International Journal of Advances in Engineering and Management (IJAEM)
Volume 3, Issue 3 Mar. 2021, pp: 370-377 www.ijaem.net ISSN: 2395-5252

Customer Behavior Forecasting Using Big Data Analytics

M P Geetha [1], B Darshini[2], M Karthikeyan[3], V Chinnathambi[4]

[1]Assistant Professor, Department of CSE, Sri Ramakrishna Institute of Technology,
[2,3,4] Student, Department of CSE, Sri Ramakrishna Institute of Technology, Coimbatore, Tamil Nadu, India

Submitted: 01-03-2021 Revised: 09-03-2021 Accepted: 12-03-2021

ABSTRACT: From an enormous dataset to enhance the accuracy and efficiency of forecasting the Data mining techniques are the very effective tools in extracting the hidden knowledge. Intelligent Decision Analytical System needs integration of decision analysis and predictions. The accuracy in sales forecast provides a huge impact in business. Many of the business organizations are mostly depend on a knowledge base and demand prediction of sales trends. The detailed study and analysis of comprehensible predictive models to improve the future sales predictions are to be carried out in this proposed work. These problems could be overcome by using various data mining techniques. The concept of sales data and sales forecast is briefly analyzed, in this project. The various techniques and measures for the sales predictions are identified. On the basis of a performance evaluation, an apt predictive model is suggested for the sales trend forecast.

Keywords: consumer behaviour, data analytics, random forest, linear regression, sales prediction.

I. INTRODUCTION

Customer Behavior, in simple words, it means how customers behave in the market.

It defines the process by which customers make a purchase decision to satisfy their needs and wants. It consists of likes and dislikes of customers and which controls their buying decision. It is a concept which includes many stages from arising needs to the purchase decision. Every customer state of mind is not the same; they all differ from each other. Thus, every business requires to understand its customers. It helps the organizations to fulfill customer's demand and desire.Organizations utilize customer relationship management should have the knowledge about their customers properly.It is a database which collects more data about their customers.

A customer behavior analysis is a quality and quantity observation of how customers interact with your company. Customers are first divided into buyer personas based on their common characteristics. Then, each group is observed at the stages on your customer journey map to analyze how the personas in touch with your company.

A customer behavior analysis provides intelligence into the different variables that controls an audience. It offers you an idea of the motives, priorities, and decision-making methods being considered during the customer's journey. This analysis helps you understand how the customers feel about of your company, as well as if that perception aligns with their core values.

Data mining techniques are very effective in tuning high volume of data into useful information for cost prediction and sales forecast, Customer behavior models are based on the data mining of customer data, it is the basic of sound budgeting. At the organizational level, forecasts of sales are necessary inputs to many decision making activities in various functional areas such as operations, marketing, sales, production and finance. In order to serve an organization’s internal resources effectively, predictive sales data plays an important role in businesses when looking for acquiring investment capital. The studies proceed with a new perspective that focuses on how to choose an appropriate approach to forecast sales with high degree of precision. Initial dataset considered in this research had a large number of entries, but the final dataset used for analysis having much smaller size compared to the original due to the riddance of non-usable data, redundant entries and irrelevant sales data.

II. LITERATURE SURVEY

Sunitha Cheriyan,et.al, in“ Intelligent Sales Prediction Using Machine Learning Techniques” says that the older method of sales forecasting were difficult to predict accuracy in big data so as a solution for this here they used several techniques like Generalized Linear Model(GLM), Decision Tree (DT) and Gradient Boost Tree(GBT).Generalized linear model provided estimate of the regression coefficients and estimated asymptotic standard errors of the coefficients. Gradient boosting is a machine learning technique for regression and classification problem was built on a principle that a collection of weak learners combined together can produce a strong learner by using boosting process. Root Mean Square Error, Mean Square Error, Absolute error ,average of the error were calculated to incorporate in the classification algorithm. They were used in generating classification accuracy. After certain predictions they concluded that Gradient Boost Algorithm shows 98% accuracy and Decision Tree Algorithms shows 71% overall accuracy and Generalized Linear Model shows 64% accuracy. Finally, they compared based on the empirical evaluation of those three chosen algorithm and the best was Gradient Boosted Tree. Its accuracy rate reached upto 100%, but this GBT model achieved approximately 98% of accuracy. Machine learning approaches highlighted in this research paper will be able to provide an effective mechanism in data tuning and decision making. In order to be competent in business, organizations are required to equip with modern approaches to accommodate different types of customer behaviour by forecasting attractive sales turn over. In their studies, they used almost 85,000 records for the comparison of algorithms. At the same time, fields and attributes, used in this analysis were insufficient for the further analysis. It was the major challenge they faced during the research. However, they had thoroughly weighed the works by implementing efficient ML techniques for prediction and forecasting. The current studies can be expedited by using Big Data as a tool for the predictive analytics in sales forecasting.[1]

