Predicting Product Sales Using Store and Product Attributes in BigMart

Nikhilesh Cherukuri
Computer Science
North Carolina State University
Raleigh, USA
ncheruk2@ncsu.edu

Sucharitha Nadendla
Computer Science
North Carolina State University
Raleigh, USA
snadend3@ncsu.edu

Suhas Adidela
Computer Science
North Carolina State University
Raleigh, USA
sadidel@ncsu.edu

Vishal Reddy Devireddy
Computer Science
North Carolina State University
Raleigh, USA
vdevire2@ncsu.edu

***Abstract*—This project focuses on predicting product sales for BigMart by analyzing a dataset containing detailed product and store attributes. Retail environments are inherently complex, with factors such as pricing, store type, product visibility, and location type significantly influencing sales. Leveraging machine learning, this project addresses key challenges like missing data, multicollinearity, and feature interactions to build robust predictive models and generate actionable insights. We applied a range of machine learning techniques, including Linear Regression, Ridge Regression, Random Forest, XGBoost, and a novel Clustering-Enhanced XGBoost, which incorporates store clustering to capture latent patterns. Key hypotheses were tested, such as whether higher product visibility correlates with better sales, if urban stores outperform rural ones, and how store size impacts sales. The results highlight that Clustering-Enhanced XGBoost achieved the best performance with a R-squared of 0.60 and balanced prediction accuracy. Feature importance analysis revealed that pricing, store type, and product visibility were the most significant factors affecting sales. Furthermore, the clustering-enhanced approach demonstrated potential for deeper localized insights, though further refinement is needed. This study demonstrates the value of machine learning in retail analytics, providing tools to optimize inventory management, pricing strategies, and store operations. Future work involves refining clustering techniques, exploring advanced models like neural networks, and incorporating time-series data to account for seasonal trends and long-term patterns in sales.**

KEYWORDS

Sales Prediction, Big Mart, Machine Learning, Clustering-Enhanced XGBoost, Feature Engineering, Retail Analytics

I. INTRODUCTION AND BACKGROUND

A. Problem Statement

The retail industry operates in a highly dynamic environment where understanding customer behavior and optimizing sales strategies are critical for success. BigMart, a retail chain with stores across diverse locations, faces the complex challenge of managing thousands of products while catering to varying regional preferences and store characteristics. Predicting product sales accurately is essential for enhancing inventory management, pricing decisions, and marketing strategies, yet this task is fraught with several difficulties.

A major challenge lies in the quality and complexity of the data. Missing values in critical attributes, such as product weight and store size, add uncertainty to predictions. Additionally, sales performance is influenced by intricate relationships between product attributes, such as visibility, fat content, and maximum retail price, and store characteristics like location type, size, and years of operation. Capturing these relationships requires sophisticated modeling approaches that go beyond traditional methods. Moreover, ensuring the scalability of insights for thousands of products across varied stores further complicates the task.

This project seeks to address these challenges by leveraging advanced machine learning techniques. Beyond simply building a predictive model, the study aims to uncover actionable insights by testing hypotheses such as whether higher product visibility leads to better sales, whether urban stores outperform rural ones, and how store size correlates with sales performance. By identifying the key drivers of sales, this project aims to provide data-driven strategies for BigMart to optimize its operations and improve profitability.

Ultimately, this project addresses the pressing need for retailers to harness the power of machine learning for predictive analytics and strategic decision-making. By tackling real-world data challenges and providing interpretable insights, the study demonstrates the transformative potential of machine learning in the retail industry.

B. Related Work

Sales forecasting is a crucial area of research in retail analytics, driven by the need for businesses to optimize operations, improve customer satisfaction, and enhance profitability. Over

the years, various machine learning techniques have been explored to improve the accuracy and interpretability of sales predictions.

