Journal of Ubiquitous Computing and Communication Technologies (ISSN: 2582-337X)
www.irojournals.com/jucct/

Predictive Analytics with Data Visualization

Satheeshkumar Palanisamy

Assistant Professor, Coimbatore Institute of Technology, Coimbatore, India

E-mail: satheeshkumar.p@cit.edu.in

Abstract

There has been tremendous growth for the need of analytics and BI tools in every organization, in every sector such as finance, software, medicine and even astronomy in order to better overall performance. C-factor Computing has the same vision of empowering their existing products through data analysis and forecasting to better suit the need of customers and decision making of stakeholders. The project involves 5 key aspects in Analytics - Data Acquisition, Big data or data Storage, Data Transformation (Unstructured to Structured), Data Wrangling, Predictive Modeling / Visualization. Data Acquisition involves gathering existing transactional and search data of customers and travel aggregators who use the product. This data is used to create powerful dashboards capable of predictive analytics which help the company make informed choices. The key aspects mentioned can be achieved through various tools available but requires testing at every stage in order to realize the appropriate software for the data present in the company. Hence the project deals with studying and implementing selected tools in order to provide the right framework to achieve an interactive dashboard capable of predictive analytics which can also be integrated into the existing products of the company.

Keywords: Data analysis, data wrangling, forecasting, data mining, predictive modeling

1. Introduction

1.1 Overview

'Machine Learning' was first coined in the year 1956. The immediate question is why it has taken so long for it to become a cutting-edge technology that everyone seems to be interested in. The answer lies in the tremendous increase in data consumption and expulsion. An estimated, 44 zeta bytes of data are to be generated in 2020 which is 44 billion terabytes

DOI: https://doi.org/10.36548/jucct.2022.2.003
Received: 20.05.2022, received in revised form: 26.06.2022, accepted: 11.07.2022, published: 21.07.2022

of data. Such increase in data has propelled the application of Machine Learning and Data Analytics in every field imaginable starting from Finance to pharma and even astronomy. This trend has invoked the need for many start-ups and established organizations alike to use the data present in their companies to make informed decisions through data analytics and most importantly predict future patterns through predictive algorithms.

1.2 Importance of Data Science and ML

The increasing capabilities of machines and rise in data generated has made data science and machine learning key elements of decision making in today's world of technology and production. The use of various techniques to identify insights from the data present within the company and relevant data outsourced from the internet can lead to significant changes that put the company in the path of making profitable decisions. For example, assume X had to choose one hotel from a hundred hotels to invest in. This can be done in two ways. First where X and his teams sit down and manually check the financial statements of the hotels and then come to a decision. The second option is to use data analytics to study and find meaningful insights for decision making, forecast the growth of all hotels in the list, narrow options to top three hotels and then choose the company to invest in. The second option helps in reducing risk while making data driven decisions which can shape the future of the company.

1.3 Project Objective

To develop a dashboard capable of predictive analytics through implementation of Statistical Analysis and Machine Learning Algorithms on company specific semi-structured and structured data.

1. Convert unstructured data to Analysis ready structured data
2. Combine transformed data and pre-existing structured data present in Company database
3. Perform Data Wrangling
4. Explore the data in depth
5. Study and implement various Data visualization tools available (both open source and paid) and determine the right tool for company specific data

6. Create Dashboard capable of dynamicity and interactivity
7. Study an implement various classification, regression and forecasting techniques in-order to predict future trends
8. Combine trained predictive models and the interactive visualizations to create dashboard capable of forecasting

1.4 Opportunity Statement (Company)

To use the power of Data Analytics for two reasons

1. To provide stakeholders insights about how, where and when the company must invest to increase profitability.
2. To use the power of statistical visualizations, integrated into the company product environment providing the user a better experience while using the product.

1.5 Opportunity Statement (Student)

1. Obtain real time exposure of how industry data looks like which differs from pre structured data provided in courses present online.
2. Hands on experience, learning various tools required for analysis, data visualization, data wrangling, big data storage, DBMS, machine learning and deep learning algorithms. Learning these tools for a start-up enables an aspiring data scientist to study and implement a wide variety of software which is not the case in an established company (where the technology to be used has been pre-decided)

2. Literature Review

The difference between the supervised and unsupervised training algorithms is that supervised training algorithms aim at predicting future values from prior or known information whereas unsupervised training algorithms are tasked with the process of classification of input data. The paper discusses the efficiency of supervised machine learning algorithms in terms speed of learning, accuracy, risk of over fitting measures and complexity. The journal aims at providing an overview of state of art machine learning algorithms [2].

Classification techniques in data mining focuses on the three main components of data mining – Clustering or Classification, Association rules and Sequence analysis. The journal provides the insight that data mining task is the automatic or semi-automatic analysis of large quantities of data which is used to extract previously unknown patterns [7]. The concept of classification being a data mining technique used to predict group membership for each individual data instance is highlighted. Insights in the Six common tasks in data mining - Anomaly detection, Association rule learning, Clustering, Classification, Regression, Summarization is discussed. Majorly classified as decision trees, K-nearest neighbor, Bayesian networks is discussed.

Data Modeling Educational Data Classification and Comparative Analysis of Classifiers Using Python informs its readers about supervised and unsupervised learning algorithms using python specifically to derive insights from huge accumulated unstructured data [14].

In the world of time series forecasting, there exists varied models and methods to analyze and forecast. ARIMA is one among many popular methods which enable time series analysis. The importance of using trend based ARIMA over basic and wavelet based ARIMA can be studied while predicting temperature time series data. It also provides an insight into how the different parameters work in an ARIMA model [3].

Time series analysis and forecasting also plays an important role in the world of finance and economics. Intra-day traders and long-term investors often outsource applications or analysts who can predict the growth and depreciation of stocks and commodities in the market. Results suggest that the usage of ARIMA has incredible potential in predicting short term fluctuations using public data from NYSE. This enables analysts understand the complexities involved in prediction through real-world time stamped data. [1] NASDAQ data used from websites like Yahoo have also been used with seasonal ARIMA models to establish effective ways in forecasting financial data [17].

