INSTACART ORDER PREDICTION SYSTEM

JOSH ZHU

SPRINGBOARD DATA SCIENCE CAREER TRACK

CAPSTONE PROJECT
2021

THE PROBLEM STATEMENT

- Instacart: a grocery ordering and delivery app
- Goal: Use anonymized data on customer orders over time to predict which previously purchased products will be reordered
- This is a classification problem
- Possible clients:
  - Instacart
  - Users of Instacart

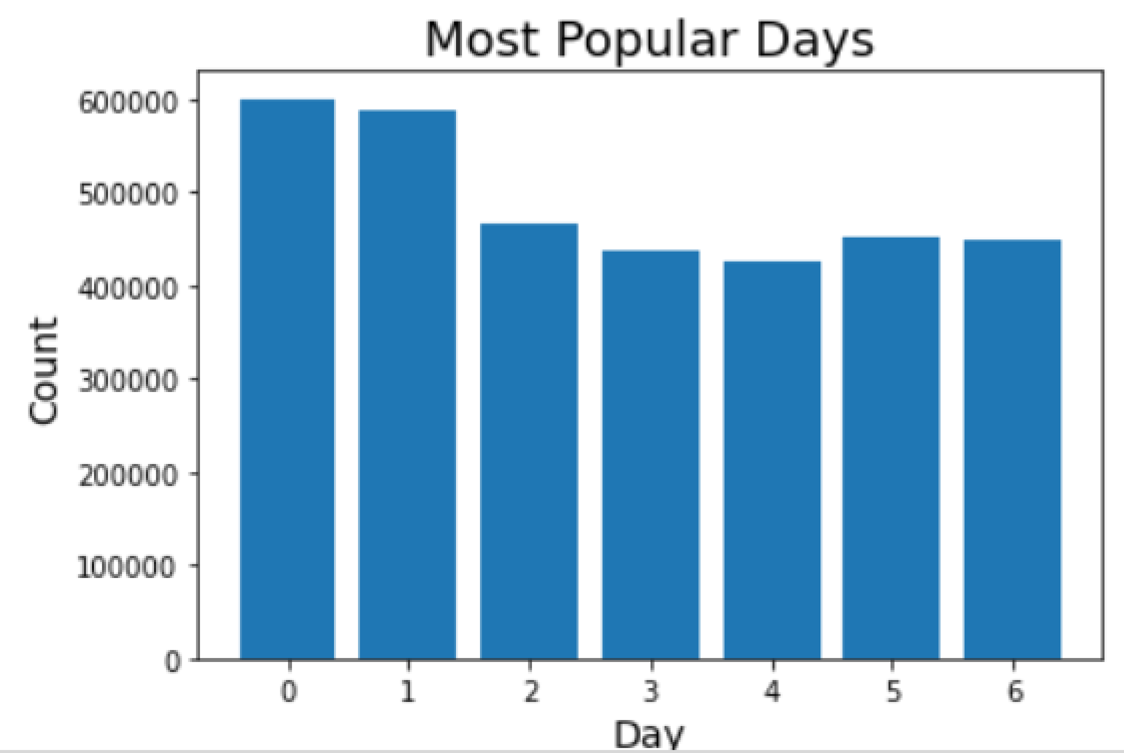
THE DATASET

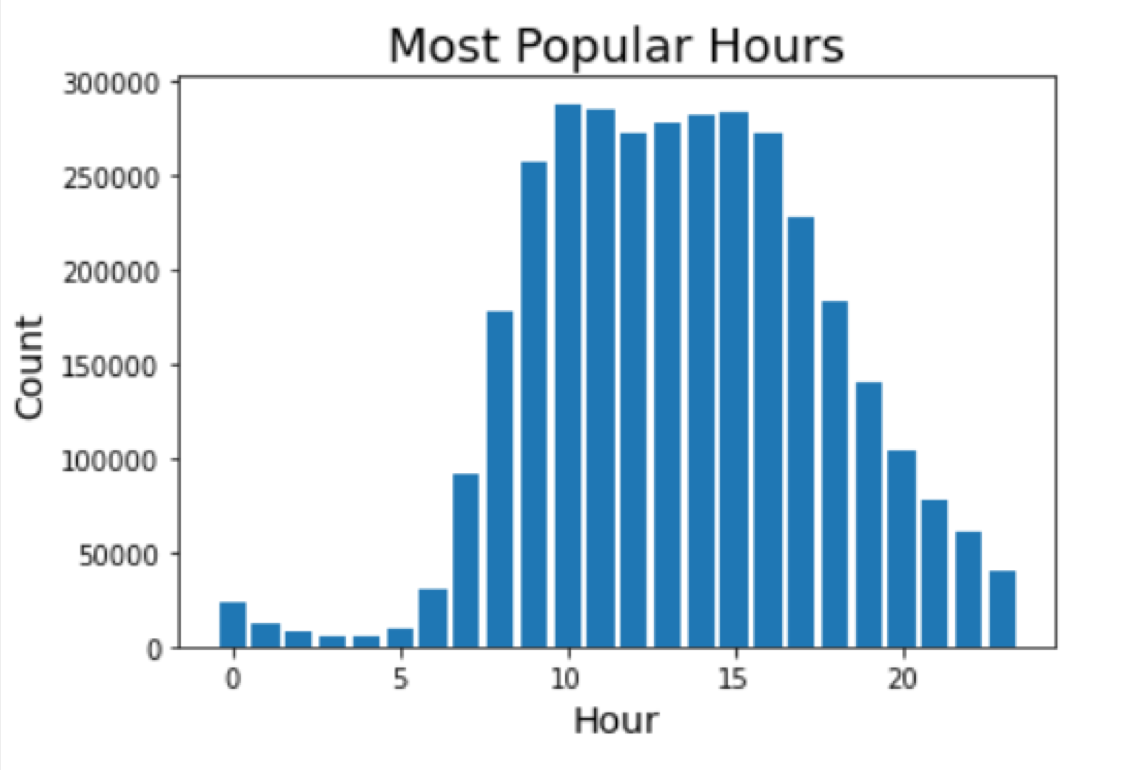
Available on Kaggle

- Downloaded as 6 CSV files from:
- https://www.kaggle.com/c/instacart-market-basket- analysis/data
- CSV files:
  - Aisles
  - Departments
  - Orders
  - Products
  - Prior Orders
  - Train Orders

