RESEARCH ARTICLE OPEN ACCESS

Sales Analysis and Prediction Using Python

Palak Mittal*, Sujay**, Simran***, Krishan Kumar****, Pronika Chawla*****
**(Department of CSE, MRIIRS, Faridabad*
***(Department of CSE, MRIIRS, Faridabad*
****(Department of CSE, MRIIRS, Faridabad*
*****(Department of CSE, MRIIRS, Faridabad*
******(Department of CSE, MRIIRS, Faridabad*

ABSTRACT

These days shopping centers and Big Marts maintain record of their selling details for all the persons to forecast the customer's potential demand and even monitor the inventory control. In a data center these data warehouses essentially comprise a vast amount of consumer details and individual object attributes. In fact, deviations and repeated variations are identified by removing data from the data warehouse. The resulting results will be used to forecast potential revenue figures for retailers like Big Mart using numerous machine learning techniques. In this paper, we build a predictive model using machine learning algorithms for predicting the sales of a company and find which model performs better. The models are compared to find out which model performs better in terms of performance.

***Keywords*:** Data Analytics, Machine Learning, Linear Regression, Random Forest, Python

Date of Submission: 13-05-2020 Date of Acceptance: 26-05-2020

I. INTRODUCTION

As the internet is growing rapidly, we have switched from utilizing standard data such as texts, documents etc to the more diverse types of data consisting of a huge amount of high-quality audio, images, photographs, interactive charts, position data and much more. Each single second the data is becoming bigger and bigger. It is of no use to have big data if it is not being utilized for taking decisions.[1]

Today, data analytics is being used across various fields for making predictions. One of the applications of data is in the government sector. For the government sector the analysis of big data has proved very important. Analysis using big data proved instrumental in Barack Obama's successful 2012 re-election campaign. For BJP and its ally's big data processing was primarily responsible for securing a widely competitive win. Various methods are being used by the government of India to assess how the population of India is reacting to political intervention, as well as policy-increase proposals.

Another area where data analytics is being used is in the field of social media analytics. The rise of social networking has caused a large data explosion. Numerous tools have been developed by various organizations like IBM to evaluate social network behavior. These tools are Cognos Consumer Insights, which is an application that runs on BigInsights Big Data Platform. These tools can be used to better understand the mood of the people about a certain activity that is going on in their region and the world.[2]

The data can be of various types such as structured, semi-structures or unstructured. Structured data is the type of data that is in the forms of rows and columns. These are basically tables of data in a database. Structured data requires minimum processing and is the easiest to analyze and it can directly be fed to the model for finding patterns, learning from the data, and then making analysis and showing the trend. Semi-structured data is the data about data. It is basically the metadata.

Unstructured data is the type of data that is in no specific format and is difficult to analyze. It requires a lot of pre processing to bring the data in a form so that it can be used for analysis. It is a very complex form of data and consists of data from all the nontraditional sources. This data can be in the form of audio, video, graphs, plots, power point presentation, instant messaging, and collaboration software.

Fig 1: structured and unstructured data

In the data era, sizeable quantities of statistics have come to be reachable on hand to decision makers. Big data refers to datasets that are now not only big, however additionally high in range and velocity, which makes them challenging to take care of using normal tools and techniques. Due to the speedy boom of such data, options need to be studied and supplied in order to take care of and extract price and expertise from these datasets. Furthermore, decision makers want to be in a position to obtain treasured insights from such varied and unexpectedly changing data. Such fee can be furnished using huge records analytics, which is the utility of advanced analytics techniques on big data. There are a number of tools that can be used for storing and analyzing data. Some of the popular tools for storing data are as follows:

- Apache Hadoop: It can be used to store enormous amount of data in a cluster. It is a java-based framework. It can run in parallel on a cluster and is capable of allowing users to process data across all nodes. This provides replication of data resulting in high availability of data.
- Hive: It's a distributed data management for Hadoop. It can be used for data mining purpose as it supports query operation like HiveSQL for accessing the big data.
- Apache Cassandra: It is a NoSQL database. It is scalable, and has high performance distributed database tohandle large amounts of data. We can store and retrieve data other than tabular relations with the help of a NoSQL database. The qualities of this database are that it is schema free, has a simple API, is consistent, supports easy replication, and can handle large amounts of data. [1]

