Sales Forecasting using Machine Learning

Harsh Goel[1], Himanshi Dwivedi[2], Prithvi Krishna Prasad[3], Rohan M[4], Vidya R[5]

[1] Harsh Goel, CS&E, Bangalore Institute of Technology, Karnataka, India
[2] Himanshi Dwivedi, CS&E, Bangalore Institute of Technology, Karnataka, India
[3] Prithvi Krishna Prasad, CS&E, Bangalore Institute of Technology, Karnataka, India
[4] Rohan M, CS&E, Bangalore Institute of Technology, Karnataka, India
[5]Vidya R, CS&E, Bangalore Institute of Technology, India

ABSTRACT

This paper presents a comprehensive approach for customer churn prediction and segmentation in the context of data- driven marketing. The proposed methodology involves multiple modules, including data processing, churn prediction using machine learning algorithms, customer segmentation through K-means clustering, and result analysis. The data processing stage ensures the quality and consistency of the dataset by performing data transformation, cleaning, and normalization. In cases of data imbalance, the Synthetic Minority Over-sampling Technique (SMOTE) is applied. The churn prediction module utilizes bagging tree, extra trees, and random forest algorithms for accurate identification of potential churners, followed by k-fold cross-validation and feature selection for model evaluation and variable importance determination. The customer segmentation module employs EDA techniques to gain insights from the data, and K-means clustering is used to group customers based on their similarities and behaviors. The results are then analyzed to assess the effectiveness of the churn prediction and segmentation models. This paper offers valuable insights for businesses seeking to proactively manage customer churn and implement targeted marketing strategies for different customer segments, ultimately leading to improved business performance and customer satisfaction.

Keyword : - *Machine Learning approach*

1. INTRODUCTION

In the retail sector, sales forecasting is crucial due to competition from e-commerce sites and changing market dynamics. To meet these challenges, machine learning algorithms are utilized to analyze data and predict sales of specific products and their impact on store sales. Sales forecasting involves anticipating consumer purchases based on product features and sale conditions. Accurate sales predictions are important for investors to make informed decisions about investments in new ventures, while managers rely on sales forecasts to determine plant size, inventory, staffing, advertising, pricing, and salaries for salespeople. Profitability depends on the accuracy of sales and cost forecasts, confidence in the forecast, and proper utilization of the forecast in the planning process.

Objectives of this paper are ,sales forecasting helps businesses determine production volumes and arrange necessary facilities.It forms the basis for production, sales, and natural budgets. Sales forecasting provides a commitment level for the sales department to achieve within a specified timeframe. It leads to the successful accomplishment of pre-established targets

It facilitates informed decision-making by providing relevant market information.

Accurate sales forecasting assists in deciding on plant expansion, production mix variations, or resource diversion for specific products. It aids in preparing schedules for production activities and materials purchase. Sales forecasting guides key business activities such as production and marketing.

The project aims to create a precise sales forecasting model using XGBoost, a potent machine learning algorithm. The central goal is to utilize data-driven techniques to make accurate predictions about sales and offer valuable insights for decision-making processes within the organization. The development of an efficient sales forecasting model is intended to enhance inventory management, resource allocation, and strategic planning, leading to maximum profitability and operational efficiency.

Scope of this project involves implementing a sales forecasting model using XGBoost. We'll analyze historical sales data, perform feature engineering to extract relevant information, and train the model with the XGBoost algorithm. We'll then evaluate the model's accuracy and robustness using appropriate metrics and validation techniques. The project will be completed within the allocated timeframe and resources, focusing on delivering a reliable and

effective sales forecasting solution. This project has broad applicability across various industries and organizations that rely on sales forecasting to inform their decision-making processes. This includes businesses operating in the retail, e-commerce, manufacturing, and service sectors. With accurate predictions of future sales, our project offers practical applications in inventory management, supply chain optimization, demand planning, and financial forecasting. The developed sales forecasting model provides valuable insights to stakeholders, such as managers, executives, and analysts, enabling them to make
informed decisions based on data.

