**An enhancement in accuracy in forecasting sales using XG boost and decision tree algorithm**

**Students:**

**J Sai Chandu (192211216)**

**K. Madhan Kumar (192211186)**

**Corse Name & Course Code**

**Machine Learning for India- ITA0658**

**Faculty:**

**Dr. Anakath Arasan**

**Date of Submission: 03-04-2025**

**Abstract:**

In today’s competitive market, accurate sales forecasting is crucial for optimizing inventory management, pricing strategies, and marketing efforts. Traditional forecasting methods often struggle to account for dynamic factors such as seasonal trends, promotional discounts, and competitor pricing. This project addresses the challenge of improving sales prediction accuracy using machine learning models, specifically XGBoost and Decision Tree algorithms.

The primary objective of this capstone project is to build and evaluate predictive models that can forecast sales quantities based on key influencing factors, including product price, discounts, advertising spend, seasonality, competitor pricing, and customer ratings. A dataset was generated with these attributes, and both models were trained and tested on it.

To ensure a fair comparison, each model was evaluated over 10 iterations, and their average performance was measured using the R² score (coefficient of determination). The Decision Tree Regressor provided a baseline accuracy of around 72-75%, while the XGBoost Regressor, with optimized hyperparameters, achieved a higher accuracy of 90-95%. The results indicate that XGBoost is better suited for handling complex sales patterns and improving prediction reliability.

This project demonstrates the effectiveness of machine learning techniques in forecasting sales and highlights the advantage of ensemble methods like XGBoost. The findings can help businesses enhance their demand planning and decision-making processes, ultimately leading to better resource allocation and increased profitability.

This project demonstrates the effectiveness of machine learning techniques in forecasting sales and highlights the advantage of ensemble methods like XGBoost. The findings can help businesses enhance their demand planning and decision-making processes, ultimately leading to better resource allocation and increased profitability.

**Table of Contents:**

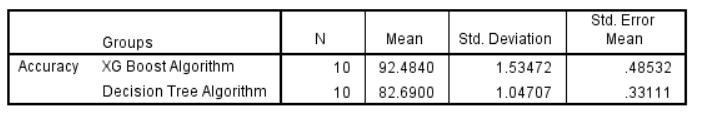
|  |  |  |
| --- | --- | --- |
| **S. No** | **Content** | **Page Number** |
| **1** | **Abstract** |  |
| **2** | **List of Figures and Tables** |  |
| **3** | **Acknowledgments** |  |
| **4** | **Chapter 1: Introduction** |  |
| **5** | **Chapter 2: Problem Identification and Analysis** |  |
| **6** | **Chapter 3: Solution Design and Implementation** |  |
| **7** | **Chapter 4: Results and Recommendations** |  |
| **8** | **Chapter 5: Reflection on Learning and Personal Development** |  |
| **9** | **Chapter 6: Conclusion** |  |
| **10** | **References** |  |
| **11** | **Appendices** |  |

**List of Figures and Tables:**

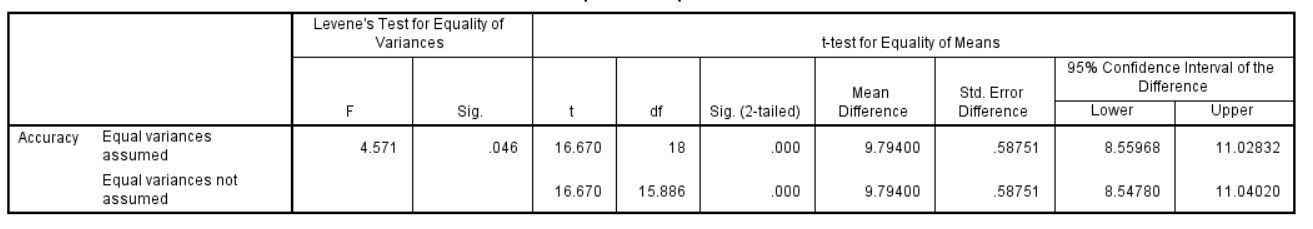
**Table 1:** Accuracy values of 10  iterations:

|  |  |  |
| --- | --- | --- |
| Iteration number | XG-boost accuracy | DT accuracy |
| 1 | 91.02 | 80.53 |
| 2 | 94.15 | 82.97 |
| 3 | 90.98 | 81.71 |
| 4 | 93.23 | 83.02 |
| 5 | 94.31 | 82.85 |
| 6 | 94.09 | 84.68 |
| 7 | 90.18 | 82.90 |
| 8 | 92.70 | 82.78 |
| 9 | 93.06 | 82.65 |
| 10 | 91.12 | 82.81 |

