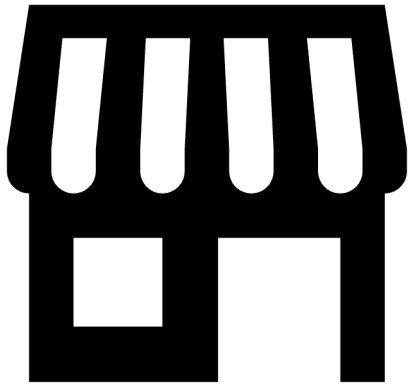
Smart Cart Analytics Instacart

Dariia Vasylieva

1/7/24



INDEX

- Problem statement
- The solution & its Potential impact
- Overview of the dataset and pre-processing
- Important findings from EDA
- Baseline models and evaluation metrics
- Model comparison and interpretation
- A product demo

Problem statement

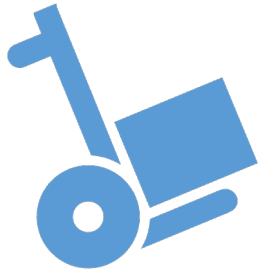
Lack of customer loyalty

Predicting Future Purchase Behavior:

Customer habits are constantly evolving, making it challenging to rely solely on past purchase data for predicting future behavior.

Demand Forecasting and Inventory Management:

Industry growth, changing customer loyalty, and evolving habits can introduce unpredictability into demand forecasting and inventory management.



Maximizing Order Value:

While customers may be receptive to product suggestions, their purchasing patterns can also be unpredictable.

Solution

Efficient Cross-Selling process with recommendation engine.

A well-executed cross-selling strategy can potentially increase revenue by 15% to 30%.

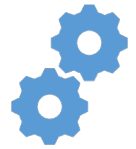
Potential impact of solution

Benefit business owners, investors, and stakeholders within the grocery and industry

Advantages for the business:

Virtual store layout decisions, and promotional activities

Enhanced sales forecasting and resource allocation

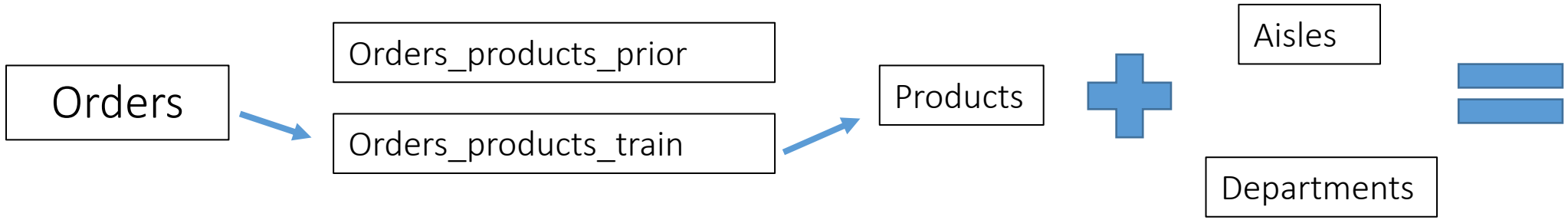


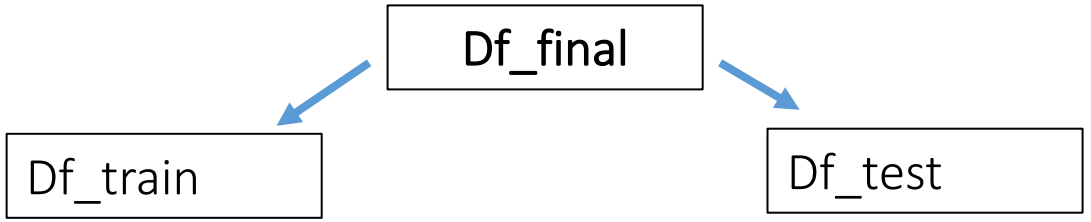
Optimized inventory management, reducing costs related to stockouts or overstock

