Group Members:
Tulsi Patel, cjw6jz, Chetu Barot, zrz9ry, Esha Sharma, exy8eb, Celestine Nguyen, haw3wk, Jennifer Vo, phb8pt, Shruti Bala, aen6ju

Github: https://github.com/CelestineNguyen/DS3001_SalesProject/tree/main

Final Paper

1. Abstract/Executive Summary: ~300 words summarizing your results for people who won't actually read the paper. (tulsi)

For our project, we decided to analyze what products affect sales in retail markets using data we sourced from a Kaggle competition for store sales time series forecasting. The data contains daily sales records for different products across multiple stores along with translation counts, oil prices, and store specific details. We used K-nearest neighbors (KNN) and Linear regression to develop and analyze sales patterns.

The KNN model outperformed the linear regression model with a relatively high R2 value of 0.89. It captured non-linear relationships by making accurate predictions for lower sales values and helped us look into some of the more irregular sales patterns. Residual analysis showed small systematic errors but the model struggled where we had fewer data observations for higher sales ranges due to fewer relevant neighbors. The linear regression model, which had a lower R2 score of 0.56, showed trends for lower sales but couldn't predict with accuracy the complexity of the higher sales. Recurring residual patterns showed missing features or unaccounted interactions, limiting its ability to generalize across more diverse sales ranges.

The analysis confirms that product type significantly impacts sales but is also dependent on other variables like store locations, local promotions, and seasonality of products. While KNN provided a more comprehensive framework for predicting sales using product families and

understanding these influencing variables for the more complex higher sales, linear regression offers more insights into the linear relationships within the data.

2. Introduction: Two or three pages that summarize your findings for people who will read the paper, so they know what is coming and why. (tulsi)

3. Data: First submission, cleaned up to be read as part of a paper (tulsi)
The Stores Sales Kaggle competition came with 7 csv files: samplesubmission.csv, train.csv, test.csv, stores.csv, oil.csv, holidays_events.csv, and transactions.csv.
4. Methods: Pre-analysis plan submission, cleaned up to be read as part of a paper (Chetu)

The objective of this pre-analysis plan, or methods, is to outline the methodology and strategies for analyzing how different types of products sold affect store sale rates. The plan also addresses additional questions regarding the prediction of future sales, the impact of promotions, and the effects of holidays and events on sales trends. To begin, the dataset includes critical observations such as the types of products sold, their associated product families, the date of data collection, the store identifier, and the number of sales predicted for specific product families. These observations form the foundation of the analysis.

This study employs supervised learning, specifically regression techniques, to predict the factors influencing sales. The primary target variable is the sales column, and the goal is to analyze how changes in various predictors impact this value. The analysis involves developing multiple regression models, including K-Nearest Neighbors (KNN) and Linear Regression, to compare their performance and identify the most influential factors on sales. By handling both

linear and non-linear relationships, these models aim to provide comprehensive insights into sales trends.

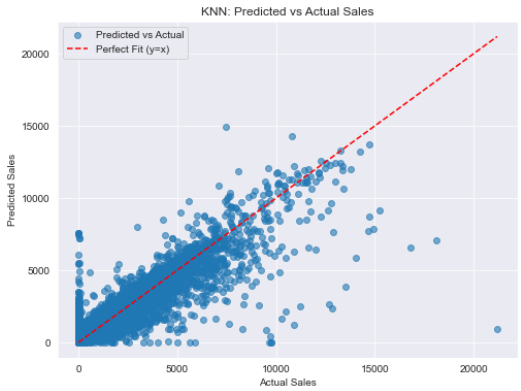
The evaluation metrics for this project include R-squared (R^2) values and Root Mean Squared Error (RMSE). The best-performing model will exhibit the highest R^2 value and the lowest RMSE, ensuring strong predictive accuracy. Residual analysis will further help assess systematic errors and refine feature selection or model design. Success will be defined by the ability to predict sales accurately and to understand the effects of variables such as product type, store location, and seasonal trends.

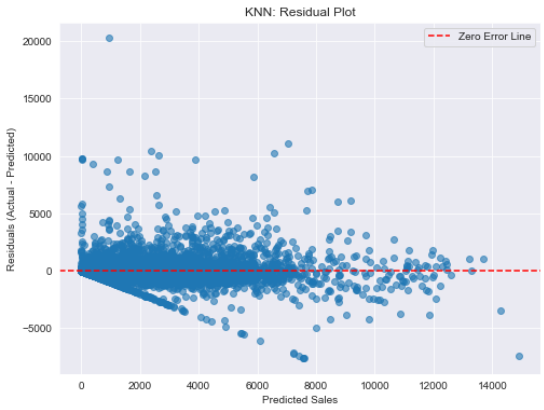
Several challenges are anticipated in the analysis. High dimensionality and noise in the dataset may necessitate careful feature selection or dimensionality reduction techniques like Principal Component Analysis (PCA). Additionally, poor model performance may require iterative refinement of the feature matrix (X) to focus on impactful variables. Sparse data in higher sales ranges could also lead to inaccuracies, which might be addressed through stratified sampling or data weighting strategies.

Feature engineering plays a critical role in preparing the dataset for analysis. Categorical variables such as the ‘family’ column will be one-hot encoded, while the ‘date’ column will be transformed into numerical features like day of the week, month, and year to capture temporal effects. Integration of additional datasets, such as store-specific details and promotion information, will be conducted by merging data based on store number and date.

The results of this analysis will provide insights into the effectiveness of the chosen models and their ability to capture complex relationships in the data, setting the stage for a detailed evaluation of their performance.

5. Results: Results submission, cleaned up to read as part of a paper (Chetu)

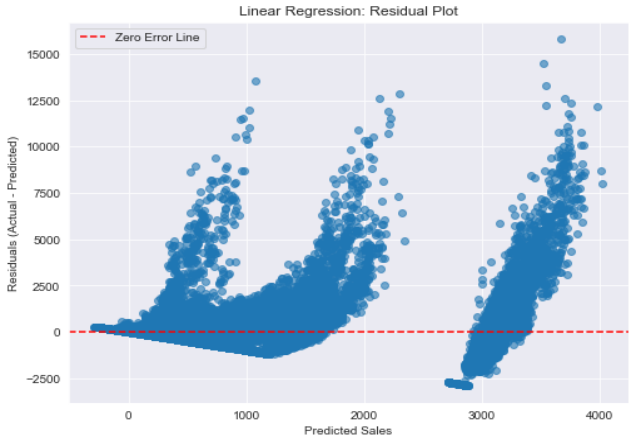




The KNN model demonstrates a strong ability to predict sales rates, particularly for lower sales ranges. In the scatter plot comparing predicted vs. actual sales, the points are closely aligned with the red "perfect prediction" line, suggesting that KNN predictions are generally accurate. However, as 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.



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 cannot 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.

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. The 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.

6. Conclusion: Two or three pages that summarize your findings for people who have read the paper. In particular, describe extensions, complications, problems, limitations that you ran into that could form the basis of future work (turn the weaknesses of your paper into results/features, rather than flaws). (Chetu)

The analysis conducted on the Kaggle “Store Sales Time Series Forecasting” dataset offers actionable insights into the relationship between product families and sales rates, underscoring key trends and areas for improvement. By utilizing K-Nearest Neighbors (KNN) and Linear Regression models, we gained a better understanding of how different products influence overall store performance. While the study demonstrates the use of machine learning models in retail analytics, it also reveals several challenges and opportunities for future exploration.

7. References/Bibliography