JETIR.ORG ISSN: 2349-5162 | ESTD Year : 2014 | Monthly Issue
JOURNAL OF EMERGING TECHNOLOGIES AND INNOVATIVE RESEARCH (JETIR)
An International Scholarly Open Access, Peer-reviewed, Refereed Journal

ASSEMBLING A MODEL FOR MARKET ANALYSIS AND SALES PREDICTION

[1]Pinnamaraju.T.S.Priya,[2]Toleti.Yamuna Satya Bhadravathi

[1]Assistant Professor,[2]MCA 2nd year,
[1]Master Of Computer Applications,
[1]Sanketika Vidya Parishad Engineering College, Visakhapatnam, India

ABSTRACT
These days, Big Marts and shopping centres gather sales information for each individual item in order to estimate future customer demand and modify inventory control. These data stores in a warehouse contain a sizable number of customer records and specific item information. With data mining, anomalies and recurring patterns are also found in data warehouses' data storage. The generated data can be used by companies like Big Mart to forecast future sales volume using a variety of machine learning techniques.In this study, we proposed the use of linear regression and random forest analogies to provide an effective analysis and prediction of big-mart data. For data virtualization, we employ the most recent machine learning techniques, Dtale and Pandas profiling. Last but not least, hyper parameter tweaking is used to help you choose important factors that will make analogarthim shine and produce the greatest outcomes. We also employ a web interface method to quickly access and forecast consumer product sales.There is no sales. And to get a better predictive, hyper-tuning strategy to model performance, we employ the ensemble technique.Future sales forecasting is a crucial component of any organisation. Effective forecasting of future sales enables businesses to create and enhance company strategies and get relevant market knowledge. Only because of the rapid expansion of international malls and online shopping is the competition between various malls and large supermarkets becoming more serious and fierce every day. Every shopping centre or market tries to give customised limited-time deals to draw in more clients based on the day so that the volume of sales for each item can be forecasted for inventory management of the organisation, logistics, and transport services, etc.

Keywords: Big Marts,shopping centers,machine learning,Dtale,linear regression,Pandas profiling, hyper parameter tweaking,ensemble technique, sales forecasting, inventory control, customer demand, market knowledge, customised deals,logistics.

I.INTRODUCTION

Large shopping centres like malls and marts keep track on sales of goods, ..as well as their numerous dependent and independent characteristics.[1]2.as an essential step in predicting future demand and inventory management. in today's modern world[2] Several dependent and independent factors were used to construct the dataset[3]Is a synthesis of item attributes, customer-gathered data, and inventory management data stored in a data warehouse[4] The data is then further refined to produce precise forecasts and collect fresh, intriguing findings that add to our understanding of the task's data. Through the use of machine learning techniques like random forests, this can then be utilised to estimate future sales.[18]Every mall or store tries to give unique, limited-time deals to draw in more people based on the day[5]so that the amount of sales for each item can be forecasted for inventory management of [19]the company, logistics, and transport service, etc.[20] We are dealing with the issue of This project involves predicting big mart sales of a particular item based on anticipated customer demand across several big mart stores in diverse regions and products.[7]. A more accurate forecast is usually helpful in creating and upgrading [8]the company strategies related to the industry, [21]as well as increasing market comprehension. A traditional sales prediction study can help by carefully examining previous occurrences or circumstances[9] and then drawing conclusions about customer acquisition, financial inadequacies, and strengths.[22]Grow since the information is so valuable in the present[10]. Both supervised and unsupervised types of tasks are dealt with in machine learning.[11] and a categorization type problem often counts as a resource for knowledge discovery.[12] Regression is used to produce resources and make accurate predictions regardingIn the future.[13] the main focus will be on creating systems self-efficient[14] so they can perform calculations and analyses on their own and produce answers [15]that are considerably more exact and precise[23] Data can be transformed into knowledge[16]applying statistical and probabilistic algorithms.Sampling distributions are utilised in statistical inference as a theoretical[17].In this project, we suggested the methods of linear regression and random forest, which offer an effective Big Mart Analysis and Prediction. For data virtualization, we employ the most recent machine learning techniques, like Dtale and pandas profiling. Last but not least, hyperparameter tuning is used to help you make Hyperparameters that are pertinent and enable the algorithm to perform well and deliver the best outcomes. We also employ a web interface method to quickly access and forecast consumer productThere is no web development technology in the current system, and only developers have access to and control over product sales. Additionally, we employ ensemble approaches to create a more accurate predictive, hyper-tuning model.

