Product Range Prediction, Comparison and Analysis using Random Forest Algorithm

Sheena H. Unnithan[1], Soumya Ganesh[2], Varna R. Sunil[3], Lekshmi Krishna Deepthi[4], Gisha G.S[5]

[1,2,3,4] Computer Science and Engineering (Pursuing), Dept. of Computer Science and Engineering, LBS Institute of Technology for Women, Thiruvananthapuram, Kerala, India.
[5] Assistant Professor, Dept. of Computer Science and Engineering, LBS Institute of Technology for Women, Thiruvananthapuram, Kerala, India.

---***---

Abstract - -*In the rapidly growing age of technology, Market research plays an important role in designing and launching a new product, improving existing services or when a company is looking forward to leap ahead of its competitors. It identifies the major competitors and potential customers in the market. It proves to be a crucial step in developing marketing strategies needed for making better decisions. Therefore, we intend to aid the market research process by developing a web-based application which directly get reviews from the customers and determine the current market demand using random forest algorithm, a supervised machine learning technique to produce the expected results. The companies conducting market research can view results as pictorial representations and simple reports from which conclusions like selling potential of the product, comparisons with rival company products, etc. could be drawn. Effective market analysis can help in getting valuable insights into customer requirements, competitor analysis and ongoing market trends.*

***Key Words*: Random Forest, Data Mining, Machine Learning, Web Application**

1. INTRODUCTION

The market analysis is very important to anybody looking to start or continue their work at a business as the main purpose of the market survey is to obtain critical information about their consumers so that existing customers can be retained and new ones can be got onboard. Furthermore, it will help to find the weaknesses in its competitor's approach which we can utilize to gain more customers. It will provide all the information that need to make a better business decision.

There are number of survey sites available now-a-days which carry out surveys for various companies, businesses and individuals for different purposes. The major drawback of such survey is that they may not be able to meet the actual customer expectation and their purpose. So the objective of this paper is to propose a web-based application which directly extracts reviews from customers and the genuine ones are identified to determine the current market demand using Random Forest algorithm and also the comparison with products manufactured by different companies of the similar category or type.

This paper consists of five sections. Second section explains the literature surveys Third section explains the methodology in the proposed system. Fourth section discusses about the implementation and results obtained. Finally in the fifth section conclusions are drawn and discussed.

2. BACKGROUND AND RELATED WORK

The market trend and the consumer behaviour go hand in hand. In market analysis, analysing the consumer behaviour can help in understanding the categories of the customers and their willingness to buy. Such information can work wonders while coming up with innovative marketing strategies. This relationship is discussed by Harsh Valecha, Aparna Varma, Ishita Khare, Aakash Sachdeva and Mukta Goyal (2018)[1] based on the online and offline surveys conducted by them. Machine learning techniques were used in the prediction of the consumer behaviour which gave good accuracy. Loraine Charlet M.C and Ashok Kumar D (2012)[2] presented the discussion about Market Basket analysis done to determine the placement of goods, designing sales and promotions for different categories of customers to improve the customer satisfaction and hence the profit of the supermarket using K-Apriori algorithm and generating association rules. Be it online or offline surveys, missing data is one of the major concerns. Marvin L. Brown and John F. Kros (2003)[3] addressed the impact of missing data on the data mining operation of the Knowledge Discovery process. Imane Ezzine and Laila Benhlima (2018)[4] have discussed about the various data handling methods for big data and fixing the data in order to improve the data quality and hence improving the resulting analysis. Machine learning algorithms are applied on the processed dataset and decisions are made. Different techniques have different performances. Rana Alaa El-Deen Ahmeda, M. Eleman Shehaba , Shereen Morsya, Nermeen Mekawiea (2015)[5] carried out a performance study of classification algorithms for consumer online shopping attitudes and behaviour using data mining. They comparatively tested eleven data mining techniques to find the best classier fit for online consumer shopping attitudes and behaviour. The results showed that decision table classifier and filtered classifier gives the highest accuracy. Feature Selection identifies removes attributes that are of minor importance to the classifiers or possibly detrimental, thus avoiding cognitive overload of decision makers. Stefan Lessmann and Stefan Voß (2009)[6] discussed about Feature Selection and several approaches have been proposed in their literature. Mohammed Zakariah (2014)[7] discussed the various applications where Random Forest algorithm is used to classify the data set, step-wise implementation of the algorithm and the results along with the features of Random Forest algorithm. For effective learning and classification of Random Forest, no. of trees needs to be reduced. Pruning is the technique used to achieve this. Vrushali Y Kulkarni and Dr. Pradeep K. Sinha (2012)[8] discussed about pruning and presented a systematic survey of pruning efforts of Random Forest classifier.