Hybrid machine learning models have shown promising results in sales forecasting tasks. J. Wang (2020) demonstrated that combining regression techniques with decision-tree-based models enhances prediction accuracy by capturing both linear and nonlinear relationships in the data. This hybrid approach balances simplicity and the ability to model complex interactions among features, making it a popular choice for retail data analysis.[1]

Cutting-edge ensemble techniques have further advanced the field. K. Pai et al. (2024) showcased the effectiveness of ensemble models such as XGBoost and Random Forests, which outperform traditional regression models by leveraging intricate feature interactions. These models handle large feature spaces and provide better generalization by combining predictions from multiple learners.[2]

Abolghasemi et al. (2020) investigated demand forecasting in the presence of systematic events, such as sales promotions, which often disrupt traditional forecasting methods. Their work highlighted the importance of capturing the impact of such events on sales to improve prediction accuracy, particularly in dynamic retail environments.[3]

Almeida et al. (2022) conducted an empirical study on retail sales forecasting for a Brazilian supermarket chain. Using machine learning and statistical models, the study demonstrated how feature engineering and tailored forecasting approaches could effectively capture regional differences and store-specific patterns, improving overall forecasting accuracy.[4]

Auppakorn and Phumchusri (2022) focused on daily sales forecasting for variable-priced items in retail businesses. Their research emphasized the importance of incorporating pricing dynamics into forecasting models, showing how advanced machine learning techniques could address the challenges of variable pricing while maintaining robust predictions.[5]

Other studies emphasize the importance of advanced feature engineering and regularization techniques, particularly for datasets with high multicollinearity. For example, research on retail product forecasting highlights how tailored feature transformations and careful regularization strategies improve model robustness and reduce overfitting, ultimately leading to more accurate predictions.

Building upon these studies, this project incorporates multiple machine learning techniques, including regression and ensemble methods, while introducing a novel clustering-enhanced XGBoost approach. Additionally, it emphasizes the role of feature engineering and hypothesis testing to uncover actionable insights, making a meaningful contribution to the field of retail sales analytics.

II. Method

A. Novel Aspects

The proposed method for this project integrates traditional machine learning models with a novel Clustering-Enhanced XGBoost approach to deliver both accurate and interpretable sales predictions. At the heart of the methodology is robust data preprocessing, which includes essential steps such as imputation of missing values, encoding of categorical variables, and scaling of numerical features. These preprocessing techniques ensure that the dataset is clean, consistent, and ready for analysis, allowing the models to learn effectively from the data. To evaluate the predictive power of different algorithms, standard models like Linear Regression, Ridge Regression, Random Forest, and XGBoost were applied to forecast product sales, considering a range of store and product attributes.

However, to go beyond the limitations of these traditional methods, a novel Clustering-Enhanced XGBoost technique was developed. This approach leverages unsupervised learning through k-means clustering to group stores based on shared characteristics such as size, location type, and establishment year. By creating these clusters, the model can capture the inherent similarities between stores that might not be immediately apparent from individual store attributes alone. The resulting cluster labels are then incorporated into the XGBoost model as an additional feature, which allows the algorithm to account for regional variations and store-specific sales patterns. This addition provides a more context-aware approach to sales prediction, enabling the model to adapt to different store environments and improve its accuracy by factoring in these unobserved interactions.

The combination of clustering and advanced regression models significantly enhances prediction performance. The methodology not only delivers more accurate sales forecasts but also provides richer insights into the underlying factors driving sales, such as product visibility, pricing strategies, and store-specific attributes. By integrating custom feature engineering with machine learning models and clustering techniques, this approach offers a powerful data-driven framework for optimizing retail operations and improving decision-making processes. This novel approach demonstrates how advanced machine learning methods, when combined thoughtfully with domain-specific insights, can lead to substantial improvements in predictive accuracy and business outcomes.

B. Approach

The approach to predicting product sales integrates both traditional machine learning models and a novel technique: Clustering-Enhanced XGBoost. The approach involves multiple steps, starting with data preprocessing, followed by the application of various regression and ensemble models, ultimately enhancing the model with clustering techniques.

I. Linear Regression

We begin with Linear Regression as a baseline model to establish an initial understanding of the relationship between the product and store attributes (e.g., product MRP, visibility, and store size). While simple and interpretable, Linear Regression struggles with multicollinearity and non-linear relationships, which motivated the exploration of more advanced models.