Pollution has become a global crisis. People have come to the bitter realization that planet earth cannot be polluted at the present rate. The ARIMA models also find purpose in predicting air pollution in order to curb the increasing rate of pollution in cities and select spaces. Experimental results from Seasonal ARIMA or SARIMA models have been shown to forecast air pollution [16]. Apart from ARIMA models, there are numerous forecasting methods such as the Holt-Winters method. The functionality of such models can be analyzed

while studying its use case in predicting aircraft failure rate. The Holt-Winters seasonal model consists of three smoothening equations which are interdependent as they form the forecast equation. The mathematics behind the models and its feasibility in effective prediction of aircraft failure help in understanding the various processes behind establishing an efficient forecasting model.

Apart from forecasting models, there are methods which are used to model the general level of time series data. EWMA and SMA can also forecast in certain cases. EWMA is used instead of SMA in-order to assign relative weight age to observations that have happened recently. Fault detection strategies using PCA based EWMA control schemes gives an insight into how EWMA can be used in multiple fields to detect small changes [6].

The error a model generates while testing is analyzed to better efficiency. Gradient descent algorithms are often used to determine optimal values which are then used to minimize the cost of an algorithm. Normal equation which is also used to find global minimum can prove to be more effective in reducing the cost of a Linear Regression algorithm [5]. This process of minimizing the cost function is important as it determines the efficiency of an algorithm.

Decision tress algorithm is one of the most popular classifier algorithms in the ML field. Developing decision trees which are accurate is a skill that is developed with practice. For this reason, in the early stages of an ML engineer's path, it is easy to over fit the algorithm there by making it inefficient. When these algorithms are trained using large volumes of data, the decision trees generated become very difficult to understand. Hence the process of generating optimal decision trees which are smaller but are capable of valid solutions is imp [4]. The concept of pruning where the complexity decreased whereas efficiency is increased also plays an important role in creating optimal decision trees. Once a Decision tree is pruned it is analyzed based on unbiased evaluation metrics [9].

An extension of decision trees which solves the problems such as over fitting is known as random forests. Random forests have been used to effectively classify grades of air pollution.[10] It uses data bagging and feature selection techniques to train all the decision trees present in the random forest differently. Random forests is also greatly known for its integrated learning ability and generalization ability which solves problems flexibly [8]. KNN Classifiers have also become popular as they are easy to understand and training time for such classifiers are less. Here the distance is measured using techniques such as Manhattan

and Euclidean distances. When such distances are weighted, they are known as weighted KNN classifier. It's importance can be studied in various use cases such as radar outlier rejection performance [11]. KNN can also become biased if the training data is biased. For this reason, it is important to validate our predictions. When the parameters of our model are changed based on testing data, we are over fitting the model based on testing data. This practice will lead to faulty algorithms. Hence the technique of cross validation was introduced which will help in determining the important parameters of an algorithm without using testing data [13].

Dashboards have been used from time to time in order to create interactive visualizations which help in analyzing data better. The process of decision making becomes data-driven when stakeholders can visualize data. This concept can also extend to medical practitioners who are looking at ways to make patient follow up more efficient [15]. The concept of performing analytics in dashboards can also be extended to time series analysis and forecasting. Different models can be used to do this in order to create powerful dashboards capable of visualizing and analyzing time series data [12].

3. Proposed Methodology

The project involves the process of research, conceptual understanding and implementation at every stage of the project path. As a start-up there is no technology that has been invested on greatly in the field of Data science and Machine Learning. Hence as s a data analyst intern it was my responsibility to test various technologies and present the various key parameters which decide its relevance and fit. For if the company has bought a set of services from AWS then it is appropriate to choose an Amazon product as it can be integrated into the existing product environment. Each of the below steps is expanded in the chapters to come.

3.1 Flowchart

Every company today has data in some format. This data is either structured or unstructured in nature. Unstructured data is any data that isn't analysis ready or data that cannot be read easily. The data can be text, JSON, XML etc. It is important to understand the data and the source. This data must be converted to structured data. XML, JSON are response files which are generated when the user interacts with the product interface. The server in turn produces a response file which holds information about the client, what he/she searched

for and transactional details. All this data can be used for the benefit of the company if converted to analysis ready structured data.

Structured data can also be obtained from using DBMS where the data is already structured. For example, all the data in relational database is tabular in nature. Most times data which is related is found in a variety of sources, hence it is important to bring them together to get the best out of our visualization / dashboard / predictive algorithm. When data from varied sources is taken, the risk of data being corrupted also increases. Therefore, data must be cleaned or processed before any form of analysis is carried out.

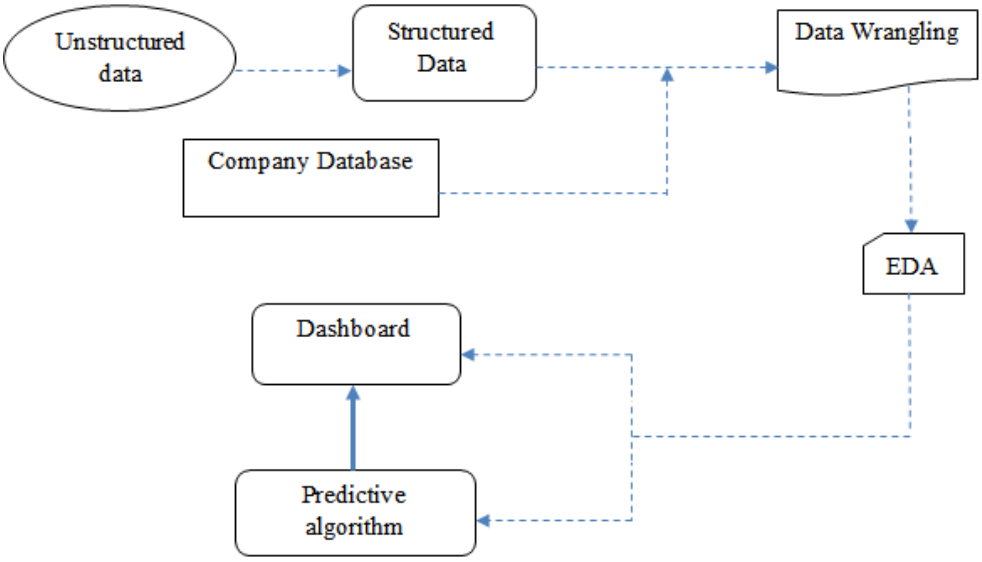

Figure 1. Flow chart for proposed method

Data Wrangling / Data Mining is a process of cleaning, enriching or even restructuring of data (data shaping and data manipulation) such that it is made ready for analytics in general, and to train predictive algorithms. This step is important as it determines the efficiency of how the dashboard functions. If the data fed isn't right, the dashboard won't be able generate meaningful insights into company data. Large amounts of data can be interpreted well when it is cleaned and organized properly.