Increased revenue through effective cross-selling, improving customer loyalty and satisfaction

Overview of the dataset and preprocessing

- Instacart 2017 order purchase data from over **130K** users
- Over **3,421,083 orders**
- Slightly imbalanced data – 60%/40% reordered items





Data:

- User ID
- Order number
- Order of item added to the cart
- Days since last order
- Number of orders
- Product Name
- Aisle & Department Number

-- Features

Target Variable:

- Reordered item
- Product Basket items

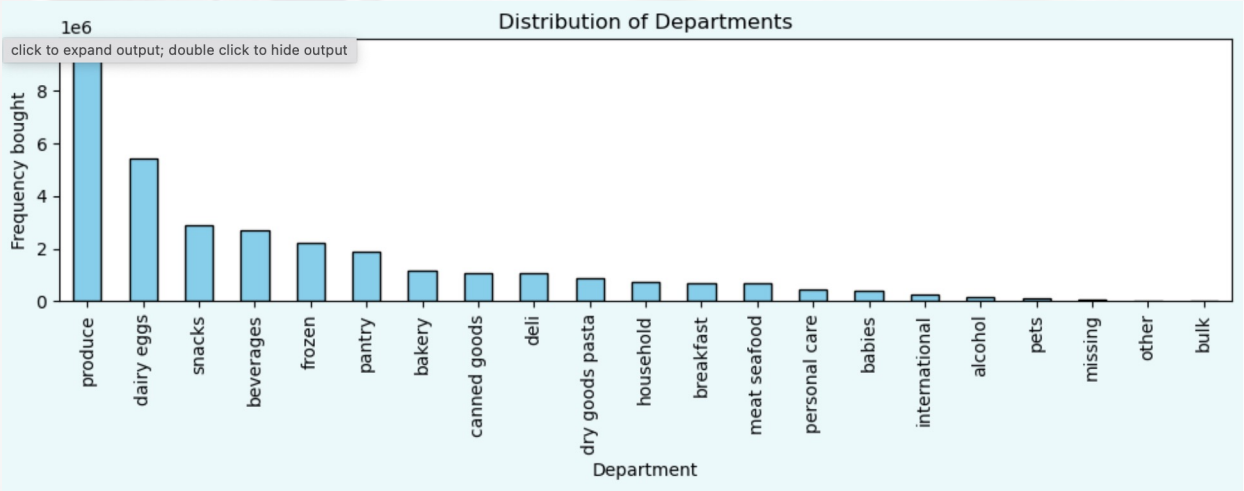
