International Research Journal of Engineering and Technology (IRJET) e-ISSN: 2395-0056
Volume: 07 Issue: 05 | May 2020 www.irjet.net p-ISSN: 2395-0072

PREDICTION OF CONSUMER PURCHASE TARGET USING SOCIAL MEDIA DATASET

Gunavathy R[1], Valarmathy M[2], Mrs Mathangi[3]

[1,2] UG Student, Department of Information Technology, Meenakshi Sundararajan Engineering College, Chennai, India
[3] Assistant Professor, Department of Information Technology,Meenakshi Sundararajan Engineering College,Chennai,India

---

Abstract - *Digital marketing is considered the preferred method comparing to traditional marketing. It is useful to both practitioners and academics of social media marketing and purchase intention. The research provides some initial insights into consumer perspectives of social media ads and online purchase behavior. Business, academician, researchers all are share their advertisements, information on internet so that they can be connected with people fast and easily to survey on searchable product websites by web scrap. Web scraping is an automated method used to extract large amounts of data from websites and the data on the websites are unstructured. To prevent this problem, web scraping helps collect these unstructured data and store it in a structured form. Hence, customer price and rating of product evaluation and prediction has become an important research area. The aim is to investigate given dataset using machine learning based techniques for product rating forecasting by prediction results in best accuracy. The analysis of dataset by Support vector classifier (SVM) to capture several information's like, variable identification, uni-variate analysis, bi-variate and multi-variate analysis, missing value treatments and analyze the data validation, data cleaning/preparing and data visualization will be done on the entire given dataset. Our analysis provides a comprehensive guide to sensitivity analysis of model parameters with regard to performance in prediction of product ratings with price details by finding accuracy calculation. So, our proposed work from the given e-commerce dataset with application of web user interface.*

1. INTRODUCTION

It is an website that helps the customer to visualize and analysis the product. The reviews of the products are analyzed in this website by using php and python. It save the people time by seeing alot of comments. So the customer can easily visualize the review within an second. If the user can give their specification our website will display the product with their specification and also visualize the comments in the graphics format. Now a days most of the website i.e amazon , flipkart will only display their products details and their rating and comments. Along with some of those specification in amazon flipkart we add some features in our Website, so that customer can easily view their product comments.

1.1. RELATED WORKS

To study the neural influence of conflict between product appearance and performance on consumer decision when consumers' cognitive resources are limited. Comparing with Han et al, consumers' affirm ratio and reaction time are higher in our study, which means that when consumers' cognitive resources are limited, they need much more time to decide, and are easier to make a "buy" decision[1]. Coffeehouse chain is booming in Vietnam, the competition between coffee beverage suppliers in this section is also stiff currently. However, the Vietnamese market is evaluated as a great potential market, thus, in order to meet coffee drinker demand as well as expand the market share, suppliers of the coffeehouse chain should be taken more consideration on promotions activities and the taste of coffee beverages[2]. With the advent of the digital age, consumers nowadays are gradually becoming more disposed towards making online purchases for fashion apparel and related products[3].

2. EXISTING SYSTEM

It review makes important contributions to customer integration theory. It eludes to differences in user type definitions, RNP operationalization, and customer integration perspectives explaining inconsistencies in prior empirical findings. It reconciles key findings across empirical studies to derive at propositions on customer integration success. It indicates a gap in our knowledge on the role of customer integration in the acceleration phase of RNP projects and defines research avenues .The impact of customer integration on radical new product (RNP) innovation has been extensively investigated. To presents inconsistent empirical findings that must be converged. It literature review addresses these inconsistencies by taking a consolidated view of customer integration's effects on the development of RNPs. it provide the primary reasons for inconsistent findings by scrutinizing the operationalization of customer types (i.e., current

customers, potential customers, ordinary users, or users with domain-specific skills) and RNPs (i.e., technological innovativeness, or both technological and market innovativeness), as well as the different perspectives on customer. To present a synthesized view on factors in the sphere of the innovating company and the customer that influence customer integration's success along the radical innovation development process (i.e., discovery, incubation, and acceleration). It present avenues for future research and discuss managerial implications of our synthesized view

Drawback:

- It have not done customer integration for radical innovations
- It only direct integrated between customer and shop.
- To encourage other researchers to formulate interesting future research questions but also provide useful and important insights for innovation managers.
- It is not inspire and fuel contributed to our understanding of the conditions that nurture successful customer integration for radical innovations.