EXPLORATORY DATA ANALYSIS

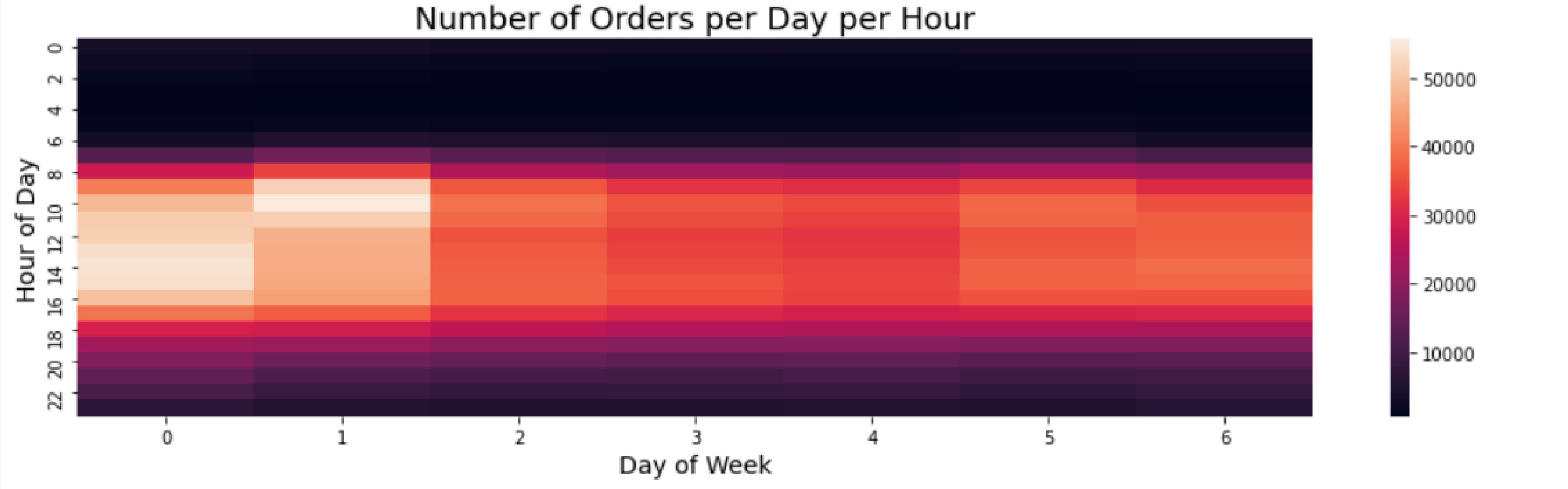
- Orders by day of week and hour of day





EXPLORATORY DATA ANALYSIS

- Orders by day of week and hour of day



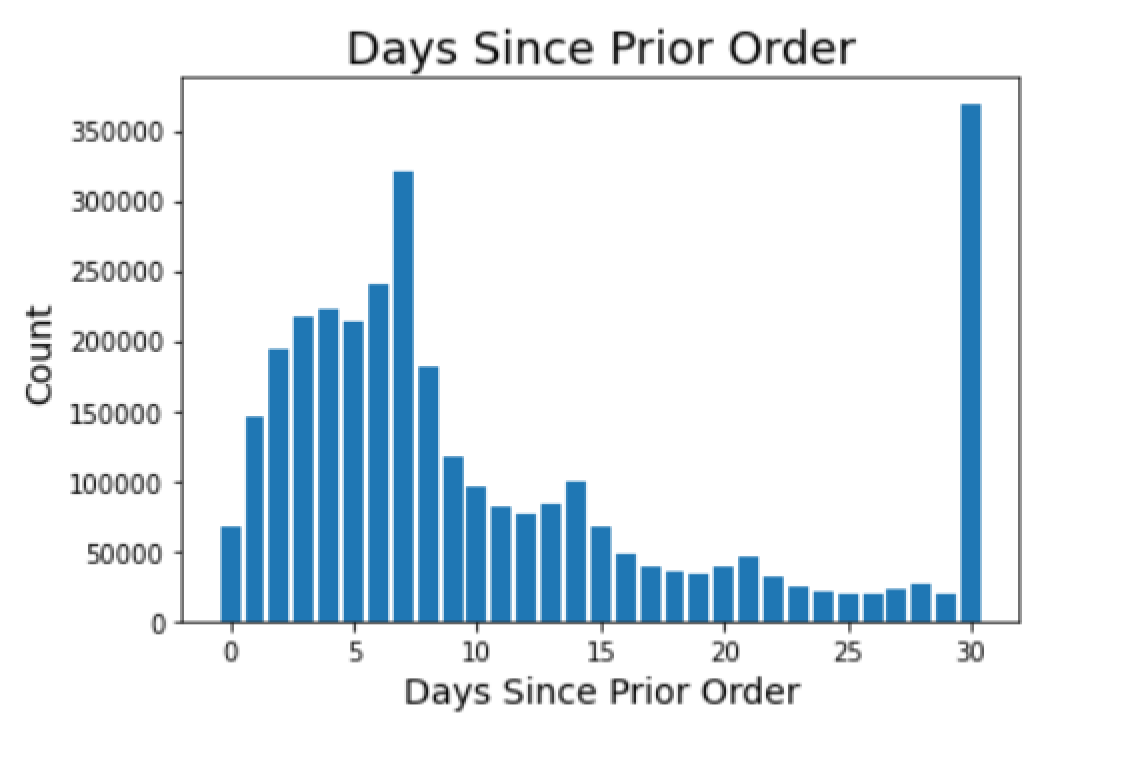
- Very few orders at beginning of day (12AM to 6AM)
- Days 0 and 1 (Sat and Sun/weekends) have the most orders