Data processing steps

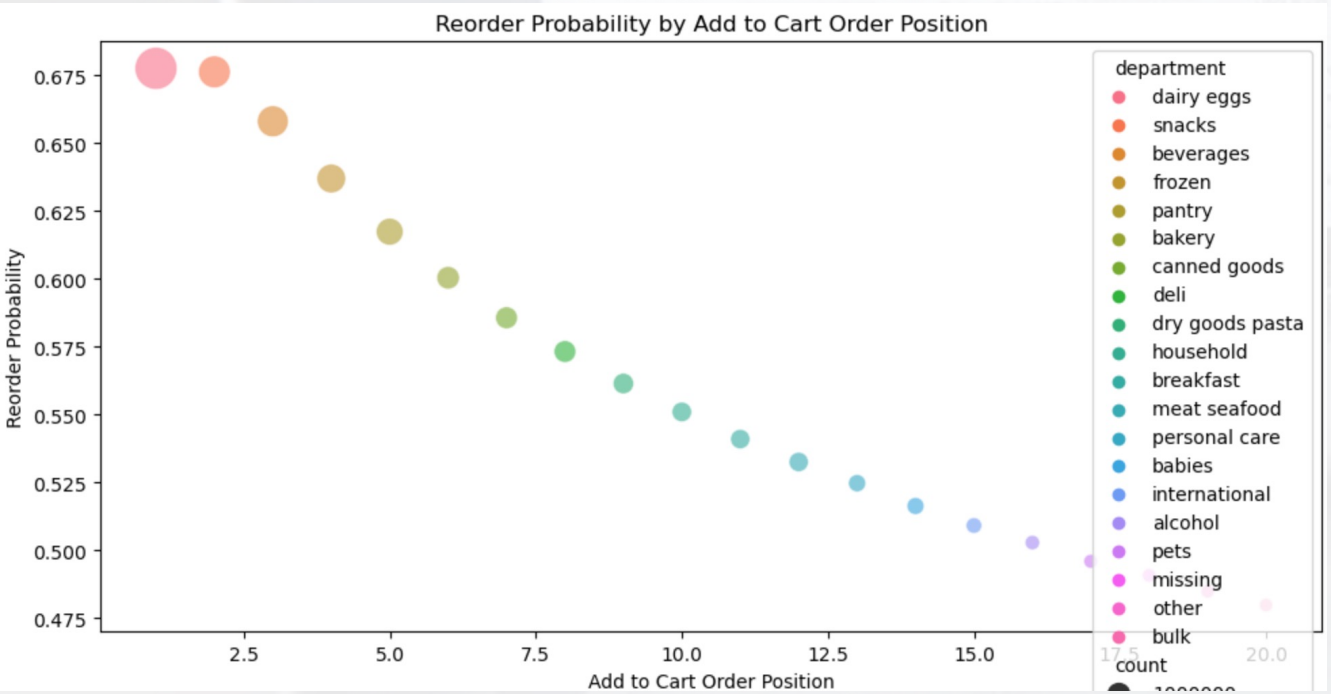
Preprocessing:

- Dropping column not needed for the modelling- aisle_id', 'department_id','aisle','department', 'product_id','user_id'
- Y – Reordered variable
- Vectorization of the Product column using Vectorizer - Bag of words method

Modeling:

- Logistic Regression model, Decision Tree
- KNN Classifier, RandomForest and SVC
- Hyperparameter optimization, ML Pipelines – Grid searchCV, pipeline design
- Market Basket analysis with Apriori algorythm