3. PROPOSED SYSTEM

Data Wrangling

In this section of the report will load the data, check for cleanliness ,and then trim and clean given dataset for analysis. Make sure that the document steps carefully and justify for cleaning decisions.

Preprocessing

The data which was collected might contain missing values that may lead to inconsistency. To gain better results data need to be preprocessed so as to improve the efficiency of the algorithm. The outliers have to be removed and also variable conversion need to be done.

Building the classification model

The predicting the consumer intension by supervised machine learning like decision tree algorithm prediction model is effective because of the following reasons: It provides better results in classification problem.

- It is strong in preprocessing outliers, irrelevant variables, and a mix of continuous, categorical and discrete variables.
- It produces out of bag estimate error which has proven to be unbiased in many tests and it is relatively easy to tune with.

Construction of a Predictive Model

Machine learning needs data gathering have lot of past data's. Data gathering have sufficient historical data and raw data. Before data pre-processing, raw data can't be used directly. It's used to preprocess then, what kind of algorithm with model. Training and testing this working model and predicting correctly with minimum errors. Tuned model involved by tuned time to time with improving the accuracy.

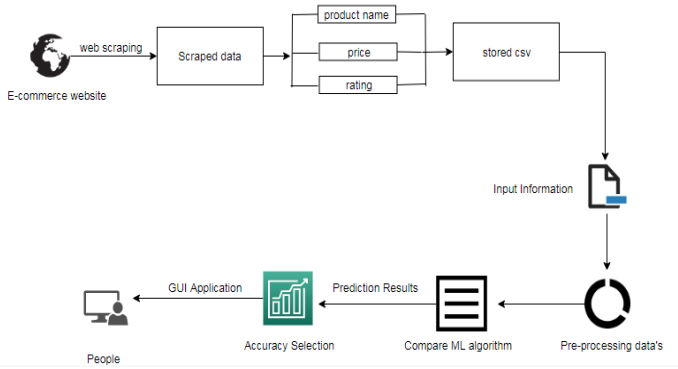


Fig -1: Design of Architecture

4. IMPLEMENTATION

MODULES:

- Scraping data for consumer intension (Module-01)
- Data validation and Preprocessing technique (Module-02)
- To train a model by given attribute with visualization (Module-03)
- Performance measurements of Support vector classifier (Module-04)
- Web based application of customer intension by php (Module-05)

MODULE SEPARATION:

PHP Part:

- ✓ Module-01
- ✓ Module-05

Machine Learning Part: (Python)

- ✓ Module-02
- ✓ Module-03
- ✓ Module-04

MODULE-01:

We have to scrape Flipkart website using dataminer extension to extract the Price, Name, Rating of Laptops and etc. So, we inspect the page to see, under which tag the data we want to scrape is nested. To inspect the page extract the Price, Name, and Rating which is nested in the ID tag respectively. After extracting the data, you might want to store it in a format. It will store the extracted data in a CSV (Comma Separated Value) format.

MODULE-02:

Data Validation/ Preparing Process:

Importing the library packages with loading given dataset. To analyzing the variable identification by data shape, data type and evaluating the missing values, duplicate values. A validation dataset is a sample of data held back from training your model that is used to give an estimate of model skill while tuning model's and procedures that you can use to make the best use of validation and test datasets when evaluating your models. Data cleaning / preparing by rename the given dataset and drop the column etc. to analyze the uni-variate, bi-variate and multi-variate process. The steps and techniques for data cleaning will vary from dataset to dataset. The primary goal of data cleaning is to detect and remove errors and anomalies to increase the value of data in analytics and decision making.