II.EXISTING SYSTEM

A sales prediction team employed to research and evaluate historical data and reports in an effort to predict future sales for a business or organisation using complex mathematical calculations and algorithms. This manual approach might not be able to produce the required results. precise accuracy, and it frequently fails in the first few quarters. Higher management must wait until the end of the quarter's sales before they can determine whether or not the report produced by the team responsible for forecasting sales is accurate. It is also a very efficient method.

III.PROPOSED SYSTEM

We have suggested one machine learning-based sales prediction solution to get over the limitations of manual sales predicting team outcomes. This technique might be able to anticipate sales for a business or organisation with greater accuracy. This system is capable of producing comprehensive sales reports.with the demand for the requisite resources and manpower. This means that one may simply determine how much people and how much resources are needed to meet the sales target by using this sales forecasting system machine learning project. Sales forecasting is based on data processing and analytics, with machine learning algorithms later being applied to the processed data to produce precise results. We also provide a link to a website that can assist with obtaining machine learning algorithm are applied on analyzed data.

IV.REQUIREMENTS ANALYSIS

a)FUNCTIONAL REQUIREMENTS

An algorithm should be able to forecast a future sale of a product.
The dataset is first downloaded from the Kaggle website.
Preprocessing is required to check the raw data for missing values and to encode the categorical values.
The data must be divided into two categories, with 80% of the data trained and the remaining 20% tested.
Using a light gradient boosting machine, we must train the data.
The next step is to build a website utilising the Flask API. When we enter characteristics into the website, Flask loads the machine learning model and returns the desired outcome.

b)NON-FUNCTIONAL REQUIREMENTS

i.Availability: Any systems that have Python installed can access the sales forecasting programme.

ii.Performance: The throughput and prediction of the system's performance are calculated.model. Accuracy is used to measure the performance of the model.

iii.Cost: Building and maintaining a property are fairly inexpensive. The method for cost optimization uses all necessary open sources.

iv.Reliability: The virtual procedure makes it more dependable and effective. High level, very dependable languages are used for the model building.

v.Scalability: Most of the recent technological innovation has been driven by the requirement for scalability. The sector has created new software languages, design methodologies, and communication and data transfer techniques.

V.SYSTEM ARCHITECTURE

Our project's initial stage is to determine the features, the values utilised in each feature, and what each value in the supplied attribute means. The data is visualised using the characteristic that determines the status of the solution in the following stage. Using preprocessing techniques, we must locate the non-available variables in the second stage and eliminate any that could destroy our model. Third is where stage, we carry out feature extraction by numerically transforming the data. We separate the dataset into train and test data after applying normalisation to the data. We build a model using machine learning methods, and that model is then trained using the train data. The model is later evaluated by implementing the model on test data.