2. WORK IN THIS AREA

This section offers the detailed depiction of review on various existing techniques employed so far. The authorin the paper [1] proposes a methodology that combines deep neural networks with SHapley Additive exPlanations (SHAP) techniques for sales forecasting. It involves preprocessing the data, training a deep neural network model, and using SHAP analysis to interpret the model's predictions and feature importance.The Paper [2] proposes light gradient-boosting machine (LightGBM) framework, to realize WalMart sales forecasting,feature engineering is also performed in this paper. First they removed some features that are not related to the model input. Then the features are extracted and classified, and the mean, standard deviation and other statistics of some features are obtained. Experiments results show that that this method has an Root Mean Sqaure Error (RMSE) of 0.641, which is significantly better than Logistic Regression (0.803) and Support Vector Machines(SVM) (0.732). In addition, this paper also shows the 20 top feature importance. This is of great significance for guiding the company's sales. The paper [3 proposes To build an efficient and accurate sales forecasting model using machine learning model eXtreme Gradient Boosting (XGBoost) for forecasting the future sales amount. The paper [4 proposes ForeXGBoost Algorithm, a sales forecast system based onlarge-scale datasets with comprehensive information including the vehicle's brand_ID, model, power, and displacement.The paper [5] proposes a method based on a multi-layer LSTM(Long Short Term Memory) network by using the grid search approach. The proposed method searches for the optimal hyperparameters of the LSTM network. Generally, time series forecasting techniques fall into the two main categories of statistical and computational intelligence methods. Widely-used statistical time series forecasting methods such as autoregressive integrated moving average(ARIMA) suppose that the time series contains only linear components. However, most real-world time series data consist of nonlinear components too. The paper [6] proposes to combine XGBoost algorithm and feature engineering processing on Walmart sales dataset competition platform. Experiments show that this method can effectively mine features of different dimensions and performs better than Logistic regression algorithm and Ridge algorithm. The paper [7] proposes the forecast of the sales of truck components using machine learning algorithms like Support Vector Machine Regression(SVMR), Ridge Regressor, Random Forest(RFR) and Gradient boosting(GBR)The study has been conducted on sales data which belonged to the truck sales of Sweden region and it cannot be assured that the similar results will be obtained from the study conducted on the sales data belonged to the other region as the sales may vary in other regions. The paper [8] proposes a novel deep neural framework for sales forecasting in ECommerce, named deep sales forecasting framework (DSF). Among many heterogeneous features relevant to sales forecasting, promotion campaigns held in E-Commerce and competing relation between substitutable products would greatly complicate the matter. Unfortunately, these factors are usually overlooked in the existing literature, since the conventional time series analysis based techniques mainly consider the sales records alone.This paper[9] proposes a short-term demand prediction method for online car-hailing services.LS- SVM model is used. Short-term demand prediction in online car-hailing services based on a Least Squares Support Vector Machine (LS-SVM) involves data collection and preprocessing, training the LS-SVM model with optimized parameters, evaluating its performance, and using it for real-time demand predictions based on input variables such as time, weather, and location.This paper [10] proposes Prediction of sales of a product from a particular outlet is performed via a two-level approach that produces better performance compared to any of the popular single model predictive learning algorithms.

3. PROPOSED METHODOLOGY

This section is a detailed explanation of the proposed system implemented. The proposed system is divided into 3 modules i.e., Data Preparation, Feature Selection, Model training and Evaluation
Acquiring datasets relevant to Sales is the initial step in Sales Forecasting. The dataset used in this paper is an Adidas Sales Dataset. An Adidas sales dataset is a collection of data that includes information on the sales of Adidas products. This type of dataset may include details such as the number of units sold, the total sales revenue, the location of the sales, the type of product sold, and any other relevant information. Adidas sales data can be useful for

a variety of purposes, such as analyzing sales trends, identifying successful products or marketing campaigns, and developing strategies for future sales. It can also be used to compare Adidas sales to those of competitors or to analyze the effectiveness of different marketing or sales channels.

3.1 Data Preparation

The data preprocessing step is crucial to ensure the quality and reliability of the Adidas sales dataset. It involves handling missing values and duplicate values. Missing values are addressed through techniques such as mean, median, or mode imputation for numerical features. Duplicate values are identified by comparing records across all features and can be removed, retaining the most relevant record based on criteria such as recency or completeness. These preprocessing steps enhance the dataset's integrity, minimizing biases and ensuring more accurate and reliable results for subsequent sales forecasting modeling.

3.2 Feature Selection

Feature selection is performed to identify the most relevant features for sales forecasting modeling. This is achieved by analyzing the correlation between features using techniques such as a correlation heatmap. Highly correlated features are identified and redundant ones are eliminated to reduce dimensionality and focus on the most informative variables. The selection of features based on the correlation heatmap helps in improving model performance, interpretability, and efficiency.

3.3 Model training and Evaluation

In this step, the selected models are trained using the preprocessed dataset. The dataset is divided into training and testing sets, with a designated time period for testing to ensure the validity of the forecasting results. The models are trained on the training set, and the R2 score is calculated to evaluate their performance on the testing set. The R2 score measures the proportion of the variance in the dependent variable (sales) that can be explained by the independent variables in the model.