Anindita A Khade et al, in this paper “Performing Customer Behavior Analysis using Big Data Analytics” defines the proposed system for distributed implementation of C4.5 algorithm using MapReduce framework along with the customer data visualization. They first discussed many things about the big data techniques and customer behaviour on various fields.Then quoted many contents on key concepts of analysis which included venn diagrams, data profiling, forecasting, mapping, association rules and decision trees. There are many data visualization tools like Poly maps, Flot, D3.js, SAS visual analytics. Here the database technique used was Apache hadoop and HDFS. For implementation Map reduce model was used which had two phases like map phase and reduce phase and each had key value pair given by the programmer. The output will be sorted groups based on key value pair. Their flow of work was like importing dataset from HDFS and implementing C4.5 classification algorithm, then map phase and reduce phase. The entropy, information gain and gain ratio were calculated. The decision rules were generated and stored in HDFS. The system can also accept new test data from web UI and decides the data category and visualization is done as bar graphs, pie charts etc. After calculating entropy and gain value using C4.5 data is visualized using D3.js. The proposed future works included the use of fast and real time database systems like Apache HBase or MongoDB can be incorporated with this system. In addition to this, we can use distributed refined algorithms like Forest Tree implemented in Apache Mahout to increase performance and scalability.[2]

Abhijit Raorane,et.al, in “Data Mining Techniques: a source for consumer behavior analysis” mentions about various psychology of consumer like how, what, when, howmuch, through which platform the products are bought frequently. Based on which he comes out with a good definition and application of customer behaviour. Then he took literature survey on eight papers and came out with certain findings. Then he drills through data mining concepts and has clearly explained its methodology for approaching consumer.Then he briefly explains about different classification of data mining based on 3 categories such as knowledge mined, application adopted, techniques utilized. Then the author chose Janata Bazzar , a super market in Kolhapur city for his experiment. Then he chooses association rule technique for grouping certain products in the shop. Market basket transaction was taken to know about the purchased products and found out the products that were bought more than once. Using association rule he found the support value and confidence value with two metrics-people who bought item a has also bought item b.Using those value the results shows almost 80% of accuracy. He did not leave any suggestions for future works. [3].

Zahid Ullah,et.al, in “Efficient Implementation od data mining:Improve customer behaviour”has used CRM technique that is Customer Relationship Management technique and also some data mining techniques. They analysed how data mining techniques can be employed to get knowledge from large amount of datasets. As a result the business can be improved to higher levels. To perform this task along with CRM, rule induction process were used on clustered data from customer database based on their queries. To predict the values of certain fields data retention, data distillation and logical pattern distillation approaches were employed. They also used a methodology called if-then for exploring the logic. In applying this process the following steps were taken: customer enquiry, where the customer queries are evaluated and found where it has to be sent next. Next clustering customers here the available data like customer profile, background, history of purchase, payments are divided into many clusters that have same characteristics. The next step is Rule induction engine to generate rules from those clusters. The results out of this can be either numerical or non numerical. This further gives way to many hypothesis like sales gain, customer satisfaction and improved marketing strategies. This method is solid solution for customer understanding. Using this CRM along with rule induction any organization can reach its peak.[4]