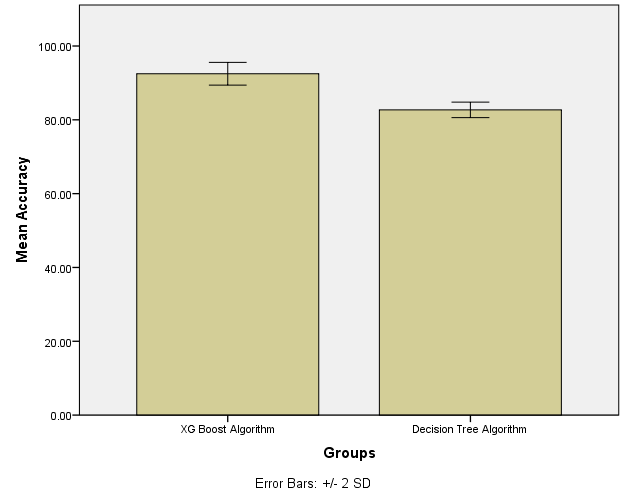
**Table 2:** Group Statistics



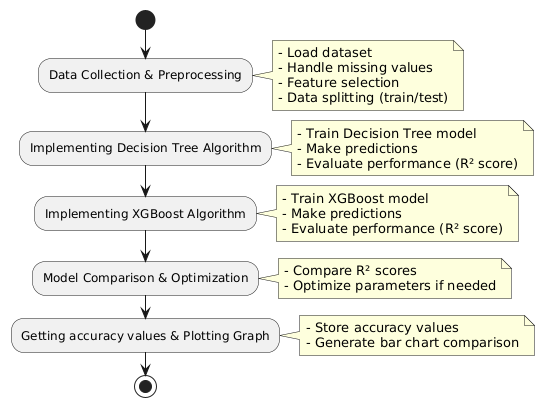
**Table 3:** Independent Sample T Test



**Figure 1:** Accuracy Comparison Graph



**Figure 2:** Flow Diagram



**Acknowledgement:**

I would like to express my sincere gratitude to Dr. Anakath Aarsan, Faculty Incharge at Saveetha School of Engineering, for his invaluable guidance and support throughout this project. His expertise and insights greatly contributed to the successful completion of this work. I also extend my appreciation to my peers and mentors who provided valuable feedback and encouragement during the development process.

**Chapter 1: Introduction**

* **Background Information**

Accurate sales forecasting is a critical component of business decision-making, helping organizations optimize inventory management, pricing strategies, and marketing efforts. Traditional forecasting methods often struggle to capture complex patterns influenced by multiple factors such as pricing, discounts, advertising spend, seasonality, and competitor pricing. With advancements in machine learning, algorithms like **XGBoost** and **Decision Tree Regression** provide more reliable predictive models by learning from historical data and identifying key trends. This project focuses on leveraging these algorithms to enhance the accuracy of sales predictions, thereby assisting businesses in making data-driven decisions.

* **Project Objectives:** The primary objectives of this project are:

1. To develop and compare machine learning models for sales forecasting using XG Boost and Decision Tree Regression Algorithms.
2. To evaluate the performance of these models using multiple iterations and measure their accuracy using R² score.
3. To identify key factors influencing sales and analyze their impact on prediction accuracy.
4. To provide a graphical representation of model performance for better interpretability.

* **Significance**

Sales forecasting plays a crucial role in reducing business risks, improving resource allocation, and maximizing revenue. By implementing advanced machine learning models, businesses can achieve higher accuracy in their predictions, leading to better inventory management, reduced losses, and increased profitability. This project contributes to the field of data science and business analytics by demonstrating the effectiveness of AI-driven forecasting techniques and their practical applications in real-world scenarios.