It is a good practice to understand your data before it is modeled or even used for visualizations in any dashboard. For this purpose, we use EDA (Exploratory Data Analysis). It is a method used to interpret or summarize the main characteristics of a dataset mainly through visual methods as it is easier to understand data that is represented visually. This data can use in two ways. Firstly, this data can be used to create interactive visualizations that can be used to build a dashboard. These dashboards can be used by business executives to make

data driven decisions, increase the revenue generated by integrating the dashboard to their application framework. Secondly, data can be used to train predictive models / forecasting models which can predict the future trend and seasonality based on the size of data that is fed into the model.

The predictive model can now be combined with the dashboard, thus enabling it of predictive analytics. This increases the use case of the product as clients can also view their data visually which won't be just restricted to past transactions but an insight into the future.

Thus, by enabling stakeholders to make data-oriented decisions, integrating cutting edge ML dashboards into their own product and enabling clients to view data (past and future) greatly increases the chance of profitability.

3.2 Dashboard Implementation

3.2.1 Data

Data is simply the collection of facts, numbers, and measurements, calculated values that may or may not contain information that can be worked on. There is a distinct difference between data and information. Data need not be useful; they are in merely in a raw and an unprocessed form. When this data is further processed and filtered to remove discrepancies, organized and presented specifically for a problem statement, it can be termed as information.

3.2.1.1 Unstructured Data

Eighty Five percent of corporate data can be categorized into unstructured data. Such data has neither pre-defined model nor a schema. It may some internal structure but cannot be considered to have the structure required to perform analysis. This data maybe text, sensory values, videos, PDF or images. It can also be classified into human generated unstructured data or machine generated unstructured data or even human generated unstructured data.

3.2.1.2 Semi-Structured Data

Semi-structured data enables information grouping and hierarchies through internal tags and markings that can identify separate data elements. This data does not conform to a data model rather it has some structure. This data cannot be stored in rows / columns. Similar entries in the data are grouped together. Though Semi-Structured data is estimated to be around 10-15% of company data it has critical use cases in decision making. Since there is no

separation of schema and data, interpreting the relationship between data becomes quite difficult.

3.2.1.2 (A) Extensible Markup Language

XML cannot be considered human readable as it requires time and effort to go through even an average XML file. It can be considered as a semi-structured document language, encoding rules which are both machine and human readable. XML files hold value as their tag driven data structure is incredibly flexible and programmers adapt it to store and transport on the web. XML generally comprises of sender information, receiver information with a header and a body. XML is essentially a lot of information about a transaction wrapped in tags.

3.2.1.2 (B) Java Script Object Notation

Though XML was the only available choice for data interchange, there has been a lot of transformation in the space of data sharing. There are many reasons why JSON has become a popular alternative. Some of the reasons are

Less words: XML uses far more words than what is necessary.

Speed: XML software is difficult to parse as its slow. The cost of parsing XML files can use large amounts of memory.

Code: The XML format is least intuitive thereby making it hard to represent it in code format.

Structure: JSON data structure is a map which makes it easier to interpret whereas XML data structure is a tree thus making it difficult to interpret. Also, the structure of JSON data is intuitive , enabling it to be mapped directly into domain objects using any programming language.

Though JSON seems to be a better choice it is essential to learn XML and other related technologies to do more processes apart from data exchange and fast processing.

3.2.1.3 Structured Data

Data that follows a well-defined structure or a data model which can be easily accessed is known as structured data. This structured data is generally stored in databases

which have a well-defined schema. Structured data accounts for 20-40% of company data. These databases are handled by RDBMS (relational Database Management Systems). SQL is used to query required information or add new information. Some important advantages of using structured data are:

Storage: When structured data is utilized, it can easily undertake Business Intelligence (BI) operation such as Data Warehousing.

Scalability: Another advantage of using structured data is that they are easily scalable if there is an increase in data.

CRUD operations: Operations which include deleting, updating, creating can be easily carried out due to its well-defined structure. Since there is a structure to the data stored, they can be indexed in order to make operations hassle-free.

Data Mining: The process of extracting useful information from chunks of data for a specific task can be carried out easily when structured data is used.

Security: Ensuring security for the data stored becomes relatively easily when compared to unstructured data.

Analytics: Analysis of data can easily be performed when the data being evaluated has some structure to it. It helps in avoiding errors and most importantly enhancing the chances of finding meaningful insights

Data Manipulation and Data Shaping: Both these processes are important while cleaning the dataset which is to be used for analysis or training predictive algorithms. Structured data helps the programmer perform these processes efficiently without errors.

3.3 Data Transformation

The merits of using structured data are a clear indication that we need to transform unstructured / semi – structured data to structured data. The data dealt with was XML. Hence the process includes a conversion of XML to JSON first. This is because JSON is relatively easier to deal with and store using NoSQL DBMS such as Mongo DB. Once XML has been converted into JSON, the data must be stored in its entirety using the DBMS known as Mongo DB. This enables the source data to be untouched while the analyst can retrieve data

that is needed. Now, after the data is stored, the required data is collected using querying statements.

