Sales Forecasting Using Prediction Analytics Algorithm

A. Gokilavani[1], T. P. Banupriya[2], S. Bhagavathi[3], A. Divya Bharathi[4], T. Tamilselvan[5]

[1]*Associate Professor, Jai Shriram Engineering College, Tamil Nadu, India*

[2,3,4,5]*Student, Jai Shriram Engineering College, Tamil Nadu, India*

Abstract: **"Sales forecasting using prediction analytics algorithm" is planned for providing a complete analysis of sales forecasting. Sales forecasting is an important aspect of different companies engaged in retailing, logistics, manufacturing, marketing and wholesaling. It allows companies to efficiently allocate resources, to estimate achievable sales revenue and to plan a better strategy for future growth of the company. In this project, prediction of sales of a product from an outlet is performed via a two-level approach that produces better predictive performance compared to any of the popular single model predictive learning algorithms. The approach is performed on Departmental store. The proposed approach was organized into six stages, first is data collection, which includes collecting data and dataset, second is hypothesis definition, which used to analyse the problems, third is data exploration which used to explore the uniqueness of the data, fourth is data cleaning, which is used to detect and correct the inaccurate dataset, fifth is data modelling, which is used to predict the data using machine learning techniques, sixth is feature engineering, which is used to import the data from machine learning algorithm.**

I. INTRODUCTION

Sales is a life blood of every company and sales forecasting plays a vital role in conducting any business. Good forecasting helps to develop and improve business strategies by increasing the knowledge about the marketplace. A standard sales forecast looks deeply into the situations or the conditions that previously occurred and then, applies inference regarding customer acquisition, identifies inadequacy and strengths before setting a budget as well as marketing plans for the upcoming year.

In other words, sales forecasting is sales prediction that is based on the available resources from the past. An in-depth knowledge of the past resources allows to prepare for the upcoming needs of the business and increases the likelihood to succeed irrespective of external circumstances. In this project the approach has been done under six stages. In first stage, data is collected from dataset. In second stage, problems are analysed from the data collection. In third stage, uniqueness of the data is explored. In fourth stage, data cleaning is done to detect and correct the dataset.

In fifth stage, data modelling techniques is used to predict the data. In sixth stage, the feature engineering is used to import the data from the machine learning algorithm. Sales prediction is done accurately by using machine learning algorithms.

II. RELATED WORK

The degree of interest in each concept has varied over the year, each has stood the test of time. Each provides the software designer with a foundation from which more sophisticated design methods can be applied. Fundamental design concepts provide the necessary framework for "getting it right". During the design process the software requirements model is transformed into design models that describe the details of the data structures, system architecture, interface, and components. Each design product is reviewed for quality before moving to the next phase of software development.

The design of input focus on controlling the amount of dataset as input required, avoiding delay and keeping the process simple. The input is designed in such a way to provide security. Input design will consider the dataset should be given as input,dataset should be arranged, methods for preparing input validations.

A quality output is one, which meets the requirement of the user and presents the information clearly. In output design, it is determined how the information is to be displayed for immediate need. Designing computer output should proceed in an organized, well thought out manner; the right output must be developed while ensuring that each output element is designed so that the user will find the system can be used easily and effectively.

Database design contains the attributes of the dataset which are maintained in the database table. The dataset collection can be of two types namely train dataset and test dataset. Data flow diagrams are used to graphically represent the flow of data in a business information system. DFD describes the processes that are involved in a system to transfer data from the input to the file storage and reports generation. Data flow diagrams can be divided into logical and physical. The logical data flow diagram describes flow of data through a system to perform certain functionality of a business. The physical data flow diagram describes the implementation of the logical data flow.

DFD graphically representing the functions, or processes, which capture, manipulate, store, and distribute data between a system and its environment and between components of a system. The visual representation makes it a good communication tool between User and System designer. The objective of a DFD is to show the scope and boundaries of a system. The DFD is also called as a data flow graph or bubble chart. It can be manual, automated, or a combination of both.

It shows how data enters and leaves the system, what changes the information, and where data is stored. In this project we have performed sales forecasting for stores using different data mining techniques. The task involved predicting the sales on any given day at any store, in order to familiarize ourselves task we have studied previously.

III. RESEARCH METHODOLOGY

The main purpose of this research is to evaluate and analyze the use of data mining techniques for sales forecasting, to produce models which are comprehensive and reliable.

A. Dataset Collection

A data set is a collection of data. Departmental store data has been used as the dataset for the proposed work. Sales data has Item Identifier, Item Fat, Item Visibility, Item Type, Outlet Type, Item MRP, Outlet Identifier, Item Weight, Outlet Size, Outlet Establishment Year, Outlet Location Type, and Item Outlet Sales.