3. METHODOLOGY

We have adopted the following methodology to perform market analysis.

Step 1: **Data Collection:** Market analysis can be performed to analyze market trends of various products. Among the multiple different products available, Mobile phones are one of the widely used gadget that have become an integral part of human life, hence chosen for conducting market analysis. Dataset is collection of specifications of various mobile phones manufactured by different companies. It includes features like mobile name, model number, processor, memory, display, battery, front camera, back camera, color, size, weight, sim type, resolution, clock speed, flash, Bluetooth support, GPS etc.

Step 2: **Collecting Reviews from Customer***:* A market survey has been conducted to collect the reviews of the customer. The survey form included information like their personal details, Goods and Service Tax (GST) number, bill number, specifications of the presently used mobile phone, individual feature preferences and their expected features. The billing details like GST number and Bill number are checked for verifying that an actual purchase has taken place and to ensure that the customer has used the product or is familiar with the product before submitting the review. This is in focus to get reviews from customers who know the real product than blind reviewers. A dummy GST database is used for the verification process as the access to actual database is restricted.

Step 3: **Data Pre-processing***:* Dataset after data collection may contain missing values, erroneous values, irregularities in values etc. which makes it imperfect to be given as input for machine learning algorithm. Thus, pre-processing methods such as data cleaning, data imputing, representation transformation, data balancing and data partitioning are applied. Different machine learning algorithms have different approaches for handling missing values.

Step 4: **Feature Extraction:** The companies conducting market research can analyze the selling potential of their existing product and identify their competitors in the market for which their product is compared with various mobile phones manufactured by different companies. The product sales value is calculated based on the reviews provided by the customer. Choosing relevant features is an important step.

Step 5: **Application of Random Forest Algorithm:** The dataset is split into training and testing set to build the model. Random forest algorithm extracts different features during each iteration to build multiple decision trees in order to improve the predictive accuracy. The outputs of decision trees undergo majority voting process to focus onto producing a single outcome.

Step 6: **Outcome:** The customer can analyze their existing product providing the product specifications. The class of that particular product, its product sales value, the results of its comparison with top level products, is graphically represented and a small report of the features that are most preferred in high level products which makes them the best sellers among the customers, are listed. This helps in improving the selling potential of products and growth of the company adhering to the customer requirements.

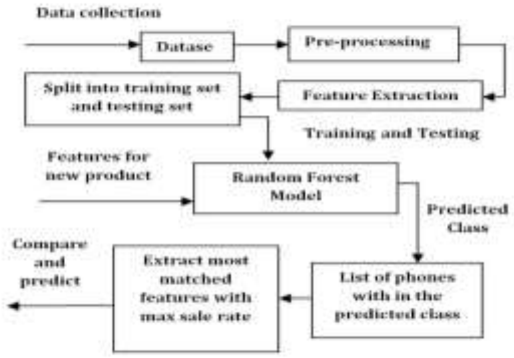


Fig -1: Data flow of the model

4. Implementation and Results

The detailed methodology adopted focussing on the computational tools and algorithm to perform the analysis is as follows:

1. **Dataset Collection:** A product specific survey was designed for the collection of customer reviews. The various hardware and software features of similar product i.e., mobile phones in the market, were collected from sources like official websites which constituted one of the datasets apart from the customer reviews.

2. **Pre-processing datasets:** The dataset is then processed to remove missing values, erroneous values, irregularities in values etc. Methods like Naïve Bayes deal with missing values seamlessly as it is linear and features are treated independently. Others, particularly non-linear methods such

as random forest of decision trees, may not allow for missing values using the mean of other data in the same feature. The dataset consists of categorical, numerical and binary data which are converted to numeric format using one-hot encoding technique.

3. **Feature Selection:** It is the process of selecting relevant features automatically or manually which contribute more to the prediction variable or the output we are interested in. Having irrelevant features decreases the accuracy of the model and makes it learn from irrelevant features. Feature selection reduces over-fitting, improves accuracy and reduces training time. We have selected "Processor", "Display", "Ram", "Internal Storage", Front Cam", "Back Cam", "Battery" as the features for the analysis.