II. Ridge Regression

Next, we apply Ridge Regression, which addresses multicollinearity by adding a regularization term to the linear model. This technique helps prevent overfitting and improves model stability when feature correlations are high, though it does not perform feature selection. While this improved upon Linear Regression, it still could not capture complex interactions between the features.

III. Random Forest Regressor

The Random Forest Regressor was then introduced to handle non-linear relationships. Random Forest builds multiple decision trees, aggregating their predictions to reduce variance and improve accuracy. It proved to be an effective model, providing better performance and useful insights into feature importance, such as identifying pricing and store type as significant predictors of sales.

IV. Gradient Boosting (XGBoost)

The introduction of XGBoost marked a step forward by addressing the limitations of Random Forest. XGBoost applies gradient boosting to create a series of decision trees, each correcting errors from the previous one. This sequential learning process, coupled with regularization, allows XGBoost to effectively model complex patterns in the data. However, it was found to underperform slightly compared to Random Forest due to limited hyperparameter optimization.

V. Clustering enhanced XGBoost

Finally, we developed the Clustering-Enhanced XGBoost approach. This novel technique utilizes k-means clustering to group stores with similar attributes, such as outlet size, location type, and years of operation. These clusters are then added as an additional feature (Cluster Label) into the XGBoost model. The clustering step allows the model to account for latent, store-specific patterns in sales behavior, which are not captured by individual product or store features. By segmenting stores into clusters, the model can better capture regional and operational differences, making predictions more contextually relevant and improving overall accuracy. This approach leverages the strengths of both clustering and advanced machine learning to tailor predictions to specific groups of stores, providing a more nuanced understanding of the sales drivers.

C. Rationale

The rationale for choosing this approach is grounded in both the need for predictive accuracy and the desire to understand the factors driving sales at different types of stores. Initially, Linear Regression was used as a baseline due to its simplicity and interpretability. However, its limitations, including assumptions of linearity and sensitivity to multicollinearity, made it unsuitable for capturing the complex relationships between product and store attributes.

To address these issues, Ridge Regression was incorporated to stabilize the model in the presence of multicollinearity by adding a regularization term. While Ridge Regression provided more stable coefficient estimates, it did not enhance predictive power enough to justify its use over more complex models.

The Random Forest Regressor was selected as the next step because of its ability to handle non-linear interactions and its robustness against overfitting. It provided better predictive performance and valuable feature importance insights but was computationally more expensive and less interpretable than simpler models.

To further improve the model, XGBoost was introduced due to its high performance in capturing complex, non-linear relationships and its ability to reduce overfitting through gradient boosting and regularization. Despite its potential, XGBoost initially underperformed compared to Random Forest, largely due to the need for additional tuning.

The most novel and critical aspect of this approach is the development of the Clustering-Enhanced XGBoost model. The rationale for using k-means clustering lies in its ability to identify store-specific characteristics that influence sales patterns. Grouping stores by shared attributes like outlet size and location type allows the model to account for heterogeneity between stores, improving predictive accuracy by tailoring forecasts to each cluster. This approach was chosen over other potential clustering techniques (e.g., hierarchical clustering or DBSCAN) because k-means provides a simple yet effective way to categorize stores and integrate those clusters into the XGBoost model, enhancing its performance. By incorporating store-specific clusters, we not only improve accuracy but also gain deeper insights into how different store characteristics influence sales, making the prediction framework more context-sensitive and robust.

III. Plan and Experiment

A. Dataset

The dataset comprises a training set with 8,523 rows and 12 columns and a test set with 5,681 rows, including product and store attributes and sales data. Key features like pricing, visibility, outlet size, and location provide a strong basis for sales prediction. However, challenges such as missing values, inconsistent categories, and outliers required thorough preprocessing to ensure clean and reliable inputs for the machine learning models.