B. Hypothesis Definition

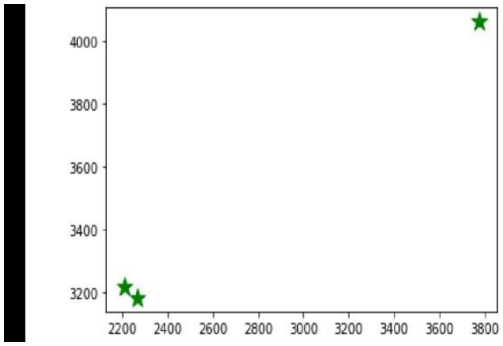
This is a very important step to analyse any problem. The first and foremost step is to understand the problem statement. The idea is to find out the factors of a product that creates an impact on the sales of a product. A null hypothesis is a type of hypothesis used in statistics that proposes that no statistical significance exists in a set of given observations. An alternative hypothesis is one that states there is a statistically significant relationship between two variables.

C. Data Exploration

Data exploration is an informative search used by data consumers to form true analysis from the information gathered.Data exploration is used to analyse the data and information from the data to form true analysis. After having a look at the dataset, certain information about the data was explored. Here the dataset is not unique while collecting the dataset. In this module, the uniqueness of the dataset can be created.

D. Data Cleaning

In data cleaning module, is used to detect and correct the inaccurate dataset. It is used to remove the duplication of attributes. Data cleaning is used to correct the dirty data which contains incomplete or outdated data, and the improper parsing of record fields from disparate systems. It plays a significant part in building a model.



Euclidean Distance Calculator to calculate the distance of the cluster

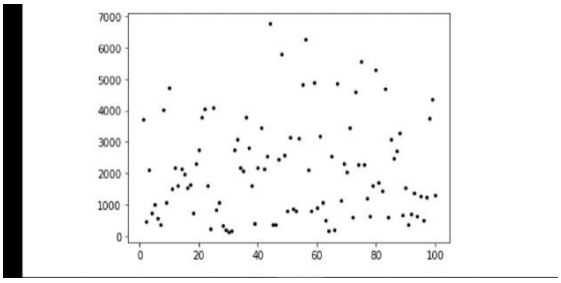
E. Data Modelling

In data modelling module, the machine learning algorithms were used to predict the sales. Linear regression and K-means algorithm were used to predict the sales. The user provides the ML algorithm with a dataset that includes desired inputs and outputs, and the algorithm finds a method to determine how to arrive at those results. **Linear regression algorithm** is a supervised learning algorithm. It implements a statistical model when relationships between the independent variables and the dependent variable are almost linear, shows optimal results. This algorithm is used to show the sales prediction with increased accuracy rate.

```
In [14]: #Linear Regression Model
         # Fitting Multiple Linear Regression to the training set
         from sklearn.linear_model import  LinearRegression
         regressor = LinearRegression()
         regressor.fit(X_train, y_train)

         # Predicting the test set results
         y_pred = regressor.predict(X_test)

In [17]: import warnings
         warnings.filterwarnings('ignore')
         # Measuring Accuracy
         from sklearn.metrics import accuracy_score, r2_score, mean_squared_error
         from sklearn.model_selection import cross_val_score
         from sklearn import metrics

         lr_accuracy = round(regressor.score(X_train,y_train) * 100,2)
         lr_accuracy

Out[17]: 79.00

In [16]: r2_score(y_train, regressor.predict(X_train))

Out[16]: 0.795563256726
```

Model Building using Linear Regression and shows accuracy rate, mean-squared value

K-means algorithm is an unsupervised learning algorithm. It deals with the correlations and relationships by analysing available data. This algorithm clusters the data and predict the value of the dataset point. The train dataset is taken and are clustered using the algorithm. The visualization of the clusters is plotted in the graph.



K-means Clustering visualization

F. Feature Engineering

In the feature engineering module, the process of using the import data into machine learning algorithms to predict the accurate sales. A feature is an attribute or property shared by all the independent products on which the prediction is to be done. Any attribute could be a feature, it is useful to the model.

III. IMPLEMENTATION

Implementation is the most crucial stage in achieving a successful system and giving the user's confidence that the new system is workable and effective. Implementation of a modified application to replace an existing one. This type of conversation is relatively easy to handle, provide there are no major changes in the system. Each program is tested individually at the time of development using the data and has verified that this program linked together in the way specified in the programs specification, the computer system and its environment is tested to the satisfaction of the user. The system that has been developed is accepted and proved to be satisfactory for the user. A simple operating procedure is included so that the user can understand the different functions clearly and quickly.