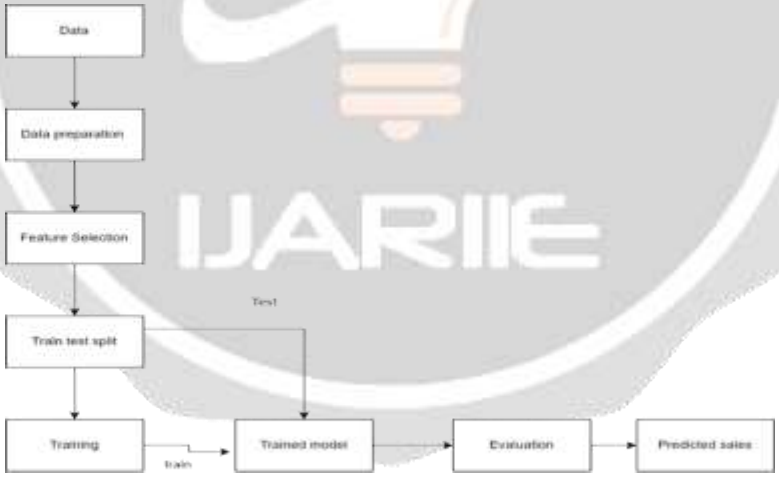

Fig -1: Proposed Methodology

4. RESULTS

This section presents the results obtained from the trained models and discusses their performance in terms of the R2 score. The selected machine learning models were trained using the preprocessed dataset, and their performance was evaluated on the testing set. The R2 score, a commonly used metric for regression tasks, measures the proportion of the variance in the dependent variable (sales) that can be explained by the independent variables in the model.

The sales forecasting models achieved an impressive R2 score of 0.99, indicating that approximately 99% of the variability in sales can be explained by the selected features. This high R2 score suggests a strong relationship between the independent variables and the sales outcome, demonstrating the effectiveness of the models in capturing and predicting sales patterns.

The exceptional R2 score implies that the selected models successfully capture the underlying trends, seasonality, and other factors influencing Adidas sales. The accurate sales forecasting achieved through machine learning techniques can provide valuable insights for Adidas, enabling them to make informed business decisions, such as optimizing inventory management, planning marketing strategies, and forecasting revenue.

5.CONCLUSION

To summarize, our project focused on developing a sales forecasting model using the XGBoost algorithm. Our aim was to create a model that could provide accurate insights for decision-making. We were able to achieve excellent results through extensive analysis and experimentation. Our XGBoost model demonstrated its effectiveness in capturing the patterns and relationships present in the sales data, leading to a high R2 score of 0.999. This score indicates that our model can explain almost all of the variance in sales. However, it is important to further evaluate the model's ability to generalize to future or unseen data to ensure its robustness. Overall, our project has successfully developed a powerful sales forecasting tool that can aid decision-making processes.

6. REFERENCES

[1] .Chen, J. et al. (2021) 'Sales Forecasting Using Deep Neural Network and SHapley Additive exPlanations (SHAP) Techniques', 2021 IEEE 2nd International Conference on Big Data, Artificial Intelligence and Internet of Things Engineering (ICBAIE) [Preprint].

[2] .T. Deng, Y. Zhao, S. Wang, and H. Yu, "Sales forecasting based on lightgbm," 2021 IEEE International Conference on Consumer Electronics and Computer Engineering (ICCECE), 2021.

[3] .X. dairu and Z. Shilong, "Machine learning model for sales forecasting by using XGBoost," 2021 IEEE International Conference on Consumer Electronics and Computer Engineering (ICCECE), 2021.

[4] .Z. Xia et al., "Forexgboost: Passenger car sales prediction based on xgboost," Distributed and Parallel Databases, vol. 38, no. 3, pp. 713–738, 2020.

[5] .H.. Abbasimehr, M. Shabani, and M. Yousefi, "An optimized model using LSTM network for demand forecasting," Computers & Industrial Engineering, vol. 143, p. 106435, 2020.

[6] .Y. Niu, "Walmart sales forecasting using XGBoost algorithm and feature engineering," 2020 International Conference on Big Data & Artificial Intelligence & Software Engineering (ICBASE), 2020.

[7] .V. Sai Vineeth, H. Kusetogullari, and A. Boone, "Forecasting sales of truck components: A machine learning approach," 2020 IEEE 10th International Conference on Intelligent Systems (IS), 2020.

[8] .Y. Qi et al., "A Deep Neural Framework for sales forecasting in e-commerce," Proceedings of the 28th ACM International Conference on Information and Knowledge Management, 2019.

[9] .S. Jiang, W. Chen, Z. Li, and H. Yu, "Short-term demand prediction method for online car-hailing services based on a least squares support vector machine," IEEE Access, vol. 7, pp. 11882–11891, 2019.

[10] .K. Punam, R. Pamula, and P. K. Jain, "A two-level statistical model for Big Mart Sales prediction," 2018 International Conference on Computing, Power and Communication Technologies (GUCON), 2018.