Data Pre-processing:

Pre-processing refers to the transformations applied to our data before feeding it to the algorithm. Data Preprocessing is a technique that is used to convert the raw data into a clean data set. In other words, whenever the data is gathered from different sources it is collected in raw format which is not feasible for the analysis. To achieving better results from the applied model in Machine Learning method of the data has to be in a proper manner. Some specified Machine Learning model needs information in a specified format; for example, Random Forest algorithm does not support null values. Therefore, to execute random forest algorithm null values have to be managed from the original raw data set. And another aspect is that data set should be formatted in such a way that more than one Machine Learning and Deep Learning algorithms are executed in given dataset.

MODULE-03:

To train a model by given attribute with visualization:

Data visualization is an important skill in applied statistics and machine learning. Statistics does indeed focus on quantitative descriptions and estimations of data. Data visualization provides an important suite of tools for gaining a qualitative understanding. This can be helpful when exploring and getting to know a dataset and can help with identifying patterns, corrupt data, outliers, and much more. With a little domain knowledge, data visualizations can be used to express and demonstrate key relationships in plots and charts that are more visceral and stakeholders than measures of association or significance. Data visualization and exploratory data analysis are whole fields themselves and it will recommend a deeper dive into some the books mentioned at the end.

Data Visualization After the classification and regression process the predicted results are visualized in graphical or tabular format for better understanding of the users. This process is called as Data Visualization. We can also get the summary of the results in numerical format.

Sometimes data does not make sense until it can look at in a visual form, such as with charts and plots. Being able to quickly visualize of data samples and others is an important skill both in applied statistics and in applied machine learning. It will discover the many types of plots that you will need to know when visualizing data in PHP and how to use them to better understand your own data.

- How to chart time series data with line plots and categorical quantities with bar charts.
- How to summarize data distributions with histograms and box plots.
- How to summarize the relationship between variables with scatter plots.

Many machine learning algorithms are sensitive to the range and distribution of attribute values in the input data. Outliers in input data can skew and mislead the training process of machine learning algorithms resulting in longer training times, less accurate models and ultimately poorer results.

Even before predictive models are prepared on training data, outliers can result in misleading representations and in turn misleading interpretations

of collected data. Outliers can skew the summary distribution of attribute values in descriptive statistics like mean and standard deviation and in plots such as histograms and scatterplots, compressing the body of the data. Finally, outliers can represent examples of data instances that are relevant to the problem such as anomalies in the case of fraud detection and computer security.

It couldn't fit the model on the training data and can't say that the model will work accurately for the real data. For this, we must assure that our model got the correct patterns from the data, and it is not getting up too much noise. Cross-validation is a technique in which we train our model using the subset of the data-set and then evaluate using the complementary subset of the data-set.

The three steps involved in cross-validation are as follows:

1. Reserve some portion of sample data-set.
2. Using the rest data-set train the model.
3. Test the model using the reserve portion of the data-set.

Advantages of train/test split:

1. This runs K times faster than Leave One Out cross-validation because K-fold cross-validation repeats the train/test split K-times.
2. Simpler to examine the detailed results of the testing process.

Advantages of cross-validation:

1. More accurate estimate of out-of-sample accuracy.
2. More "efficient" use of data as every observation is used for both training and testing.

Training the Dataset:

- The first line imports iris data set which is already predefined in sklearn module and raw data set is basically a table which contains information about various varieties.
- For example, to import any algorithm and train_test_split class from sklearn and numpy module for use in this program.
- To encapsulate load_data() method in data_dataset variable. Further divide the dataset into training data and test data using train_test_split method. The X prefix in variable denotes the feature values and y prefix denotes target values.
- This method divides dataset into training and test data randomly in ratio of 67:33 / 70:30. Then we encapsulate any algorithm.
- In the next line, we fit our training data into this algorithm so that computer can get trained using this data. Now the training part is complete.