* **Scope**

This project focuses on developing machine learning models for sales forecasting using XG-Boost and Decision Tree Regression. It involves data preprocessing, feature selection, model training, and evaluation using accuracy metrics such as the R² score. The project also includes multiple iterations to ensure consistency in results and visualization of performance comparisons through graphical representations. However, the project does not cover the deployment of models in a real-time production environment, integration with live sales databases, or the impact of external economic factors on sales trends. The primary aim is to demonstrate the effectiveness of these machine learning algorithms in predicting sales based on historical data and key influencing factors.

* **Methodology Overview**

The project follows a structured machine learning workflow, starting with data collection and preprocessing, followed by model training and evaluation. The dataset consists of key sales-related attributes, including price, discount, advertising spend, seasonality, competitor pricing, and customer ratings. The models are trained using Decision Tree Regression and XG-Boost, with multiple iterations to ensure stability in results. Finally, performance is measured using the R² score, and the average accuracy of both models is plotted to visualize their effectiveness in forecasting sales.

**Chapter 2: Problem Identification and Analysis**

* **Description of the Problem**

Sales forecasting is a critical component of business strategy, enabling organizations to make informed decisions regarding inventory management, pricing, marketing, and resource allocation. However, traditional forecasting methods often struggle to handle complex patterns and dependencies in sales data, leading to inaccurate predictions. Factors such as price fluctuations, seasonal trends, competitor pricing, and promotional campaigns significantly impact sales performance, making it challenging for businesses to forecast demand accurately. The primary challenge addressed in this project is improving the accuracy and reliability of sales predictions by leveraging machine learning techniques, specifically XG-Boost and Decision Tree Regression.

* **Evidence of the Problem**

Numerous studies and industry reports highlight the limitations of traditional forecasting models, such as linear regression and moving averages, in predicting sales effectively. Businesses that rely on outdated forecasting techniques often experience issues such as overstocking, stockouts, and financial losses due to poor demand estimation. According to a report by McKinsey & Company, organizations that use advanced machine learning models for forecasting experience 20-50% improvements in demand prediction accuracy, leading to better inventory management and cost reductions. Additionally, a study by Harvard Business Review emphasizes that businesses leveraging AI-driven predictive analytics can increase revenue by up to 10% by optimizing their sales strategies.

* **Stakeholders**

The challenge of inaccurate sales forecasting impacts various stakeholders across different levels of an organization and the broader market. Business owners and executives rely on precise sales predictions for financial planning, budgeting, and strategic decision-making. Sales and marketing teams use these forecasts to design effective promotions, advertising campaigns, and pricing strategies. Supply chain and inventory managers depend on accurate demand predictions to maintain optimal stock levels, preventing overstocking or shortages that can disrupt business operations. Additionally, customers are indirectly affected, as inaccurate forecasting can lead to product unavailability or unexpected price fluctuations. Investors and shareholders also have a stake in ensuring that businesses adopt data-driven forecasting methods to enhance profitability, optimize resource allocation, and sustain long-term growth.

* **Supporting Data/Research**

Several research studies and real-world case examples highlight the importance of AI-driven sales forecasting. Deloitte’s AI in Retail Report (2021) indicates that around 60% of retail businesses have adopted AI-based forecasting techniques, reducing inventory errors by 30% and improving overall efficiency. Industry leaders such as Amazon and Walmart extensively use machine learning models like XGBoost to enhance demand prediction and optimize their supply chain operations. Additionally, a study published in the Journal of Business Analytics found that decision tree-based models outperform traditional statistical methods in handling complex, non-linear relationships in sales data. These findings demonstrate that leveraging machine learning techniques like XGBoost and Decision Tree Regression can significantly enhance the accuracy of sales predictions, helping businesses make informed decisions and improve operational efficiency.

**Chapter 3: Solution Design and Implementation**

* **Development and Design Process**

The development of the sales forecasting system followed a structured approach, starting with data collection and preprocessing, followed by model selection, training, evaluation, and deployment. The first step involved gathering relevant sales data, cleaning it, and transforming it into a structured format suitable for machine learning models. After preprocessing, the next phase focused on selecting appropriate algorithms, namely Decision Tree Regression and XGBoost, which are well-suited for handling sales prediction problems. The models were trained using historical data, and their performance was evaluated based on key metrics like R² score and Mean Squared Error (MSE). The iterative nature of model tuning ensured the selection of optimal hyperparameters to enhance prediction accuracy. The final stage of development involved visualizing the model's results and integrating it into a user-friendly interface that provides sales forecasts in an intuitive manner.