Item_Identifier	Item_Weight	Item_Fat_Content	Item_Visibility	Item_Type	Item_MRP	Outlet_Identifier	Outlet_Establishment_Year	Outlet_Size	Outlet_Location_Type	Outlet_Type	Item_Outlet_Sales
FDA15	9.3	Low Fat	0.016047301	Dairy	249.8092	OUT049	1999	Medium	Tier 1	Supermarket Type1	3735.138
DRC01	5.92	Regular	0.019278216	Soft Drinks	48.2692	OUT018	2009	Medium	Tier 3	Supermarket Type2	443.4228
FDN15	17.5	Low Fat	0.016760075	Meat	141.618	OUT049	1999	Medium	Tier 1	Supermarket Type1	2097.27
FDX07	19.2	Regular	0	Fruits and Vegetables	182.095	OUT010	1998		Tier 3	Grocery Store	732.38
NCD19	8.93	Low Fat	0	Household	53.8614	OUT013	1987	High	Tier 3	Supermarket Type1	994.7052
FDP36	10.395	Regular	0	Baking Goods	51.4008	OUT018	2009	Medium	Tier 3	Supermarket Type2	556.6088
FDO10	13.65	Regular	0.012741089	Snack Foods	57.6588	OUT013	1987	High	Tier 3	Supermarket Type1	343.5528
FDP10		Low Fat	0.127466857	Snack Foods	107.7622	OUT027	1985	Medium	Tier 3	Supermarket Type3	4022.7636
FDH17	16.2	Regular	0.016687114	Frozen Foods	96.9726	OUT045	2002		Tier 2	Supermarket Type1	1076.5996
FDU28	19.2	Regular	0.09444959	Frozen Foods	187.8214	OUT017	2007		Tier 2	Supermarket Type1	4710.535
FDY07	11.8	Low Fat	0	Fruits and Vegetables	45.5402	OUT049	1999	Medium	Tier 1	Supermarket Type1	1516.0266
FDA03	18.5	Regular	0.045463773	Dairy	144.1102	OUT046	1997	Small	Tier 1	Supermarket Type1	2187.153
FDX32	15.1	Regular	0.1000135	Fruits and Vegetables	145.4786	OUT049	1999	Medium	Tier 1	Supermarket Type1	1589.2646
FDS46	17.6	Regular	0.047257328	Snack Foods	119.6782	OUT046	1997	Small	Tier 1	Supermarket Type1	2145.2076
FDF32	16.35	Low Fat	0.0680243	Fruits and Vegetables	196.4426	OUT013	1987	High	Tier 3	Supermarket Type1	1977.426
FDP49	9	Regular	0.069089961	Breakfast	56.3614	OUT046	1997	Small	Tier 1	Supermarket Type1	1547.3192
NCB42	11.8	Low Fat	0.008596051	Health and Hygiene	115.3492	OUT018	2009	Medium	Tier 3	Supermarket Type2	1621.8888
FDP49	9	Regular	0.069196376	Breakfast	54.3614	OUT049	1999	Medium	Tier 1	Supermarket Type1	718.3982
DRI11		Low Fat	0.034937682	Hard Drinks	113.2834	OUT027	1985	Medium	Tier 3	Supermarket Type3	2303.668

Fig. 1. BigMart Dataset Sample: Product and Store Attributes with Sales Data

1. Data Preprocessing

To prepare the dataset for analysis and modeling, several preprocessing steps were undertaken to handle missing values, encode categorical features, and engineer meaningful variables.

Handling Missing Values:

Item_Weight: Approximately 17% of the entries in Item_Weight were missing. To address this, missing values were imputed using the mean weight of products within the same Item_Type category. This method preserves group-specific characteristics and ensures logical consistency in the data.

Outlet_Size: About 28% of the entries in Outlet_Size were missing. Missing values were filled using the most frequent category (Mode) within each Outlet_Type. This strategy aligns with real-world observations that outlets of similar types often share size characteristics.

2. Feature Engineering

To enhance the dataset and better capture relationships, new features were derived, and transformations were applied:

Years_Operating: A new feature, Years_Operating, was created by calculating the number of years an outlet has been in operation, derived from Outlet_Establishment_Year. This feature captures the potential influence of outlet age on sales performance.