Important EDA findings

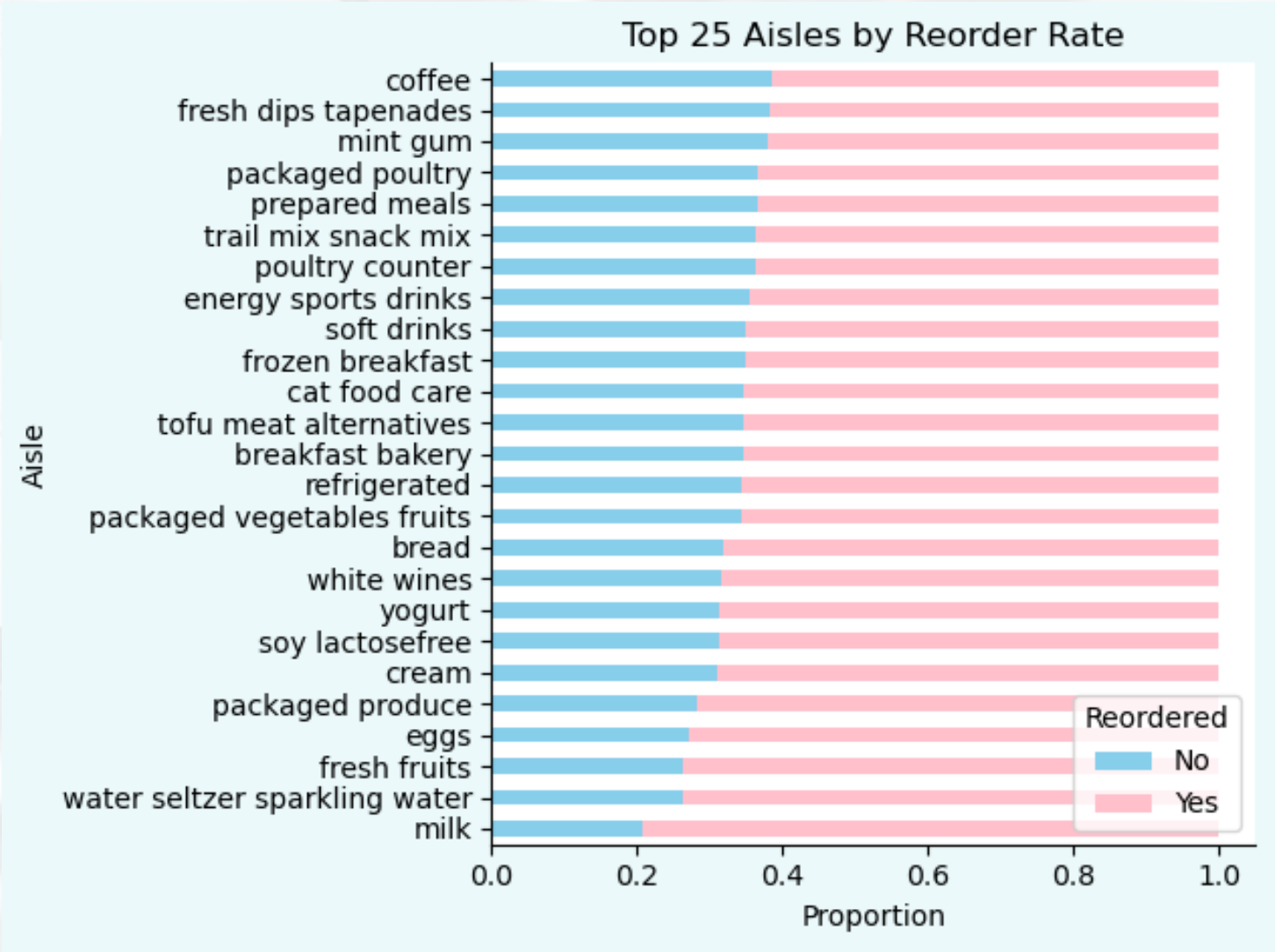
Top-Selling Products:

