Sales Component Analysis & Prediction Using Linear Regression

Abhay Mishra [1], Mohd Hamd [2], Anubhav Yadav [3], Shubham Tiwari [4]
[1,2,3,4]*Department of Computer Engineering, Raj Kumar Goel of Institute of Technology, Ghaziabad*

Abstract **- Industrial engineering is about enhancing a process and improving the return of investment both to make more profit. Such is the background behind the Sales Mart Data analysis, the purpose being the determination of the properties of products and stores that help increase the sales. The data analysis was part of a competition launched by the American stop-shop Big Mart with the aim of building a predictive model that could predict the sales of the following year for each of the 1559 products in the 10 different stores of Big Mart. The aim is to build a predictive model and find out the sales of each product at a particular store. Create a model by which Big Mart can analyse and predict the outlet production sales. Motivation came into the mind with the idea of developing excellent Business Strategies. To predict the future of a particular product whether it is in demand or not. The main objective is to understand whether specific properties of products and/or stores play a significant role in terms of increasing or decreasing sales volume. To achieve this goal, we will build a predictive model and find out the sales of each product at a particular store. We will use Data science and ML on Python to create the predictive models that allowed us to have a better understanding of the client's behaviour and an**
people estimation of the store's future sales. It turned out that customers tend to prefer a product with a high MRP because the negotiation margin is bigger and because a high price tends to be associated with a better quality. With this information the corporation hopes we can identify the products and stores which play a key role in their sales and use that information to take the correct measures to ensure success of their business.

I.INTRODUCTION

With the rapid development of global malls and stores chains and the increase in the number of electronic payment customers, the competition among the rival organizations is becoming more serious day by day. Each organization is trying to attract more customers using personalized and short-time offers which makes the prediction of future volume of sales of every item an important asset in the planning and inventory management of every organization, transport service, etc. Due to the cheap availability of computing and storage, it has become possible to use sophisticated ML algorithms for this purpose. In this paper, we are providing forecasts for the sales data of big mart in a number of big mart stores across various location types which is based on the historical data of sales volume. According to the characteristics of the data, we can use the method of multiple linear regression analysis and random forest to forecast the sales volume.

II. LITERATURE REVIEW

The author of this paper has main motive to analyse the electricity consumption data of Iran for the year 1991 to 2013 to predict the electricity consumption rate from 2014 to 2020.After research author found that the growth rate of consumption from year 1999 to 2006 it was73.53% whereas from year 2006 to 2013 it was 28.41%. After using a regression model the result of prediction shows that the electricity consumption rate would increase by 22.28% by 2020 as compared to 2013. Pawan Pandey analyse two advanced machine learning algorithms known as ANN and SVM on soil dataset to find hidden facts, information, various patterns etc. Author [3] basically interprets what is data analysis and how we can do it efficiently? In this paper the author recommends Python for data analysis because of its tremendous capability of data exploration, several inbuilt packages, easy to implement several ML algorithms etc. As we know that Python is a statistical language as well as a programming language which helps in effective model prediction and better visualization techniques. So, after survey authors found that with Python data analysis is much more efficient. To build this application, various machine learning aspects are

used, such as Supervised Learning task, Regression task, and Plain Batch learning. Supervised learning will help you to understand the flow of data and know the sale prices, etc. The regression task uses algorithms to predict sales prices. Batch learning will help you to study the data in batches and improvised the results. The sales are to be kept in mind to predict the results, and the sales depend on the location of the store, population around the store, brand popularity, etc. One should also know about the city in which the store is located, either it is in an urban area or rural. Population statistics around the store also affect the sales, then store capacity should also be considered, and many more things. The brands which are being sold at the Big mart also play an essential role in predicting sales. The product sallies vary through the utility of the product, display area, advertising of the product, and many more aspects. The data set is quite big, and it needs to be decoded through algorithms. The first step will be declaring variables that will do the calculations of data. The variables should be declared for Item visibility, Item type, Outlet size, Outlet location type, Outlet type, and Item outlet sales. The data is categorized, and the first step will be to the correction of irregularities through data pre-processing. The variation of data is a real tough task as there are around 1562 unique items in a single store. The second step is to combine the outlet type through various parameters such as item visibility, years of operation, etc. Then create a broad category for item type using many item identifiers. Then the algorithm of ML will study the variations. A generic function that makes the model and performs cross- validation should be made. The next step will be the model making of the application, which will comprise the linear regression model, ridge regression model, decision tree model to decide the results, etc. The data fed to the application will go through sorting and arrangements which will be efficiently performed by Machine Learning. A sales analysis report shows the trends that occur in a company's sales volume over time. In its most basic form, a sales analysis report shows whether sales are increasing or declining. At any time during the fiscal year, sales managers may analyse the trends in the report to determine the best course of action. Managers often use sales analysis reports to identify market opportunities and areas where they could increase volume. For instance, a customer may show a history of increased sales during certain periods. This data can be used to ask for additional business during these peak periods. A sales analysis report shows a company's actual sales for a specified period a quarter, a year, or any time frame that managers feel is significant. In larger corporations, sales analysis reports may only contain data for a subsidiary, division or region. A small-business manager may be more interested in breaking sales down by location or product. Some small, specialized businesses with a single location are compact enough to use general sales data. A sales analysis report may compare actual sales to projected sales. Linear regression and logistic regression are the best machine learning models for this kind of problem where we can easily fit a line of high sale and low sale product, quarters and zone for a product. Also we need a huge amount of data for the training of the model which we can collect from the sales data of any product or company of the last 1 or 2 years for any live project. However, for this research project, the description of the data set which we are going to use for this project is provided in the data set portion of the experimental setup section.