EXPLORATORY DATA ANALYSIS

- Days since prior order

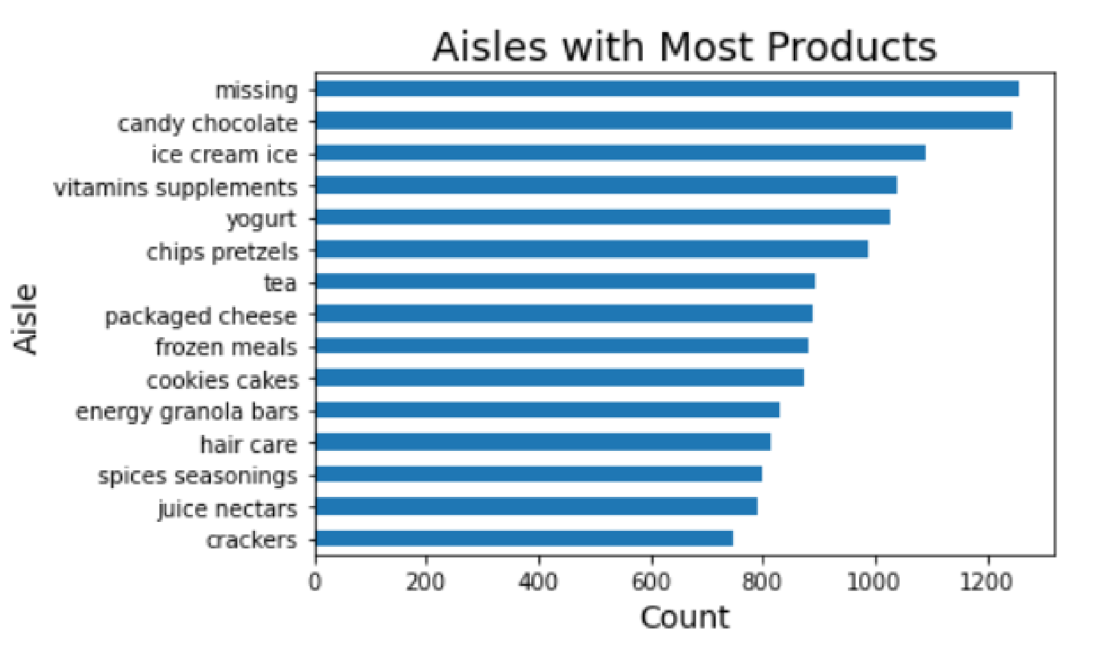
- The majority is 30, which is most likely due to any number of days over 30 being assigned to the 30 category

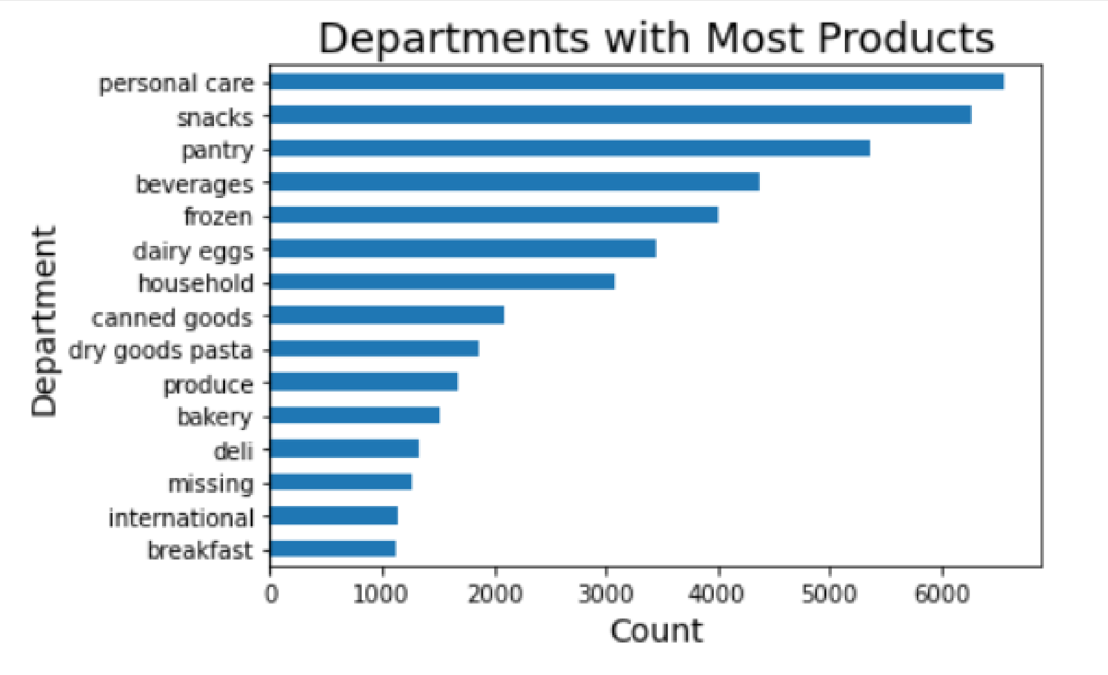
- Many customers order weekly, as the bars for 7, 14, and 21 days are local maxim



EXPLORATORY DATA ANALYSIS

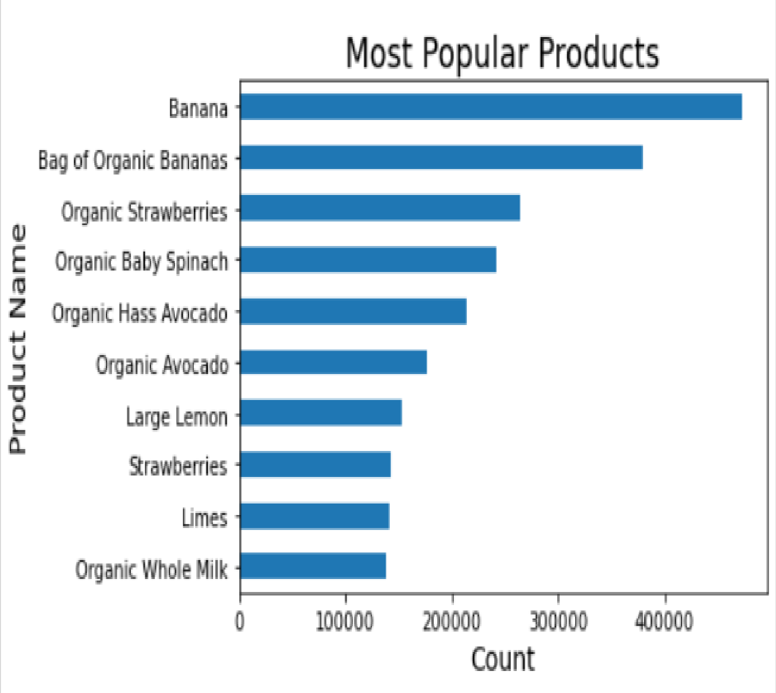
- Inventory: aisles and departments with most products

