Project Data Analysis 2

Name	ID
Amany Saeed Fallatah	44411940
Nahla Mohammad AL osaimi	44412006

COURSE PRESENTER

"DR. Omaima Fallatah "

Course Number: DS3114

Bachelor of science in Data Science

Kingdom of Saudi Arabia - Umm Al-Qura University

Introduction

This report outlines the steps taken to perform data preprocessing, model training, and market basket analysis using multiple datasets from a retail store. The main goals are to predict reordered products and discover association rules between items bought together. We use logistic regression for prediction and the Apriori algorithm for market basket analysis.

The datasets used include:

Departments: Information about product departments.

Order Products (Train & Prior): Details of products ordered in prior and training datasets.

Orders: Data about customer orders.

Products: Product details.

Aisles: Information about the aisles where products are located.

2. Data Loading and Cleaning

We begin by loading multiple CSV files containing order and product data using pandas. The datasets are merged to create a comprehensive view of the orders and products.

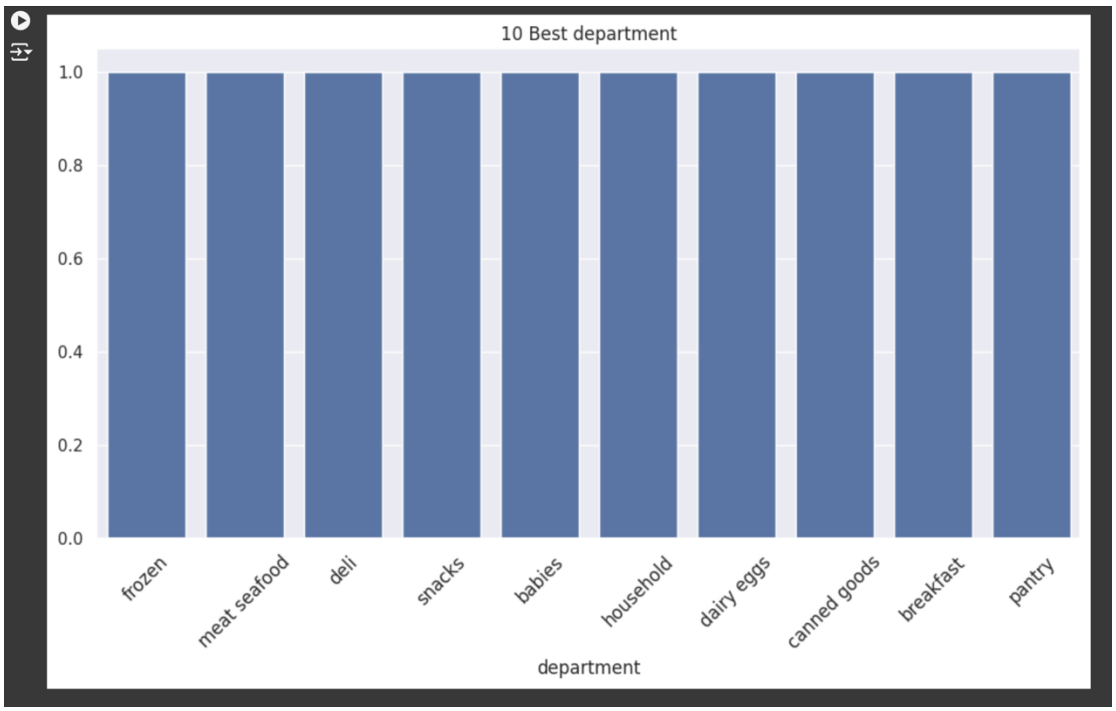
Data Exploration and Cleaning:

The training dataset is created by merging the orders marked as "train" with the training order products data.

We check for missing values and clean the data accordingly. For instance, missing values in the column days_since_prior_order are replaced with the median.

Additional features are added, such as the number of times a product is bought by a user and the number of unique users who bought a product.

Data Visualization:



The chart displays a representation of the top 10 departments in the data using a bar plot. The horizontal axis (x) shows departments such as frozen, meat seafood, deli, etc., while the vertical axis (y) represents the values associated with the count of each department.

3. Feature Engineering and Data Preparation

To prepare the dataset for modeling:

We drop columns that are not useful for the logistic regression model, such as order_id, user_id, and product_id.

The target variable (reordered) is separated from the training data.

4. Handling Outliers

Outliers can affect the performance of machine learning models. We detect and handle them using the Interquartile Range (IQR) method.

5. Model Training and Evaluation Using Logistic Regression

We employ a logistic regression model to predict whether a product will be reordered.

Data Splitting:

The training data (x_train, y_train) and the cleaned testing data (x_test) are prepared.

Handling Missing Values:

Any rows with missing values in y_train are removed before fitting the model.

Training the Logistic Regression Model:

We fit the logistic regression model on the cleaned training data and predict on both training and testing datasets.

Model Evaluation:

Metrics such as accuracy, precision, recall, and F1-score are used to evaluate the model's performance on the training set.

A similar evaluation is performed on the testing set.

6. Multicollinearity Check

We use the Variance Inflation Factor (VIF) to check for multicollinearity among features, which can distort model interpretation.

7. Market Basket Analysis Using the Apriori Algorithm

To identify frequent itemsets and generate association rules, we use the Apriori algorithm from the mlxtend library.

Transform Data for Basket Analysis:

The dataset is transformed into a basket format, where each row represents a transaction, and each column represents an item.

Find Frequent Itemsets:

We apply the Apriori algorithm with a minimum support threshold to find frequent itemsets.

8. Conclusion and Recommendations

Model Performance: Logistic regression was used for prediction, but further hyperparameter tuning and advanced models (e.g., Random Forest, Gradient Boosting) could improve the results.

Insights from Market Basket Analysis: The Apriori algorithm identified significant relationships between products, which can be used for cross-selling and personalized recommendations.

Future Work: Consider using different machine learning techniques for prediction and exploring more sophisticated algorithms for market basket analysis.