Air Conditioner Sales Prediction Using CTGAN, XGBoost and SHAP

Vinay Redkar[1], Kunj Shah[2], Akanksha Sonone [3], Ayushi Mishra [4], Anushree Deshmukh[5]

1 BE Student, Department of Information Technology

2 BE Student, Department of Information Technology

3 BE Student, Department of Information Technology

4 BE Student, Department of Information Technology

5 Assistant Professor, Department of Information Technology

1,2,3,4 MCT's Rajiv Gandhi Institute of Technology

---***---

Abstract - In today's modern world, air conditioners (ACs) have become a vital part of daily life due to rising temperatures and increasing disposable incomes, leading to a surge in AC sales. With numerous companies competing in this industry, businesses face challenges in selecting the most suitable AC brand for their needs, managing inventories across multiple brands. To address this issue, this research paper presents a machine learning model designed to recommend the optimal AC brand based on size, price, and predicted sales potential. Leveraging data from Amazon, the leading AC retailer in the Indian market, the model facilitates informed decision-making for businesses seeking efficient AC procurement strategies.

***Key Words*:** Sales Prediction, Amazon, AC, XGBoost, CT-GAN, SHAP, KNN

1.INTRODUCTION

In today's modern era, air conditioning (AC) has transformed from a luxury into a necessity, playing a vital role in both residential and commercial environments. However, the process of acquiring AC units can be complex, especially in commercial settings where a diverse range of brands and models are managed by intermediaries. This complexity stems from the vast array of available options, making it challenging to identify the most suitable and profitable AC units for businesses.

To tackle this challenge, this project introduces an innovative Machine Learning Model developed using data sourced from Amazon, India's leading retail platform. The dataset utilized in this project comprises over 300,000 product details across various categories, with a specific focus on the AC category encompassing more than 600 product details and 5 crucial attributes. These attributes include product names, customer ratings, discount prices, actual prices, and sales figures. Additionally, the dataset has been enriched with new attributes such as ton, star, material composition (copper or non-copper), inverter technology, and split type, all aimed at enhancing the model's predictive capabilities.

The project methodology employs advanced data processing techniques, starting with the use of K-Nearest Neighbors (KNN) to handle missing values effectively. Furthermore, Conditional Generative Adversarial Networks (CTGAN) are utilized for data augmentation, a critical step given the limited size of the original dataset. Subsequently, the model is trained using the powerful XGBoost algorithm, renowned for its ability to handle structured data and produce robust predictions. Additionally, SHAP (SHapley Additive exPlanations) is leveraged for model interpretation, parameter tuning, and overall accuracy improvement, especially considering the utilization of synthetic data in the training process.

One significant advantage is the time-saving aspect for retailers. With streamlined inventory management, retailers can avoid the need for extensive selection processes when choosing AC units. By automating processes, retailers can streamline operations, saving time and enabling a greater focus on customer service and other critical areas.

Furthermore, efficient inventory management enables businesses to stock only specific brand AC units that are in high demand or have proven to be popular among customers. This targeted stocking approach helps in optimizing inventory levels, reducing the risk of overstocking or understocking, and ensuring that the right products are available when customers need them.

Reduced maintenance costs are another benefit of optimized inventory management. By storing fewer types of spare parts and components, businesses can minimize the costs associated with maintaining inventory. This includes storage costs, inventory handling costs, and the costs of managing and tracking multiple types of spare parts. Additionally, with a more focused inventory, businesses can negotiate better deals with suppliers and streamline their procurement processes.

Labor costs are also positively impacted by efficient inventory management practices. When businesses stock specific brand AC units, employees tasked with maintenance, repair, or installation tasks don't need to learn the structures of multiple AC unit types. This reduces training time and costs, improves employee productivity, and ensures that maintenance tasks are carried out more efficiently.