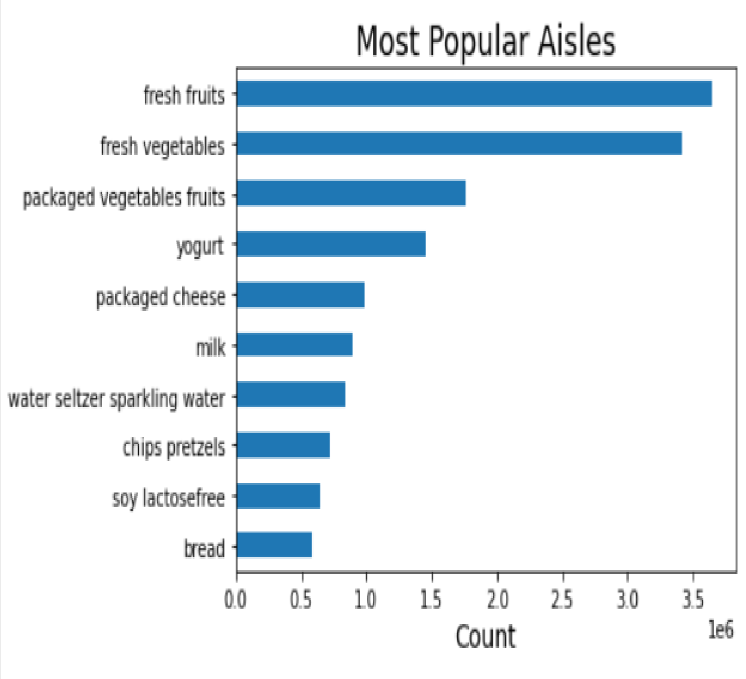


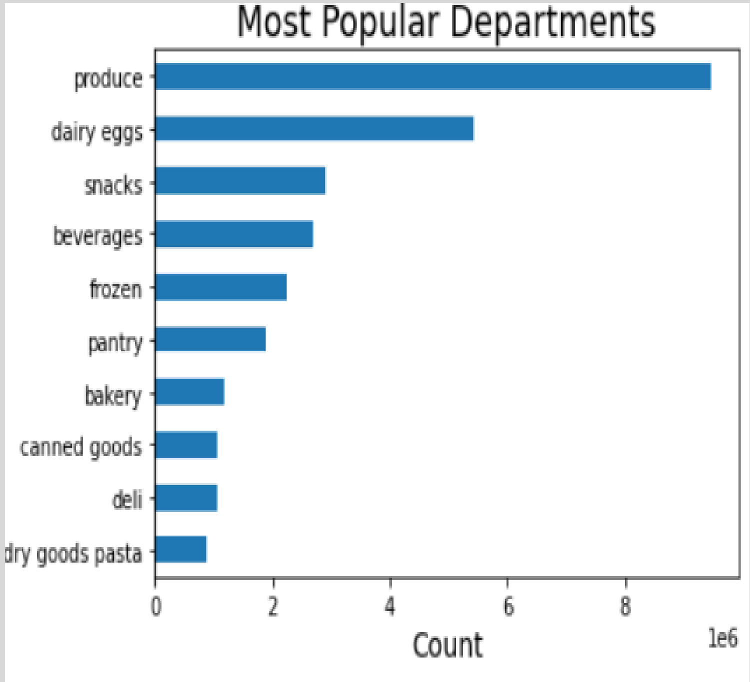


EXPLORATORY DATA ANALYSIS

- Most popular products, aisles, and departments
- Fruits and vegetables, specifically organic, are the most popular