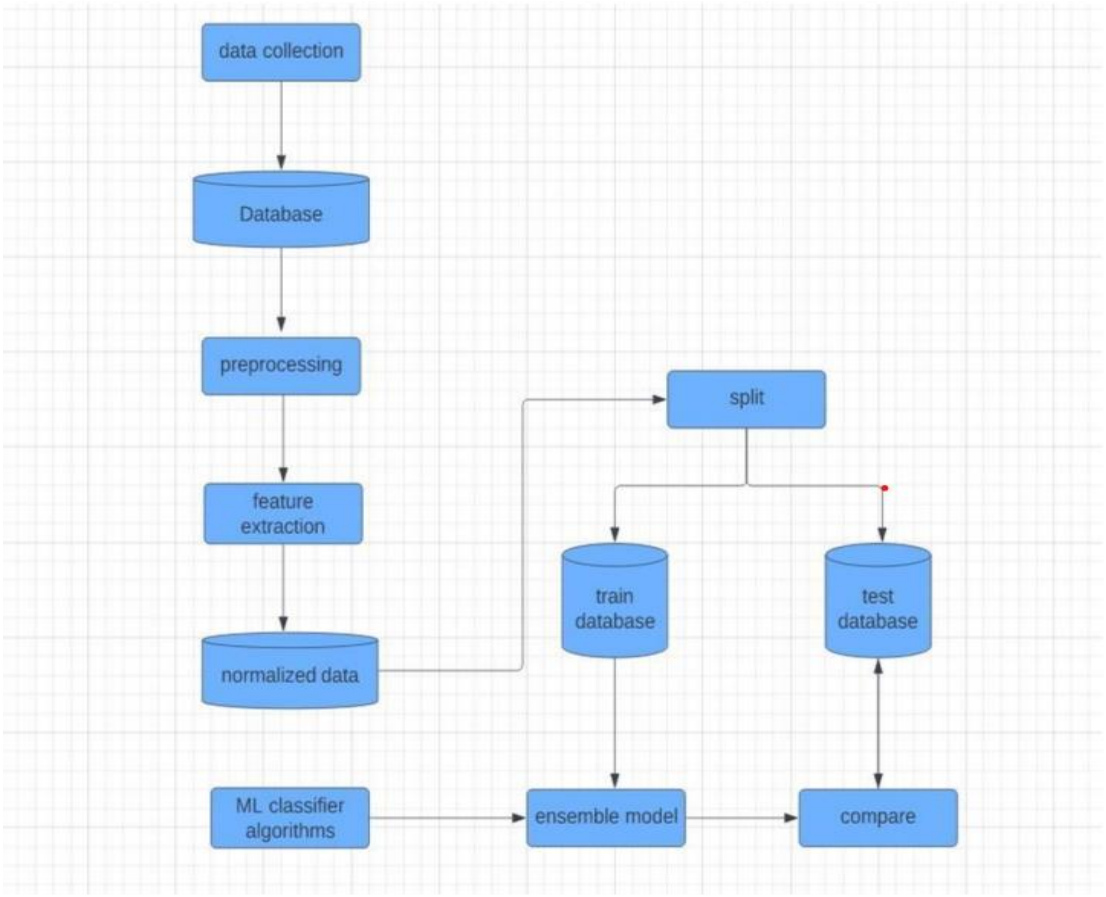

Figure 1: Overall System Design For market analysis and sales prediction.

VI.UML DESIGN

Unified Modeling Language is known as UML. UML diagrams have been created using the SRS document of analysis as the input for the design phase. One component of the software development methodology is the UML. Although the UML is process neutral, it is best employed in a process that should be driven, architecture-centered, gradual, and iterative. The UML is a language for visualising, specifying, building, and documenting the components of a system that is heavily dependent on software. A language's vocabulary and rules that are centred on the conceptual and actual representations of the system are known as modelling languages. So, a modelling language like the UML serves as a common language for software design.A graphical language called the UML contains all interesting systems. Moreover, there are several architectures.

Interaction Diagram

i.Sequence Diagram

State chart Diagram

Activity Diagram

VII.USE CASE DIAGRAM

A use is an outside view of the system that depicts a possible action the user might take to finish a task. Use cases and actors are the two fundamental elements of a use case diagram.A use case is a lengthy explanation of a whole process, usually with numerous steps or transitions. Use cases are hypothetical situations used to comprehend system needs.An actor is a user who assumes a position within the system. To identify the right use cases, the actor is crucial. Several use cases can be carried out by a single actor. An actor could be an external system that requires data from the existing system. The connections between actors and use cases are depicted in a use case diagram. A usage, in its simplest form user perspective.

