Midproject Update:

InstaCart

michael fedell - 2019-05-15

Some highlights from this sprint

Wrangling

3.5 million orders used to define profiles for 206k customers. Data integrated and analyzed across 5 different tables and used to engineer two final datasets for model.

Targets

Orders were classified as one of 11 archetypes (using GMM). Each user's final order in dataset was held out as the prediction target and aligned with user profiles.

Automation

Project infrastructure established and automated pipelines built for data acquisition, processing, and ingestion. Project management pipelines built as well.

Review of Progress

Data Processing

Data was aggregated across various sources and cleaned, analyzed, and processed for modeling.

Order Classification

Various clustering approaches were evaluated against the set of past orders to classify each basket under an archetype.

User Profiles

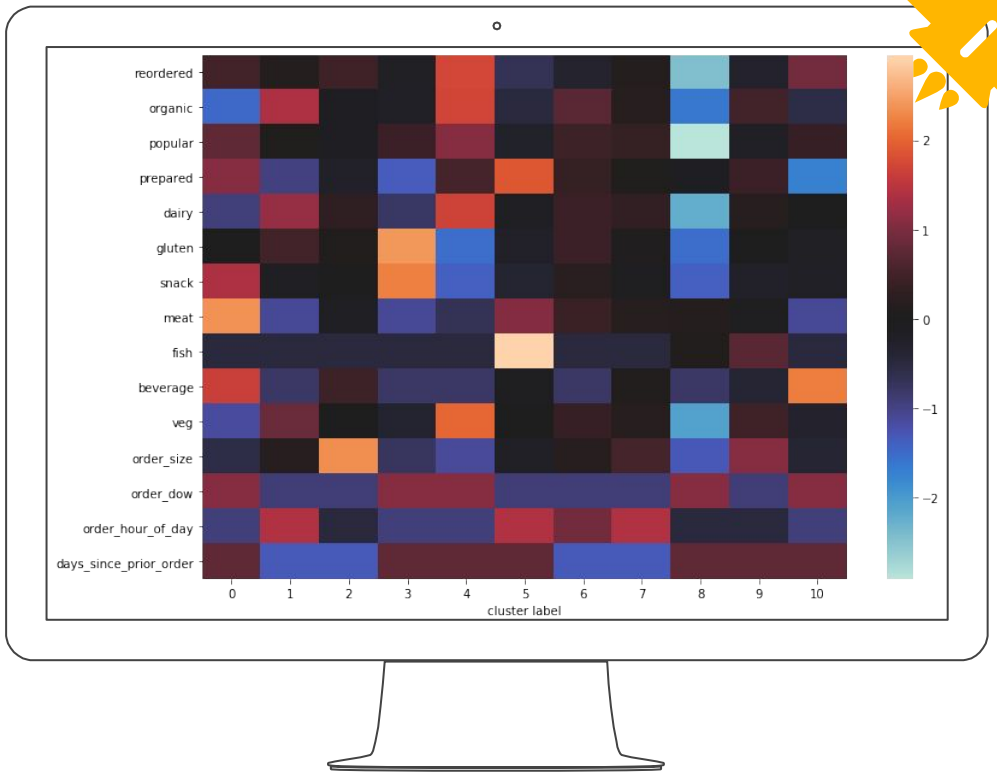
All users in the dataset were profiled according to a relevant, but different set of attributes as compared to order class.

Infrastructure

All the necessary infrastructure was set up to facilitate further development (S3, RDS, EC2, and automation)

Demo analysis

A Gaussian Mixture Model was fit to all orders in the data and then each order_type cluster analyzed based on its characteristics. Given a customer's history we will predict what type of order they will place next.



Lessons learned

- Use of Makefiles to manage pipeline
- File acquisition and management via **`curl`**, **`tar`**, etc
- Programmatic file search with python **`glob`**
- AWS Security groups and IAM setup
- Yaml parsing to set configurations and **`kwarg`** expansion
- Setting and reading environment variables for script execution and resource configuration

Recommendations - what's next

Classifier Training

Several model types and variations will be evaluated for predicting order type of user's next purchase.

Evaluation/Scoring

Process will be developed for scoring the model on a single record as well as evaluating the model's performance against a test set.

PCA

PCA will be explored as a means of simplifying the user-profile feature space. This simplification will allow mapping user input to the model

Interface

A sleek user interface will be developed to allow visitors to describe themselves intuitively and see the corresponding user-profile and next-order basket-type prediction

Visualization

Model results, user profiles, and basket types will be further visualized to augment the usefulness and accessibility of the application

Optimization

Once standing, the web app will be examined fully for optimization. Infrastructure may be modified for purpose of speeding up user interactions