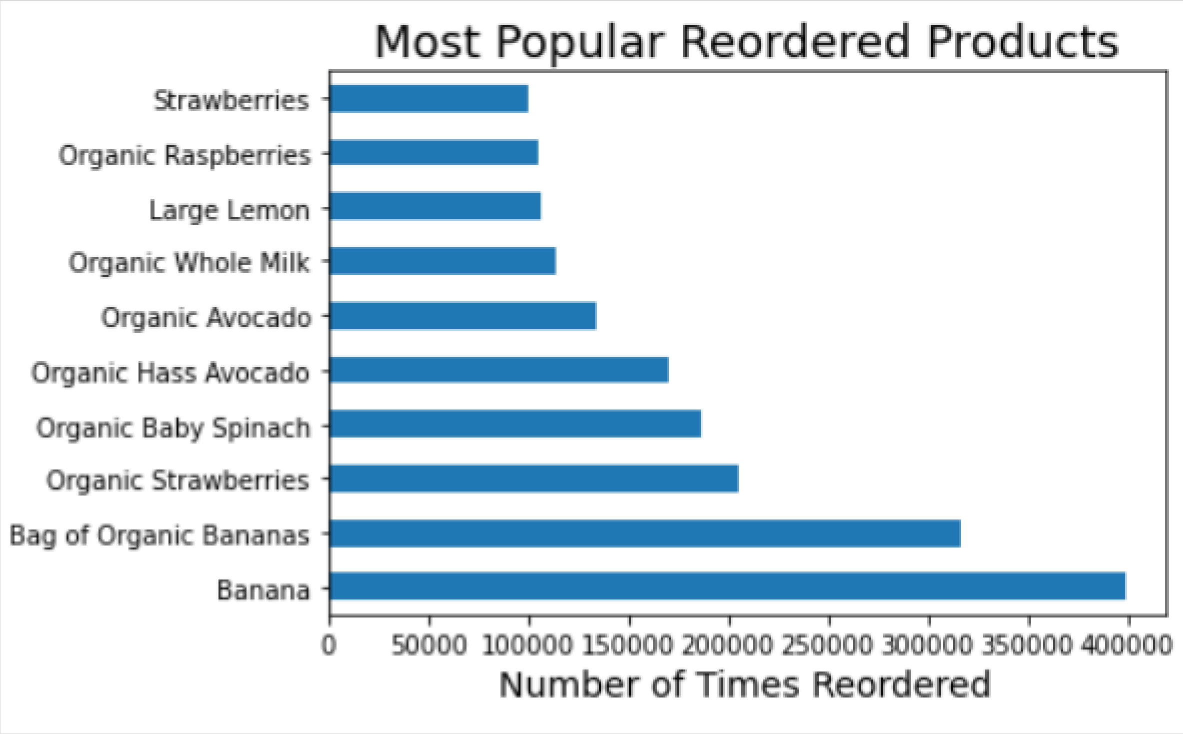






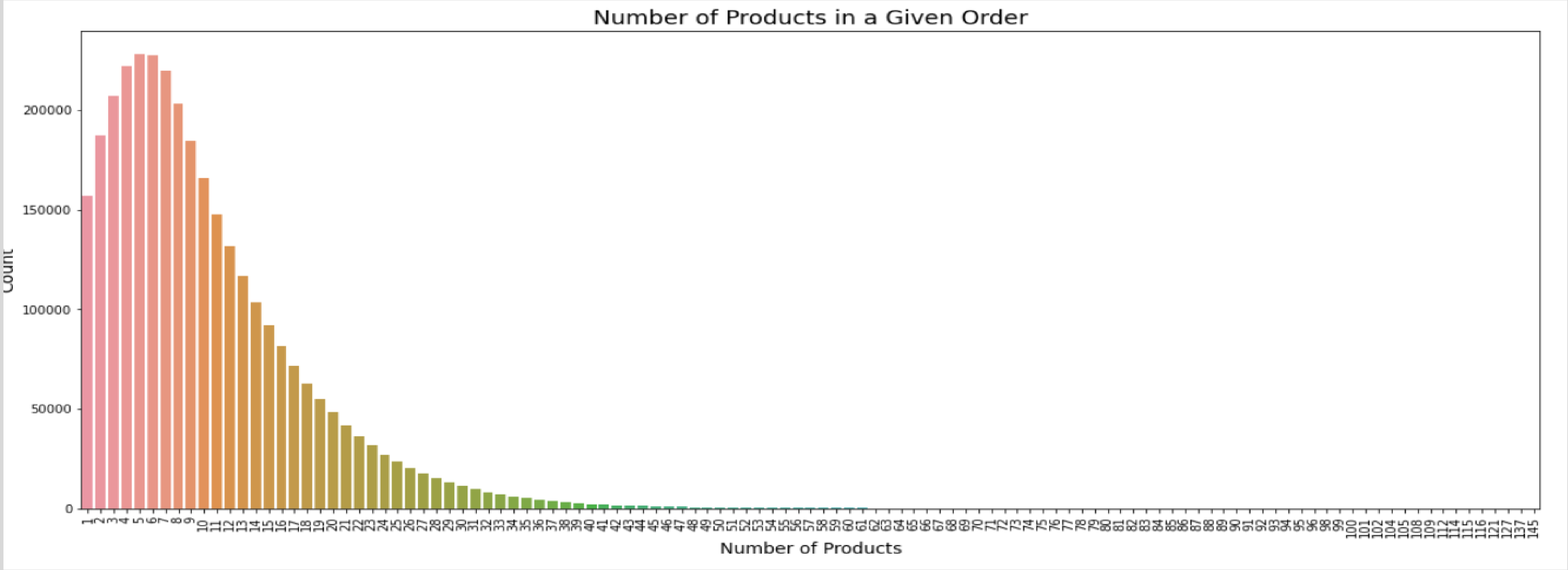
EXPLORATORY DATA ANALYSIS

- Most popular reordered products
- Again, majority are fruits and vegetables
- 59% of the products in the orders were reordered products
- 12% of the orders had no reordered products