Daqing Chen,et.al, in“Data mining for the online retail industry: A case study of RFM model-based customer segmentation using data mining” performs customer-centric business intelligence for an online retailer shop. They used RFM technique. R denotes recency, F denotes frequency, M denotes monetary. Recency calculates how recently a product is bought by the customer and frequency is measure of how frequently they are buying. Monetary value is how much they spend. Along with this k means clustering algorithm and decision tree induction was used to segmenting. The available dataset was preprocessed for applying algorithm. So now they applied k means on the dataset to segment all those into various meaningful groups. This process was done using cluster node in SAS enterprise miner. This algorithm is very efficient in dataset having unequal variables of different magnitudes. After these process they found out that monetary value had some difference over recency and frequency. So finally a case study was done on online retail shop these outcomes are very valuable to increase the profit gain and understand customer mentality in terms of several attributes like preference, affordable, useful etc. The two main tasks for that is data cleaning and model prediction if this was done properly then they can easily find out the market strategies. The future work mentioned in this paper is to use association techniques to predict the patterns of products that means the major combinations are products that are often bought. And to analyse which segment of people buys which product, enhancing the websites of the company so as to attract people. Frequently tracking people and predicting accurately will improve the business. Offering discounts and comparing the lifetime value of the customer will enrich our growth.[5]

Roung-ShiunnWua,et.al, in “Customer segmentation of multiple category data in e –commerce using a soft-clustering approach” took research on electronic commerce market. He preferred this because online shopping will have multiple kinds of datas of customer purchasing and their characteristics. That too as it was depending on internet resources, the information was taken easily and marketing strategies were framed according to the expectation. In this paper they use soft clustering method which uses a latent mixed class clustering approach to classify customers based on categories. To create customer segment latent Drichlet model was used. To generate estimates from segmentation variation approximation was taken. This results were better than hard clustering and finite mixture model. They continued the process using soft clustering approach and selected 5 segments which had 25 components per segment. Then they divided the segments into two parts based on mean value of segment. After certain preprocessing they came up with the sample of 2329 partitions. They found out the results of mean and SD of buying frequency and money spent. And also they took customer satisfaction survey .The segment 1 and 2 had frequent and high shopping customer who also used high internet, 3,4,9 segment customers are found that they spent little money because they were not satisfies with the service, in 5 and 6 segment the customers were reluctant and 7 ,8 was the most efficient shoppers. They also had a table in which the age, gender, income, marital status of customers were mentioned. The major problem in this method was selecting number of latent classes and choosing the cutoff percentage for the membership-score difference. This paper gave the solution for above mentioned issue. [6].

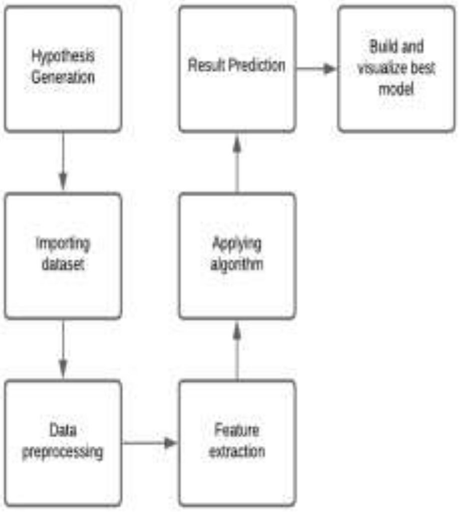
III. EXISTING SYSTEM

Customer Behavior forecasting is something like creating mathematical model to depict the common behaviors observed among particular groups of customers in order to predict how similar customers will behave under similar situations. The existing customer behavior models depends on data mining techniques, and each model is designed to answer one question at one time. For instance, this model can be used to predict the behavior of particular group of customer with a effect of marketing strategy. If that model was effective then the marketer will follow the same strategy to attract more and more group of customers. But the existing system were difficult and more expensive, because the mathematical techniques and tools used by the experts were very complex and highly costly. Even after building the costly model that were very expensive in manipulating and processing to find what exactly marketer has to do for bringing customers to their company. But on the other hand many models were very simple and predictive because they left out some features that were hindrance for them. From the above literature survey we could observe that based on Machine learning algorithm Generalized linear model, Decision tree, Gradient boost tree were taken as combination and they procceded. Their optimal solution had almost 64% accuracy with 85,000 dataset. In next paper the combination of C4.5 classification algorithm and map-reduce model was used for data mapping and grouping. Here major big data concepts like apache hadoop and HDFS were ruling. They also had a feature of data visualization using D3.js. Their efficiency and scalability was comparatively good. Another paper consist of data mining technique along with association rule . This is for generating patterns of customer preference by formulating association rules. He worked on a supermarket dataset and found out the results using support and confidence value. Their accuracy reached upto 80%. In the other paper he has used CRM technique that is Customer Relationship Management technique and also some data mining techniques. They analysed how data mining techniques can be employed to get knowledge from large amount of datasets. To perform this project along with CRM, rule induction process were used on clustered data from customer database based on their queries. They gave importance to data retention, data distillation, queries. With the answers of those queries many hypothesis were predicted. In next paper they used RFM along with k-means algorithm. This process was done using cluster node in SAS enterprise miner. This algorithm is very efficient in dataset having unequal variables of different magnitudes. After these process they found out that monetary value had some difference over recency and frequency. Next paper work was on e-commerce dataset and used soft clustering. They used soft clustering method which uses a latent mixed class clustering approach to classify customers based on categories. The main diffrencein this was it was totally internet dependent. They used Drichlet method segmented the results into many categories and the results was better than hard clustering and finite mixture model.