The streamlined selection process resulting from optimized inventory management contributes significantly to increased profitability. Businesses can allocate resources more effectively, focus on promoting high-demand AC units, and capitalize on market trends more efficiently. This targeted approach leads to higher sales volumes, improved customer satisfaction, and ultimately, higher profits.

In conclusion, efficient inventory management is a key driver of success for businesses in the AC industry. It enables time savings, targeted stocking, reduced maintenance costs, optimized labor utilization, and increased profitability, ultimately contributing to enhanced operational efficiency and competitive advantage in the market.

2. LITERATURE SURVEY

In order to develop our project, we conducted a comprehensive literature survey analyzing various research papers. This survey enabled us to gain insights into the strengths and limitations of each paper, providing a valuable foundation for our subsequent actions. Below, we provide a brief overview of the research papers studied, highlighting key findings and implications for our project.

The study described in [1] examines how machine learning techniques can be utilized to predict sales outcomes for businesses. Its goal is to boost revenue by using these methods to study consumer purchasing behaviors and predict upcoming sales trends. The research employs clustering models and data mining to assess how well different algorithms can forecast sales. The results underscore the importance of machine learning in developing business strategies that are informed by consumer behavior and historical sales data, helping businesses to customize their approaches to enhance sales and remain competitive in the market.

This research paper [2] explores the application of various machine learning algorithms for sales prediction across retail stores. The authors focus on building accurate models using historical data to forecast future sales. They compare the effectiveness of techniques like neural networks, decision trees, and regression analysis. The study highlights the importance of advanced analytics in sales management and forecasting, suggesting that machine learning can improve sales prediction accuracy. The research details the process of data preparation, feature engineering, model selection, and performance evaluation. The goal is to develop a reliable predictive model that considers factors like seasonality, promotions, economic conditions, and competitor activity. The findings demonstrate the value of these sophisticated analytical methods for improving sales forecasting and management strategies in the business sector.

The [3] paper examines the essential role of precise sales forecasting within the e-commerce sector, especially highlighting its rapid expansion in China. This growth is significantly influenced by national policies and the impacts of the COVID-19 pandemic. The research identifies critical factors influencing e-commerce sales predictions and organizes them into three main categories: the attribute characteristics of products or businesses, customer reviews and derived metrics, and online product search data. Moreover, the paper conducts an extensive evaluation of various predictive models including linear, machine learning, and deep learning approaches, assessing their ability to accurately predict sales volumes in the e-commerce realm. The importance is given to study of integrating additional data like calendar events and pricing strategies to enhance the sales forecasts accuracy.

A study paper [4] explores how data mining techniques combined with machine learning algorithms can significantly improve the accuracy of sales forecasting for businesses. The research emphasizes the importance of accurate sales trend predictions for effective planning and decision-making. By analyzing three years of data from an e-commerce fashion store, the study explores various data analysis stages like data cleaning, initial exploration, outlier identification, and ultimately, forecasting future sales trends. It compares the effectiveness of three machine learning models: Generalized Linear Models, Decision Trees, and Gradient Boosted Trees. The study finds that Gradient Boosted Trees outperforms the others, achieving an impressive accuracy rate of 98%. These findings highlight the significant advantage of using modern data mining and machine learning techniques to achieve highly precise sales forecasts. This, in turn, allows businesses to make better decisions based on reliable data-driven insights.

The study [5], building on the value of sales forecasting, explores machine learning for product-level sales prediction in retail stores. This research investigates various regression methods (Multiple, Polynomial, Ridge, and Lasso Regression) alongside boosting algorithms like AdaBoost and Gradient Tree Boosting. Their focus on product-level forecasting underscores the importance of precise predictions for optimal resource allocation and improved customer satisfaction. Similar to the previous study [4], the research evaluates different algorithms using a sales dataset. Gradient Tree Boosting emerges as the most accurate predictor, achieving the lowest Root Mean Square Error (RMSE) and highest coefficient of determination (R^2) values. These findings solidify the value of sales forecasting and the effectiveness of machine learning in enhancing prediction accuracy within the retail landscape.