Some of the popular tools for analyzing data are as follows:

- RapidMiner: RapidMiner can include any number of information source types, which include Microsoft SQL, Sybase, IBM SPSS, Excel, Oracle, MySQL, Access, Tera data, IBM DB2, Ingress, Dbase. The tool is very effective and can generate analysis primarily based on real life records transformation settings.
- Tableau Public: It's an intuitive and simple tool that offers interesting insights by data visualization. One can inspect a hypothesis, discover the data, and cross-check their insights.
- Jupyter Notebook: It is an accessible tool for performing end to end data science workflows – information cleansing, statistical modeling, building and training machine learning models, and visualizing data. [3]

Among all the different fields where data analytics can be used for making predictions and thereby gaining insights for making decisions one of the fields is sales. We have used Big Data Analytics to analyze and predict the sales of a product using various different models like linear regression and decision trees. We compare the two models to understand which of these performs better to obtain the best results. The language used for implementation is Python. The platform used for implementation is Jupyter Notebook.

II. DATA SET

Collection of data is termed as a dataset. Dataset refers to numerous database tables in the case of data in the form of a table. The row of the table gives information about the data set's record whereas the column gives the information about the particular variable in a table. The data set gives the complete values that are stored in the database in the form of variables for all data set members. Every value present in the database is termed as a datum. These may also consist of a large number of files and document.

There are many different characteristics that define a dataset such as the attributes and variables present in the dataset as well as their numbers and types and the numerous statistical measures applied to the dataset. There are a number of popular built-in datasets in the Python libraries used for analysis. Few examples of such built-in databases are:

- Iris flower dataset: It is a dataset which was introduces by Robert Fisher in 1936. It is a multivariate dataset.
- MNIST database: It is used for text classification, clustering, and image processing. It consists of the images of handwritten digits. [4]

The dataset that we have used is the sales dataset which is acquired from Kaggle. This dataset contains two files namely train and test. Both of these files are csv files. The aim is to predict the sales of a product using the test data set.

The dataset consists of 11 fields in the dataset namely: Item_Identifier, Item_Weight, Item_Fat_Content, Item_Visibility, Item_Type, Item_MRP, Outlet_Identifier, Outlet_Establishment_Year, Outlet_Size, Outlet_Location_Type, and Outlet_Type fields.
The description of the fields mentioned above are as follows:

- Item_Identifier: This field consists of the unique product ID of the item. It is an ID variable.
- Item_Weight: This fields consists of the weight of the product. This is not considered in hypothesis.
- Item_Fat_Content: This field tells whether the product has low fat or not. More than any other items the low-fat items are preferred. This particular field is linked to the 'Utility' hypothesis.
- Item_Visibility: This field tells us about the area assigned to a particular product with respect to the percent of the total display area of all products. It is used for the hypothesis of the 'display area'.
- Item_Type: This field tells about the category of the product. To derive more knowledge about the utility this field can be used.
- Item_MRP: This field tells about the MRP of the product. This field is not important for analysis and hence is not considered for the hypothesis.
- Outlet_Identifier: This field consists of the unique store ID. It is an ID variable.
- Outlet_Establishment_Year: This field gives information about the year in which the store was established. It is not considered in the hypothesis.
- Outlet_Size: This field tells about the ground area that the store covers. This field is linked to 'store capacity' hypothesis.
- Outlet_Location_Type: This field tells us about the location that is the type of city where the store is located. This field is linked to the 'city type' hypothesis.
- Outlet_Type: This field tells about whether the store is a supermarket or a small store. This field is also connected to the 'store capacity' hypothesis.
- Item_Outlet_Sales: This field is the outcome variable that is being predicted. It tells about the sales of the product in a store. This field is the desired outcome variable.[5]

III. PYTHON FOR DATA ANALYTICS

Python is a programming language that has a very easy syntax and semantics and is an interpreted language and high-level language. It takes less effort to create applications using this language thus making the job easier to perform. Other programming languages are harder than Python. Python has emerged to be one of the favorite languages of the programmers. One that is widely used for developing various applications as well as performing data analytics.