IV. IMPLEMENTATION:

To make some difference from the existing system, a new combo of algorithms is suggested in this report. Here mainly Linear regression and random forest algorithm was implemented. And the results of those methodologies were compared to find out the high accuracy.

The basic modules of this process include,
1.Hypothesis Generation
2.Importing dataset
3.Data preprocessing
4.Feature extraction
5.Applying algorithm
6.Result Prediction
7.Build and visualize best model.



1.HYPOTHESIS GENERATION:

This is the first stage for proceeding a project in this step, the problem statement is analyzed keenly and various hypothesis are generated. Hypothesis is something that shows relationship between various data. So we have to create hypothesis by asking many questions within us and we have to refer to many sources and formulate those hypothesis, after which we have to refine it and bring a solution for it. In our case, the task is to predict a best suited method to understand customer behavior and find out how they will behave in various circumstances. In this stage several articles were referred and with the help of many sources we understood the prevailing problem. Many queries were also generated and found solution for it. While considering our problem, we have to observe some information like from which type of city the product is sold if that was urban area then the sales will be high. Then we considered capacity of the shop similarly big shops will have more sales. If marketing strategies were better and if there was no competing shops then we can expect good sales. If a company gives attractive advertisement and discounts then sales will be high. These are various hypothesis generated for our project.

2. IMPORTING DATASET:

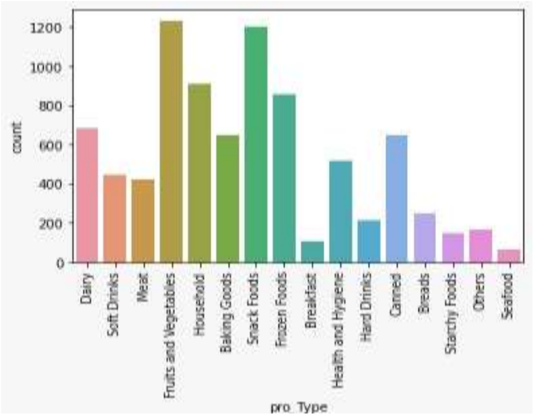
The dataset chosen here was a dataset that was obtained from various stores which have all the products used in household activities precisely we can tell a supermarket dataset. This was collected in 2013 and it has about 1560 products from 10 cities. We fetched this dataset from Kaggle website. The attributes of those products were mentioned in that. The attributes include-Product code which denotes the unique identity of the item, then weight denoted the weight of the product, fat level is another attribute from which we can get the fat content of the item , product visibility shows the % of total display area allocated to that item in the store. Product type denotes the category to which that product belongs to, MRP gives the cost of the product. Outlet shows the unique store ID, year shows the year in which shop was established. Outlet size is size of shop in terms of area covered. Location shows the type of city in which store is located whether urban /rural . Outlet type denotes whether the outlet is just a grocery or sort of departmental store. Sales shows sales of product in particular store. Here we used python for programming and using those libraries dataset were imported .

