**Big Sales Prediction using Machine Learning**

**Submitted for**

**Statistical Machine Learning CSET211**

Submitted by:

**(E23CSEU1862) PRAKHAR SWAROOP**

**(E23CSEU1863) KARTIK CHAUHAN**

Submitted to

**Dr. Shakshi Sharma**

**July-Dec 2024**

**SCHOOL OF COMPUTER SCIENCE AND ENGINEERING**

A close-up of a logo

Description automatically generated

**INDEX**

|  |  |  |
| --- | --- | --- |
| S. No. | Content | Page No. |
| 1 | **Abstract** | 3 |
| 2 | **Introduction** | 3 |
| 3 | **Related Work (If Any)** | 4 |
| 4 | **Methodology** | 7 |
| 5 | **Hardware/Software Required** | 7 |
| 6 | **Experimental Results** | 8 |
| 7 | **Conclusions** | 9 |
| 8 | **Future Scope** | 9 |
| 9 | **GitHub Link** | 10 |

1. **Abstract**

**This project develops a machine learning model to predict outlet sales, leveraging Random Forest Regressor and XGBRF Regressor algorithms. Using historical sales data and features such as item type, location, and seasonality, the model provides accurate sales forecasts to support inventory and marketing decisions, enhancing overall operational efficiency and profitability.**

**2.Introduction**

**Sales prediction is a critical task in retail and business analytics, as it helps organizations optimize inventory, plan marketing strategies, and improve revenue forecasting. This project focuses on developing a sales prediction model that leverages machine learning techniques to predict the sales of items in an outlet based on historical data and relevant features.**

**The model employs two powerful regression algorithms: the XGBRF Regressor (eXtreme Gradient Boosting with Random Forest) and the Random Forest Regressor. Both algorithms are well-suited for handling complex relationships in data and work effectively with datasets that exhibit non-linearity and feature interactions.**

* **The XGBRF Regressor combines the strengths of gradient boosting and random forests, balancing feature selection with ensemble diversity. It is particularly useful for capturing subtle patterns in the data while mitigating overfitting.**
* **The Random Forest Regressor uses an ensemble of decision trees to provide robust predictions and offers high accuracy through bagging and averaging mechanisms.**

**By comparing these models, the project aims to identify the most accurate and efficient algorithm for predicting sales, enabling data-driven decision-making for retail optimization.**

**3.Related Work**

**Random Forest Regressor and XGBRF Regressor: Applications and Limitations**

**Random Forest Regressor**

**Random Forest Regressor (RFR) is a robust machine learning algorithm widely used for predicting continuous outcomes. It works by averaging predictions from an ensemble of decision trees, ensuring accuracy and reducing overfitting.**

**Applications:**

1. **Finance: Used for stock market predictions, credit scoring, and risk assessment by analyzing historical data and financial indicators.**
2. **Healthcare: Helps in disease prognosis and drug discovery by processing patient data and molecular interactions.**
3. **Environment: Predicts climate variables, pollution levels, and species habitats using large-scale environmental data.**
4. **Marketing: Facilitates demand forecasting and churn prediction by analyzing customer behavior.**
5. **Manufacturing: Supports predictive maintenance and defect detection to optimize production processes.**

**Drawbacks:**

* **Overfitting: Susceptible to noise, requiring preprocessing and parameter tuning.**
* **Interpretability: Complex to explain; tools like SHAP or LIME help but are limited.**
* **Computational Costs: High memory and time requirements for large datasets.**
* **Hyperparameter Sensitivity: Performance depends on careful tuning of parameters.**
* **Extrapolation Issues: Poor performance on data outside the training range.**

**XGBRF Regressor**

**XGBRF (XGBoost Random Forest) combines the gradient boosting power of XGBoost with the ensemble diversity of Random Forest, offering a hybrid approach for improved predictive accuracy.**