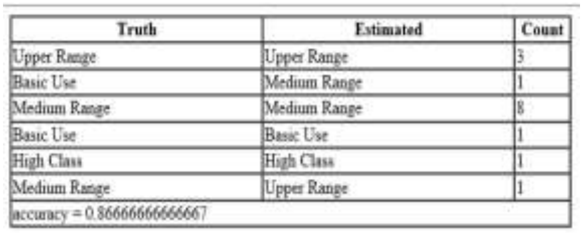
4. **Analysis:** We have implemented the following analysis methods:

a) Predicting the range of the product: A classifier model is built and trained to predict the range of the product as "Upper Range", "Middle Range", "Basic Use", or "High Class" based on its specifications. After Feature Selection, the dataset is divided into training and testing sets for training and testing the model respectively. The values of the dataset are adjusted and the process is repeated till the required accuracy is attained. Once this is achieved the dataset is taken as a whole and the final model is trained to make predictions for new data. Random forest algorithm is the classification algorithm used to train the model and the steps are as follows:

Algorithm steps

Classification:
Step 1: Randomly select "K" features from total "m" features where k << m.
Step 2: Among the "K" features, calculate the node "d" using the best split point.
Step 3: Split the node into daughter nodes using the best split.
Step 4: Repeat the 1 to 3 steps until "l" number of nodes has been reached.
5. Build forest by repeating steps 1 to 4 for "n" number times to create "n" number of trees.
Prediction:
Step 1: Takes the test features and use the rules of each randomly created decision tree to predict the outcome and stores the predicted outcome (target).
Step 2: Calculate the votes for each predicted target.
Step 3: Consider the high voted predicted target as the final prediction from the random forest algorithm.

The confusion matrix obtained during the testing is as follows:



Truth	Estimated	Count
Upper Range	Upper Range	3
Basic Use	Medium Range	1
Medium Range	Medium Range	8
Basic Use	Basic Use	1
High Class	High Class	1
Medium Range	Upper Range	1
accuracy = 0.8666666666666667		

Fig -2: Confusion matrix obtained

Based on the confusion matrix, recall and precision values of various categories of mobile phones can be calculated to be as follows:

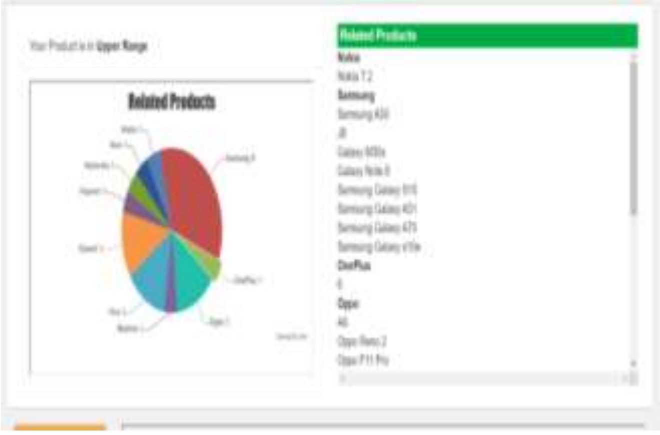
Table -1: Calculated values of Precision and Recall

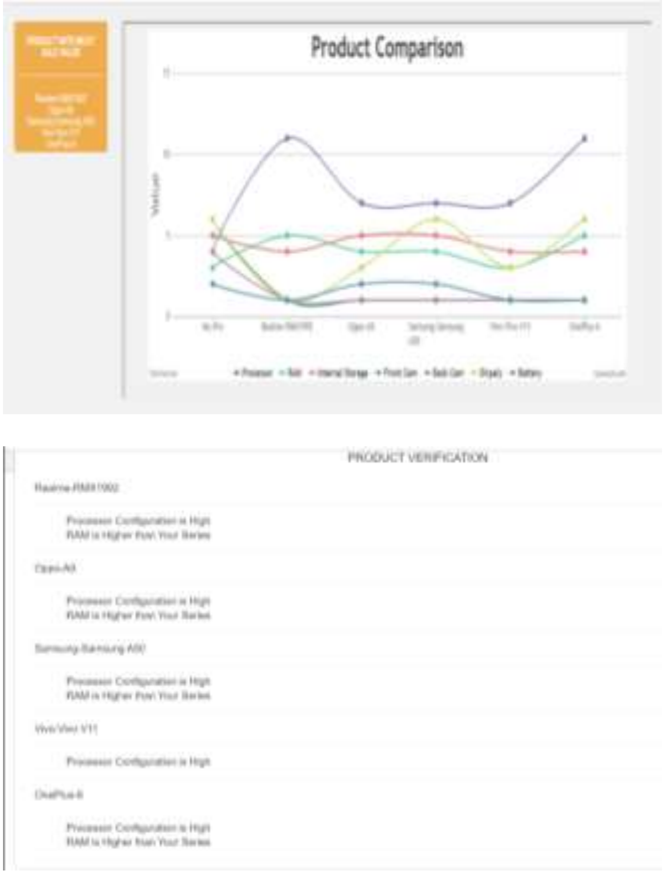
Product Category	Precision	Recall
High Class	1	1
Upper Range	0.75	1
Medium Range	0.89	0.89
Basic Use	1	0.5