3.1 Features of Python

Python can achieve better productivity with less amount of code. However, it is not as fast as some of the other programming languages. The features of this language are:

- High-level: it has components of natural language that people use for communication. It is easy to understand what task the code is performing.
- Interpreted: Debugging errors is easy and efficient as the code is compiled line by line. This makes the Python programming language slow than other languages.
- Easy syntax: Indentations are used instead of braces in Python to determine which code block is under a certain class or function. This makes the code easy to read.
- Dynamic Semantics: There is no need to initialize anything before using. This process is done automatically in Python.
- Portable: There is no need to make changes in the code to run it on different systems. This makes it easy to work on a task.
- Open Source: It is free and can be used and modified by anyone as per their preference.
- Object-Oriented Language: It helps simulate real-world scenarios and provides security to get a well-made application.
- Simplicity: By understanding only indentations one can code any application in less lines of code.
- Embedding Properties: It is powerful and versatile and allows embedding of code from other languages like C.
- Library Support: It supports various libraries that can make obtaining solutions easy and fast.

3.2 Usage of Python

- Frameworks like Django and Flask are used for developing web applications.
- Creating workflows for the software.
- Modifying files and data in Databases.
- Complex calculations and scientific and analytic calculations.

3.3 History of Python

Python programming language was developed approximately 30 years ago in 1990's by

Guido van Rossum and first came into being in the year 1991. The main aspect of this programming language is its code readability and the usage of large enough to be noticed whitespace. It uses the multi programming paradigm. It also makes the usage of functional, imperative, object-oriented, structured, and reflective paradigm.

There are about 8 different implementations of Python programming language namely: CPython, PyPy, Stackless Python, MicroPython, CircuitPython, IronPython, Jython, RustPython. Python language is influenced by a number of other languages namely: ABC, Ada, ALGOL 68, APL, C, C++, CLU, Dylan, Haskell, Icon, Java, Lisp, Modula-3, Perl, Standard ML. There are languages whose development is influenced by Python. These languages are: Apache Groovy, Boo, Cobra, CoffeeScript, D, F#, Genie, Go, JavaScript, Julia, Nim, Ring, Ruby, Swift.

3.4 Scope of Python

There are a number of applications for Python which are as follows:

- Web and Internet development: Python has a vast collection of libraries and packages of internet protocols to make the task of developing web applications easier. Few of the libraries are: IMAP, FTP, image processing. Few of the packages present are: Feedparser, Beautifulsoup, Requests etc. frameworks such as Django, and Flask are also available.
- Desktop GUI: One can draft a user interface using binary distributions of Python shipped with Tk, which is a standard library for GUI.
- Scientific and Numeric Applications: Python is a powerful programming language and scientific and numeric applications is one of the most popular applications of this language. There are a number of libraries that allow to perform these tasks such as Numpy, Pandas, SciPy.
- Software Development Application: Python programming language can be used as a support language (for testing, build-control and management) for software development applications by software developers. Few of the examples are: SCons, Buildbot Apache Group etc.

IV. PROPOSED SYSTEM

The method to solve the problem at hand is given below. The unprocessed data at the Big Mart is collected. This raw data has to be pre-processed to obtain the missing data, outliers and the anomalies. We train two different machine learning algorithms namely linear regression and random forest on the raw data that is collected to obtain a model. This model helps us to predict the final outcome.

ETL refers to Extract, Transform and Load. This is the tool which will combine all three of the functions. It is fed the data from a particular database and the tool transforms the input data into a suitable format. The raw data is transformed to an understandable format by using data mining techniques that is data preprocessing. Data processing is a very important step as the data collected from real sources may be incomplete or inconsistent.

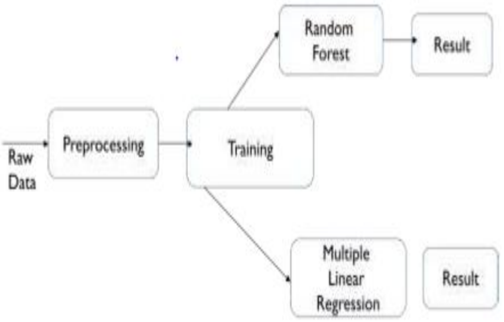

Fig 2: block diagram of the system

4.1 Linear Regression

It finds the relationship between the dependent variable (Y) and one or more independent variables (X) using one straight line which is the best fit line also termed as the regression line. The equation representing this line is:

Y=a+b*X + e

In the above equation:

a is intercept,

b is the slope of the line,

e is the error term.