**Applications:**

1. **Finance: Predicts credit risk, loan defaults, and stock trends using historical and transactional data.**
2. **Healthcare: Analyzes clinical and genetic data for disease prediction and personalized medicine.**
3. **Retail: Powers recommendation systems and customer segmentation by leveraging transaction patterns.**
4. **Energy: Forecasts energy demand and schedules maintenance by analyzing consumption and sensor data.**
5. **Environment: Predicts air quality and ecological changes using integrated environmental datasets.**

**Drawbacks:**

* **Computational Complexity: Training is resource-intensive, especially for large datasets.**
* **Memory Usage: High memory requirements due to ensemble size.**
* **Interpretability: Difficult to explain; tools like SHAP and LIME offer partial insights.**
* **Overfitting: Prone to capturing noise in small or unbalanced datasets.**
* **Hyperparameter Sensitivity: Requires careful tuning to optimize performance.**
* **Extrapolation Issues: Struggles with predictions outside the training range.**

**4.Methodology:**

**Two machine learning models were implemented in this study using the scikit-learn library.The models used include Random forest regressor and XGBoost Regressor. The Random Forest Regressor works by creating multiple decision trees on different data subsets, averaging their predictions to improve accuracy and reduce overfitting. XGBoost Regressor uses gradient boosting, building trees sequentially with each tree correcting the previous errors, optimizing performance through regularization and efficient computation.**

**5.Hardware/Software Required**

* **Google Collab**
* **Scikit Learn**
* **Pandas**
* **MatplotLib**
* **Numpy**
* **seaborn**

**6.Experimental Results**

**The assessment on how well each of the model generalizes was done by finding the cross validation score for each of the two models.**

|  |  |
| --- | --- |
| **Model Name** | **Cross Validation Score** |
| **Random Forest Regressor** | **0.555040442702387** |
| **XGBRF Regressor** | **0.5951283683315813** |

**The predictions produced by each model are evaluated using mean absolute error, mean squared error and R-Squared error.**

**Mean Absolute Error (MAE): Measures the average magnitude of errors in predictions, without considering their direction. It’s calculated as the average of the absolute differences between predicted and actual values, giving a clear measure of prediction accuracy.**

**Mean Squared Error (MSE): Computes the average of the squared differences between predicted and actual values. Squaring amplifies larger errors, making MSE sensitive to large outliers, and thus useful for understanding overall error impact.**

**R² Score (Coefficient of Determination): Indicates the proportion of variance in the target variable explained by the model. It ranges from 0 to 1, where higher values represent better fit, and is used to assess model performance.**

|  |  |  |  |
| --- | --- | --- | --- |
| **Model Name** | **Mean Absolute Error** | **Mean Squared Error** | **R-squared Error** |
| **Random Forest Regressor** | **840.9972969527988** | **1484352.251595246** | **0.4538746876477048** |
| **XGBRF Regressor** | **714.6033522105251** | **1047393.4927614973** | **0.6146412701059387** |

**7.Conclusion**

**In this project, we applied both Random Forest and XGBRF Regressors to predict outlet sales, comparing their performance. XGBRF outperformed Random Forest by capturing complex sales patterns more accurately, thanks to its boosting approach, making it a more effective model for forecasting in retail analytics.**

**8.Future Scope**

**The future scope of a sales prediction model leveraging Random Forest Regressor and XGBRF Regressor is highly promising, particularly in the retail and e-commerce sectors. These models can be used to optimize inventory management, dynamic pricing, and targeted marketing by accurately forecasting sales trends. Their robustness in handling non-linear relationships and feature importance insights enable businesses to identify key sales drivers and customer behaviors. Furthermore, these models can be integrated into real-time decision-making systems, helping businesses respond promptly to changing market conditions. With advancements in big data and cloud computing, these models can process larger datasets with greater complexity, paving the way for more precise and scalable solutions. This capability is crucial for businesses aiming to stay competitive in a data-driven marketplace.**

**9.Github Link of Our Project**

**https://github.com/117Kartik/Machine-Learning-Project**