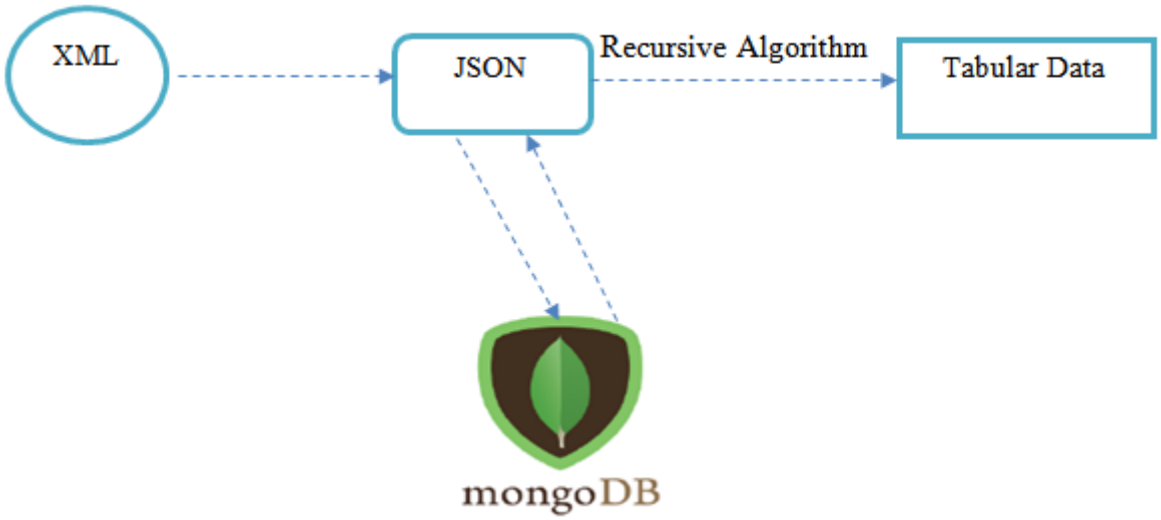

Figure 2. Data transformation

3.3.1 XML to JSON

All tasks performed programmatically is done using python in this project. Python has enormous number of libraries which can be used for a variety purposes starting from converting XML to JSON up to deep learning. The python community helps in keeping the programming language updated to meet necessities. Python holds a library called **xmltodict** which can convert xml files into dictionary like format. A dictionary is a data structure which holds elements in key- value pairs that enables accessing information easily. JSON is very similar to this data structure in Python; hence it becomes easy to convert this dictionary like format to JSON.

This is done using the built-in library available in python called **json.** The dictionary like object can be JSON-dumped. The process is now automated so that all XML files in a location can be converted into JSON files. Before converting JSON files to tabular data we need to store it.

3.3.2 Data Storage using Mongo DB

Most data analysis tasks require only a chunk of data to be retrieved from the complete source. Hence converting all available JSON data to tabular form is not optimal. Hence, we store the data in a database before it is converted. This helps in securing source

data. Since the data here is semi-structured, conventional RDBMS cannot be used such as SQL Server, PostgreSQL or MySQL. This requires a special DBMS which can store and perform CRUD operations on available data. Such a DBMS which works on no-relational data is called NoSQL. They can store data which has no schema or format. The MongoDB ecosystem is varied and vast which contains BI connectors, Atlas, MongoDB Charts, MongoDB mobile and even a GUI which enables the user to work effortlessly with the NoSQL DBMS.

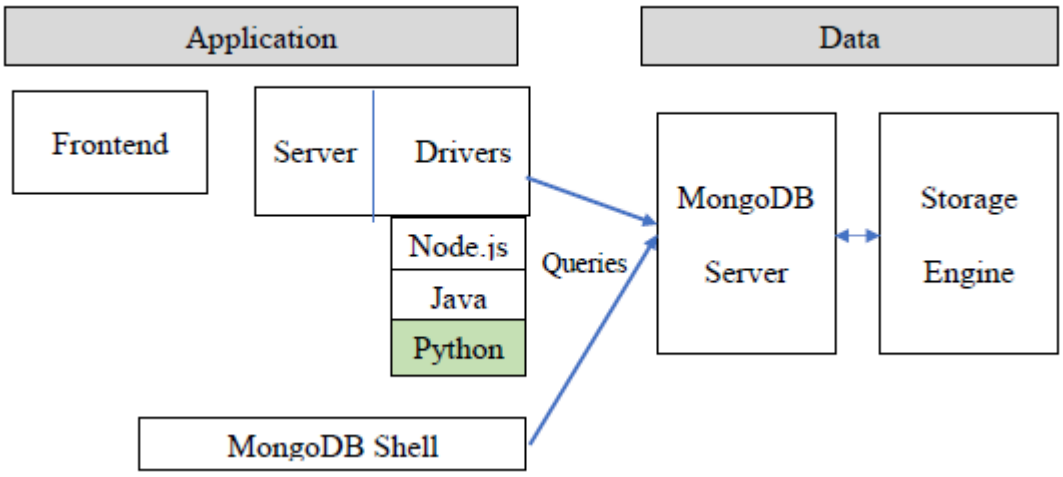

Figure 3. MongoDB ecosystem

Mongo DB has a good number of drivers such as Pymongo available which act as connectors. This helps perform all operations that can be done using Mongo DB shell. All CRUD operations that are possible using the shell is capable of being performed using Python. Mongo DB stores all the JSON data as BSON data. This helps MongoDB increase efficiency in performing operations on the data stored.

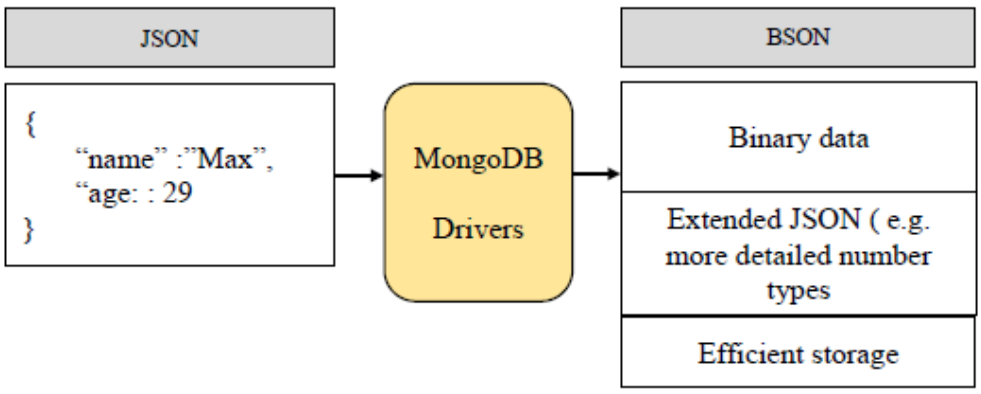

Figure 4. JSON to BSON conversion

3.3.3 Recursive Algorithm

Since the JSON files produced by the server are complex, a recursive algorithm is used to convert all information present in the file into tabular format. Recursive algorithm used to work based on PMI. Principal of Mathematical induction states that if the base case; the case which says that the process of recursion must be terminated, induction hypothesis and the induction step are taken care of, the process will be completed. Though what goes behind recursion is hard to grasp, it is easier to use PMI to get the task completed and then introspect the path recursion has taken to complete the task

3.4 Data Wrangling

In every sector governing the economics of the world, data is playing an important role. But most of this data is useless in its raw format. Seventy percent of an analyst's / data scientist's time is invested in unifying and cleaning data which make them analysis ready. In a formal sense, data wrangling is the process of mapping/cleaning raw data such that it becomes ready for advanced tasks such as data science and machine learning.