Based on this brief discussion, we present a brief literature review on lead scoring and machine learning applications in automated customer relationship management.

III. METHODOLOGY

Predictive analytics is also a process employed within the business to customer marketing to rank lead supported their activities within the study, the general recommended process from for predictive analytics in information systems research is applied. With focus of the research being on the event and evaluation of possible predictive machine learning models for automated lead scoring, data understanding focuses on examining the knowledge and identifying and correcting potential problems present in it. The calculated purchase probability can then be used by companies to resolve different business problem. within the information preparation process, the knowledge is transformed so on address missing values and outliers, and to create a variable structure utilizing feature extraction, filtering and have selection that's appropriate for further machine learning model building.

A. Data Description

The data on which analysis the goal of the information description is to record all information about the knowledge files and their contents so as that somebody can use the knowledge in a very future research and understand the information content and structure. Documentation and more specifically metadata both provide information about the information at hand. Describing your data is significant. Systematically described research data is that the key to making your data findable, understandable and reusable. Overall data quality improves with clear data description and detailed documentation and metadata.

B. Data Preprocessing

In any Machine Learning process, data pre- processing is that step within which the information gets transformed, or encoded, to bring it to such a state that now the machine can easily parse it. In other words, the features of the information can now be easily interpreted by the algorithm.

Data pre-processing may affect the way during which outcomes of the ultimate processing is interpreted. This aspect should be carefully considered when interpretation of the results may be a key point, such within the multivariate processing of chemical data.

C. Data Cleaning

Data cleaning is that the process of preparing data for analysis by removing bad data, organizing the data, and filling within the null values. Ultimately, cleaning data prepares the information for the method of knowledge mining when the foremost valuable information will be pulled from the information set.

Data cleaning is one in every of the important parts of machine learning. It plays a big part in building a model. Data Cleaning is one in every of those things that everybody does but nobody really talks about. It surely isn't the fanciest a part of machine learning and at the identical time.

IV. RESULT

A. Results based on Algorithms:

Based on the analysis of the dataset we've seen that a lot of columns aren't adding any information to the model also these columns are leading to the degradation of the accuracy of the model hence we will drop these tables for more precision. Recursive Feature Elimination [7] and Principal Component Analysis [8] are used for the reduction of the dimensionality. High dimensionality results to the matter of overfitting, therefore help within the inefficiency of the model. After reducing the ineffective columns, the models showed high accuracy.

Table

Model	Accuracy
LR	91.36
KNN	91.06
SVM	92.37
NB	87.54
RF	91.11

The above table shows that Support Vector Machine model showed the highest accuracy, it is also seen that other models are also very much close in terms of accuracy

V. CONCLUSION

After the work on "Enhancing sales strategy of e-learning platform using ML", it's concluded that it's possible to estimate the sales strategy using supervised learning algorithms on the training set data. plenty of conclusion and result were extracted during the functioning on the info given by the user and from the developer sides also. we've used some unsupervised algorithm which could be a statical process which helps to convert the correlated features observations into a collection of linearly uncorrelated features with the assistance of orthogonal transformation for betterment of the result, because it reduces the info which don't seem to be useful (unwanted data) which increases efficiency within the result.

Recursive Feature Elimination are often used for feature selection algorithm which effectively choose or work on those data column and rows which are likely to supply results of the targeting result.

The model which supplies the very best percentage of efficiency comes from using Support Vector Machine of 92.37% and also the other models like Logistic Regression (accuracy of 91.36%), K Nearest Neighbour (accuracy of 91.06%) and Random Forest (accuracy of 91.11%) all showed an accuracy very near Support vector machine. Thus, we've selected Support Vector Machine for training the info for the estimation of sales efficiency for e-learning platform.

REFERENCES

[1] Archisha Chandel, Akanksha Dubey, Saurabh Dhawale, Madhuri Ghuge "Sales Prediction System using Machine Learning", International Journal of Scientific Research and Engineering Development-– Volume2 Issue 2, Mar – Apr 2019.

[2] Gopal Behera and Neeta Nain, "A Comparative Study of Big Mart Sales Prediction".

[3] Punam, K., Pamula, R., Jain, P.K.: A two-level statistical model for big mart sales prediction. In: 2018 International Conference on Computing, Power and Communication Technologies (GUCON). pp. 617–620. IEEE (2018).

[4] https://builtin.com/data-science/step-step-explanation-principal-component-analysis