The study in [6] leverages the XGBoost machine learning model to forecast retail sales, focusing on achieving high accuracy and effectiveness. By implementing advanced techniques like feature engineering and data preprocessing, the research captures essential sales trends over various periods. Presented at a conference, the paper highlights the vital role of accurate sales forecasts in promoting business success, which aids in informed decision-making and efficient resource allocation. The robustness of XGBoost in handling complex datasets was crucial in enhancing the methodologies used for forecast of sales. Research paper illustrates transformative potential of machine learning, particularly through XGBoost usage, in revolutionizing forecast of sales in retail category and strategic business decision-making.

The [7] paper underscores the critical importance of adhering to proper document formatting standards to enhance both readability and professionalism. It advocates for the utilization of PDF files as a preferred format for disseminating information effectively. Moreover, it accentuates the significance of incorporating visual aids, such as infographics, within documents to augment engagement and comprehension levels. Additionally, it delves into the pivotal role of effective communication strategies in articulating ideas clearly and persuasively. Strategies such as data mining, content reuse, and meticulous analysis are recommended to ensure a thorough understanding and interpretation of the conveyed information.
Furthermore, the paper emphasizes the necessity of employing appropriate language and structural coherence in communication endeavors to guarantee clarity and maximize impact. In essence, the document offers invaluable insights into optimizing document formatting practices, leveraging presentation techniques, and deploying effective communication strategies to facilitate seamless business communication.

The [8] research introduces an innovative fuzzy time series model designed for enhancing forecast of sales within management of supply chain. By implementing this approach, businesses can expect improved accuracy in their demand forecasts, enabling businesses to manage inventory effectively and meet customer needs efficiently. Utilizing fuzzy trapezoidal membership functions and advanced fuzzy logic relationships, the model outperforms traditional forecasting methods by offering greater precision. Building on their prior success in applying this model to student enrollment forecasting, the authors have adapted it for sales forecasting in the supply chain context. The study refines the model by computing equally spaced intervals and using trapezoidal membership functions, resulting in superior predictive performance compared to conventional methods. This research not only highlights the importance of demand management in supply chain operations but also showcases the potential of fuzzy time series analysis in refining sales forecasting for better decision-making and efficient customer demand management.

The study [9] examines how Walmart utilizes Long Short-Term Memory (LSTM) technology, a type of machine learning, to improve sales forecasting accuracy. The research highlights the increasing importance of advanced technologies in retail for better decision-making. By leveraging LSTM, Walmart can analyze historical sales data to identify complex patterns and trends. These insights support more accurate predictions of future sales trends. The study also emphasizes the crucial role of feature engineering in enhancing the effectiveness of machine learning models. This involves tailoring the data by selecting, transforming, and creating relevant features, ultimately leading to improved forecasting accuracy. The combined power of LSTM and feature engineering allows Walmart to optimize resource allocation, refine inventory management, and tailor marketing strategies to better meet customer needs. The research validates the effectiveness of these techniques for sales forecasting and suggests broader applications of machine learning in retail analytics. The study's findings imply that strategically using LSTM and feature engineering has the potential to transform sales forecasting practices and contribute significantly to retail business success.

A 2018 study presented at the 3rd IEEE International Conference on Computational Systems and Information Technology for Sustainable Solutions [10] explored the use of machine learning for sales forecasting in retail. The research emphasizes the critical role of accurate sales predictions in managing complex retail operations, which are influenced by a combination of internal and external factors. The study aimed to improve forecasting accuracy by combining various machine learning techniques. This approach addresses the limitations of traditional statistical methods that may struggle to capture the intricacies of sales data. The research highlights the use of feature engineering techniques, such as creating new variables from existing data (e.g., product categories, store age). This structured approach to data preparation enhances its suitability for machine learning algorithms. The paper presents a comprehensive architectural framework for sales prediction. It details the process, from formulating hypotheses to data exploration. This framework emphasizes the importance of understanding product and store-specific variables that can impact sales. Overall, the study offers valuable insights on applying machine learning in retail sales forecasting. It demonstrates how advanced algorithms can significantly improve business strategies and decision-making capabilities.