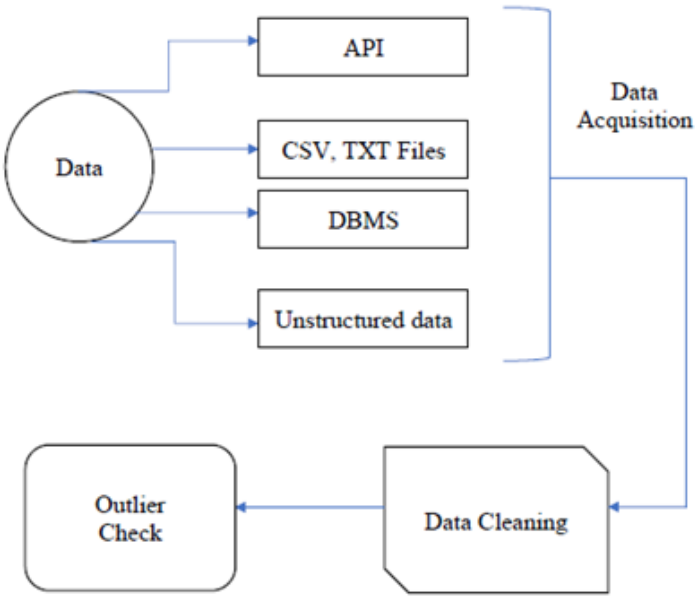

Figure 5. Data wrangling

3.5 EDA

Once the data has been cleaned, it can be used to train predictive models or create interactive dashboards. But before any of these tasks are done, we need to understand the data better. This enables an analyst to make informed choices during implementation of dashboard

visualizations or feature selection in machine learning. EDA (Exploratory Data Analysis) is a technique used to study the data and summarize its characteristics. This is generally done in visual form so that the data can be studied intuitively. There are many functions available in python which help understand the dataset. The describe () and info () functions are useful in providing summary statistics such as mean, median, mode, first quartile, second quartile etc.

3.6 Research on BI Tools

When companies accumulate large amounts of data over a period, Business Intelligence tools are used to make sense of this data. These tools are empowered with the necessary functionalities to convert raw data into actionable information that can guide the process of decision making.

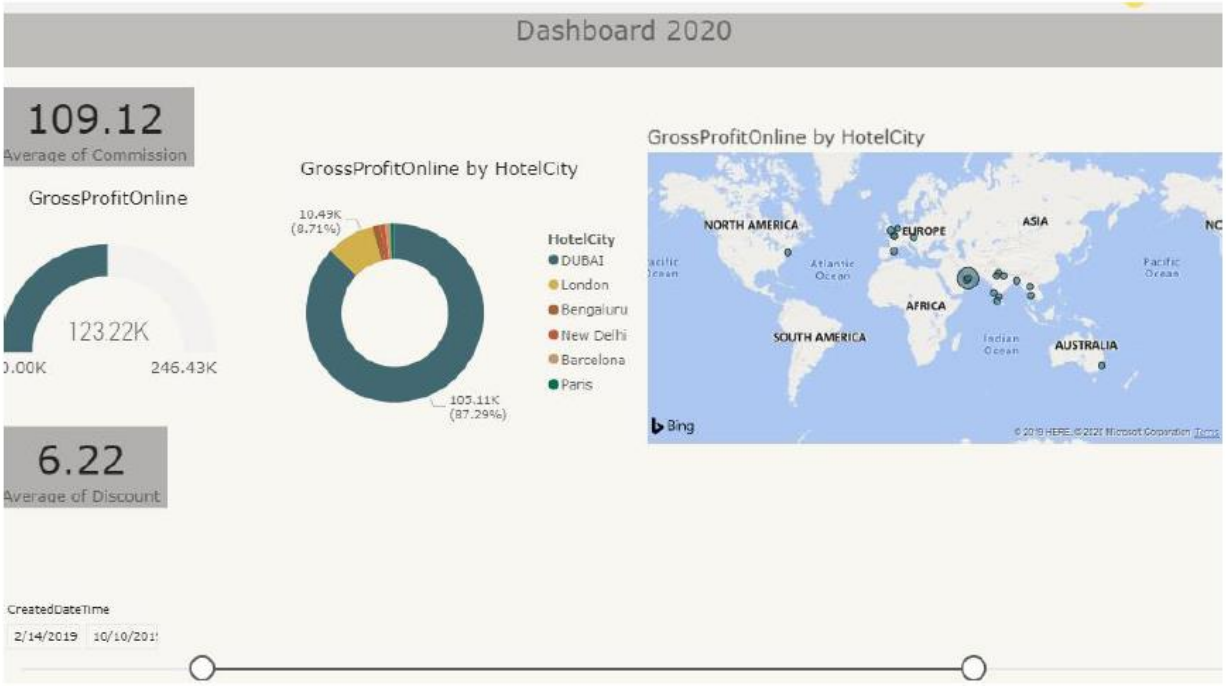

Figure 6. Power BI Dashboard

3.6.1 Tableau

Tableau is data visualization software that makes creating visualizations intuitive. The user can drag and drop the required chart/plot which is needed for visual analysis of data. There are many versions available for the software which is also priced differently. Tableau Desktop Professional edition was used for this project (Students with valid id proof can avail one-year free subscription).

3.6.2 Power BI

Power BI is a combination of connectors, services and applications which work in sync to provide visually interactive insights. Power BI is a Microsoft product and hence offers numerous services which complete the arsenal of any data analyst. Power BI allows you to connect to a number of data sources. The dashboard shown below was developed using the data that was transformed using the recursive algorithm.

3.6.3 Dash - Plotly

Dash is a productive python web application framework, written on top of flask, React.js, Plotly.js. Plotly is both the name of the company that provides services such as Dash, Chart Studio and has a graphing library called Plotly which is used to create interactive plots. Plots which are dynamic in nature and change with respect to user input are called interactive plots. Dash applications are rendered in the browser. These applications can also be deployed to servers using azure or Heroku and shared through URLs.