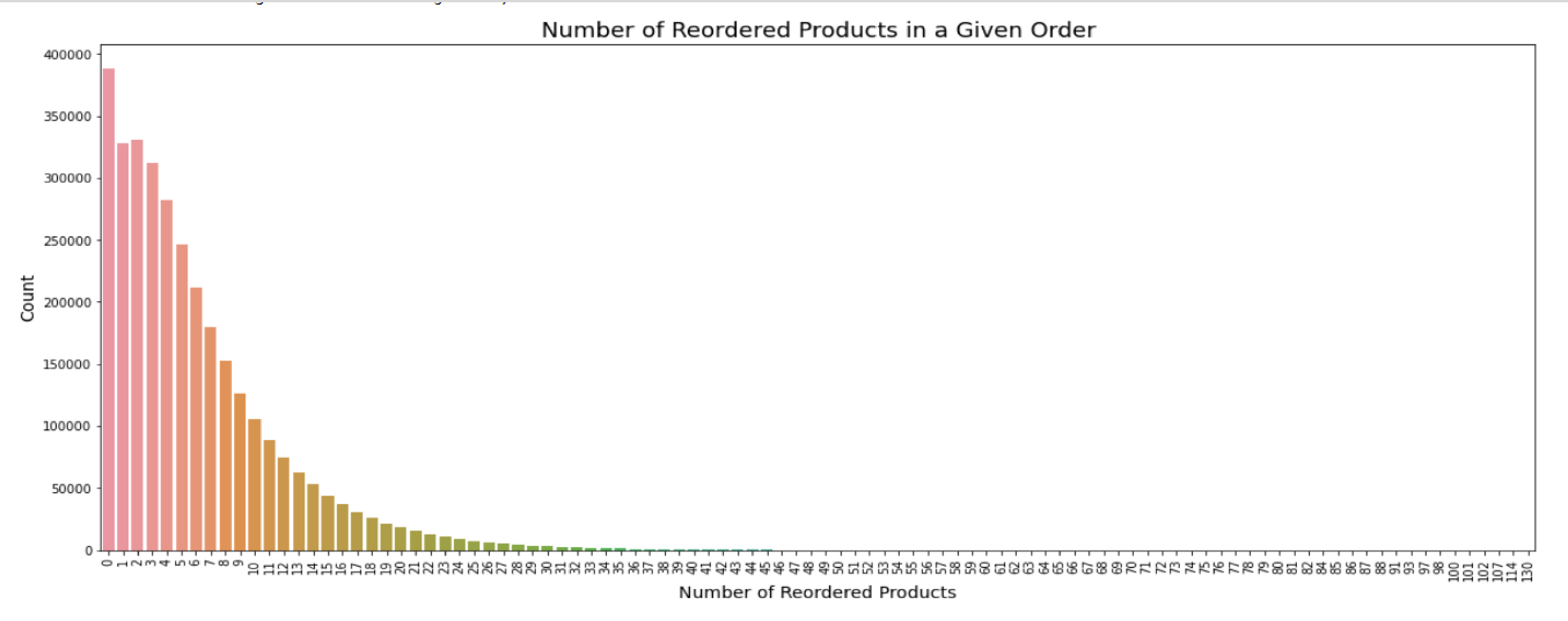
EXPLORATORY DATA ANALYSIS

- Number of products in a given order
- Majority: 5-6 products



EXPLORATORY DATA ANALYSIS

- Number of reordered products in a given order
- Majority: 0- reordered products



Feature Engineering

- User features
  - Reordered ratio
  - Total number of orders
  - Total items purchased
  - Average days since prior
- Order
  - Average basket size
  - Product features
  - Reordered ratio
  - Number of purchase
- Order features
  - Order number
  - Aisle ID
  - Department ID
  - Day of week
  - Hour of day
  - Days since prior order

Preprocessing

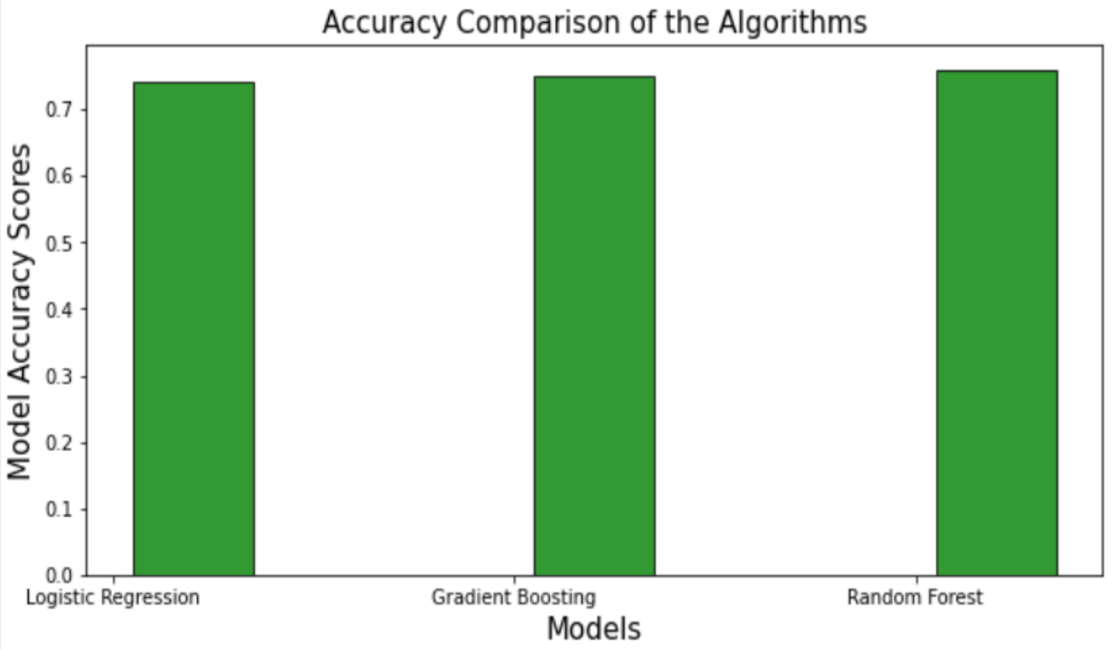
- Mean encoding: aisle ID and department ID
- Split data into training set and testing set
- Imbalanced dataset: downsampled the majority class
- Total number of 0's: 580,297
- Total number of 1's: 580,297

Modeling

- Logistic Regression
- Best value for regularization parameter, C: 1
- Gradient Boosting
- Random Forest