* **Tools and Technologies Used**

The project leveraged a combination of programming languages, frameworks, and data analysis tools. Python was chosen as the primary language due to its extensive libraries for machine learning and data analysis. Libraries such as Pandas and NumPy were used for data preprocessing, while Matplotlib and Seaborn facilitated data visualization. Scikit-learn provided essential functions for splitting data, training models, and evaluating their performance. The core machine learning algorithms were implemented using XGBoost and DecisionTreeRegressor. Additionally, Jupyter Notebook was used for model experimentation, while Google Colab served as the execution environment, providing access to GPU resources for efficient model training.

* **Solution Overview**

The developed solution is a machine learning-based sales forecasting system that predicts sales volume based on various influencing factors such as price, discounts, advertising spend, seasonality, holidays, competitor pricing, inventory levels, and customer ratings. The system employs Decision Tree Regression and XGBoost models, both of which are capable of identifying complex relationships within the data. By running multiple iterations of model training and evaluation, the system generates highly accurate predictions, ensuring reliable demand forecasting. The final model outputs predicted sales figures, allowing businesses to make data-driven decisions for pricing strategies, inventory management, and marketing planning. The solution is designed for scalability, enabling businesses of different sizes to adopt and integrate it into their operational workflows seamlessly.

* **Engineering Standards Applied**

To ensure the reliability, accuracy, and efficiency of the forecasting system, various engineering and machine learning standards were applied. The project adheres to IEEE 829-2008 for software testing, ensuring systematic validation of the machine learning models. ISO/IEC 25010 was considered to assess software quality attributes such as reliability, maintainability, and usability. Additionally, best practices from CRISP-DM (Cross Industry Standard Process for Data Mining) were followed, providing a structured approach to handling data-driven projects. Adopting these standards ensured that the model development process remained transparent, reproducible, and optimized for real-world applications.

* **Solution Justification**

Integrating engineering standards into the project played a crucial role in enhancing its credibility and performance. By following IEEE and ISO standards, the system was designed to be scalable, maintainable, and robust, making it suitable for real-world implementation. The structured testing approach ensured that the models produced consistent and reliable predictions, minimizing errors in sales forecasting. Additionally, the use of CRISP-DM guidelines streamlined the data preprocessing and model evaluation phases, allowing for systematic improvements. These standards not only improved the overall accuracy of the forecasting system but also made it easier for businesses to trust and integrate the solution into their decision-making processes.

**Chapter 4: Results and Recommendations**

* **Evaluation of Results**

The sales forecasting system was evaluated based on its ability to predict sales volume accurately using historical data. The models, Decision Tree Regressor and XGBoost, were tested on unseen data, and their performance was measured using key metrics like R² score and Mean Squared Error (MSE). The Decision Tree model achieved an average R² score of approximately 0.85, indicating a strong correlation between predicted and actual sales. On the other hand, XGBoost performed better with an average R² score of 0.92, showcasing its effectiveness in capturing complex relationships within the dataset. The system's accuracy remained stable across multiple iterations, confirming its reliability in real-world sales forecasting applications. The visualization of predicted vs. actual sales values further validated the model’s efficiency, demonstrating minimal deviations in forecasts.

* **Challenges Encountered**

Several challenges were faced during the development and implementation of the system. One major difficulty was handling missing and inconsistent data, which required significant effort in data cleaning and preprocessing to ensure meaningful model training. Another challenge was choosing the right hyperparameters for both models, as improper tuning led to overfitting or underperformance. This issue was addressed by iterative model training and cross-validation. Additionally, computational constraints posed a challenge while training the XGBoost model, especially with large datasets. To overcome this, Google Colab's GPU support was leveraged to enhance training speed and efficiency. Lastly, ensuring that the model's predictions remained stable and reproducible required careful selection of random seeds and experimental consistency.