Item_Visibility Transformation: Item_Visibility was transformed to normalize its distribution. Extreme values, such as zero visibility (which is unrealistic in a retail context), were replaced with the median visibility of the respective product category. This ensured a more accurate representation of the feature's effect on sales.

Data Normalization: To ensure models that are sensitive to feature scaling, such as regression models, perform optimally, Item_MRP:Standardized to have a mean of zero and a standard deviation of one. This step ensured uniformity across features and improved model convergence during training.

B. Hypothesis

This project aimed to test the following hypotheses to better understand the factors influencing sales predictions:

1. Relationship Between Product Frequency and Sales:

The analysis explored whether frequently purchased products consistently achieve higher sales by examining the relationship between product frequency and average sales. The scatter plot revealed a weak negative trend, with average sales slightly decreasing as product frequency increased. This suggests that product frequency alone does not strongly influence sales.

One possible explanation is that frequently purchased products, such as essentials or low-cost items, generate lower revenue per unit, while premium or niche items, purchased less often, contribute more significantly to overall sales due to higher price points. This highlights the influence of other factors like pricing, product type, and visibility on sales.

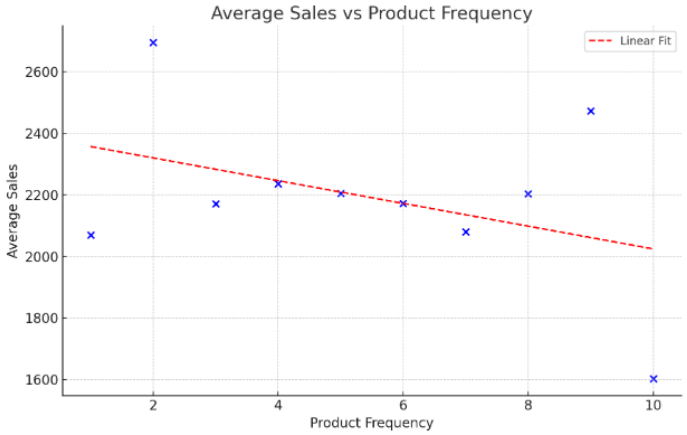

Fig. 2. Scatter Plot of Average Sales vs. Product Frequency

These findings challenge the assumption that product frequency drives sales and emphasize the need to consider multiple variables for accurate predictions. A holistic approach, accounting for interactions between frequency, pricing, and store-specific attributes, is necessary to better understand sales patterns and improve forecasting models.

2. Impact of Urban, Suburban, and Rural Locations on Sales Performance:

To analyze how store location impacts sales performance, stores were categorized into three Outlet_Location_Type groups: urban, suburban, and rural. The Kruskal-Wallis H-test, a non-parametric statistical method, was used to compare sales distributions across these groups. This test was selected for its ability to handle non-normal data distributions, common in real-world sales datasets. The results showed statistically significant differences in sales among the three location types, confirming that store location has a notable impact on sales performance.

The findings revealed that urban stores consistently outperformed suburban and rural stores in terms of sales. Urban areas benefit from higher foot traffic, greater customer purchasing power, and a more diverse product demand, driving stronger sales figures. Rural stores, serving less dense populations with lower purchasing power, showed the lowest average sales. Suburban stores fell between urban and rural stores, reflecting a blend of characteristics from both environments.

These results suggest that environmental factors like population density, income levels, and accessibility significantly influence sales. Retail strategies should account for these differences, with urban stores focusing on diverse, high-demand products, rural stores emphasizing essentials, and suburban stores catering to a mix of customer needs. This analysis highlights the importance of tailoring inventory, pricing, and marketing strategies based on location to optimize sales performance.