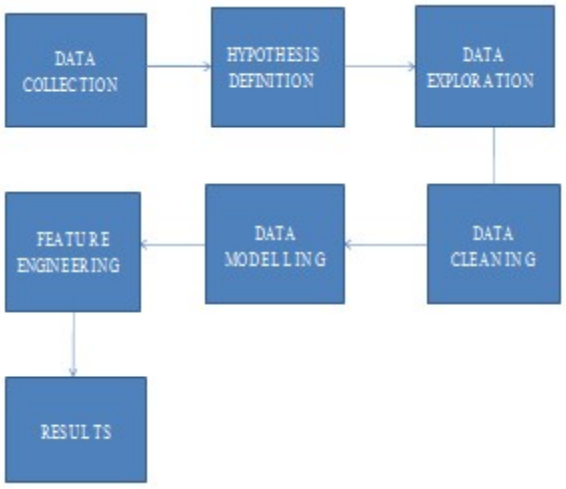
In early days' the sales value can be predicted manually by the user. The user can analyse the sales from the historical data and records. Here more paper work needed to be done by collecting the data from the historical record. The user can manually envision the sales which can be able to reach his target. And at the same time the user can able to get nostalgia result.

The processed data is used for predictive modelling so that appropriate results can be generated from it. This predictive modelling is done using a technique called Machine Learning. It is defined as a "computer's ability to learn without being explicitly programmed". Machine learning uses programmed algorithms that receive and analyse input data to predict output values within an acceptable range.

Machine learning algorithms are programs (math and logic) that adjust themselves to perform better as they are exposed to more data. The "learning" part of machine learning means that those programs change how they process data over time, much as humans change how they process data by learning. So, a machine-learning algorithm is a program with a specific way to adjusting its own parameters, given feedback on its previous performance making predictions about a dataset.

Machine Learning algorithm such as Linear regression and K-means clustering is been used to predict the sales of the departmental store and it is implemented in this project. The use of algorithm enables to increase the accuracy of the sales. The accuracy of the sales can be reached up to the value and it is plotted in the graph format.

Linear regression may be defined as the statistical model that analyses the linear relationship between a dependent variable with given set of independent variables.



The relationship can be represented by, Y=mX+b

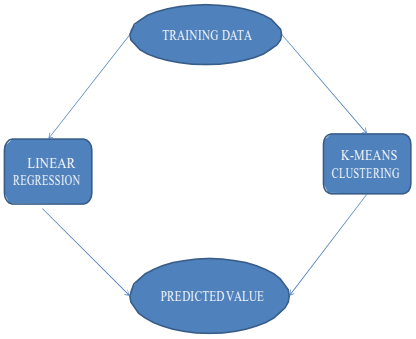
Y is the dependent variable, a sale which is to be predicted.

X is the independent variable, the dataset which is used to make predictions.

m is the slope of the regression line which represents the effect X that has on Y.

b is the line which crosses the y axis.

K-means clustering algorithm computes the centroids and iterates until it finds optimal centroid. It assumes that the number of clusters isalready known. The number of clusters identified from data by algorithm is represented by 'K' in K-means. It groups the similar data points into a cluster.



Select the number of clusters which is to be identified. Randomly select the distinct data points and assign each data point to the cluster. Measures the distance between the first data point and the selected cluster. Then the first data point is added to the nearest cluster then calculates the mean value, including new point of the first cluster. Repeat them until to get the optimal clustering the data point. K-means iterates repeatedly and until the data points within each cluster stops changing. Select the best variance out of it.

IV. CONCLUSION

Every company desires to know the demand of the customer in any season beforehand to avoid the shortage of products. As time passes by, the demand of the store to be more accurate about the predictions will increase exponentially. So, huge research is going on in this sector to make accurate predictions of sales. Better predictions are directly proportional to the profit made by the departmental store. The purpose of measuring accuracy was to validate our prediction with the actual result. In this project, an effort has been made to predict sales of the product from an outlet accurately by using a two-level statistical model that reduces the mean absolute error value. The two-level statistical model performed than the other single model predictive techniques and contributed better predictions to the departmental store dataset.

Further Enhancement

Further expansion of the system also can be done in future if needed. The application can be enhanced in the future with the needs of the food store. The database and the information can be updated to the latest forthcoming versions. Thus, the system can be altered in accordance with the future requirements and advancements. System performance evaluation must be monitored not only to determine whether they perform as plan but also to determine if they should have to meet changes in the information needed for the food store.

REFERENCES

[1]. Box G.E, Jenkins G.M, Reinsel G.C, Ljung G.M, Time Series Analysis.
[2]. Chatfield C, Time-Series Forecasting; Chapma and Hall/CRC: Boca Raton, FL, USA, 2014.
[3]. Doganis P, Alexandridis A, Patrinos P, Sarimveis H, Time series sales forecasting for short shelf-life food products based on artificial neural networks and evolutionary computing. J. Food Eng,2016.
[4]. Efendigil T, Önüt S, Kahraman C, A decision support system for demand forecasting with artificial neural networks and neuro-fuzzy models: A comparative analysis. Expert Syst. Appl,2017.