3. DATA PREPROCESSING:

A dataset is a collection of data objects, also called as a records, points, vectors, patterns, events, cases, samples, observations, or entities. Data preprocessing is that step in which the data gets transformed to bring it to a state such that now the machine can easily parse it. The features of the data can now be easily interpreted by the algorithm. The features can be either categorical or numerical .Categorial refers to features that have defined value set. Numerical refers to features that include numerical values which may be continuous or discrete. The steps involved in data preprocessing is data cleaning, data transformation, data reduction. Here we do data cleaning process to find and replace all null values available in the dataset. In this data cleaning the missing values are identified and they are counted using sum function and the replacing technique is done with mean value of those category that has missing value. In our problem, we could find null values in sales attribute and weight, outlet size. Those null values were replaced with mean values. After this step, a clear count of all item were found. Based on each and every category like item, outlet type, size, weight, item type their count was calculated.

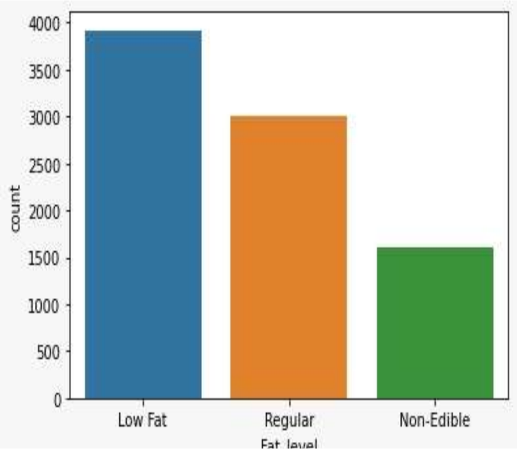
4. FEATURE EXTRACTION:

A feature is a property shared by all of the independent values on which prediction can be done. Any property could be a feature, as long as it is useful to the model. Feature engineering is the process of using domain knowledge to extract features from raw data through data mining techniques. These features can be used to improve the performance of our project. Feature engineering can be considered as important step of any project. We explored some nuances in the dataset in the feature extraction section. Now let’s resolve them and make our data ready for analysis. First lets have a clear graph for all products available in our dataset.

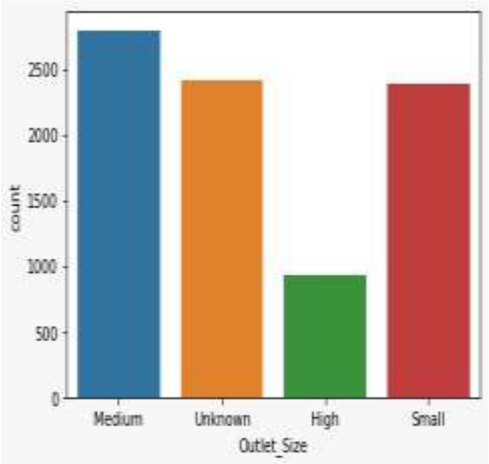


We found outlet type had some confusions so, here both supermarket type 2 and supermarket type 3 are combined together. After this combination there is some clear significant difference in that. And count of outlet type is calculated in each category. Then visibility attribute is taken into account because they have higher feature importance if a product has better visibility and kept in large area then that will have good sales. In product type we had 16 types of products they were further categorized into three major types as food, non consumable products and drinks.

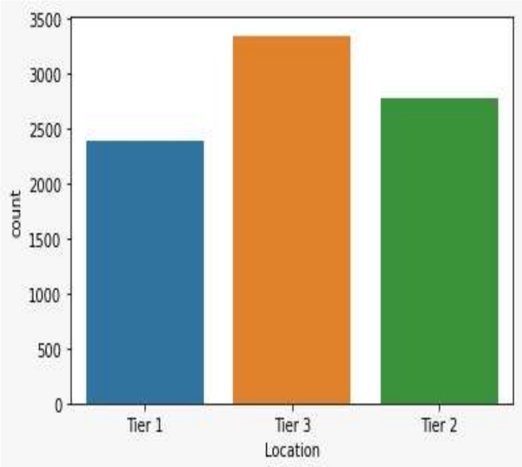
Then we are considering the fat level attribute, first for the food products the fat content is observed and divided as low fat food and regular fat food. Their count is calculated. Then further categories are modified because we have some products that does not belong to food item. So modified categories include low fat, regular fat, non edible products.