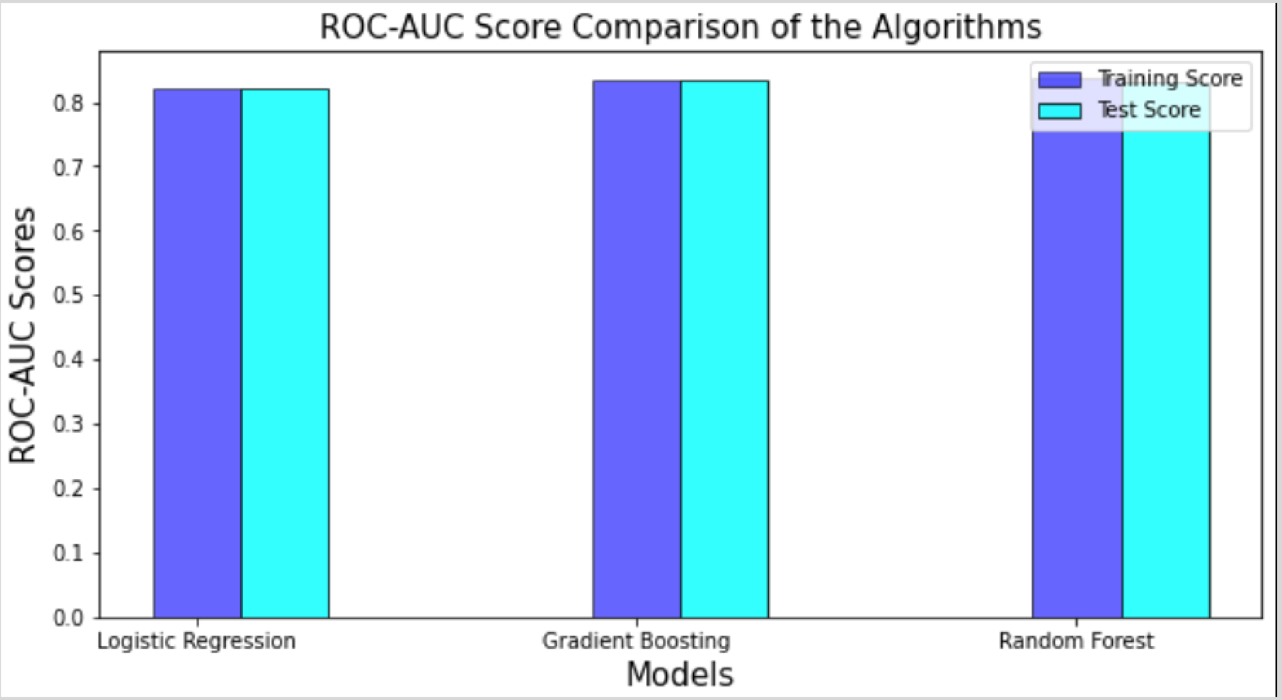
RESULTS: ACCURACY SCORES

Algorithm	Model Accuracy Score
Logistic Regression	0.741736
Gradient Boosting	0.751214
Random Forest	0.759345



RESULTS: ROC-AUC TRAIN AND TEST SCORES

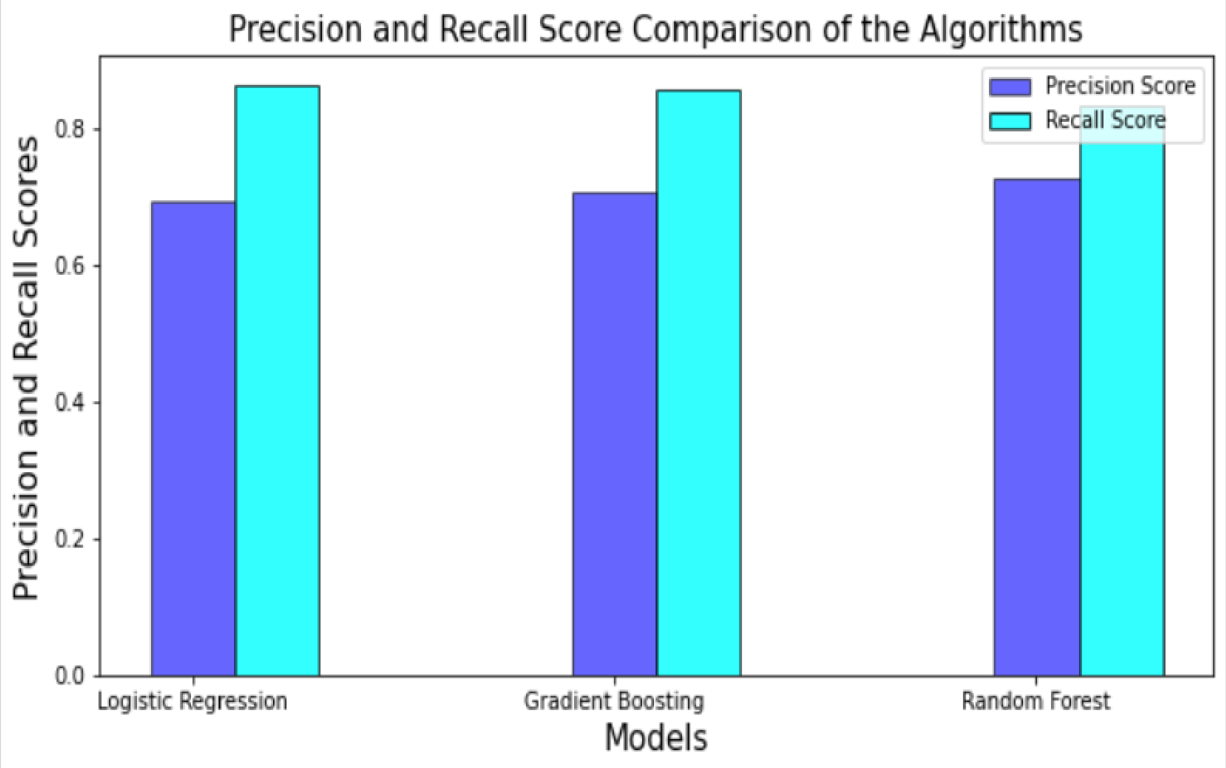
Algorithm	ROC-AUC Train Score	ROC-AUC Test Score
Logistic Regression	0.821613	0.820162
Gradient Boosting	0.834044	0.832633
Random Forest	0.837478	0.830043