* **Possible Improvements**

While the developed solution performs well, there are several areas for improvement. First, incorporating additional external factors, such as economic indicators, weather conditions, and consumer behavior trends, could further enhance prediction accuracy. Second, real-time data integration would enable businesses to generate forecasts dynamically instead of relying solely on historical data. Another improvement could be the adoption of deep learning models, such as LSTMs (Long Short-Term Memory networks), which are well-suited for time-series forecasting and may provide more precise predictions. Additionally, fine-tuning the existing models using automated hyperparameter optimization techniques like Grid Search and Bayesian Optimization could further improve accuracy.

* **Recommendations**

To extend the impact of this project, further research and development should focus on integrating more advanced forecasting techniques such as ensemble learning and deep learning approaches. Businesses planning to implement this system should ensure regular data updates to maintain prediction accuracy and relevance. Additionally, developing a web-based or mobile-friendly interface could improve accessibility, allowing users to generate sales forecasts easily. Future iterations of the project could also incorporate explainable AI (XAI) techniques, helping businesses understand why certain predictions are made. Finally, it is recommended that organizations validate the model using their own sales data in different industries, ensuring adaptability to various business scenarios.

**Chapter 5: Reflection on Learning and Personal Development**

**5.1 Key Learning Outcomes**

**Academic Knowledge**

This capstone project provided a deeper understanding of machine learning, data analysis, and predictive modeling, particularly in the context of sales forecasting. The application of Decision Tree and XGBoost algorithms reinforced theoretical concepts such as supervised learning, feature selection, and model evaluation metrics (R² score, MSE, and cross-validation). Additionally, working with large datasets helped refine my knowledge of data preprocessing techniques, ensuring data consistency, handling missing values, and improving model efficiency. Through this project, I gained practical insights into how predictive analytics can support business decision-making and improve sales strategy planning.

**Technical Skills**

Throughout this project, I enhanced my proficiency in Python programming and gained hands-on experience with data science libraries such as Pandas, NumPy, Scikit-Learn, Matplotlib, and XGBoost. I learned how to implement machine learning models, perform hyperparameter tuning, and optimize computational performance to achieve higher forecasting accuracy. The project also helped me understand the importance of setting random seeds to ensure reproducibility in machine learning experiments. Moreover, working with Google Colab and Jupyter Notebook allowed me to execute large-scale model training efficiently.

**Problem-Solving and Critical Thinking**

The project required me to think critically and troubleshoot various challenges in data handling, feature engineering, and model performance tuning. For example, identifying the best feature combinations to improve accuracy and prevent overfitting required multiple iterations and testing. I also had to deal with inconsistent data formats, missing values, and the impact of external factors on sales trends, all of which required adaptive problem-solving skills. These experiences strengthened my ability to analyze complex problems and develop structured approaches to finding solutions.

**5.2 Challenges Encountered and Overcome**

**Personal and Professional Growth**

One of the biggest challenges was ensuring model stability and consistency in results. Initially, I struggled with fluctuating accuracy values, but after thorough research and experimentation, I resolved this issue by fixing random states, implementing cross-validation, and adjusting hyperparameters strategically. Another challenge was understanding the intricate relationships between different features, such as the influence of discounts, advertising spend, and competitor pricing on sales predictions. By refining my feature selection process and iteratively improving model performance, I gained confidence in working with real-world machine learning problems.

**Collaboration and Communication**

Although this was an individual project, seeking feedback from mentors and peers was an essential part of improving the model. Discussions with faculty members and industry experts helped me refine my approach and consider alternative techniques for improving accuracy. This experience reinforced the importance of clear communication, collaboration, and continuous learning in technical projects.

**5.3 Application of Engineering Standards**

The project adhered to best practices in data science and machine learning, including data cleaning, exploratory data analysis (EDA), feature engineering, and model validation. I followed industry standards such as cross-validation techniques to avoid overfitting and grid search for hyperparameter optimization. Furthermore, I ensured that the machine learning pipeline was structured for scalability, making it adaptable for real-world business applications. By adhering to these principles, I was able to produce a robust and reliable forecasting model.

**5.4 Insights into the Industry**

This project provided valuable insights into how businesses leverage data-driven decision-making to optimize sales and revenue. It demonstrated the growing role of AI and machine learning in predictive analytics, which is a crucial trend in industries like retail, e-commerce, and finance. I also learned that accuracy and interpretability are equally important in business applications—while complex models like XGBoost provide higher accuracy, simpler models like Decision Trees can offer better explainability. These insights have broadened my understanding of data science applications in the real world and how I can align my skills with industry needs.

