countries, similar socio-economic practices prevail, and researchers may take due observance of this fact. From a managerial perspective, the paper contributes in two ways. First, the study provides product category wise Random Forest models representing the percent of customers inclined towards online and offline purchase. Second, retailers can use this model to predict the buying behavior of customers based on the location and other local conditions.

Though the findings offer some new insights to many stakeholders in India who are in business of online and offline retailing, the research has its own limitations. Only 124 observations were taken and we could not have the data for all the strata of the society. Also, the model will only work if the new data are within the limits of the distributions. We created eight datasets with eight dependent variables to run the model and the data were divided into training and validation data with 70% and 30% share respectively. A bigger and comprehensive training and validation dataset which includes respondents from all strata of life would have been resulted in the efficient and robust model. The students and academicians can use this research as a platform to further improve upon it. Also, varied product categories can be included to further improve upon the models. The factors like promotional pricing offers, quality, ease of return and brand orientation could also be considered in further studies.

References

[1] Bonson Ponte, E., Carvajal-Trujillo, E. and Escobar-Rodríguez, T. (2015) Influence of Trust and Perceived Value on the Intention to Purchase Travel Online: Integrating the Effects of Assurance on Trust Antecedents. *Tourism Management*, **47**, 286-

302. https://doi.org/10.1016/j.tourman.2014.10.009

[2] Brynjolfsson, E., Hu, Y.J. and Rahman, M.S. (2013) Competing in the Age of Omnichannel Retailing. *MIT Sloan Management Review*, **54**, 23-29.

[3] Beck, N. and Rygl, D. (2015) Categorization of Multiple Channel Retailing in Multi-, Cross-, and Omni-Channel Retailing for Retailers and Retailing. *Journal of Retailing and Consumer Services*, **27**, 170-178. https://doi.org/10.1016/j.jretconser.2015.08.001

[4] Liao, T.H. (2017) Online Shopping Post-Payment Dissonance: Dissonance Reduction Strategy using Online Consumer Social Experiences. *International Journal of Information Management*, **37**, 520-538. https://doi.org/10.1016/j.ijinfomgt.2017.03.006

[5] Shi, X. and Liao, Z. (2017) Online Consumer Review and Group-Buying Participation: The Mediating Effects of Consumer Beliefs. *Telematics and Informatics*, **34**, 605-617. https://doi.org/10.1016/j.tele.2016.12.001

[6] Gehrt, K.C., Onzo, N., Fujita, K. and Rajan, N.R. (2007) The Emergence of Internet Shopping in Japan: Identification of Shopping Orientation-Defined Segment. *Journal of Marketing Theory and Practice*, **15**, 167-177. https://doi.org/10.2753/MTP1069-6679150206

[7] Shim, S., Eastlick, M.A., Lotz, S.L. and Warrington, P. (2001) An Online Prepurchase Intentions Model: The Role of Intention to Search. *Journal of Retailing*, **77**, 397-416. https://doi.org/10.1016/S0022-4359(01)00051-3

[8] Internet World Stats (2017) http://www.internetworldstats.com/asia.htm

[9] Dhanabhakyam, M. (2017) Indian Retail Industry—Its Growth, Challenges and Opportunities. http://www.fibre2fashion.com/industry-article/printarticle/2203

[10] Thamizhvanan, A. and Xavier, M.J. (2013) Determinants of Customers' Online Purchase Intention: An Empirical Study in India. *Journal of Indian Business Research*, **5**, 17-32. https://doi.org/10.1108/17554191311303367

[11] Rigby, D. (2011) The Future of Shopping. *Harvard Business Review*, **89**, 65-76.