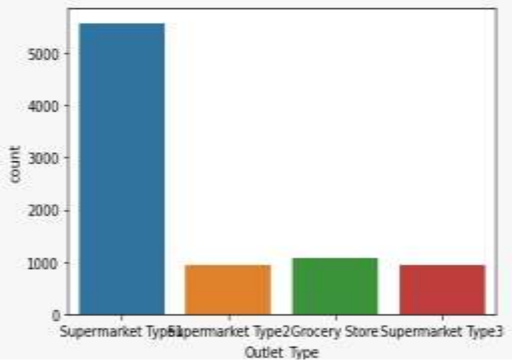
Then next feature we worked on was outlet size of the shop. This category includes medium, small and high level shops . If the outlet of the product was very big shop then the sales would be comparitively lower because there are all classes of people belonging to different locations like urban, rural and small areas which were refered as tier 1, tier 2 and tier 3.



As said above the following graph shows the count of products based on the location type.



Then based on outlet type of products the graph was generated.



So this step of feature extraction is now completed and all these extracted features will be included in the next step of applying algorithm. This features of our data will have a impact on the final results. Having this feature engineering we can create most accurate structure of data and hence creation of best model will be easier. Reducing the number of features through feature extraction ensures training the model will require less memory and computational power, which leads to shorter training times and will also help to reduce the chance of overfit . Simplifying the training data will make the model to interpret easily, which can be important when justifying real-world problems as a result of model outputs.

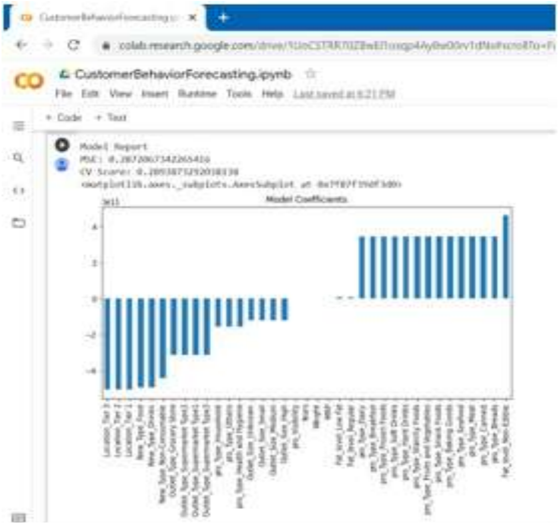
5. APPLYING ALGORITHM:

Unlike others implementation, we decided to apply linear regression and random forest algorithm. Many researchers used several algorithms like Generalized linear model, Decision tree, Gradient boost, C4.5 classification algorithm and map-reduce model, data mining technique along with association rule, CRM technique, rule induction process, RFM technique, K means algorithm.

LINEAR REGRESSION:

Linear Regression is used for predictive analysis models. This is a technique which explains the degree of relationship between two or more variables (multiple regression) using a best fit line. Simple Linear Regression is used when we have, one independent variable and one dependent variable. A scatter plot is used to fit a single line in this technique. The simple form of linear regression with one dependent and one independent variable is:
Y = aX + b

In this equation is known as linear regression equation, where Y is target variable, X is input variable and 'a' is slope and 'b' is intercept. That line which can explains the relationship better is said known as best fit line .On the other hand, the best fit line will return most accurate value of Y based on X that is causing a minimum difference between actual and predicted value of Y.


CustomerBehaviorForecasting.ipynb
File Edit View Insert Runtime Tools Help Last saved at 6:21 PM
+ Code + Text

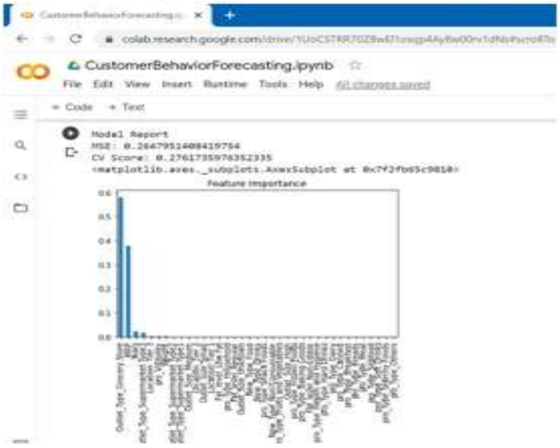
```
Model Report
MSE: 0.2872067342265416
CV Score: 0.2893873292018338
<matplotlib.axes._subplots.AxesSubplot at 0x7f87f19df3d0>
```

Model Coefficients
1e11
4
2
0
-2
-4

From this output we could see that, mean squared error value is approximately 0.287(approx) and the cross variation score is 0.289(approx).