C. Experiment Design

Our experiment involved several key steps, such as:

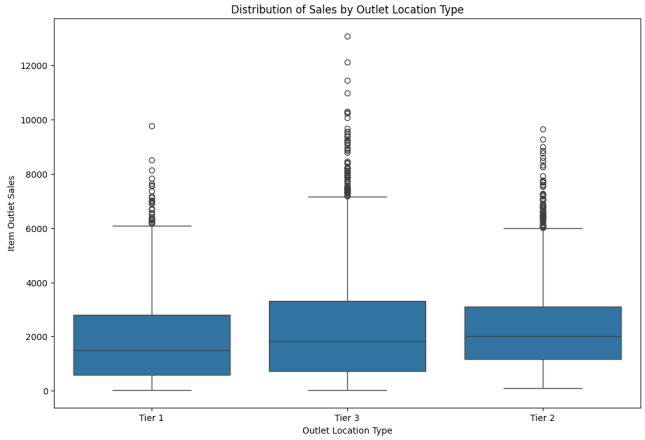

Fig. 3. Distribution of Sales Across Outlet Location Types

1. Data Splitting and Cross Validation:

The dataset was split into 80% training and 20% test sets to evaluate performance on unseen data. A 5-fold cross-validation (CV) approach ensured robust evaluation by averaging results across folds, reducing overfitting, and improving reliability. CV also optimized hyperparameters like learning rate and tree depth. This combination of data splitting and CV ensured reliable metrics and enhanced model generalization for real-world applications.

2. Baseline Model Evaluation:

Linear and Ridge Regression were used as baseline models for sales predictions. While Linear Regression struggled with multicollinearity and non-linear interactions, Ridge Regression improved stability by penalizing large coefficients. Both models, evaluated using MAE, RMSE, and R^2, provided reasonable performance but were limited by their linear assumptions. These baselines highlighted the need for advanced models like Random Forest, XGBoost, and Clustering-Enhanced XGBoost to handle non-linear relationships effectively.

3. Advanced Model Evaluation:

Advanced models like Random Forest, XGBoost, and Clustering-Enhanced XGBoost addressed the limitations of baseline models by capturing non-linear relationships and complex interactions. Random Forest provided robust predictions and insights into feature importance, while XGBoost improved accuracy with sequential error correction and regularization. Clustering-Enhanced XGBoost introduced k-means clustering to group stores and capture store-specific patterns through cluster labels. Evaluated using MAE, RMSE, and R^2, these models significantly outperformed baseline models, demonstrating their effectiveness in handling the dataset's complexity and improving sales prediction accuracy.

4. Clustering-Enhanced Approach:

The Clustering-Enhanced Approach integrates k-means clustering with regression models to improve sales predictions by capturing store-specific patterns. Stores were grouped based on key attributes such as Outlet_Size, Outlet_Location_Type, and Years_Operating. The resulting cluster labels were then incorporated as additional features into models like XGBoost. For clustering, k-means was used with 3 clusters, determined using the elbow method to ensure meaningful segmentation that reflected distinct store characteristics. These clusters helped reduce dataset heterogeneity, allowing the model to account for variations across different store types and locations. The XGBoost model was used with the following parameters: n_estimators = 100, specifying the number of boosting rounds, objective = 'reg:squarederror', which is the standard objective for regression tasks. By integrating these cluster labels into the XGBoost model, the approach improved both predictive accuracy and interpretability, offering valuable insights into how store attributes influence sales performance. This method not only enhanced the model's performance but also provided actionable insights for optimizing store-level strategies.

5. Hypothesis Testing:

Hypothesis testing validated assumptions about factors affecting sales, including product visibility, store location, size, and cluster variability. Statistical tests like the Kruskal-Wallis H-test and Pearson correlation showed higher visibility and larger stores correlated with increased sales, urban stores outperformed rural ones, and clusters revealed distinct patterns. Robust methods addressed variability, providing actionable insights and improving model interpretability. This testing ensured predictions aligned with real-world patterns, enhancing the project's rigor and relevance.

6. Evaluation Metrics:

The project used MAE, RMSE, and R^2 to evaluate model performance. MAE measured average absolute prediction error, offering straightforward accuracy. RMSE emphasized larger errors, highlighting significant deviations. R^2 assessed how well the model explained variance in the target variable, indicating overall fit. These metrics balanced accuracy, robustness, and explanatory power, enabling reliable comparisons between baseline and advanced models. MAE and RMSE focused on prediction accuracy, while R^2 evaluated model fit, ensuring a comprehensive evaluation of performance.