The accuracy cab be found out using this method. Although this model is very famous for analysis its disadvantage is that it gives less accurate results.[6]

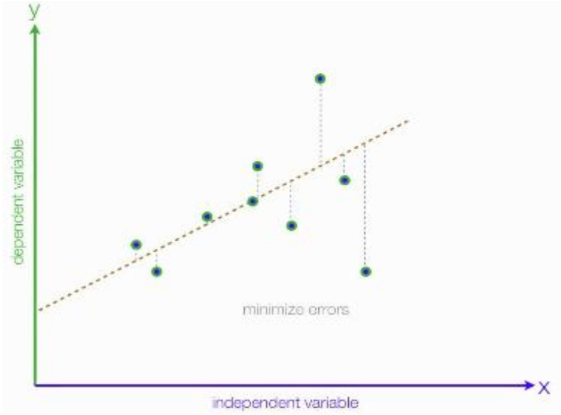

Fig 3: linear regression

4.2 Random Forest

These are also known as random decision forests. It is a machine learning algorithm that combines various tasks such as classification, regression among others. It builds multiple decision trees during the training period and output's the class that is mode of classes that is classification or mean prediction that is regression of the individual trees. It is used to overcome the disadvantage of decision trees that is overfitting.[6]

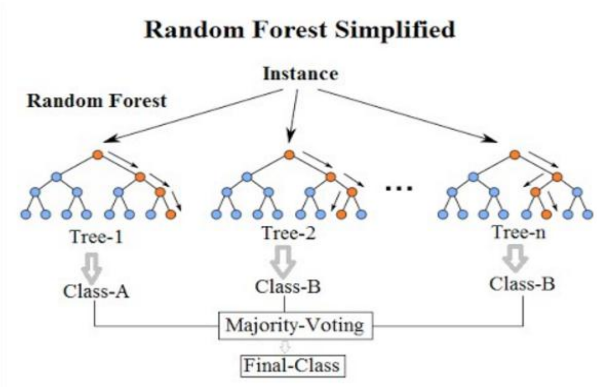

Fig 4: Random Forest

V. CONCLUSION

A software tool is proposed by us for predicting the future sales based on the historical data. With this tool, it can be found out how precise is the prediction for linear regression and random forest machine learning algorithms.

ACKNOWLEDGEMENT

The successful realization of the project is an outgrowth of a consolidated effort of people from disparate fronts. We are thankful to Dr. Krishan Kumar for his valuable advice and support extended to us without which we would have not been able to complete the project for success.

We are thankful to Ms. Pronika Chawla for her guidance and support.

Words cannot express our gratitude for all those people who helped us directly or indirectly in our Endeavour. We take this opportunity to express our sincere thanks to everyone for their valuable suggestions and also to our family and friends for their support.

REFERENCES

[1]. Palak Mittal, Mansi Sharma, Dr. Prateek Jain, A Detailed Study of Security and Privacy Concerns in Big Data,International Journal of Applied Engineering Research ISSN 0973-4562 Volume 13, Number 10 (2018) pp. 7406-7411

[2]. https://www.digitalvidya.com/blog/big-data-applications/

[3]. https://www.analyticsvidhya.com/blog/2018/05/starters-guide-jupyter-notebook/

[4]. https://en.wikipedia.org/wiki/Data_set

[5]. https://medium.com/@nr3702/bigmart-sales-data-regression-using-python-57a5155767d7

[6]. Heramb Kadam, Rahul Shevade, Prof.DevenKetkar, Mr. Sufiyan Rajguru, A Forecast for Big Mart Sales Based on Random Forests and Multiple Linear Regression, BE IT, FAMT, Ratnagiri, Assistant Professor ,IT department, FAMT, IJEDR 2018 | Volume 6, Issue 4 | ISSN: 2321-9939