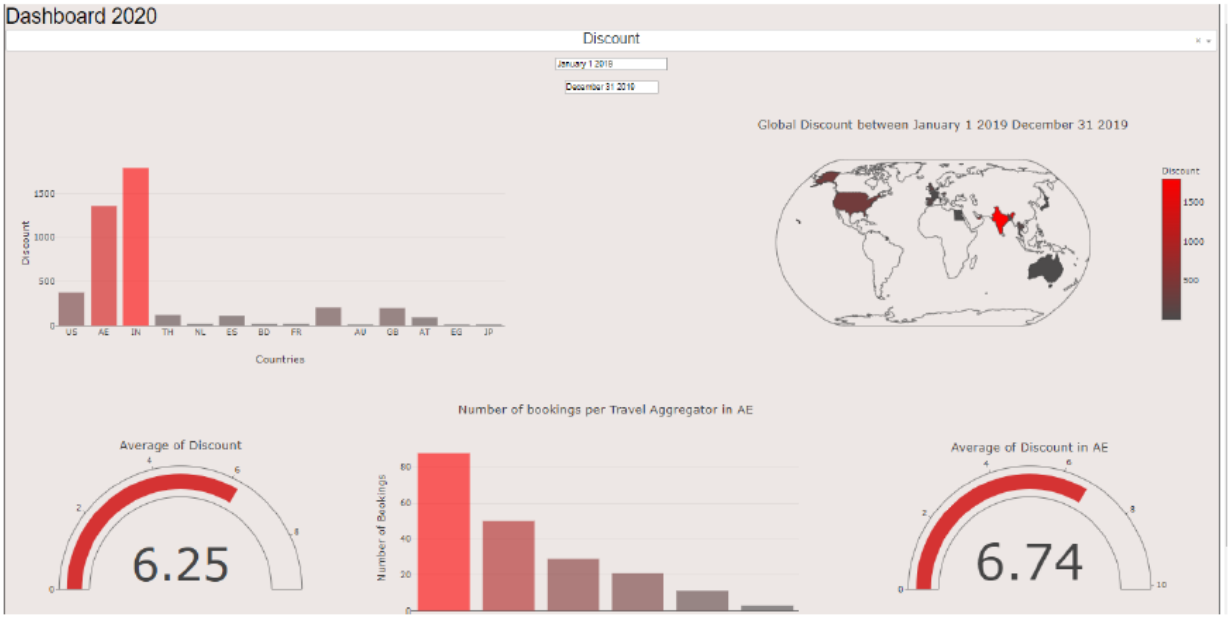

Figure 7. Dash-Plotly Dashboard

3.7 Predictive Analytics

Predictive Analytics comprises of predictive modeling, statistical techniques, and data mining which is used to forecast future trends. There are plenty of models available such as Linear Regression, Logistic Regression, Decision Trees, Random forests and time series

forecasting algorithms such as ARIMA, Holt-Winters method which are capable of prediction. As a data analyst intern, I was asked to explore all possible models, their limitations and capabilities as each model caters to a specific kind of data. As the company deals with time series data (transactional data), data that is dependent on time, it was decided that time series forecasting models will be used to predict future data. All specified models were explored, evaluated with company data before finalizing time series forecasting as method to predict future trends.

3.7.1 Machine Learning

Machine Learning can be defined as the methodology to make machines learn by providing them with past data to predict future data. Consider a task to be T, performance metric to be P and experience or past data to be E. Tom Mitchell said that a computer program is said to learn from experience E with respect to some task T on the basis of a performance metric P, if its performance P improves with experience E.

Artificial Intelligence can be divided broadly into two types based on the boundaries we set for the machines we train. When the machine is given a set of boundaries by the user to work on, it is known as a Rule-based system. Here the decision making of the machine is limited to the boundary set by the user. When a machine is given complete freedom, i.e. when the user sets no boundaries, the machine is called as self-aware or intelligent. Machine Learning can be divided based on the how the output is predicted. They are classified into two major types.

The algorithm is trained by providing it with past data. When this input has a clearly labeled output or target variable, we call it Supervised Learning. This essentially means that the user knows what he/she wants to predict. A classic example for this is a cancer dataset where the output class is clearly specified- malignant or benign. When the input has no clear output class specified, it is known as unsupervised learning. For example, if all users of Face book or Instagram were to be classified into ten groups based on popularity.

Supervised Learning can be further classified into two types based on the type of data that is being predicted. When the output data is continuous in nature, the algorithm developed is called a Regression Algorithm. When the output data is categorical in nature, the algorithm is known as a Classification algorithm or Classifier. Continuous values are values that have an infinite number of possibilities. The number of values between two numbers such as

hundred and hundred and one is infinite. On the contrary, categorical values are finite. An example for this is profit and loss. There is a third class of machine learning algorithms called reinforced learning. In this type of machine learning the model is trained based on a system of reward and punishment. Each algorithm has its own features and is used based on how we want to predict the output. This is also based on the type data used to train the algorithm. During the internship, many algorithms were tested and trained and analyzed using company data.

4. Results and Discussion

First, an interactive dashboard capable of multiple call-backs where changes in the input or selecting a point in the chart change the entire dashboard was created. This was done after extensive research of various open source and paid data visualization tools such as Dash, Power BI, and Tableau. Dash was selected to create the dashboard as it an open source tool and has similar functionalities as paid software. Secondly, a forecasting algorithm was built after developing a strong foundation on methods to study time series data and the models used to forecast into the future.