3. DATASET

In this section of the research paper, we utilized a dataset obtained from Kaggle, specifically the Amazon sales data from 2023, comprising over 300,000 products across 100+ categories. For our analysis, we focused solely on AC sales data, which amounted to a total of 600 rows representing different AC unit models.

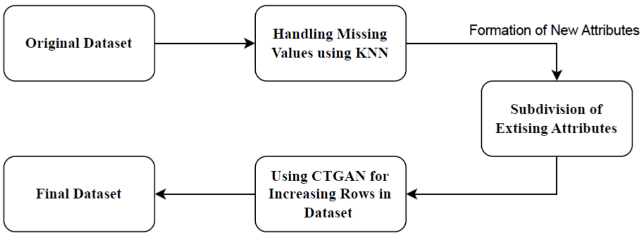
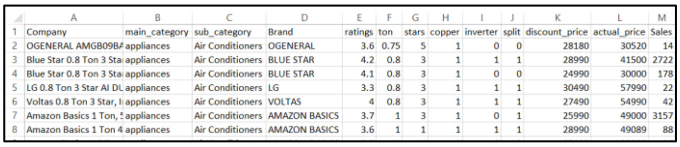

Original Dataset
Handling Missing Values using KNN
Formation of New Attributes
Subdivision of Extising Attributes
Using CTGAN for Increasing Rows in Dataset
Final Dataset

Fig -1: Flowchart of Dataset

Handling null values in the dataset was a critical task. We employed K-Nearest Neighbors (KNN) and Iterative Imputer methods to address these null values effectively. KNN Imputer and Iterative Imputer are essential techniques in machine learning that help fill in missing data points based on the proximity of neighboring data points. The Mean Squared Error (MSE) score was then calculated to assess the reliability and quality of the dataset post-null value handling. The original dataset included five key attributes: product names, discount prices, customer ratings, actual prices, and sales figures. The primary output of interest was Sales Prediction. Upon data examination, we identified opportunities to derive new attributes from the existing dataset. These new attributes included ratings, ton, material composition (copper or non-copper), star, inverter technology, and split type, adding five additional attributes and bringing the total to 10 attributes for analysis.

Due to the limitations of a relatively small dataset, we implemented data augmentation to enhance model performance during training. Conditional Generative Adversarial Networks (CTGAN) was employed to generate synthetic data that closely mirrors the existing data distribution. This process effectively increases the dataset size, leading to improved model efficiency during the training phase. Following data augmentation, we assessed the model's accuracy on the expanded dataset using Root Mean Square Error (RMSE) and Mean Squared Error (MSE) metrics.



	A	B	C	D	E	F	G	H	I	J	K	L	M
1	Company	main_category	sub_category	Brand	ratings	ton	stars	copper	inverter	split	discount_price	actual_price	Sales
2	OGENERAL AMGB09BA	appliances	Air Conditioners	OGENERAL	3.6	0.75	5	1	0	0	28180	30520	14
3	Blue Star 0.8 Ton 3 Sta	appliances	Air Conditioners	BLUE STAR	4.2	0.8	3	1	1	1	28990	41500	2722
4	Blue Star 0.8 Ton 3 Sta	appliances	Air Conditioners	BLUE STAR	4.1	0.8	3	1	0	0	24990	30000	178
5	LG 0.8 Ton 3 Star AI DU	appliances	Air Conditioners	LG	3.3	0.8	3	1	1	1	30490	57990	22
6	Voltas 0.8 Ton 3 Star, I	appliances	Air Conditioners	VOLTAS	4	0.8	3	1	1	1	27490	54990	42
7	Amazon Basics 1 Ton, 5	appliances	Air Conditioners	AMAZON BASICS	3.7	1	3	1	0	1	25990	49000	3157
8	Amazon Basics 1 Ton 4	appliances	Air Conditioners	AMAZON BASICS	3.6	1	1	1	1	1	28990	49089	88