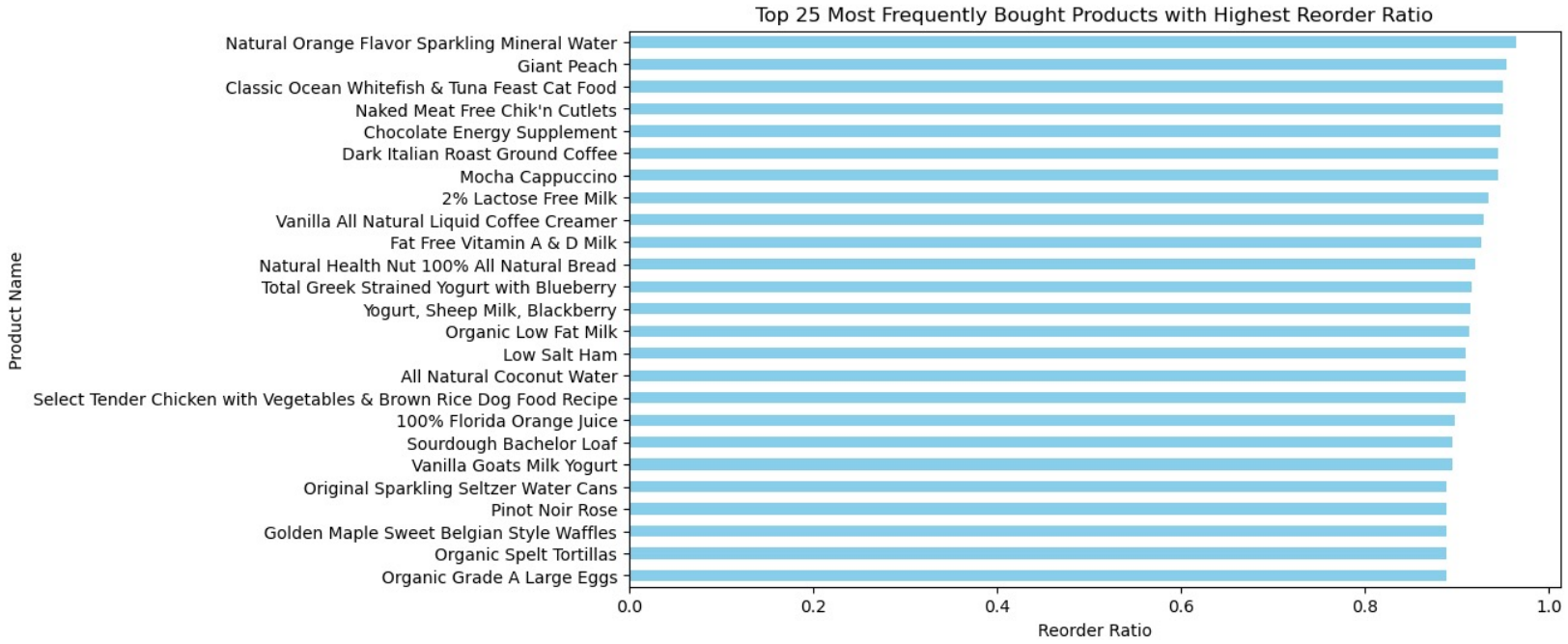
- Fresh fruits and vegetables, particularly bananas and organic bananas.
- A strong emphasis on organic options.
- A variety of berries and avocados, reflecting a preference for 'superfoods'.

Correlation between when an item is added to the cart and its reorder probability, along with the volume of purchases per department.

Important EDA findings

- **Fresh** produce aisles (vegetables and fruits) are top choices.
- **Dairy products** like yogurt, cheese, and milk are popular.
- **Snack**-related aisles (chips, pretzels) and bread cater to various dietary needs.
- Lactose-free options indicate consideration of dietary restrictions.