RESULTS: PRECISION AND RECALL SCORES

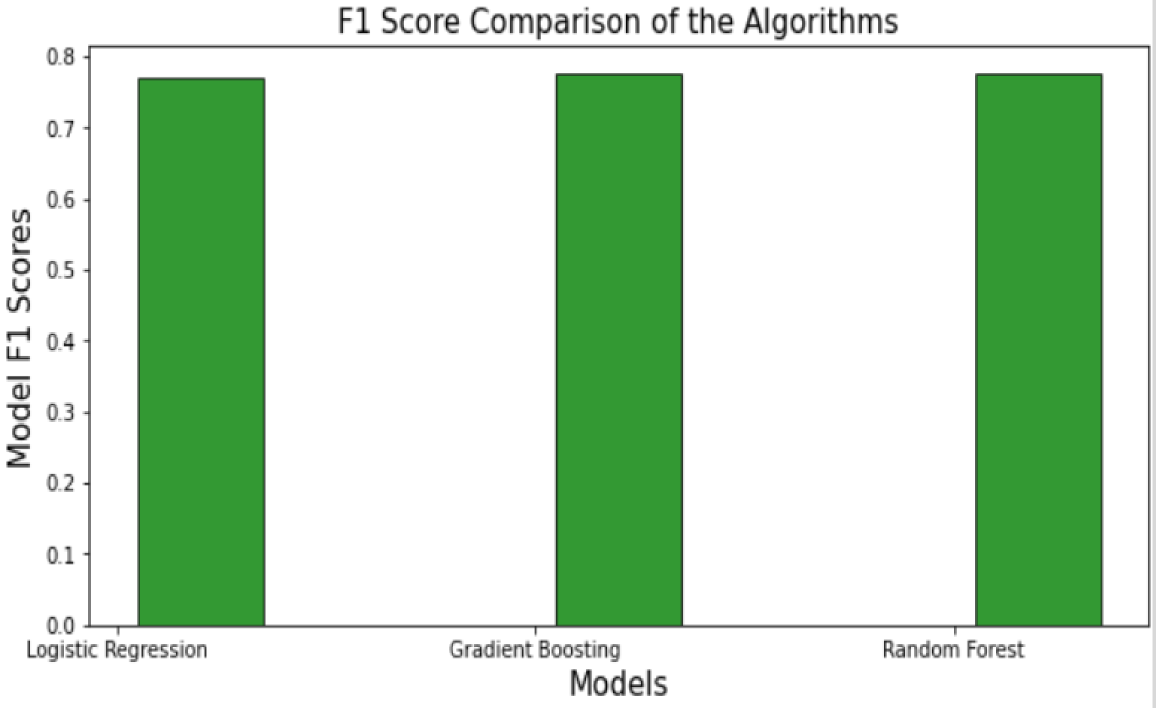
Algorithm	Model Precision Score
Logistic Regression	0.694425
Gradient Boosting	0.707842
Random Forest	0.725366

Algorithm	Model Recall Score
Logistic Regression	0.863403
Gradient Boosting	0.855551
Random Forest	0.834728



RESULTS: F1 SCORES

Algorithm	Model F1 Score
Logistic Regression	0.769750
Gradient Boosting	0.774719
Random Forest	0.776214



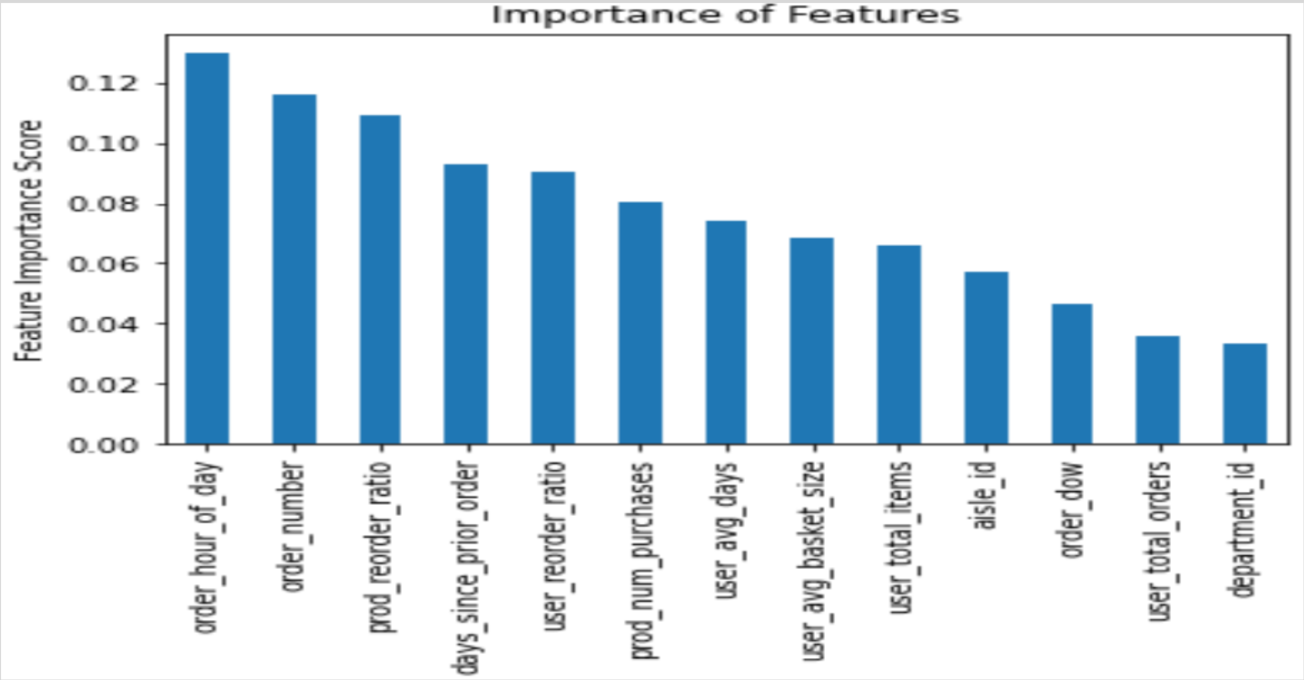
RESULTS: TRAIN AND PREDICT TIMES

Algorithm	Train Time (s)	Predict Time (s)
Logistic Regression	1.542	0.007
Gradient Boosting	154.664	0.560
Random Forest	216.206	10.848

Based on the accuracy and F1 score, the Random Forest model is the best model.

FEATURE IMPORTANCE

- Most important features: hour of day, order number, product reorder ratio
- Least important features: user’s total orders, department ID



CONCLUSION AND FUTURE WORK

- Developed a Random Forest model with 75.93% accuracy and F1 score of 0.776
- Extracted feature importance
- Future work:
- Engineer more features from the users and products combined
- For example, how many times a user bought a particular product
- Construct more advanced models