Journal of Current Research in Engineering and Science
Bi-Annual Online Journal (ISSN: 2581 - 611X)
VOLUME:8, ISSUE:2, AUGUST 2025, PAPER NUMBER:14

Precision Sales Forecasting Using LGBM Regressor

Tharun Adhithya S.S[1], Shyam K[2], Gokul Krishna Singh M[3],

Senthil Prabhu S. M.E.[4],

[1,2,3] Students,and [4] Faculty
Dept. of Computer Science Engineering,
Dr. Mahalingam College of Engineering and Technology, Coimbatore,
India.
senthilprabhus@drmcet.ac.in

Abstract: *In today's competitive market landscape, precise sales forecasting is indispensable for businesses seeking to optimize resource allocation, capitalize on market opportunities, and maintain a competitive edge. Traditional forecasting methods often fall short in accurately predicting sales figures, leading to suboptimal decision-making and resource allocation. To address this challenge, this paper proposes a framework for precision sales forecasting using the LightGBM (LGBM) regressor. The proposed framework leverages historical sales data, market trends, and external factors to generate highly accurate sales forecasts. The framework automates the forecasting process, allowing businesses to make informed decisions based on data-driven insights. Unlike conventional forecasting methods, which may overlook subtle patterns and fluctuations in sales data, the LGBM regressor excels in capturing complex relationships and non-linear trends, resulting in more accurate forecasts.To evaluate the performance of the proposed framework, extensive experimentation and validation are conducted using real-world sales data. The results demonstrate the superior accuracy and reliability of the LGBM regressor in predicting sales figures across different time horizons and business contexts. Furthermore, the framework's scalability and efficiency make it well-suited for handling large-scale datasets and real-time sales forecasting. In conclusion, precision sales forecasting using the LGBM regressor represents a significant advancement in the field of predictive analytics, empowering businesses to make informed decisions, optimize resource allocation, and achieve sustainable growth in today's dynamic market environment.*

Keywords: *LGBM Regressor, Sales Forecasting , Non-linear trends, Predictive Analysis.*

1. INTRODUCTION:

Sales forecasting plays a pivotal role in the strategic decision-making process for businesses by providing insights into future sales performance. Accurate forecasts are essential for optimizing production levels, managing inventory efficiently, devising effective marketing strategies, and overall business planning. However, traditional forecasting methods often struggle to capture the intricate nuances of modern consumer behavior and rapidly evolving market trends. To overcome these challenges, this project introduces a groundbreaking approach to sales forecasting that leverages advanced machine learning techniques, particularly focusing on the LightGBM (LGBM) Regressor. Unlike conventional methods that rely on simplistic statistical models or subjective judgment, this approach harnesses the power of data-driven insights and predictive analytics to generate precise sales forecasts.

At the core of this project lies the LGBM Regressor, a cutting-edge algorithm renowned for its speed, efficiency, and exceptional performance in handling vast datasets. By integrating historical sales data, customer demographics, market trends, and other relevant variables, this model effectively identifies patterns and relationships that significantly influence sales outcomes. By delivering precise, reliable, and actionable sales forecasts, we empower businesses to proactively navigate challenges, capitalize on opportunities, and achieve sustainable growth in an increasingly competitive marketplace.

Furthermore, this project places a strong emphasis on interpretability and transparency in sales

forecasting. We believe that providing clear insights into the factors driving sales predictions is crucial for informed decision-making within organizations.

2. DATA ACQUISITION :

The Data Acquisition module collects the sales data for the development of the model. The Sales dataset of Walmart from Kaggle is the dataset used here and it involves several steps to ensure data readiness for the sales forecasting project. The data acquisition process necessitates inspecting and understanding the dataset's structure, including the meaning and format of each column.The specific data source will depend on the chosen machine learning problem and the availability of data. The relevant data is extracted from the dataset, focusing on the features necessary for sales forecasting and analyzing.

3. DATA PREPROCESSING:

A crucial phase in the data analysis process is data preparation, which includes preparing raw data for analysis and modeling by cleaning, converting, and organizing it. Missing values resolving: Data imputation solves this problem. Imputation is the process of substituting approximated or computed values for missing data. Data Down casting: Down casting a dataset refers to the process of reducing the memory usage of a Data Frame by converting data types to their most memory-efficient counterparts while preserving data integrity. Data Melting: The process of transforming a dataset's structure from a wide format to a long format is known as data melting, data reshaping, or unpivoting. When working with datasets that have observations recorded as rows and variables stored as columns, this transformation is very helpful when the data needs to be rearranged for analysis or visualization. Feature engineering can then begin using the pre-processed data.