RANDOM FOREST ALGORITHM:

Random forest algorithm is very easy technique in which we can measure the relative importance of each and every feature that we extracted in feature engineering process. This measures the feature by calculating how much the tree nodes use that feature will reduce impurity across all trees in the forest. This also computes a value for each feature after training and results the sum of all important features. By these feature importance we can predict what are all the features necessary and what is unnecessary so the features that does not contribute can be left out.This algorithm has best performance and can work better in all data mining techniques.


CustomerBehaviorForecasting.ipynb
File Edit View Insert Runtime Tools Help All changes saved
+ Code + Text

```
Model Report
MSE: 0.2647951408419754
CV Score: 0.2761735976352335
<matplotlib.axes._subplots.AxesSubplot at 0x7f2fb65c9810>
```

Feature Importance
0.6
0.5
0.4
0.3
0.2
0.1
0.0

After applying random forest algorithm ,we could see that the obtained mean squared value is 0.264(approx) and CV score is 0.276(approx).

V. RESULTS AND DISCUSSION:

From the above obtained results we could conclude that random forest algorithm shows lesser mean squared value and CV score that is cross validation score is lesser than linear regression algorithm. By looking at the feature importance graphs of both the process, linear regression says grocery store feature has more importance and random forest says fat level has more importance. Finally the algorithm which has lesser mean squared value and cross validation score is considered to have higher accuracy and Random forest algorithm only produced that result.

VI. CONCLUSION

By predicting more and more features like this, a company can easily increase its sales and can predict the customer's behavior at different circumstances. In recent trend this technology and method plays a major role as the market's competition is going on increasing. If a company literally wants to withstand for many years then they have to concentrate and spend something on this process. A part of spending their total capital in sales prediction is too good on any case.

VII. FUTURE WORKS

In this paper we showed and discussed about several algorithms and their predictions. In future you may try with some other latest algorithms. And also we can include some different features of a dataset. Some other fields other than this market can also be used. This process will also create a great impact on government related activities like if the government is trying to implement something new then in prior, they can predict people behavior towards that project.

ACKNOWLEDGEMENTS: The authors are deeply grateful to SRI RAMAKRISHNA INSTITUTE OF TECHNOLOGY, Coimbatore for providing the necessary facilities forthe preparation of the paper.

REFERENCES

[1]. Sunitha cheriyan,ShanibaIbrahim,"Intelligent sales prediction using machine learning technique", 978-1-5386-4904-6/18/$31.00 © 2018 IEEE.

[2]. Anindita AKhade," Performing Customer Behavior Analysis using Big Data Analytics", 7thInternational Conferenceon Communication, Computing and Virtualization 2016.

[3]. Abhijit Raorane ,R.V.Kulkarni, "Data Mining Techniques: A Source For Consumer Behavior Analysis", International Journal of Database Management Systems,September 2011.

[4]. Abdullah Al- Mudimigh, Farrukh Saleem, Zahid Ullah, "Efficient Implementation ofData Mining: Improve Customer's Behaviour ",2019.

[5]. Paolo Giudici, Gianluca Passerone, "Data mining of association structures to model consumer behavior", Published on "Computational Statistics & Data Analysis", 2016. www.elsevier.com/locate/csda

[6]. Tomoharu Iwata, Shinji Watanabe, Takeshi Yamada, Naonori Ueda",Topic Tracking Model for Analyzing Consumer Purchase Behavior",2009.

[7]. Patcharin Ponyiam, Somjit Arch-int,"Customer Behavior Analysis Using Data Mining Techniques", International Seminar on Application for Technology of Information and Communication, 2018.

International Journal of Advances in Engineering and Management
ISSN: 2395-5252

IJAEM

Volume: 03 Issue: 03 DOI: 10.35629/5252

www.ijaem.net

Email id: ijaem.paper@gmail.com