Leticia Lima Huang CIS9660- Data Mining for Business Analytics Technical report

In the retail sector, identifying demand forecasting is essential for business to maintain optimal inventory levels, minimizing operational costs associated with last minute orders to suppliers, and ensuring high quality service to customers by delivering service promptly and efficiently. Inefficient forecasting can lead to stockouts, excess inventory on low-demand products, loss of revenue and increased logistical challenges. To taggle those problems, I intend to build an AI Agent using Regression models to predict surge in demand by using historical sales data. The data used was retrieved from Kaggle and originates from Favorita Grocery Store. This large dataset covers various products including food, home, and personal items. It is useful for forecasting demand spikes related to seasonality, promotions, and changes in oil prices, a predominant factor in the economy.

The preprocessing process included several steps: necessary libraries were downloaded, data was retrieved from Kaggle using an API, and seven files were merged into a single dataset. The dataset was previewed with head() and info(), and missing values were analyzed both numerically and graphically, revealing that 457.16% of the data was missing. For cleanup, a copy of the original data was made, unnecessary columns were removed, and missing values in three columns were addressed by applying the median for transactions, assigning "Not Holiday" for holidays, and using a fill method for oil. Time-based columns were added to support forecasting tasks. Next, the “sales” column was reviewed for completeness, resulting in 68.77% valid sales over five years. The timeframe was restricted to the last two years, yielding 82.41% valid sales. Duplicate values and columns were then examined and removed, unique values in each column were verified, and NaNs were eliminated before conducting data exploration analysis. During EDA, histogram and boxplot visualizations were created to assess the dataset's distribution, and the describe() function was used to compute statistics for the target variable. Correlations between variables and the target were examined with scatterplots and heatmaps. The subsequent step involved feature engineering, including creating new columns and grouping others into smaller categories, followed by one-hot encoding. Finally, the data was split into training and test sets, and feature scaling was performed.

The next step was model building and training, for which both linear and multiple regression models were selected. Linear regression was trained using the most correlated feature. The evaluation metrics for both linear models indicated limited performance, with multiple R-squared values of 36.83% for training and 37.20% for testing. The simpler model yielded even lower R-squared values, at 25.97% for training and 26.21% for testing. Due to the low performance of these models, RGBoost was selected because it is better suited for handling nonlinear data, which was reflected in the dataset. With RGBoost, model performance improved to 72.04% for training and 72.05% for testing. Comparison of the three models showed that RGBoost provided the best results for nonlinear data. To validate the accuracy of the model, 5-Fold Cross Validation was used, resulting in an average R-squared of 71.10%. The process was concluded with a business prediction scenario by adjusting input values to generate forecasts. Both promotion and no-promotion scenarios were examined to analyze their impacts on sales.

The RGBoost model explains about 72.05% of sales variance, making it effective for predicting demand in volatile retail environments. Due to its precision, we recommend using RGBoost to inform inventory planning and promotional strategies at the store level. Retrain the model regularly with updated sales and promotion data. Track feature importance and prediction errors by region, category, and time period to improve forecasts. Enhance accuracy by integrating external data such as inflation, and competitor pricing. Run scenario simulations to guide business decisions and campaign optimization. Combine model insights with expertise from marketing and supply chain teams for final decisions.