Testing the Dataset:

- Now, the dimensions of new features in a numpy array called 'n' and it want to predict the species of this features and to do using the predict method which takes this array as input and spits out predicted target value as output.
- So, the predicted target value comes out to be 0. Finally to find the test score which is the ratio of no. of predictions found correct and total predictions made and finding accuracy score method which basically compares the actual values of the test set with the predicted values.

MODULE-04:

Support vector machines (SVM):

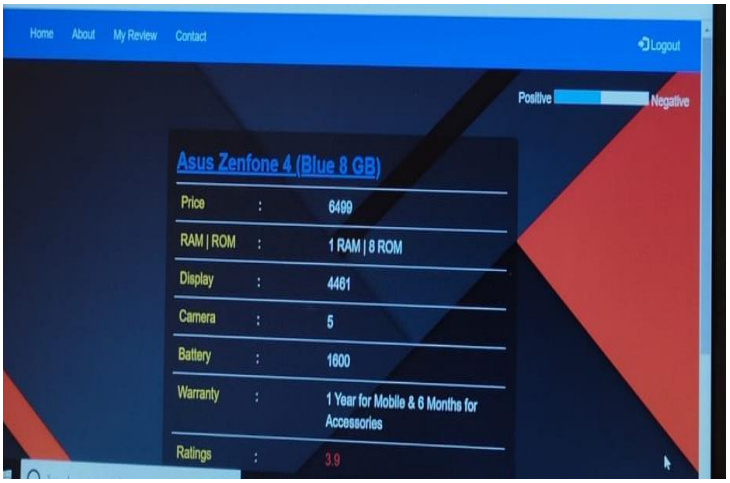
A classifier that categorizes the data set by setting an optimal hyper plane between data. I chose this classifier as it is incredibly versatile in the number of different kernelling functions that can be applied and this model can yield a high predictability rate. Support Vector Machines are perhaps one of the most popular and talked about machine learning algorithms. They were extremely popular around the time they were developed in the 1990s and continue to be the go-to method for a high-performing algorithm with little tuning.

- How to disentangle the many names used to refer to support vector machines.
- The representation used by SVM when the model is actually stored on disk.
- How a learned SVM model representation can be used to make predictions for new data.
- How to learn an SVM model from training data.
- How to best prepare your data for the SVM algorithm.
- Where you might look to get more information on SVM.

MODULE-05:

A classifier that categorizes the data set by setting a dictionary dataset between data. I chose this classifier as it is calculate count of total count used for looping function in the number of different positive & negative count that can be applied and this model can yield a high predictability rate. Support Vector Machines are perhaps one of the most popular and talked about machine learning algorithms. Sentiment analysis is the automated process that uses SVM algorithm to analyze data and classify opinions as positive, neutral or negative. Sentiment analysis enables businesses to understand how customers feel about their brand, product or service, helping them to make data-driven decisions. Whether you're a developer, a marketer, a data analyst, or you're just interested in sentiment analysis.

5. RESULT ANALYSIS



6. CONCLUSION

The analytical process started from data cleaning and processing, exploratory analysis and finally model building and evaluation.

7. FUTURE WORK

- To automate this processes by show the result in desktop application.
- To optimize the work to implement in Artificial Intelligence environment.

8. REFERENCES

[1] "Influence of Cognitive Resource Limitation on Consumer Purchasing Decision An Event-related Potentials Perspective Weiwei Han, Hua Bai", 2018

[2]"Determining the Effects of Marketing Mix on Customers' Purchase Decision Using the Grey Model GM(O,N) " Case Study of the Western style Coffeehouse Chains in Vietnam ,Yu-Chien Chai, Ying-Fang Huang, and Hoang-Sa Dang ,2017

[3] ''Online Apparel Shopping Behavior Yueh-Chin Chen, Yen-His Lee, Hsiao-Chun Wu", Yu-Chin Sung, Hung-Yi Chen ,2017