DS 3001
Esha Sharma, exy8eb
Celestine Nguyen, haw3wk
Jennifer Vo, phb8pt
Shruti Bala, aen6ju
Chetu Barot, zrz9ry
Tulsi Patel, cjw6jz

Results

Github: https://github.com/CelestineNguyen/DS3001_SalesProject/tree/main
Cleaned Dataset: all_ex_holidays
Dataset: https://www.kaggle.com/competitions/store-sales-time-series-forecasting/data

Introduction/Prediction Question:

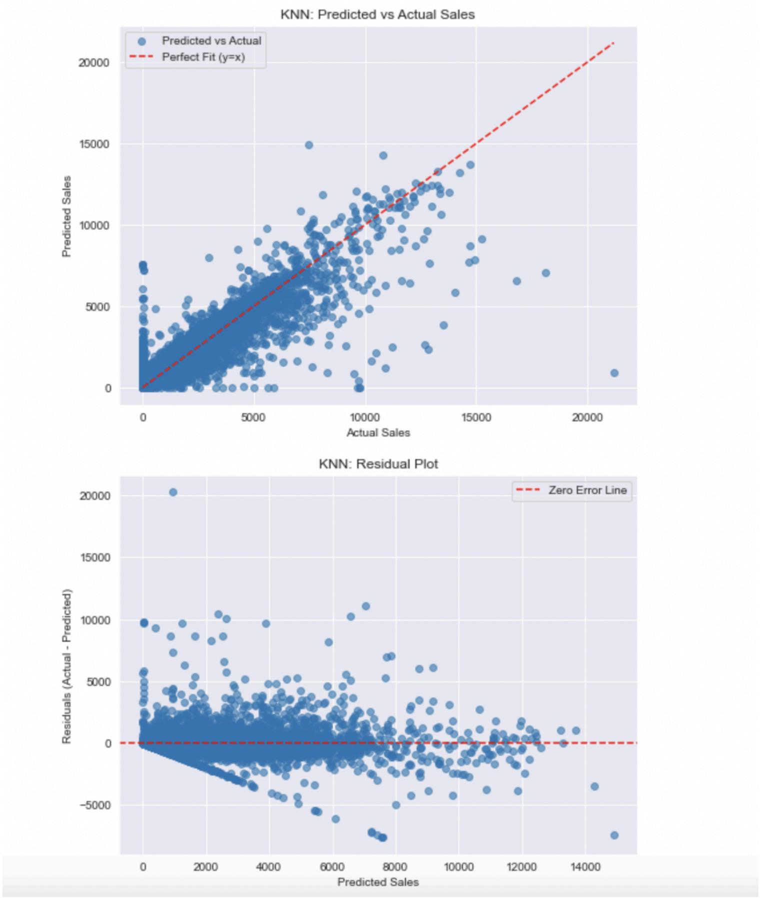
The main question that is being answered is "Which products affect sales rates at stores?". In retail, understanding which products drive sales rates is critical for effective inventory management, targeted marketing strategies, and maximizing profitability. By identifying key product families that significantly influence overall sales, store managers can optimize shelf space, adjust pricing strategies, and tailor promotional efforts to meet customer demand. This insight is particularly valuable in a competitive marketplace, where aligning product offerings with consumer preferences can enhance both revenue and customer satisfaction. Our analysis explores the relationship between product families and sales rates, aiming to provide actionable insights that can guide data-driven decisions for grocery store operations.

Information About the Dataset and Cleaning:

The dataset used for this analysis is from the Kaggle competition "Store Sales Time Series Forecasting," which contains time-series data on daily sales for various product families across multiple stores. Key features include product type, store number, promotional events, holidays, and sales figures, which provide a rich context for understanding sales patterns. Using supervised learning techniques, we developed regression models to predict sales rates and analyze the influence of product families, comparing model performance based on metrics such as R^2 and RMSE to identify the most impactful features.