4. FEATURE ENGINEERING:

The process of improving the features of an existing dataset or introducing new ones to aid in the performance of machine learning models. It involves extracting valuable information from raw data, selecting relevant features, and creating new representations that enhance the predictive power of the models. The Time series does not have the notion of input and output features. Rather, the variable to be forecasted and build all of the inputs needed for future time step predictions using feature engineering. Label Encoding: The method by which numerical representations of categorical variables are created. An integer value is assigned to each distinct category or label. The retrieved data can now be used by machine learning algorithms thanks to this change. Lag Introduction: Lag refers to incorporating past sales data as features to predict future sales. Lags represent the time delay between historical sales events and their potential impact on future sales. The sales data can be shifted by a specific number of time periods (e.g., days, weeks, months) to provide lag characteristics. Mean Encoding: A probability of the target variable, depending on each feature value, is represented using mean encoding. With its encoded value, it represents the target variable in a sense. Based on the data's logical properties, the mean encodings were determined.

5. MODEL TRAINING:

The various key steps involved in Model Training: Model Selection: An essential first step in developing a machine learning pipeline for sales forecasting is selecting a model. It entails selecting the best algorithm, or best group of algorithms, to match the dataset and fulfill the demands of the forecasting task. At this point in the system, the sales forecasting scenario problem's forecasting solution is identified as the Light Gradient Boosting Machine (LGBM) regressor machine learning algorithm. Training Process: The model selection phase results in the selection of an appropriate model for the further execution of the forecasting process. At the end of the selection phase, the training of the model is done using a certain amount of data from the prepared dataset.

The model's current training data is supplemented with preprocessed, segregated data for training at each iterative round. Phase-wise or iterative training is used to train the model. With the parameters' assistance, it uncovers the hidden pattern in the data. Following the training phase, the model’s performance is calculated using a validation dataset to evaluate its ability to forecast the upcoming sales with the help of the available data for forecasting and prediction.

6. MODEL EVALUATION:

The crucial phase in the creation and application of machine learning models is model evaluation, which gauges the effectiveness, dependability, and generalizability of the models. To determine how well the model represents patterns and relationships in the data, it is necessary to compare the model's predictions with the actual values, or ground truth. Here, the crucial evaluation metric in use is: Root Mean Squared Error (RMSE): The Root Mean Squared Error measure is frequently used to evaluate the performance of regression models. A fundamental metric called Root Mean Squared Error, or RMSE, is employed to assess the accuracy of prediction models. Projecting future sales figures using relevant historical data and other information is the aim of sales forecasting. RMSE is a numerical indicator of how well the observed sales values match the expected sales values. A lower RMSE indicates that a forecasting model is more adept at seeing underlying patterns and trends in sales data, which leads to more accurate predictions. During the phase of recurrent training and testing, the RMSE value of the model decreases.

7. FORECASTING ANALYSIS:

The results of the forecast gives insights regarding the future sales which helps the organizations and stakeholders to introduce various changes in the inventory and stock to improve the rate of sales. The feature importance analysis reveals how the importance of features changes across different steps of the forecasting process. Observing variations in feature importance can provide insights into which features are more influential at different stages of the analysis. The analysis is conducted separately for each store, allowing for the identification of store-specific patterns and preferences. Variations in feature importance across stores may indicate differences in customer behavior, market dynamics, or other factors influencing sales performance. The iterative nature of the forecasting analysis, as indicated by the multiple steps and feature engineering techniques, highlights the importance of continuous improvement. Observing changes in feature importance and model performance over time can inform ongoing refinement and iteration of the forecasting process.

8. FLOW CHART:

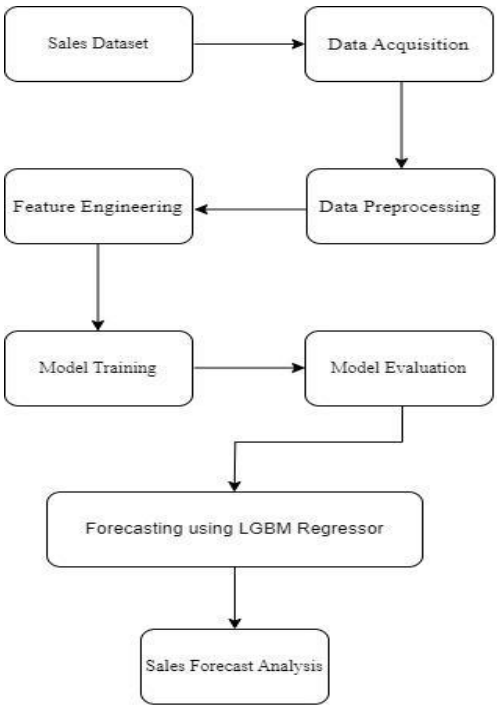

Fig. 1. Flowchart diagram for certificate automation

9. EVALUATION AND RESULTS:

The table shows model performance across each phase of the model testing and prediction. Performed well with low RMSE value which results in accurate forecasts.