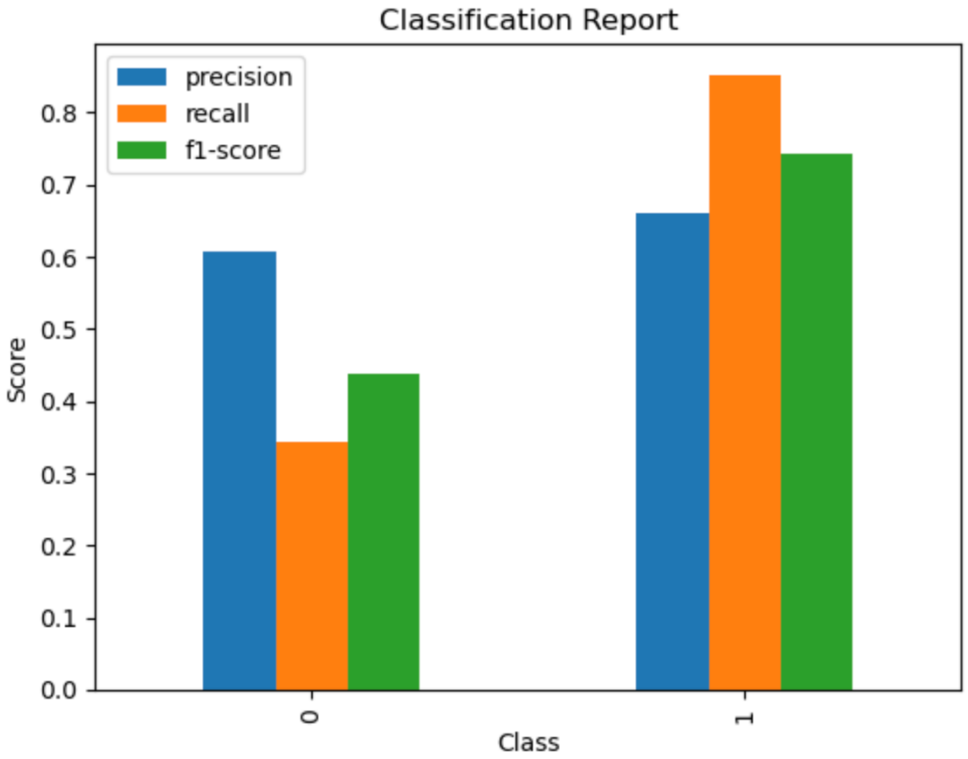
Important EDA findings

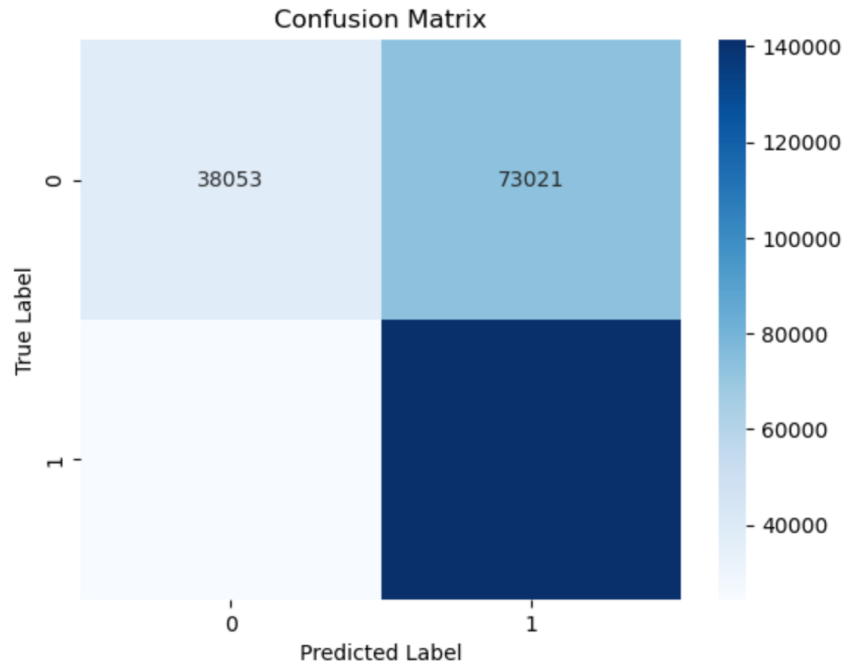
Popular and Staple Products:

- Assortment includes beverages (e.g., Sparkling Mineral Water, Peach juice, Orange Juice), dairy, dairy substitutes, and groceries.
- These items are likely staples in customers' diets, indicating regular purchases.
- Reasons could include product quality, limited substitutes, or customer habits.

