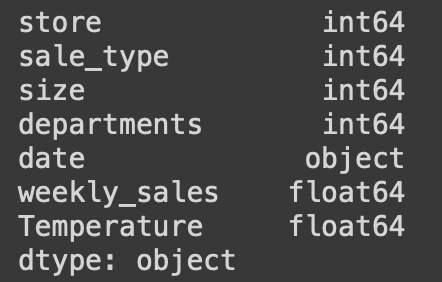
**Retail Sales Prediction Project**

**1.Problem Statement**

The goal of this project was to build a predictive model that can accurately forecast weekly retail sales for 45 stores, using historical sales data on various features such as store characteristics, sales type, and external factors like temperature. The project aims to provide insights and recommendations to optimize sales strategies and improve business performance.

**2.Data Preprocessing**

The project begins by importing and merging three datasets: `sales\_features.csv`, `sales\_pred.csv`, and `sales\_stores.csv`. These datasets contain information about stores, sale type, department, weekly sales and temperature. After merging the datasets, the following preprocessing steps were performed:

***a. Data Understanding:*** Exploratory data analysis was conducted to understand the data types, shapes, and missing values in each dataset. Figure 1 shows that almost all features are int64 or float in the merged dataset. Date was further changed to datetime64 type.

While examining the statistics of the merged data, it was discovered that there were negative values in weekly sales. Subsequently, further analysis was conducted on rows.

containing negative weekly sales values to comprehend the relationship with the dependent variable.

Table 1: Sales Dataset Datatypes

***b. Handling Missing Values:*** It was observed that there were no Null values in sales dataset.

***c. Feature Engineering:*** New features such as `day`, `Month`, `Year` and `Week` were created from the `date` column to capture temporal information.

***d. Correlation Matrix:*** In Figure 1, a weak positive correlation is observed between the number of departments in a store and weekly sales (correlation coefficient of 0.03). This suggests that stores with more departments tend to have slightly higher weekly sales. However, all other variables show no correlation with the target variable. Therefore, to enhance model training, it is necessary to either incorporate more features or seek assistance from domain experts.

A graph of heatmap and temperature

Description automatically generated

Figure 1: Correlation Heatmap

***e. Handling Outliers:***As you can see in figure 2 there are lots of data points outside the main distribution of the data. Training the dataset without removing or altering outliers and then assessing the model performance before and after outlier removal can provide valuable insights into the impact of outliers on the model.

A graph with numbers and lines

Description automatically generated

Figure 2: Outliers in Dataset for weekly sales.

***f. Encoding Categorical Variables:*** *The categorical variable `sale\_type` was encoded using label encoding.*

***3.Model Selection and Evaluation***

Following regression models were evaluated for their performance in predicting weekly sales:

- Linear Regression

- Decision Tree Regression

- Random Forest Regression

- Gradient Boosting Regression

Further the dataset was split into training, validation, and test sets and model performance was evaluated using Mean Absolute Error (MAE) and Root Mean Squared Error (RMSE) on both the training and validation sets.

A screenshot of a black and white screen

Description automatically generated

Table 2: Mean Absolute Error (MAE) and Root Mean Squared Error (RMSE) for all four models

Table 2 shows the Decision Tree model have the lowest MAE and RMSE values on the training set. However, it has a higher MAE and RMSE on the validation set compared to the other models, indicating potential overfitting.

The Random Forest model has lower MAE and RMSE values on the validation set compared to the Decision Tree model, suggesting better generalization performance. Therefore, the Random Forest model may be the best candidate for further hyperparameter tuning and consideration.

The Linear Regression model has higher MAE and RMSE values compared to the Decision Tree and Random Forest models on both the training and validation sets. This indicates that it may not capture the non-linear relationships present in the data as effectively as the tree-based models.

The Gradient Boosting model has slightly lower MAE and RMSE values on the training set compared to the Decision Tree and Random Forest models. However, its performance on the validation set is like that of the Linear Regression model, with higher MAE and RMSE values. This suggests that while the Gradient Boosting model performs well on the training data, it may also suffer from overfitting on the validation data.

To summarize, the Random Forest Regression model demonstrated the best performance on the validation set as seen in above table.

**4.Handling Outlier and Hyperparameter Tunning**

**A graph of a sales curve

Description automatically generated with medium confidence**

Figure 3: Weekly Sales with Outliers

Outliers in the weekly sales variable were addressed using two approaches:

1. Truncating Outliers: Values above a threshold of 50,000 were capped at 50,000 to mitigate the influence of extreme outliers.