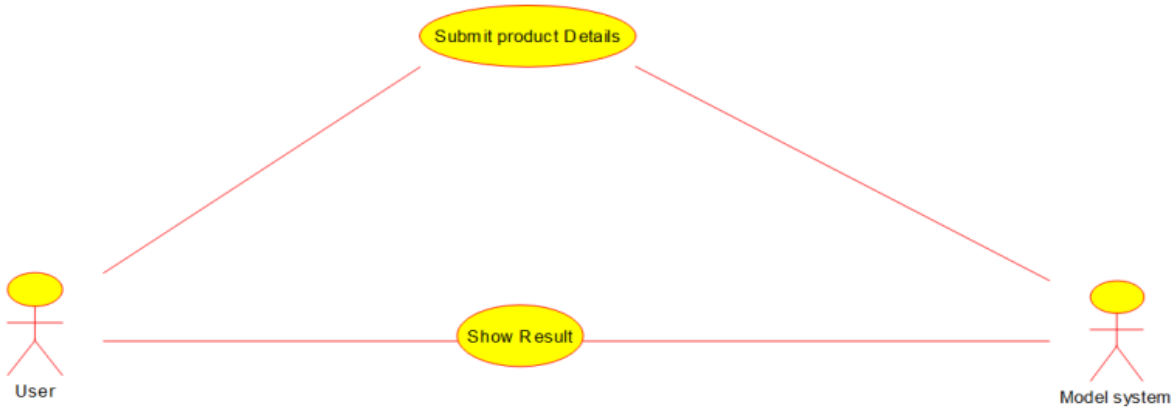

Figure 2: Use Case Diagram for Market Analysis and sales prediction.

VIII.CLASS DIAGRAM

Class diagram gives a static view of system and describes the responsibilities of the system

Figure 3: Class Diagram for Market analysis and sales Prediction.

a)

INTERACTION DIAGRAM

The Sequence and Collaboration diagrams in UML provide as representations for the Interactive diagram. The Interaction Diagram is used to represent the system's interacting behaviour.

b)SEQUENCE DIAGRAM

The sequence of interactions between AM
the system's objects is shown in the sequence diagram.

c)STATE CHART DIAGRAM

The State Chart Diagram displays the system's components and control flow at various points in time as it is being used.The transfer of control from one state to another is depicted in a state chart graphic. States are described as a situation in which an object is present and changes in response to an event. Modeling an object's lifetime from conception to termination is the main goal of a state chart diagram. the State Chart Diagram shows the flow of control and parts of system at different instances of time while utilizing the system. An actor is a user playing role with respect to the system. The actor is the key to find the correct use cases. A single actor may perform many use cases. An actor can be external system that needs some information from the current system. A use case diagram displays the relationships among actors and use cases. In its simplest form, a use case can be described as a specific way of using the system from a user perspective. the UML is thus a standard language for software blueprints. The UML is a graphical language, which consists of all interesting systems. There are also different structures that can transcend what can be represented in a programming languageby taking the dataset's different subsamples. min_samples_split is taken as the minimum number when splitting an internal node if integer number of minimum samples are considered. A split's quality is measured using mse (mean squared error), which can also be termed as feature selection criterion. This also means reduction in variance mae (mean absolute error), which is another criterion for feature selection. Maximum tree depth, measured in integer terms, if equals one, then all leaves are pure or pruning for better model fitting is done for all leaves less than min_samples_split samples.