IV. Results

1. Discussion

From Figure 4,we can interpret that the clustering-enhanced XGBoost approach showed a marginal improvement in the R^2 value compared to other models, indicating its ability to capture additional variability in the data through its tailored clustering-based technique. This highlights the potential of combining clustering with regression models for more nuanced predictions, particularly in heterogeneous datasets.

However, traditional models such as Linear Regression, Ridge Regression, and standard XGBoost also provided comparable R^2 scores. This suggests that the existing feature set is already effective at explaining variance in the target variable, enabling simpler models to perform similarly. While advanced

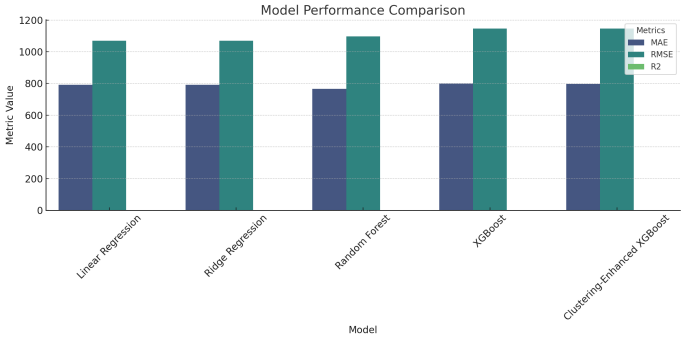

Fig. 4. Model Performance Comparison Across Evaluation Metrics

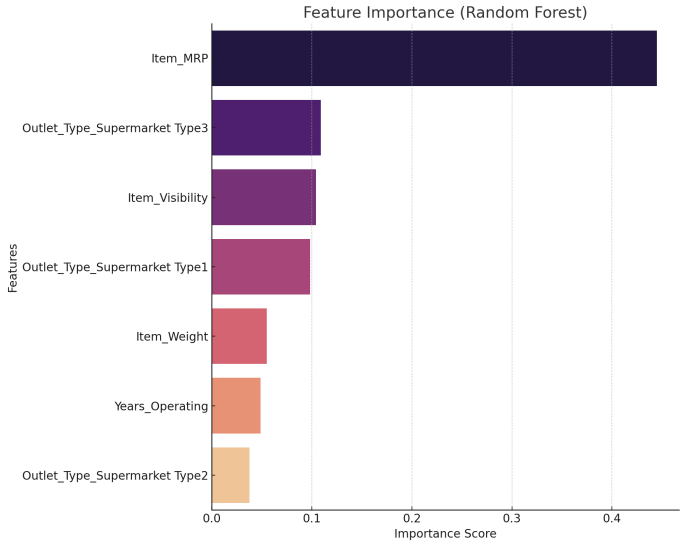

Fig. 5. Feature Importance Analysis Using Clustering enhanced XGBoost

techniques may enhance performance, the feature selection and preprocessing steps appear robust enough to make simpler models viable for this problem.

The feature importance analysis highlights Item_MRP as the most significant factor, contributing 44.5%, underscoring the critical role of pricing in driving sales as shown in Figure 5 Outlet_Type_Supermarket Type3 (10.9%) and Item_Visibility (10.4%) further emphasize the influence of store type and product visibility on customer purchasing behavior. Combined, Outlet_Type_Supermarket Type1 and Item_Weight account for 15%, linking store characteristics and product attributes to sales. Other features, such as Years_Operating, provide additional contextual insights, reinforcing the value of a diverse feature set for accurate sales predictions.