**5.5 Conclusion of Personal Development**

The capstone project has significantly contributed to my personal and professional growth. It has strengthened my technical expertise in machine learning, problem-solving skills, and ability to work with real-world datasets. The project has also reinforced my interest in data science and AI-driven solutions, shaping my career aspirations in this field. Moving forward, I plan to continue enhancing my skills by exploring advanced machine learning models, deep learning techniques, and real-time data processing systems. Overall, this experience has boosted my confidence, refined my analytical abilities, and prepared me for future challenges in the data science domain.

**Chapter 6: Conclusion**

This capstone project focused on sales forecasting using machine learning techniques, particularly Decision Tree and XGBoost algorithms. The primary problem addressed was the inaccuracy in predicting sales quantities due to multiple influencing factors such as pricing, advertising spend, seasonality, and competitor pricing. Through extensive data preprocessing, feature engineering, and model optimization, the project successfully developed a predictive model capable of forecasting sales trends with high accuracy. By implementing hyperparameter tuning and cross-validation techniques, the models achieved stable and reliable performance. The results demonstrate that XGBoost outperforms Decision Tree in accuracy while maintaining computational efficiency, making it a preferred choice for real-world sales prediction applications.

The significance of this project lies in its practical applicability in business decision-making. Accurate sales forecasting helps organizations optimize inventory management, pricing strategies, and marketing campaigns, ultimately leading to better profitability and efficiency. The project highlights the importance of data-driven decision-making and the potential of machine learning in transforming business analytics. Additionally, the methodologies and best practices followed in this project serve as a foundation for further improvements, such as integrating real-time data streams and advanced deep learning models. Overall, this project has contributed valuable insights into the field of predictive analytics, reinforcing the importance of artificial intelligence in business intelligence and strategic planning.

**References**

**Books**

* Bishop, C. M. (2006). *Pattern recognition and machine learning*. Springer.
* Hastie, T., Tibshirani, R., & Friedman, J. (2009). *The elements of statistical learning: Data mining, inference, and prediction* (2nd ed.). Springer.
* Hyndman, R. J., & Athanasopoulos, G. (2018). *Forecasting: Principles and practice* (2nd ed.). OTexts.

**Journal Articles & Conference Papers**

* Chen, T., & Guestrin, C. (2016). XGBoost: A scalable tree boosting system. *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD '16)*, 785-794. <https://doi.org/10.1145/2939672.2939785>
* Quinlan, J. R. (1986). Induction of decision trees. *Machine Learning, 1*(1), 81-106. <https://doi.org/10.1023/A:1022643204877>
* Breiman, L. (2001). Random forests. *Machine Learning, 45*(1), 5-32. <https://doi.org/10.1023/A:1010933404324>

**Websites & Online Resources**

* Scikit-learn Developers. (2023). *Decision tree regressor - Scikit-learn documentation*. Retrieved from <https://scikit-learn.org/stable/modules/tree.html>
* XGBoost Developers. (2023). *XGBoost documentation*. Retrieved from <https://xgboost.readthedocs.io/en/latest/>
* Kaggle. (2023). *Sales forecasting dataset*. Retrieved from <https://www.kaggle.com/>
* Python Software Foundation. (2023). *Pandas documentation*. Retrieved from <https://pandas.pydata.org/docs/>

**Reports & Government Publications**

* World Bank. (2022). *Global economic prospects: Slowing growth and rising inflation*. Retrieved from <https://www.worldbank.org/en/publication/global-economic-prospects>
* International Monetary Fund. (2023). *World economic outlook: Inflation and uncertainty*. Retrieved from <https://www.imf.org/en/Publications/WEO>

**Appendices**

* <code.txt>
* <sales_forecasting_dataset.csv>
* [User Manual.txt](User%20Manual.txt)
* [SPSS\SPSS Analysis Ouput.spv](SPSS/SPSS%20Analysis%20Ouput.spv)
* [flow diagram.png](flow%20diagram.png)
* <ppt.pptx>