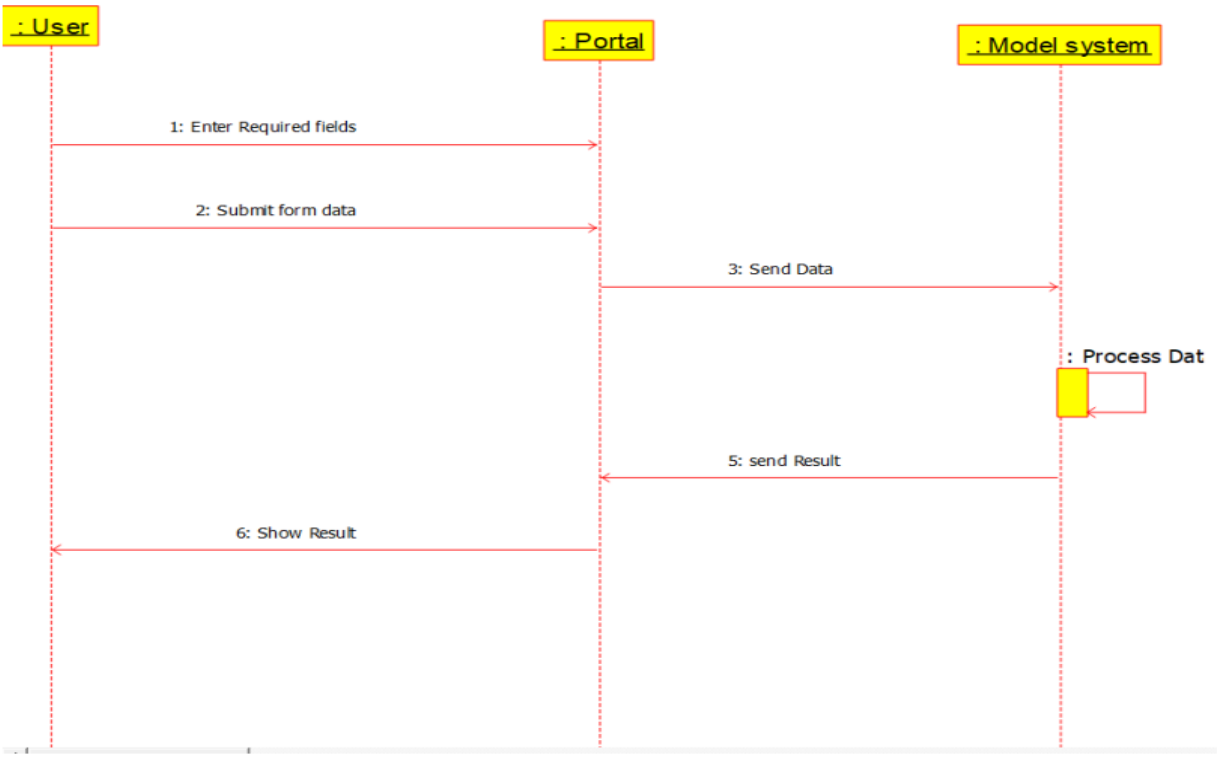

Figure 4: Sequence Diagram for Market Analysis and sales Prediction.

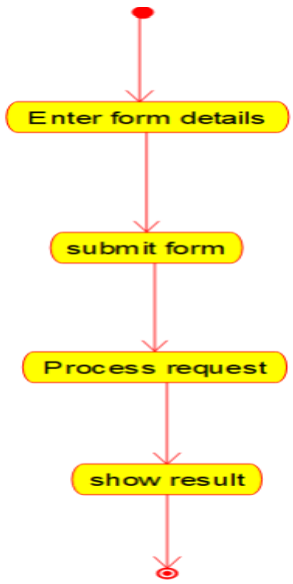

Figure 5: State Chart Diagram for Market Analysis and sales Prediction.

IX.METHODOLOGY

We are use the Kaggle dataset known as Predict Future Sales dataset, which has 1559 products spread across 10 retailers across various cities. Additionally defined are the characteristics of each product and retailer. The goal is to create a predictive model that can foretell the sales of each item at a specific retailer.The train data set has both input and output variables, and it consists of 2 million records and six characteristics (s). For the test data set, we must forecast the sales.With the help of this model, we will attempt to comprehend the qualities of the merchandise and the retail locations that are crucial to boosting sales.We must confirm whether the dataset might contain missing values.

```
train.head(10)
```

	date	date_block_num	shop_id	item_id	item_price	item_cnt_day
0	02.01.2013	0	59	22154	999.00	1.0
1	03.01.2013	0	25	2552	899.00	1.0
2	05.01.2013	0	25	2552	899.00	-1.0
3	06.01.2013	0	25	2554	1709.05	1.0
4	15.01.2013	0	25	2555	1099.00	1.0
5	10.01.2013	0	25	2564	349.00	1.0
6	02.01.2013	0	25	2565	549.00	1.0
7	04.01.2013	0	25	2572	239.00	1.0
8	11.01.2013	0	25	2572	299.00	1.0
9	03.01.2013	0	25	2573	299.00	3.0

```
shops.head(10)
```

	shop_name	shop_id
0	!Якутск Орджоникидзе, 56 фран	0
1	!Якутск ТЦ "Центральный" фран	1
2	Адыгея ТЦ "Мега"	2
3	Балашиха ТРК "Октябрь-Киномир"	3
4	Волжский ТЦ "Волга Молл"	4
5	Вологда ТРЦ "Мармелад"	5
6	Воронеж (Плехановская, 13)	6
7	Воронеж ТРЦ "Максимир"	7
8	Воронеж ТРЦ Сити-Парк "Град"	8
9	Выездная Торговля	9

X.RANDOM FOREST REGRESSOR

The random forest algorithm is the most precise algorithm available for forecasting sales. For the goal of forecasting the outcomes of machine learning activities, it is simple to use and comprehend. Random forest classifier is used to forecast sales.is utilised because its hyperparameters resemble those of a decision tree. The decision tool is the same as the tree model. The relationship between decision trees and random forests is depicted in Fig. 5. The random forest regressor class of the sklearn.ensemble package is used to resolve regression tasks of prediction by random forest. The parameter known as n estimators, which is also referred to as a random forest regressor, plays a crucial part. As a meta-estimator that fits on a number of classification-based decision trees, random forest .

