Online Grocery Recommendation System

Final Project Report

Team Members :

[Zeina Ayman 202200351 - Habiba Khalil 202200720 - Nour Helmy 202202012 - Youssef Abohendia 202202699 - Youssef Fathy 202202193]

Introduction :

This project presents a comprehensive analysis of a grocery recommendation system based on users' past purchases. The primary goal is to predict whether a user will reorder a particular product in their next purchase. Three machine learning models were evaluated: K-Nearest Neighbors (KNN), Decision Tree, and Logistic Regression.

Methodology :

1. **Data Preparation :**
    - **Dataset :** The merged dataset consists of user orders and product information with the following columns : [order_id, product_id, add_to_cart_order, reordered, product_name, aisle_id, department_id, aisle, department, user_id, eval_set, order_number, order_dow, order_hour_of_day, days_since_prior_order, user_total_orders, user_total_products, user_reorder_ratio, product_total_orders, product_reorder_ratio].
    - **Features :** Selected features for model training : [user_total_orders, user_total_products, user_reorder_ratio, product_total_orders, product_reorder_ratio, add_to_cart_order, order_dow, order_hour_of_day, days_since_prior_order].
    - **Target :** reordered (binary variable indicating whether a product was reordered).
2. **Data Splitting :**
    - **Train -Test Split :** The data was split into training and testing sets with an 80-20 ratio using train_test_split from Scikit-learn.

Model Training and Evaluation :

1. **K-Nearest Neighbors (KNN) :**

    n_neighbors = 3

    Suitable for small to medium-sized datasets.

2. **Decision Tree :**

   random_state = 42

   Easy to interpret but can overfit.

3. **Logistic Regression :**

   random_state = 42

   max_iter = 1000

Evaluation Metrics :

- **Accuracy :** Proportion of correct predictions.
- **Precision :** Proportion of true positive predictions among all positive predictions.
- **Recall :** Proportion of true positive predictions among all actual positives.
- **AUC-ROC :** Area under the ROC curve, indicating the model's ability to distinguish between classes.

Results :

1. **K-Nearest Neighbors (KNN) :**
   - **Accuracy :** 0.66
   - **Precision :** 0.70
   - **Recall :** 0.76
   - **AUC-ROC :** 0.69
2. **Decision Tree :**
   - **Accuracy :** 0.66
   - **Precision :** 0.71
   - **Recall :** 0.71
   - **AUC-ROC :** 0.65
3. **Logistic Regression :**
   - **Accuracy :** 0.66
   - **Precision :** 0.70
   - **Recall :** 0.75
   - **AUC-ROC :** 0.70

Model Selection :

Based on the evaluation metrics, the Logistic Regression model demonstrates the best overall performance with balanced accuracy, precision, recall, and AUC-ROC scores. It is also efficient in terms of prediction speed.

Conclusion :

The Logistic Regression model is recommended for predicting grocery reorders based on users' past purchases due to its high performance and efficiency. Future work could involve further optimization and exploring additional features to improve the model's predictive power.