A graph showing the sales of a sales funnel

Description automatically generated with medium confidence

Figure 4: Weekly sales after Truncating Outliers

1. Removing Outliers Based on Z-score: Data points with a z-score greater than 2 or less than -2 were identified as outliers and removed from the dataset.

A graph showing the sales of a sales funnel

Description automatically generated with medium confidence

Figure 5 Weekly sales After Removing Outliers

Evaluating the performance of the Random Forest Regression model on the modified datasets showed that removing outliers based on the z-score threshold improved the model’s performance on the validation set.

To further improve the performance of the Random Forest Regression model, hyperparameter tuning was performed.

|  |  |  |
| --- | --- | --- |
| **Handling Outlier Model** | **MAE(Validation Set)** | **RMSE(Validation Set)** |
| Truncating Outliers | 13221.35 | 16549.83 |
| Removing Outliers | 11644.84 | 14459.27 |
| Hyperparameter Tunning | 10588.63 | 12935.38 |

*Table 3: Model Performance for Handling Outliers.*

Due to limited computational power, manual tuning was performed instead of utilizing GridSearchCV or RandomizedSearchCV. The following hyperparameters were adjusted:

- `n\_estimators`: Increased to 150

- `max\_depth`: Decreased to 4

- `min\_samples\_split`: Set to 10

- `min\_samples\_leaf`: Set to 1

- `max\_features`: Set to 'sqrt'

- `bootstrap`: Set to True

- `criterion`: Set to 'friedman\_mse'

These hyperparameters were selected based on experimentation and iterative evaluation of model performance. In conclusion, the hyperparameter tuning led to marginal improvements in the results by showing validation score jump from an MAE OF 11701.93 and an RMSE of 14616.121 to MAE of 10588.63 and an RMSE of 12935.38.

**5.Findings, Insights and Recommendation**

After training the final Random Forest Regression model with the tuned hyperparameters, the following results were achieved:

1. The Mean Absolute Error (MAE) is approx. 10,500.
2. The Root Mean Squared Error (RMSE) is approx.12,900 T

These metrics indicate that while the model is not entirely accurate, it still holds potential for generating insights that can aid sales and inventory management.

A graph of a graph of a graph

Description automatically generated with medium confidence

Figure 6: Difference between predicted and actual values for average weekly sales over all years.

The figure above illustrates the disparity between predicted and actual values for all weeks across the three years. Notably, the differences are generally within a few thousand units, except for certain anomalies.

As shown in the figure 3 for the predicted target variable, we can observe spikes in February, reaching 13,520 sales units, while the actual values fall within the same range, approximately 13,200. Similarly, in June, the predicted sales hover around 13,520, aligning closely with the actual sales of around 13,600. Conversely, the predicted values exhibit a slight decline around August, at 13,480, compared to the actual sales of 13,450. Hence, we can rely on our model to make important decisions.

A graph of different colored lines

Description automatically generated with medium confidence

Figure 7:Predicted and Actual sales over months across three years

The analysis of our predictive model reveals that while it's not perfect, it offers valuable insights for sales and inventory management. We noticed some variations between predicted and actual sales, but overall, the model shows promise. To make the most of it, we suggest investigating anomalies, tweaking model parameters, and adding more relevant features for better accuracy. Regular evaluation and collaboration with experts can also enhance the model's performance. By continuously refining the model and leveraging its insights, we can make informed decisions and optimize our operations effectively for improved sales outcomes.

**7. Feature Engineering**

To enhance the predictive performance of the model, additional features were engineered to capture factors influencing sales:

1. Seasonality Features: A new `season` column was created by mapping months to seasons (winter, spring, summer, fall) using a custom function.

2. Economic Features: External economic indicators were incorporated, including:

- Consumer Price Index (CPI)

- Unemployment Rate

- Gross Domestic Product (GDP)

These features were fetched from the Federal Reserve Economic Data (FRED) and merged with the existing dataset based on the date.

As shown in above figure there is noy much correlation between target variable and other features even after adding features therefore there was not as significant improvement in the model performance. This features significantly improved the validation set result from MAE: 10588.63 and RMSE: 12935.38 to MAE: 5715.45 and RMSE: 10098.18.

**8. Conclusion**

To conclude, the collected data do not exhibit a significant relationship with the target variable. Therefore, training the data and deriving precise insights may not significantly benefit the business. Although integrating the unemployment rate and gap improved accuracy, the enhancement was not substantial. This indicates the necessity of gathering additional data to facilitate more effective training processes.