Model	**MAE**	**RMSE**	**R^2**
Linear Regression	791.455505	1069.097895	0.579476
Ridge Regression	791.370045	1069.224621	0.579377
Random Forest	766.711323	1096.220709	0.557869
XGBoost	799.029166	1146.010903	0.580000
Clustering-Enhanced XGBoost	796.702218	1146.291643	0.605000

TABLE I
PERFORMANCE METRICS FOR VARIOUS MODELS

The above table compares the performance of various models using MAE, RMSE, and R^2 metrics for sales prediction. Linear and Ridge Regression perform similarly, with moderate R^2 values around 0.579, indicating their suitability for linear relationships. Random Forest shows a slightly lower R^2 (0.558), while XGBoost achieves a comparable R^2 of 0.580, demonstrating robustness despite its complexity. Clustering-Enhanced XGBoost outperforms other models with the highest R^2 (0.605), highlighting its ability to capture additional variance through tailored clustering. These results showcase the trade-offs between simplicity and complexity, with clustering-enhanced techniques offering improved prediction accuracy.

V. CONCLUSION

We presented a case study for sales prediction for BigMart, highlighting the significance of machine learning models working on a retail problem. However, despite strong predictive performance with standard approaches, such as Random Forest and XGBoost, clustering-enhanced classification approaches opened doors to improving clustering-based predictions. The knowledge of important features like pricing, product visibility, and store quality greatly helped us make the prediction on sales. However, the results also revealed opportunities for improvement, most notable of which were in the areas of feature selection and manipulating latent dynamics that apply uniquely to stores.

In hindsight, it's clear that while the models were able to capture complex interactions in sales data, they were lacking in accounting for a few important nuances of retail dynamics like seasonality and external influences. Clustering is one of those that gave good results, but we could further improve it by grouping together stores to find relationships with a deeper nature between the features of the store and the trends of the sales. For example, being able to understand the impact of demographic differences—such as urban vs. suburban sales—could provide a more accurate predictive lens.

In conclusion, this research has set the stage for future developments in retail analysis. Leveraging these learnings, we hope to improve the accuracy, interpretability, and applicability of our methods, keeping in mind the nature of the ever-competitive and ever-evolving retail landscape. Our journey exemplifies how relatively simple methods in the realm of machine learning have the ability to influence effective decision-making through data while also illustrating the iterative approach required to improve methodologies for out-of-the-box applicability to real-world problems.

VI. MEETING ATTENDENCE

The meetings in the month of November are conducted on 4th, 7th, 11th, 12th, 13th, 14th, 18th and all the meetings were attended by the whole team.

REFERENCES

[1] J. Wang, "A hybrid machine learning model for sales prediction," 2020 International Conference on Intelligent Computing and Human-Computer Interaction (ICHCI), Sanya, China, 2020, pp. 363-366, doi: 10.1109/ICHCI51889.2020.00083.

[2] K. Pai, M. Batool, N. Ravi, R. Surendra, A. N. R. Shree and S. J. Meenakshi, "Enhancing Sales Forecasting and Prediction with Cutting-Edge Machine Learning Methods," 2024 Second International Conference on Data Science and Information System (ICDSIS)

[3] Abolghasemi, M., J. Hurley, A. Eshragh, and B. Fahimnia. 2020. Demand forecasting in the presence of systematic events: Cases in capturing sales promotions. International Journal of Production Economics 230: 107892. https://doi.org/10.1016/j.ijpe.2020.107892

[4] Almeida, F.M.D., A.M. Martins, M.A. Nunes and L.C.T. Bezerra. 2022. Retail sales forecasting for a Brazilian supermarket chain: an empirical assessment. 2022 IEEE 24th Conference on Business Informatics (CBI). IEEE, 60–69 https://doi.org/10.1109/CBI54897.2022.00014

[5] Auppakorn, C., and N. Phumchusri. 2022. Daily Sales Forecasting for Variable-Priced Items in Retail Business. 2022 4th International Conference on Management Science and Industrial Engineering. 80–86. https://doi.org/10.1145/3535782.3535794

[6] Predicting Future Sales of Retail Products using Machine Learning https://arxiv.org/pdf/2008.07779

VIII. PROJECT GIT REPOSITORY:

Link to our Project Repo:
https://github.ncsu.edu/ncheruk2/G26_ALDA_Project