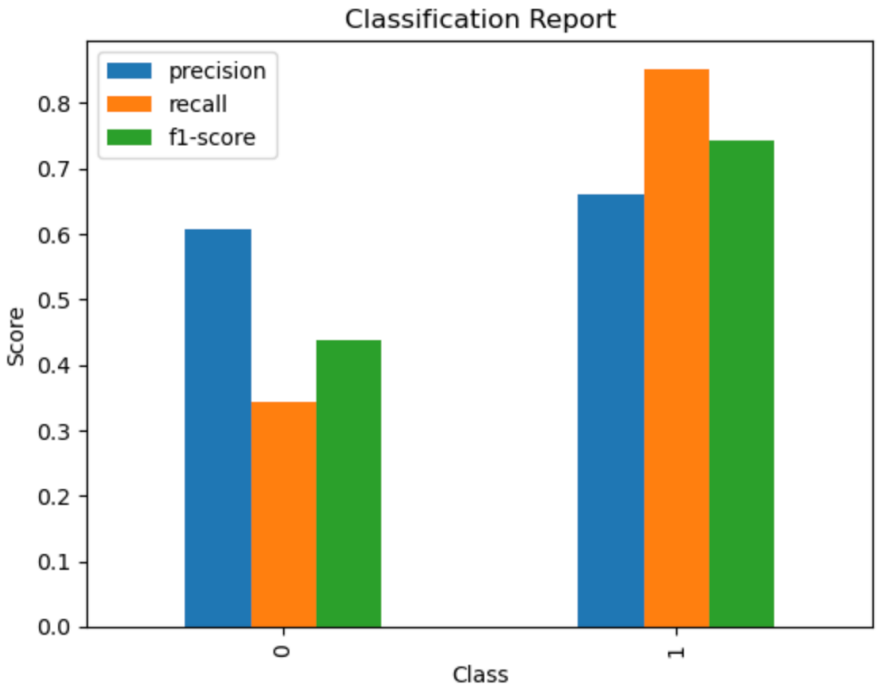
Baseline Logistic regression

The model is better at identifying products that will be reordered than those that will not. This could be because there are more reordered instances in the dataset (evidence of class imbalance), or the features are more predictive for the reordered class.
The relatively high number of False Positives (products predicted as reordered but actually not) suggests that the model might be erring on the side of predicting reorder.





The accuracy is moderately good (65%), but there is definitely room for improvement, especially in terms of precision and recall for the not reordered class.



Models comparison and evaluation metrics

The SVC model is more balanced in terms of generalization, making it a preferable choice for this particular dataset.

It shows a good balance between training and test accuracy. The close performance on both datasets suggests that the model is generalizing well without significant overfitting. This model seems more suitable for the dataset based on these results.

	Hyperparameter	Training Accuracy	Test Accuracy	Notes
Logistic regression	none	66.46%	66.39%	good performance
Logistic regression2	scaled	66.574%	66.514%	performance improved after scaling
Decision Tree	one	66.499%	66.095%	performance is relatively consistent
Decision Tree2	mdepth=11, min_s_leaf=15, min_s_split=6	66.465%	66.111%	slightly better performance for class 0
KNN Classifier	scaled	62.65%	62.65%	low performance
KNN Classifier2	manhattan, n_neighbors=9, w=uniform	76.32%	77.61%	high performance
RandomForest	none	91.42%	79.67%	overfitting, needs adjustment
SVC		82.45%	82.43%	the highest performance
RandomForest optimized	mdepth: 10,min_s_leaf: 1,min_s_split: 10,n_est...	78.22%	77.61%	overfitting, good performance