Fig -2: Final Dataset

4. ARCHITECTURE AND MODEL DEVLOPMENT

The project architecture comprises a structured flowchart outlining the process. The sequential steps include data preprocessing, feature engineering, XGBoost model training, hyperparameter tuning, model evaluation, and SHAP integration. Each step builds upon the previous one, ensuring a systematic approach to developing an accurate sales forecasting model.

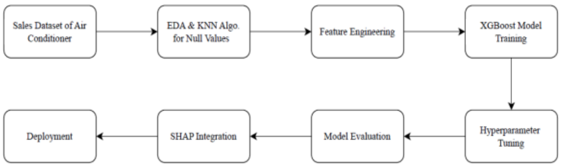


Fig -3: Process Flow

The model development phase for creating an XGBoost model for sales forecasting involves a series of meticulous steps to ensure accuracy and reliability in predictions. Initially, the data preprocessing stage focuses on cleaning and preparing the dataset. This includes handling missing values through imputation techniques, identifying and addressing outliers that could skew results, and standardizing data to ensure uniformity and comparability across features.

Following data preprocessing, the feature engineering phase becomes essential. In this stage, a combination of domain expertise and statistical methods is utilized to derive significant features from the data. This includes selecting pertinent variables, generating new features via transformations or combinations, and suitably encoding categorical variables to ensure model compatibility.

After preprocessing the data and engineering the features, the XGBoost model undergoes training with the prepared dataset. XGBoost, known for its efficacy as a gradient boosting algorithm, incrementally constructs decision trees to reduce prediction errors, which makes it adept at handling complex structured data, such as sales information. During the training phase, the tuning of hyperparameters like the learning rate, the maximum depth of the trees, and regularization parameters is critical to optimizing the model's performance.

Following the initial training, the XGBoost model is evaluated using metrics such as Root Mean Squared Error (RMSE) or Mean Absolute Error (MAE). This evaluation gauges the predictive accuracy of the model and identifies potential areas for enhancement. Hyperparameter tuning is then iteratively executed to refine the model settings, aiming for peak performance.

The incorporation of SHAP (SHapley Additive exPlanations) into the model development enhances interpretability and provides deeper insights. SHAP offers explanations for individual predictions, elucidating the impact of each feature on the model's output. This transparency not only helps in comprehending the model’s decision-making process but also facilitates further refinements and optimizations.

In summary, the development of an XGBoost model for sales forecasting entails meticulous data preprocessing, sophisticated feature engineering, comprehensive model training, precise hyperparameter tuning, thorough model evaluation, and the integration of SHAP. These steps collectively form a robust and precise model equipped to deliver well-informed sales forecasts in dynamic retail scenarios.

5. RESULTS

The initial findings highlight the importance of newly added attributes in the dataset compared to existing ones, shedding light on their significance. For instance, in Figure 4, the comparison between the Ton attribute (a new addition) and sales (an existing attribute) provides insights into the most preferred AC sizes. The analysis reveals that 1.5 Ton AC units are the highest-selling, followed by 1 Ton, 2 Ton, and 0.8 Ton units in descending order. This comparison aids in understanding customer preferences regarding Air Conditioner sizes and their corresponding sales trends.