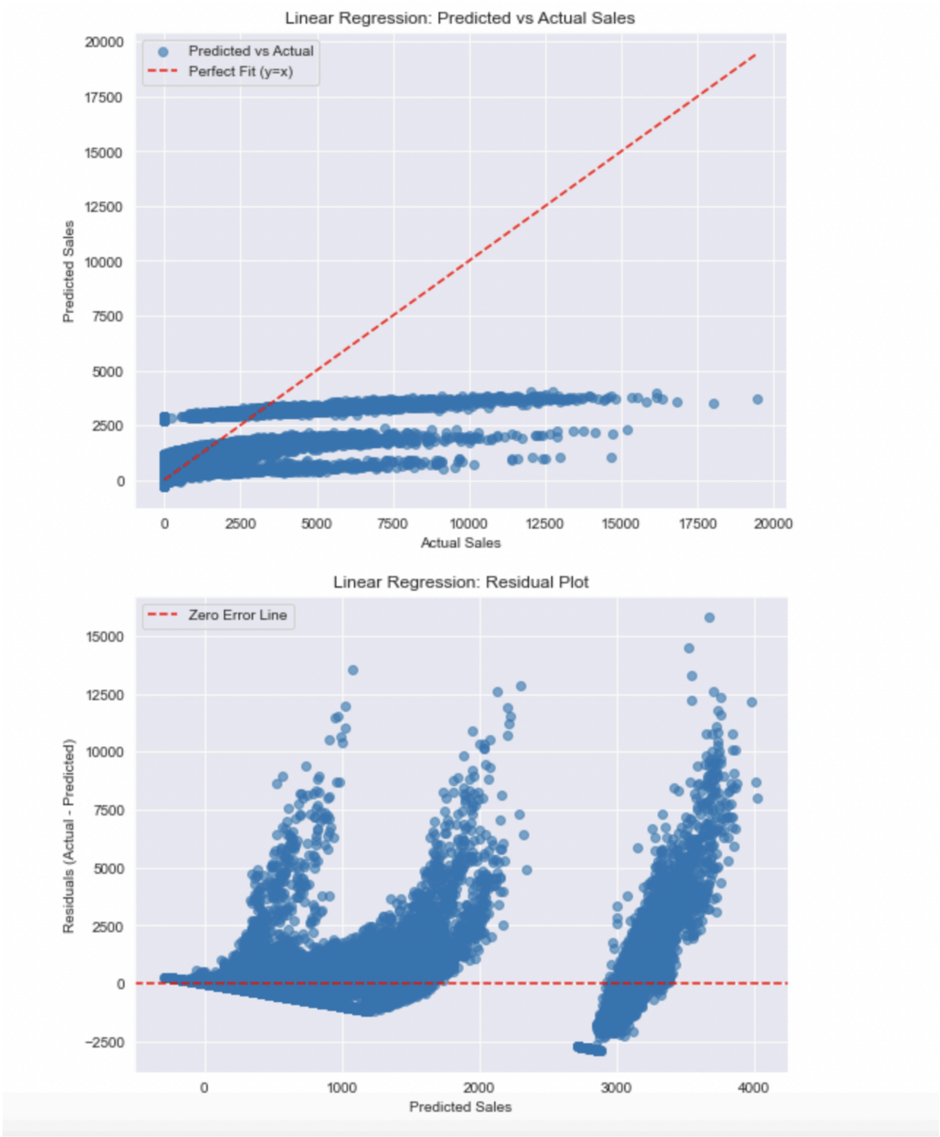
To prepare the dataset for analysis, three additional datasets were merged with the primary training dataset to enrich the information available for modeling. The transactions dataset, which records the number of transactions for each store on a given date across all product types, was first joined with the training data using store number and date as keys. This introduced some missing values in the transactions column, as not all product categories were purchased at every store on all days. These missing values were filled with 0, reflecting the absence of sales on those days. The oil dataset, providing daily oil prices, was then merged by date, but the initial absence of 43 days of oil price data resulted in 928,422 rows with missing values after the merge. Finally, the store dataset, which includes each store's city, state, type, and cluster information, was merged by store number, adding contextual information about the stores. This comprehensive cleaning and integration process ensured that the data was consolidated and ready for analysis, despite challenges with missing values.