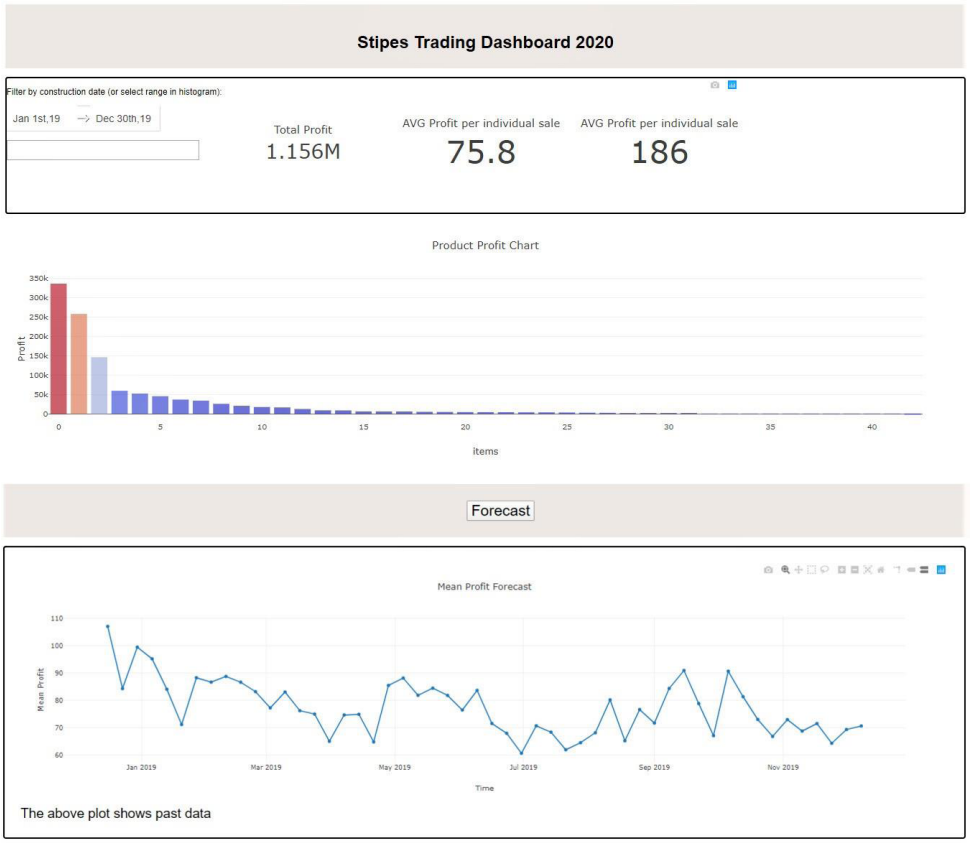

Figure 8. Interactive Dashboard

Finally, the forecasting algorithm is integrated into the dashboard thereby making it capable of predictive analytics. The dashboard consists of two parts. The first half of the dashboard takes care of analytics and the second half consists of the forecasting option. The analytics window consists of a bar chart and few metrics such as total profit and average profit. The dashboard consists of a date range picker and a text box at the top-left corner where client id can be added. The objective of the dashboard is to perform client-specific predictive analytics.

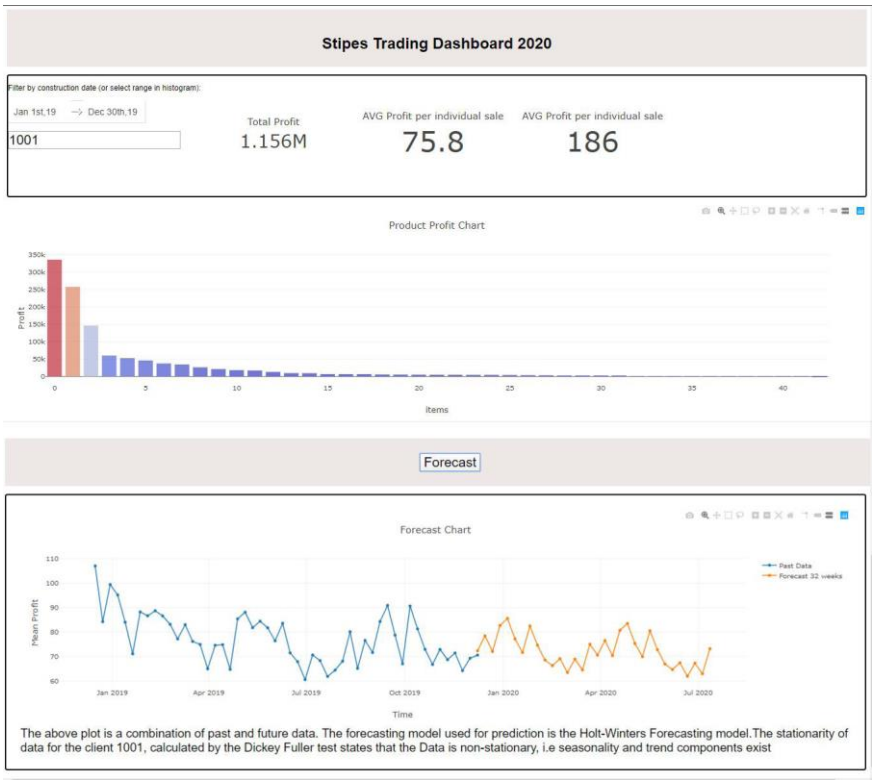

Figure 9. Interactive Dashboard with predictive analytics

The first step is to select a time range. The data is filtered based on the time range provided and a bar chart is developed. Here each bar represents total profit generated by one commodity over that time frame in descending order. Also, total profit generated from all commodities is calculated. Two average profit values are calculated. The first average profit value is calculated for all commodities and the second average value is calculated for the commodity the user chooses. As each bar represents one commodity, the user only needs to click the bar of his/her choice and the average profit value will be created for that commodity. If the time frame were changed, all the values including the bar chart will also change.

The forecast window consists of a forecast button and a line chart with markers. The data used for the line chart is re-sampled where each data point represents weekly average profit. The chart also has a description which changes when a forecast is made. In Time series forecasting, the testing data plays an important role. If the testing data involves sixteen weeks of data, then it is safe to predict sixteen weeks into the future. A forecasting model can predict even ten years into the future, but the accuracy will be low. The next step is to add client id. As the company works with various clients, the dashboard can be used to understand how the performance has been for a client in the stipulated period selected in the dashboard. Once client id is entered, the forecast button is clicked. The predictive model forecasts four months into the future based on the clients data of weekly profit. The blue line indicates past data whereas the orange line indicates future data forecasted by the Holt-Winters model. A description of the nature of data is also provided in the form of text. The description includes knowledge on whether the data is stationary or not, based on the dickey fuller test. As dash uses the Plotly graphing library to create interactive plots, the graphs created can be zoomed in and individual values can also be seen by hovering over the graph. If the client id is removed from and forecast button is clicked, the line chart will revert to the original graph which consists of only past data.