Association Rules — Market Basket Analysis

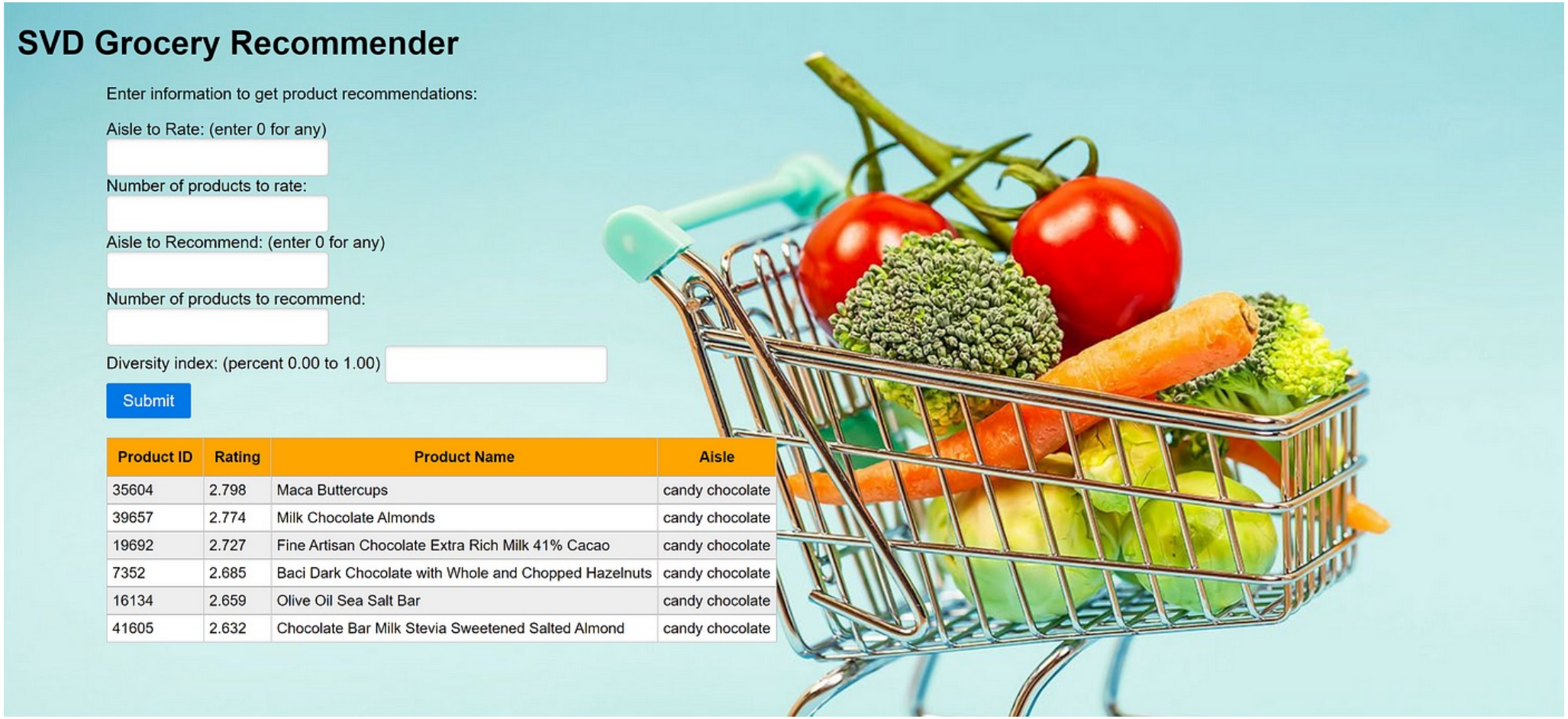
Rule1: (Bag of Organic Bananas) => (Organic Strawberries)

- Support: 2.34% (both items appear together in 2.34% of transactions)
- Confidence: 19.86% (if Bag of Organic Bananas is bought, there's a 19.86% chance of Organic Strawberries being bought too)
- Lift: 2.39 (the likelihood of Organic Strawberries being bought is 2.39 times higher when Bag of Organic Bananas is already in the basket)

Rule2: (Organic Hass Avocado) => (Bag of Organic Bananas)

- Support: 1.84%
- Confidence: 33.18%
- Lift: 2.81

Product demo – recommender system



Product ID	Rating	Product Name	Aisle
35604	2.798	Maca Buttercups	candy chocolate
39657	2.774	Milk Chocolate Almonds	candy chocolate
19692	2.727	Fine Artisan Chocolate Extra Rich Milk 41% Cacao	candy chocolate
7352	2.685	Baci Dark Chocolate with Whole and Chopped Hazelnuts	candy chocolate
16134	2.659	Olive Oil Sea Salt Bar	candy chocolate
41605	2.632	Chocolate Bar Milk Stevia Sweetened Salted Almond	candy chocolate


Thank
You