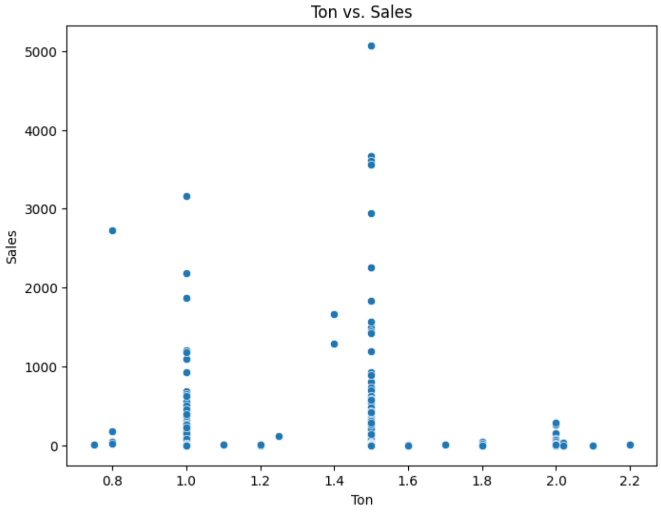


Fig -3: Ton vs. Sales

6. CONCLUSIONS

In conclusion, this project provides predictive values based on input specifications of AC units, aiding AC retailers in making informed decisions on whether to purchase a particular AC model. The predictive values offer insights into potential sales, contributing to increased profitability and operational ease for businesses. From a technical standpoint, the project employs K-Nearest Neighbors (KNN) to handle null values and Conditional Generative Adversarial Networks (CTGAN) to augment the dataset, especially considering its initial size of 500 rows. New attributes are derived from existing ones, enhancing data efficiency and precision. The XGBoost algorithm is utilized for model training, followed by SHAP for further refinement, resulting in a precise and deployable model for execution.

REFERENCES

[1] Purvika Bajaj, Renesa Ray, Shivani Shedge, Shravani Vidhate, Prof. Dr. Nikhilkumar Shardoor, "Sales Prediction Using Machine Learning Algorithms". International Research Journal of Engineering and Technology (IRJET)

[2] Garud Akshada Anil, Chavan Ritambara Shankar, Bobade Prachi Santosh, Gorad Akshada Rajendra, Prof. B.D. Thorat, "Sales Forecasting Using Machine Learning Techniques". International Research Journal of Modernization in Engineering Technology and Science

[3] Zixuan Huo, "Sales Prediction based on Machine Learning". International Conference on E-Commerce and Internet Technology (ECIT).

[4] Sunitha Cheriyan, Shaniba Ibrahim, Saju Mohanan, Susan Treesa, "Intelligent Sales Prediction Using Machine Learning Techniques". IEEE conference

[5] Akshay Krishna, Akhilesh V, Animikh Aich, Chetana Hegde, "Sales-forecasting of Retail Stores using Machine Learning Techniques". IEEE International Conference on Computational Systems and Information Technology for Sustainable Solutions 2018.

[6] Xie dairu, Zhang Shilong, "Machine Learning Model for Sales Forecasting by Using XGBoost ". 2021 IEEE International Conference, DOI: 10.1109/ICCECE51280.2021.9342304

[7] Prabhat Sharma, Shreyansh Khater, Vasudha Vashisht,"Sales Forecast of Manufacturing Companies using Machine Learning navigating the Pandemic like COVID-19". 2021 2ND International Conference on Computation, Automation and Knowledge Management (ICCAKM Amity University).

[8] S. M. Aqil Burney, Syed Mubashir Ali," Sales Forecasting for Supply Chain Demand Management – A Novel Fuzzy Time Series Approach". 2019 13th International Conference on Mathematics, Actuarial Science, Computer Science and Statistics (MACS)

[9] XINJE LI, Jiakai Du, Yang Wang, Yuan Cao,"Automatic Sales Forecasting System Based On LSTM Network". 2020 International Conference on Computer Science and Management Technology (ICCSMT)

[10] A. Krishna, A. V, A. Aich and C. Hegde, "Sales-forecasting of Retail Stores using Machine Learning Techniques," 2018 3rd International Conference on Computational Systems and Information Technology for Sustainable Solutions (CSITSS).