[12] Hong, I. and Cha, H. (2013) The Mediating Role of Consumer Trust in an Online Merchant in Predicting Purchase Intention. *International Journal of Information Management*, **33**, 927-939. https://doi.org/10.1016/j.ijinfomgt.2013.08.007

[13] Close, A.G. and Kukar-Kinney, M. (2010) Beyond Buying: Motivations behind Consumers' Online Shopping Cart Use. *Journal of Business Research*, **63**, 986-992. https://doi.org/10.1016/j.jbusres.2009.01.022

[14] Kamarulzaman, Y. (2007) Adoption of Travel e-Shopping in the UK. *International Journal of Retail & Distribution Management*, **35**, 703-719. https://doi.org/10.1108/09590550710773255

[15] Lobel, I., Patel, J., Vulcano, G. and Zhang, J. (2015) Optimizing Product Launches in the Presence of Strategic Consumers. *Management Science*, **62**, 1778-1799. https://doi.org/10.1287/mnsc.2015.2189

[16] Gallino, S. and Moreno, A. (2014) How to Win in an Omnichannel World. *MIT Sloan Management*, **56**, 49-53.

[17] Gao, F. and Su, X. (2016) Omnichannel Retail Operations with Buy-Online and Pick-Up-in-Store. *Management Science*, **63**, 2478-2492. https://doi.org/10.1287/mnsc.2016.2473

[18] Cachon, G.P. and Feldman, P. (2015) Price Commitments with Strategic Consumers: Why It Can Be Optimal to Discount More Frequently than Optimal. *Manufacturing Service Operations Management*, **17**, 399-410.

https://doi.org/10.1287/msom.2015.0527

[19] IBEF (2017) Indian Brand Equity Foundation. https://www.ibef.org/industry/retail-india.aspx

[20] Ushavaidehi, P. (2014) Factors Influencing Online Shopping Behavior of Students in Engineering Colleges at Rangareddy District. *Sumedha Journal of Management*, **3**, 50-62.

[21] Khare, A. and Rakesh, S. (2011) Antecedents of Online Shopping Behavior in India: An Examination. *Journal of Internet Commerce*, **10**, 227-244. https://doi.org/10.1080/15332861.2011.622691

[22] Geetha, V. and Rangarajan, K. (2015) A Conceptual Framework for Online Shopping Behavior. *Sona Global Management Review*, **10**, 9-23.

[23] Kiran, R., Sharma, A. and Mittal, K.C. (2008) Attitudes, Preferences and Profile of Online Buyers in India: Changing Trends. *South Asian Journal of Management*, **15**, 56-73.

[24] Ganguly, B., Dash, S.B. and Cyr, D. (2009) Website Characteristics, Trust and Purchase Intention in Online Stores: An Empirical Study in the Indian Context. *Journal of Information Science and Technology*, **6**, 22-44.

[25] Huang, W. and Swaminathan, J.M. (2009) Introduction of a Second Channel: Implications for Pricing and Profits. *European Journal of Operational Research*, **194**, 258-279. https://doi.org/10.1016/j.ejor.2007.11.041

[26] Cattani, K., Gilland, W., Heese, H.S. and Swaminathan, J. (2005) Boiling Frogs: Pricing Strategies for a Manufacturer Adding a Direct Channel that Competes with the Traditional Channel. *Production and Operations Management*, **15**, 40-56.

[27] Ancarani, F. and Shankar, V. (2004) Price Levels and Price Dispersion within and across Multiple Retailer Types: Further Evidence and Extension. *Journal of the Academy of Marketing Science*, **32**, 176-187. https://doi.org/10.1177/0092070303261464

[28] Panda, R. and Swar, B.N. (2013) Online Shopping : An Exploratory Study to Identify the Determinants of Shopper Buying Behaviour. *International Journal of Business Insights & Transformation*, **7**, 52-59.

[29] Ajzen, I. (1991) The Theory of Planned Behavior. *Organizational Behavior and Human Decision Processes*, **50**, 179-211. https://doi.org/10.1016/0749-5978(91)90020-T

[30] Baishya, K., Samalia, H.V. and Joshi, R. (2016) Factors Influencing E-District Adoption: An Empirical Assessment in Indian Context. *International Review of Management and Marketing*, **7**, 1-7.