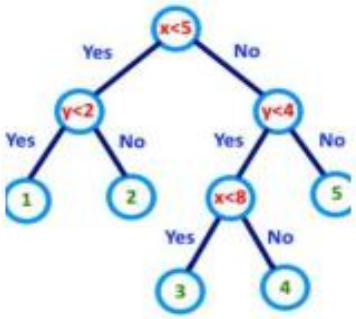

FIGURE :7 Random Forest Regression

XI.WEB APPLICATION WITH FLASK

In my project, web development is used to quickly access and forecast consumer product sales. There is no web development technology in the current system, and only developers have access to and control over product sales. In addition, we employ ensemble techniques to produce more accurate predictions and fine-tune model performance. The server side processing will be handled by code that we write. Requests will be sent to our code. It will determine the subject matter and nature of the requests. Also, it will decide what kind of reply to deliver to the user.

XII.DATA PREPROCESSING

Preparing raw data to be used with a machine learning model is known as data pre-processing. In order to build a machine learning model, it is the first and most important stage.It is not always the case that we come across the clean and prepared data when developing a machine learning project. Also, any time you work with data, you must clean it up and format it. So, we use a data pre-processing task for this.It contains missing values in the context of our dataset, which causes the dataset to be out of balance.So, we can clean the dataset by applying the pre-processing procedure to it. BigMart's data scientists have gathered 2013 sales information for 1559 products from 10 stores located in various cities. Additionally defined are the characteristics of each product and retailer. The data analysis reveals that the attributes Item Weight and Outlet Size contain missing values step for visualisation. Data must be pre-processed in order to be utilised in a machine learning model, which increases the model's effectiveness, by filling in any missing values. The missing values for Outset Size were filled by using the mode of the outlet size of a particular type of outlet, while the blank values for Item Weight were filled by averaging the weight of the particular item.

XIII.IMPLEMENTATION

Python is a high-level, general-purpose, interpreted language that is now popular for dealing with domain issues as opposed to system issues. The phrase "batteries included programming language" is another name for it. A number of third-party libraries are available to aid with problem solving, in addition to a variety of libraries for scientific purposes and enquiries.This project made use of the Python libraries Numpy for scientific computation and Matplotlib for 2D graphing. Moreover, data analysis has been done using the Python Pandas tool. The random forest regressor is used to complete tasks by assembling the random forest technique. Due to Jupyter Notebook's genius in "literate programming," where human-friendly code is written, it has been used as a development tool. Data visualisation revealed that the smallest sites created the lowest sales. Yet, in other instances, it was discovered that a medium-sized location—type 3; there are three types of super markets, for example—produced the most sales.Supermarkets of types 1, 2, and 3 are preferred to the location with the largest size. More locations should be converted to Type 3 Supermarkets in order to improve the product sales of Big Mart in a specific outlet.

However compared to other models, the suggested approach provides more accurate forecasts for future sales across all sites. Figure 20 illustrates the relationship between item MRP and outlet sales, for instance. Moreover, Figure 20 demonstrates a high correlation between Item Outlet Sales and Item to the Medium.

XIV.ALGORITHMS

a)LIGHT GRADIENT BOOSTING MACHINE

A gradient boosting framework called Light Gradient Boosting Machine uses decision trees to maximise model performance while using less memory.It employs two cutting-edge methods: Gradient-based Exclusive feature and one-side sampling.All GBDT (Gradient Boosting Decision Tree) frameworks use bundling (EFB), which addresses the shortcomings of histogram-based algorithms. The properties of the LightGBM Algorithm are formed by the two GOSS and EFB approaches that are detailed below. Together, they enable the model to function effectively and provide it an advantage over competing GBDT frameworks.

b)EXCLUSIVE FEATURE BUNDLING TECHNIQUE FOR LIGHTGBM

As high-dimensional data are frequently quite sparse, we may create a practically lossless method of reducing the amount of features. Particularly, many features in a sparse feature space are mutually exclusive, that is, they never assume nonzero value simultaneously. It is safe to combine the unique features into a single feature, known as an exclusive feature bundle. As a result, the difficulty of creating a histogram shifts from O(#data #feature) to O(#data #bundle), where bundle > feature. Thus, the training framework's speed is increased without compromising precision.