Progressive Data Accumulation	RMSE
Step 1	166.592288
Step 2	136.861063
Step 3	18.613120
Step 4	6.470619
Step 5	6.355152
Step 6	4.598721

Table. 1. Performance of model progressively

10. CONCLUSION:

In conclusion, the project presents a groundbreaking framework for precision sales forecasting using the LightGBM(LGBM) regressor, which represents a significant advancement in the field of predictive analytics. We have demonstrated the effectiveness of our approach through extensive experimentation and validation using real-world sales data. The results showcase the superior accuracy, reliability, and scalability of the LGBM regressor in predicting sales figures across various time horizons and business contexts. The progressive data accumulation results, as shown in Table I, highlight the continuous improvement in forecasting accuracy as our framework evolves through different steps. From an initial RMSE of **166.592288** in **Step 1**, we achieve a remarkable reduction to **4.598721** in **Step 6**, indicating the effectiveness of the model in capturing intricate linkages and non-linear patterns in sales data. Proposed framework not only automates the forecasting process but also provides businesses with data-driven insights for making informed decisions. By using the historical sales data, trends in market, and other available external factors, we enable organizations to optimize resource allocation, capitalize on market opportunities, and maintain a competitive edge in today's dynamic market environment. Furthermore, our emphasis on interpretability and transparency in sales forecasting ensures that stakeholders understand the factors driving sales predictions, building trust and enabling well-informed decision-making in organizational settings.

11. REFERENCES:

[1] Xie dairu and Zhang Shilong, "Machine Learning Model for Sales Forecasting by Using XGBoost", 2021 IEEE International Conference on Consumer Electronics and Computer Engineering, Guangzhou, China, DOI: 10.1109/ICCECE51280.2021.9342304.

[2] Jun-Ki Hong,"LSTM-based Sales Forecasting Model", KSII TRANSACTIONS ON INTERNET AND INFORMATION SYSTEMS VOL. 15, NO. 4, Apr. 2021, Seoul, Korea, DOI: 10.3837/tiis.2021.04.003.

[3] Mohit Gurnani, Yogesh Korkey, Prachi Shahz, Sandeep Udmalex, Vijay Sambhe and Sunil Bhirudk, "Forecasting of sales by using fusion of Machine Learning techniques", 2017 International Conference on Data Management, Analytics and Innovation (ICDMAI) Zeal Education Society, Feb. 2017, Pune, India, DOI: 10.1109/ICDMAI.2017.8073492.

[4] JingyiDing, Li Xiaolong, Ziqing Chen, Baoxin Lai, "Sales Forecasting Based on CatBoost", 2020 2nd International Conference on Information Technology and Computer Application (ITCA), May 2021, Guangzhou, China, DOI: 10.1109/ITCA52113.2020.00138.

[5] Aly Megahed, Peifeng Yin and Hamid Reza Motahari Nezhad, "An Optimization Approach to Services Sales Forecasting in A Multi-Staged Sales Pipeline", 2016 IEEE International Conference on Services Computing, Sep 2016,San Francisco CA USA, DOI : 10.1109/SCC.2016.98.

[6] Giuseppe Nunnari and Valeria Nunnari, "Forecasting Monthly Sales Retail Time Series: A Case Study", 2017 IEEE 19th Conference on Business Informatics, Aug 2017, Thessaloniki Greece, DOI: 10.1109/CBI.2017.57

[7] Akshay Krishna, Akhilesh V, Animikh Aich, Chetana Hegde, "Sales-forecasting of Retail Stores using Machine Learning Techniques" , 3rd IEEE International Conference on Computational Systems and Information Technology for Sustainable Solutions, Dec 2018, Bengaluru, India, DOI: 10.1109/CSITSS.2018.8768765.

[8] Venishetty Sai Vineeth, Huseyin Kusetogullari, Alain Boone, "Forecasting Sales of Truck Components: A Machine Learning Approach", Proceedings of 2020 IEEE 10th International Conference on Intelligent Systems, Sep 2020, Varna, Bulgaria, DOI: 10.1109/IS48319.2020.9200128.

[9] Balpreet Singh, Pawan Kumar, Dr. Nonita Sharma, Dr. K P Sharma, "Sales Forecast for Amazon Sales with Time Series Modeling ", 2020 First International Conference on Power, Control and Computing Technologies (ICPC2T), 20 April 2020, Raipur, India, DOI: 10.1109/ICPC2T48082.2020.9071463.

[10] Yiyang Niu, "Walmart Sales Forecasting using XGBoost algorithm and Feature engineering", 2020 International Conference on Big Data & Artificial Intelligence & Software Engineering (ICBASE), April 2021, Bangkok, Thailand, DOI: 10.1109/ICBASE51474.2020.00103.