Key Tables:



The KNN model demonstrates a strong ability to predict sales rates, particularly for lower sales ranges. In the scatter plot comparing predicted vs. actual sales, the points are closely aligned with the red "perfect prediction" line, suggesting that KNN predictions are generally accurate. However, as sales increase, predictions begin to deviate slightly, reflecting the model's limitations in handling extreme values. This behavior indicates that while KNN captures non-linear relationships in the data better than linear regression, it still struggles with sparsity in higher sales regions, likely due to fewer relevant neighbors in these ranges.

The residual plot for KNN shows well-distributed errors with fewer discernible patterns compared to Linear Regression. The residuals remain closer to zero overall, indicating that the model does not systematically over- or under-predict sales across most ranges. However, there is some widening of residuals for higher sales, suggesting that KNN underestimates sales at these values. The high R^2 score of 0.89 underscores the model's effectiveness in explaining the variance in the data and its ability to generalize well for most sales predictions.


Linear Regression: Predicted vs Actual Sales

Linear Regression: Residual Plot

The regression model offers moderate predictive capabilities but falls short when handling higher sales values. The scatter plot comparing predicted vs. actual sales reveals that most predictions cluster near the correct values for lower sales. However, as actual sales increase, the predictions drift further from the red "perfect prediction" line, indicating that the model struggles to capture the complexity of the data in higher ranges. This issue is further highlighted by the residual plot, which exhibits a systematic curve and increasing variance in errors as predicted sales grow. Ideally, residuals should scatter randomly around zero with no visible structure, but the observed patterns suggest the model lacks the ability to generalize well for higher sales.

The R^2 score for Linear Regression, at 0.56, reflects a moderate ability to explain the variance in the data. While the model captures some trends in the data, significant gaps in the scatter and residual plots confirm its limitations. This aligns with the observation that Linear Regression predicts lower sales reasonably well but struggles with higher values.

Analysis of Key Results:

The graphs provide valuable insights into the predictive question, "Which products affect sales rates at stores?" The models demonstrate that product type is a significant factor influencing sales rates, as indicated by the variations captured in both the KNN and regression results. The scatter plots from both models show that lower sales values are predicted more accurately, which aligns with the idea that sales patterns for commonly sold products are easier to model due to their frequent occurrence in the dataset. The divergence from the perfect prediction line at higher sales values suggests that the influence of product type becomes more complex or interacts with other variables, such as store location or promotional events.

The KNN model, with its high R^2 score of 0.89, captures non-linear relationships more effectively, indicating that product types with irregular sales patterns might benefit from considering nearest-neighbor relationships based on sales trends and store-specific factors. On the other hand, the regression model, with a moderate R^2 of 0.56, highlights the limitations of linear assumptions in explaining the variance in sales rates. Residual analysis further underscores that neither model fully accounts for the complexity in higher sales, potentially pointing to interactions between product type and other features like promotions or seasonality.

Overall, the analysis confirms that products significantly impact sales rates, but their effects are intertwined with other variables. While KNN provides a more detailed and adaptable framework for predicting sales across diverse products, both models suggest the need for further feature engineering or advanced modeling to fully capture the dynamics of product influence on sales rates.

Conclusion:

When considering the predictive question, "Which products affect sales rates at stores?" the analysis shows the strengths and weaknesses of the models. KNN performs better than Linear Regression overall, handling non-linear relationships and yielding more accurate predictions, especially for lower sales values. The high R^2 score and well-distributed residuals suggest that KNN provides a more robust understanding of the factors influencing sales.

On the other hand, Linear Regression offers insights into the linear relationships within the data but fails to account for the complexity of higher sales. The systematic patterns in the residuals highlight missing features or interactions that the model cannot capture. These findings suggest that while both models provide valuable insights, KNN is better suited for predicting sales rates and identifying influential factors across diverse sales ranges. However, further refinement or additional modeling techniques might be necessary to address the challenges of higher sales predictions.