XV.ARCHITECTURE

In contrast to previous boosting algorithms that develop trees level-by-level, LightGBM divides the tree leaf-wise. It selects the leaf with the greatest delta loss for growth. The leaf-wise algorithm has less loss than the level-wise algorithm since the leaf is fixed.The complexity of the model could rise as a result of leaf-wise tree growth, which could also result in overfitting in limited samples.The diagrammatic representation of Leaf-Wise Tree Growth is shown below:

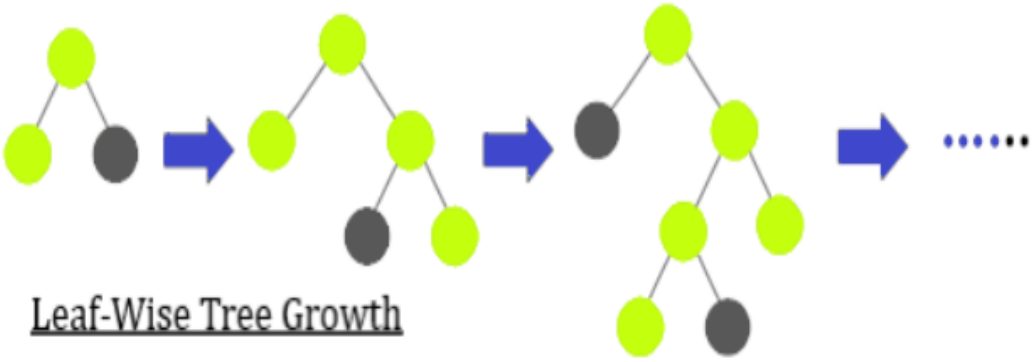

Figure 8 :Lgbm leaf wise tree growth image

XVI.WEB APPLICATION WITH FLASK

In this project, we make advantage of web development to quickly access and forecast consumer goods sales.There is no web development technology in the current system, and only developers have access to and control over product sales. • Moreover, in order to achieve superior prediction, hyper-tunning,develop code that will handle the server-side processing as a means to simulate performance. Requests will be sent to our code.It will determine the subject matter and nature of the requests. Also, it will determine the appropriate reply to send to the user.

XVII.CONCLUSION

This research aims to forecast future sales using machine learning methods and data from the prior year. In this study, we investigated the generation of several machine learning models utilising a variety of strategies, including Light Gradient Boosting Machine, Ada Boost, and XG Boost techniques. Considering the We chose the Light Gradient Boosting Machine Algorithm because of its accuracy and performance. The forecasting of sales results has been done using this method. We discussed in detail the techniques used to estimate the outcome and how the noisy data was removed. Based on the accuracy that many models predicted, we get to the conclusion that the random forest technique is the best model. Many retailers use our forecasts to adjust their techniques and strategies that enable them to increase earnings.