[31] Venkatesh, V., Morris, M.G., Davis, G.B. and Davis, F.D. (2003) User Acceptance of Information Technology: Toward a Unified View. *MIS Quarterly*, **27**, 425-478. https://doi.org/10.2307/30036540

[32] Berry, L.L. (1969) The Components of Department Store Image: A Theoretical and Empirical Analysis. *Journal of Retailing*, **45**, 3-20.

[33] Dennis, C., Harris, L. and Sandhu, B. (2002) From Bricks to Clicks: Understanding the E-Consumer. *Qualitative Market Research*: *An International Journal*, **5**, 281-290. https://doi.org/10.1108/13522750210443236

[34] Ajzen, I. and Fishbein, M. (1980) Understanding Attitudes and Predicting Social Behavior. Prentice-Hall, Englewood Cliffs.

[35] Park, S. and Lee, D. (2017) An Empirical Study on Consumer Online Shopping Channel Choice Behavior in Omni-Channel Environment. *Telematics and Informatics*, **34**, 1398-1407. https://doi.org/10.1016/j.tele.2017.06.003

[36] Fortin, D.R., Dholakia, R.R. and Dholakia, N. (2002) Introduction to Special Issue: Emerging Issues in Electronic Marketing: Thinking Outside the Square. *Journal of Business Research*, **55**, 623-635. https://doi.org/10.1016/S0148-2963(00)00202-2

[37] Roth, A.V. and Jackson, W. (1995) Strategic Determinants of Service Quality and Performance: Evidence from the Banking Industry. *Management Science*, **41**, 1720-1733. https://doi.org/10.1287/mnsc.41.11.1720

[38] Roth, A.V. and Menor, L.J. (2003) Insights into Service Operations Management: A Research Agenda. *Production and Operations Management*, **12**, 145-164. https://doi.org/10.1111/j.1937-5956.2003.tb00498.x

[39] Joshi, R., Kakoty, S. and Dwivedi, R. (2015) Community Based Agri-Chain Network: Sustainable Alternate Pathway towards Development in India. *International Journal of Indian Culture and Business Management*, **13**, 415-449. https://doi.org/10.1504/IJICBM.2016.079812

[40] Breiman, L. (2001) Random Forests. *Machine Learning*, **45**, 5-32. https://doi.org/10.1023/A:1010933404324

[41] Strobl, C., Malley, J. and Tutz, G. (2009) An Introduction to Recursive Partitioning: Rationale, Application and Characteristics of Classification and Regression Trees, Bagging and Random Forests. Technical Reports, Department of Statistics.

[42] Alvarez, S., Diaz-Uriarte, R., Osorio, A. and Barroso, A. (2005) A Predictor Based on the Somatic Genomic Changes of the BRCA1/BRCA2 Breast Cancer Tumors Identifies the Non-BRCA1/BR Tumors with BRCA1 Promoter Hypermethylation. *Clinical Cancer Research*, **11**, 1146-1153.

[43] Hastie, T., Tibshirani, R. and Friedman, J. (2001) The Elements of Statistical Learning. Springer-Verlag, New York. https://doi.org/10.1007/978-0-387-21606-5

[44] Muchlinski, D., Siroky, D., He, J. and Kocher, M. (2015) Comparing Random Forest with Logistic Regression for Predicting Class-Imbalanced Civil War Onset Data. *Political Analysis*, **24**, 87-103. https://doi.org/10.1093/pan/mpv024

[45] Shellman, S.M. (2004) Time Series Intervals and Statistical Inference: The Effects of Temporal Aggregation on Event Data Analysis. *Political Analysis*, **12**, 97-104. https://doi.org/10.1093/pan/mpg017

[46] Joshi, R., Banwet, D.K., Shankar, R. and Gandhi, J. (2012) Performance Improvement of Cold Chain in an Emerging Economy, Production Planning and Control. *The Management of Operations*, **23**, 817-836. https://doi.org/10.1080/09537287.2011.642187