4. Conclusion and Future Enhancement

The dashboard created will be instrumental in providing clients as well as stakeholders the needed support to make data-driven decisions. As the years pass the amount of data generated by the company shall also increase at an exponential rate. Hence Big data technologies must be explored. Big data is high on veracity, volume and velocity. The Hadoop Ecosystem, Spark and other big data technologies will enable users to systematically process large and complex data. Customer behavior analytics, fraud detection and social media analytics are few examples of big data applications which can open new realms of revenue generation for a company.

References

[1] Ariyo, A. O. Adewumi and C. K. Ayo (2014) ‘Stock Price Prediction Using the ARIMA Model’, UKSim-AMSS 16th International Conference on Computer Modelling and Simulation, Cambridge, pp. 106-112, doi: 10.1109/UKSim.2014.67.

[2] Singh, N. Thakur and A. Sharma (2016) ‘A review of supervised machine learning algorithms,’ 3rd International Conference on Computing for Sustainable Global Development (INDIACom), New Delhi, pp. 1310-1315.

[3] V. Patil and R. S. Bichkar, (2006) ‘A Hybrid Evolutionary Approach To Construct Optimal Decision Trees With Large Data Sets,’ IEEE International Conference on Industrial Technology, Mumbai, pp. 429-433, doi: 10.1109/ICIT.2006.372250.

[4] F. F. Lubis, Y. Rosmansyah and S. H. Supangkat, (2014) ‘Gradient descent and normal equations on cost function minimization for online predictive using linear regression with multiple variables,’ International Conference on ICT For Smart Society (ICISS), Bandung, pp. 202-205, doi: 10.1109/ICTSS.2014.7013173.

[5] F. Harrou, M. Nounou and H. Nounou, (2013) ‘A statistical fault detection strategy using PCA based EWMA control schemes,’ 9th Asian Control Conference (ASCC), Istanbul, pp. 1-4, doi: 10.1109/ASCC.2013.6606311.

[6] G. Kesavaraj and S. Sukumaran, (2013) ‘A study on classification techniques in data mining,’ Fourth International Conference on Computing, Communications and Networking Technologies (ICCCNT), Tiruchengode, pp. 1-7. DOI: 10.1109/ICCCNT.2013.6726842

[7] H. Lan and Y. Pan, (2019) ‘A Crowdsourcing Quality Prediction Model Based on Random Forests,’ IEEE/ACIS 18th International Conference on Computer and Information Science (ICIS), Beijing, China, pp. 315-319, doi: 10.1109/ICIS46139.2019.8940306.

[8] H. Xie and F. Shang, (2014) ‘The study of methods for post-pruning decision trees based on comprehensive evaluation standard,’ 11th International Conference on Fuzzy Systems and Knowledge Discovery (FSKD), Xiamen, pp. 903-908, doi: 10.1109/FSKD.2014.6980959.

[9] H. Yi, Q. Xiong, Q. Zou, R. Xu, K. Wang and M. Gao, (2019) ‘A Novel Random Forest and its Application on Classification of Air Quality,’ 8th International Congress on Advanced Applied Informatics (IIAI-AAI), Toyama, Japan, pp. 35-38, doi: 10.1109/IIAI-AAI.2019.00018.

[10] Jing Chai, Hongwei Liu and Zheng Bao,(2009) ‘A W-KNN classifier to improve radar outlier rejection performance,’ IET International Radar Conference, Guilin, pp. 1-4, doi: 10.1049/cp.2009.0106.

[11] Kumar, P. Satheesh, P. Chitra, and S. Sneha. "Design of Improved Quadruple-Mode Bandpass Filter Using Cavity Resonator for 5G Mid-Band Applications." Future Trends in 5G and 6G: Challenges, Architecture, and Applications (2021): 219

[12] P. Guleria and M. Sood, (2018) 'Predictive Data Modelling: Educational Data Classification and Comparative Analysis of Classifiers Using Python,' Fifth International Conference on Parallel, Distributed and Grid Computing (PDGC), Solan Himachal Pradesh, India, pp. 740-746. DOI: 10.1109/PDGC.2018.87457.

[13] R. De Croon, J. Klerkx and E. Duval, (2015) 'Interactive proof-of-concept dashboard to explore patient follow-up in general practice,' 9th International Conference on Pervasive Computing Technologies for Healthcare (PervasiveHealth), Istanbul, pp. 233-236, doi: 10.4108/icst.pervasivehealth.2015.258991.

[14] Satheesh Kumar P., Jeevitha, Manikandan (2021) Diagnosing COVID-19 Virus in the Cardiovascular System Using ANN. In: Oliva D., Hassan S.A., Mohamed A. (eds) Artificial Intelligence for COVID-19. Studies in Systems, Decision and Control, vol 358. Springer, Cham. https://doi.org/10.1007/978-3-030-69744-0_5.

[15] W. Wang and Y. Guo,(2009) 'Air Pollution PM2.5 Data Analysis in Los Angeles Long Beach with Seasonal ARIMA Model,' International Conference on Energy and Environment Technology, Guilin, Guangxi, pp. 7-10, doi: 10.1109/ICEET.2009.468.

[16] W. Wang and Z. Niu, (2009) 'Time Series Analysis of NASDAQ Composite Based on Seasonal ARIMA Model,' International Conference on Management and Service Science, Wuhan, pp. 1-4, doi: 10.1109/ICMSS.2009.5300866.

[17] Yan-ming Yang, Hui Yu and Zhi Sun, (2017) 'Aircraft failure rate forecasting method based on Holt-Winters seasonal model,' IEEE 2nd International Conference on Cloud Computing and Big Data Analysis (ICCCBDA), Chengdu, pp. 520-524, doi: 10.1109/ICCCBDA.2017.7951969.

Author's biography

SatheeshKumar Palanisamy was born in Krishnagiri, Tamilnadu, India. He completed his B. E in Electronics and Communication Engineering and M.E-(Communication Systems) from Anna University, Chennai, Tamil Nadu in the year 2009 and 2012 respectively. He has been in the teaching profession for the past 9 years. Presently, he is working towards his doctorate degree from Coimbatore Institute of Technology (India). His areas of interest include RF systems, Antenna Design and Electromagnetics. He is a professional member in

Institution of Telecommunication Engineers(IETE), International Association of Computer Science and Information Technology (IACSIT) and also a member in International Association of Engineers (IAENG). Currently, holding the post of Executive Council Members of IETE Coimbatore Centre. He has published 11 papers in International Journals, 4 papers in International conferences and 4 papers in National Conferences in the area of signal processing and Communication systems.