XVIII.REFERENCES

[1]An article Reference of their numerous dependent and independent characteristics
https://www.sciencedirect.com/science/article/abs/pii/S0263237300000438
[2]A web of Reference in predicting future demand and inventory management
https://onlinelibrary.wiley.com/doi/full/10.1111/jbl.12010
[3]A book of independent factors were used to construct the dataset.
https://books.google.co.in/books?hl=en&lr=&id=zB_G2u0pN7kC&oi=fnd&pg=PA2&dq=info:Y8M37KXb4_0J:scholar.google.com/&ots=N6cYrWG_Jl&sig=pbdG9lSLhrH20E-hYaHJar2DYcI&redir_esc=y#v=onepage&q&f=false
[4]An article of Reference inventory management data stored in a data warehouse
http://dspace.univ-msila.dz:8080/xmlui/handle/123456789/1918
[5]A web of Reference to estimate future sales.
https://ieeexplore.ieee.org/abstract/document/5590249/
[6]A book of Reference Every mall or store tries to give unique, limited-time deals
https://books.google.co.in/books?hl=en&lr=&id=9bSSVY7_QiAC&oi=fnd&pg=PP2&dq=info:ewjdd-xgeSgJ:scholar.google.com/&ots=5OkkaErqUH&sig=CJhGpHJU8x5IsjP7p9nV9OBFKig&redir_esc=y#v=onepage&q&f=false
[7]A web of Reference This project involves predicting big mart sales
https://jicet.org/index.php/JICET/article/view/53
[8]A Book of Reference A more accurate forecast is usually helpful in creating and upgrading
https://www.tandfonline.com/doi/abs/10.1057/palgrave.jors.2600567?journalCode=tjor20
[9]An article of Reference A traditional sales prediction study can help by carefully examining previous occurrences or circumstances
https://journals.sagepub.com/doi/abs/10.1177/002224298805200103?journalCode=jmxa
[10]A Web of Reference the information is so valuable in the present
https://www.sciencedirect.com/science/article/abs/pii/S0045790617315835
[11An article of t Reference ypes of tasks are dealt with in machine learning
https://www.sciencedirect.com/science/article/abs/pii/S0378426606000926
[12] A book of Reference a categorization type problem often counts as a resource for knowledge discovery.
https://books.google.co.in/books?hl=en&lr=&id=aaDbBwAAQBAJ&oi=fnd&pg=PP10&dq=info:3jA2Bt8V-b8J:scholar.google.com/&ots=iwOn5V-Edg&sig=6h1AF_ebvo_RXVRYaxTjWUDCIag&redir_esc=y#v=onepage&q&f=false
[13] A web of Reference Regression is used to produce resources and make accurate predictions
https://link.springer.com/article/10.1007/s10586-008-0052-0
[14]A book of Reference the main focus will be on creating systems self-efficient
https://journals.plos.org/plosone/article?id=10.1371/journal.pone.0239424

[15]An article of Reference they can perform calculations and analyses on their own and produce answers
https://link.springer.com/article/10.1007/s11192-017-2300-7
[16]A web of Reference Data can be transformed into knowledge
https://ieeexplore.ieee.org/abstract/document/6733221
[17]A book of Reference Sampling distributions are utilised in statistical inference as a theoretical
https://books.google.co.in/books?hl=en&lr=&id=ZxaRC4I2z6sC&oi=fnd&pg=PP6&dq=info:I3IW0wydl1cJ:scholar.google.com/&ots=oRn8Fp8UsI&sig=NWaaxT4t_V2_6827P8nkm4N7K9g&redir_esc=y#v=onepage&q&f=false
[18]An article of Reference Through the use of machine learning techniques
https://www.sciencedirect.com/science/article/abs/pii/S0957417417308333
[19] A web of Reference The amount of sales for each item can be forecasted for inventory management
https://www.inderscienceonline.com/doi/abs/10.1504/IJMTM.2000.001329
[20]A book of Reference the company, logistics, and transport service
https://books.google.co.in/books?hl=en&lr=&id=3uZpMDb4j_kC&oi=fnd&pg=PR7&dq=info:52mQLRPnYocJ:scholar.google.com/&ots=3ine-xwnbe&sig=wl5AJ2KDEvJz1VSbHh2ZeoSEdyY&redir_esc=y#v=onepage&q&f=false
[21]An article of Reference the company strategies related to the industry
https://www.sciencedirect.com/science/article/pii/S0263237320300190
[22]A web of Reference then drawing conclusions about customer acquisition, financial inadequacies, and strengths
https://journals.sagepub.com/doi/abs/10.1002/dir.10032?journalCode=jnma
[23]A book of Reference that are considerably more exact and precise.
https://link.springer.com/article/10.1186/1465-6906-12-s1-p11

BIBILOGRAPHY

T. Yamuna Satya Bhadravathi is studying his 2nd year Master of Computer Applications in Sanketika Vidya Parishad Engineering College, Visakhapatnam, A.P With her interest in Python,machine Learning and as a part of academic project she chose Image processing using Python. The article have been evolved from an idea to understand the flaws in conventional reporting and keeping time consistency, quality report generation in privacy preserving. A full fledged project along with code has been submitted for Andhra University as an Academic Project.

Pinnamaraju.T.S.Priya working as Assistant professor in Master of computer application(MCA) in Sanketika Vidya Parishad Engineering College, visakhapatanam Andhra pradesh. with 6 years of experience in Masters of Computer Applications (MCA) , accredited by NAAC. With her area of interests in C,DBMS,Computer Organization, Software Engineering,IOT