The model accounts for an accuracy of 86.67%.

b) Product Comparison: The top 5 products with most sales value was found based the sales count. Sales count is the no. of product sold in normal operation of the company in a specified period of time. We considered the period between the launch date of the product and the local date when the analysis was carried out. And the no. of users who gave the corresponding product review was taken as the equivalent no. of product sold for calculation purposes.

c) Comparison of the product being analyzed with the top 5 products with most sales value: The features of the product being analyzed is compared with the corresponding feature of each product in the top 5 product with most sales value and the comparison is shown as a multi-chart and a simple textual report. The Multi-chart is displayed using the CanvasJS API. Also we have displayed the products coming in the Upper Range in our dataset as a pie chart using the CanvasJS API for a quick reference.





PRODUCT VERIFICATION

Realme-RMX1902

Processor Configuration is High
RAM is Higher than Your Series

Oppo-A9

Processor Configuration is High
RAM is Higher than Your Series

Samsung-Samsung A50

Processor Configuration is High
RAM is Higher than Your Series

Vivo-Vivo V11

Processor Configuration is High

OnePlus-6

Processor Configuration is High
RAM is Higher than Your Series

Fig -3: Output of Analysis obtained

5. Conclusion and Discussion

Market research opens up the door to information from consumers and market which would help in product design, developments and improvement services or when a company is looking forward to leap ahead of their competitors and hence the key component behind their success. Market surveys are easy and affordable way of gathering information from the target market. The reliability of the surveys could be improved with innovative measures as we have discussed. Responses of customers who have genuinely used the product increase the effectiveness. In this paper we have presented our work on developing a web based application which collects such reviews as one of the datasets along with the details about similar products available in the market as the second and performs product comparison, analysis and range prediction using machine learning technique. The outcome is given in graphical representation and simple text formats. Currently we have included “Mobile phones” as the category of product for analysis and as future work we propose to include more product categories and provide a more improved and detailed report which would include suggestions like common feature expectations of the customers, elaborate customer category details, etc.

REFERENCES

[1] Harsh Valecha, Aparna Varma, Ishita Khare, Aakash Sachdeva and Mukta Goyal, “Prediction of Consumer Behavior using Random Forest Algorithm”, 5th IEEE Uttar Pradesh Section International Conference on Electrical, Electronics and Computer Engineering (UPCON), 2018

[2] Loraine Charlet M.C and Ashok Kumar D, “Market Basket Analysis for a Supermarket based on Frequent Itemset Mining”, IJCSI International Journal of Computer Science Issues, Vol. 9, Issue 5, No 3, September 2012

[3] Marvin L. Brown and John F. Kros, “Data Mining and impact of missing data”, Industrial Management & Data Systems 103/8 [2003] 611-621

[4] ImaneEzzine and Laila Benhlima, “A study of handling missing data methods for big data”, Institute of Electrical and Electronics Engineers (IEEE), 2018

[5] Rana Alaa El-DeenAhmeda, M. ElemanShehaba , Shereen Morsya, NermeenMekawiea, “Performance study of classification algorithms for consumer online shopping attitudes and behaviour using data mining”, Fifth International Conference on Communication Systems and Network Technologies, 2015

[6] Stefan Lessmann and Stefan Voß, “Feature Selection in Marketing Applications”, R. Huang et al. (Eds.): ADMA 2009, LNAI 5678, pp. 200–208, 2009

[7] Mohammed Zakariah, “Classification of large datasets using Random Forest Algorithm in various applications: Survey”, International Journal of Engineering and Innovative Technology (IJEIT) Volume 4, Issue 3, September, 2014

[8] Vrushali Y Kulkarni and Dr. Pradeep K. Sinha, “Pruning of Random Forest Classifiers: A Survey and Future Directions”, International Conference on Data Science & Engineering (ICDSE),2012

[9] Kamran Kowsari, Kiana Jafari Meimandi, MojtabaHeidarysafa, Sanjana Mendu, Laura Barnes, Donald Brown, “Text Classification Algorithms: A Survey”, Information Journal, 2019

[10] Rajashree S. Jadhav, Prof. Deipali V. Gore, “A New Approach for Identifying Manipulated Online Reviews using Decision Tree”, (IJCSIT) International Journal of Computer Science and Information Technologies, Vol. 5 (2), 1447-1450, 2014