Final report for the Revenue Index Prediction pre-screening exercise

1. Introduction

This project focused to develop a predictive model for estimating the revenue index of a retail company using historical transaction data. The process involved data exploration, preprocessing, feature engineering, model training, and evaluation. The final model was assessed based on its ability to generalize to unseen data and provide meaningful insights for business planning.

2. Data Collection and Description

The dataset consisted of transactional records, including variables such as order numbers, spending indices, and user activity metrics. The raw data was analysed to understand its structure, trends, and missing values. The primary datasets used included orders data, which contained records of daily orders placed by customers. Transactions data provided financial metrics such as total spending and active user indices. Revenue data recorded the revenue index for different periods. Understanding these datasets helped in designing a model that could effectively capture revenue patterns.

3. Data Preprocessing

To ensure data accuracy and consistency, preprocessing steps were applied. The order_number column was adjusted to correct inconsistencies, data was drop if the order number didn’t increase because there is no way to know if the data correspond to the before or after it’s a independent variable but this decision would be consulted in the company to corroborate this hypothesis, for that purpose the anomalous data is stored in another file. Additionally, date formats were standardized for uniformity across datasets. Missing values were either imputed or removed to maintain data integrity. Aggregating daily records into quarterly summaries helped reduce noise and improve interpretability. Feature engineering introduced a new variable, spend_per_user, to better capture customer spending behavior. These transformations allow that the dataset was clean and well-structured for model training.

4. Model Development

A linear regression model was selected to predict the revenue index. The training process incorporated key features, including total orders, total spend index, weekly active users index, and spend per user. To ensure reliable performance assessment, the dataset was split into training and test sets, reserving data from the last quarter of 2022 for model evaluation. This approach allowed the model to be tested on previously unseen data, providing a more accurate estimate of its predictive power.

5. Model Evaluation

The model’s performance was assessed using standard regression metrics. The R^2 score was 0.7952, indicating that the model explained nearly 80% of the variance in the revenue index. The mean absolute error was 50.74 units, suggesting the average prediction deviation from actual values. The mean squared error stood at 3604.59, while the mean absolute percentage error was 20.15%. These results confirmed that the model was effective in capturing revenue trends with reasonable accuracy. The average prediction error of approximately 51 units demonstrated the model’s capability in identifying revenue fluctuations.

6. Visual Evaluation

To further analyse the model’s performance, visualizations were utilized to examine the distribution of errors, residual trends, and feature importance. Residual plots helped identify any patterns or biases in the predictions, ensuring that the model was making fair estimations. A comparison of predicted versus actual values illustrated the model’s ability to align with real-world revenue behavior. Additionally, feature importance analysis provided insights into which variables had the strongest influence on revenue predictions, reinforcing the significance of key drivers such as spending indices and active user engagement.

7. Insights and Business Implications

The findings highlighted the importance of high-quality data and feature engineering in predictive modelling. The model successfully identified key revenue drivers, making it a valuable tool for business decision-making and financial planning. With further refinements, such as incorporating external economic indicators or experimenting with more advanced models like decision trees or neural networks, predictive accuracy could be enhanced. Automating the data cleaning and feature creation processes could improve efficiency and scalability.

8. Conclusion

This project successfully developed a revenue index prediction model using historical data. The structured approach, which involved thorough data exploration, transformation, and evaluation, provided valuable insights into revenue trends and business performance.