
Predicting Optimal Meal Kit Choices: a Comparison of Methods

By Robert N. Nakano
October 2, 2020

Committee Members

Olga Korosteleva, Ph.D. (Chair)
Kagba N. Suaray, Ph.D.
Alan Safer, Ph.D

Outline: Predicting Optimal Meal Kit Choices

Meal Kits

Survey

- Survey Design
- IRB Process
- Descriptive Statistics

Algorithms

- Collaborative Filtering
- Content-based Filtering
- Deep Learning Approaches
- Results

Future Work

Background

Meal Kit Services



Robert Nakano

Applied Statistics, CSULB
Business Economics, UCI
Film (Assistant Editor)
Middle School Math Teacher
Teaching Associate Math 104
Hobbyist Web Developer

Aaron Yoshitake

Pick a Kit Founder

Computer Science, UCB
Web Developer
Meal Kit Aficionado

What is a mealkit?

Meal kits are boxes containing premeasured and packaged ingredients for one or more recipes that are delivered to a buyer's address, oftentimes on a subscription basis.



Gobble

martha
&
MARLEY SPOON

 **HelloFRESH**

SUN BASKET


HOME CHEF

PICK A KIT =



PICK A KIT

[Menus](#)[Discounts](#)[Meal Kit Services ▾](#)[FAQs](#)[Blog](#)[Survey](#)

Find your perfect meal plan.

Compare menus and prices for [Blue Apron](#), [HelloFresh](#), and more

[Affiliate Disclosure](#)

Menus for Oct 4 to Oct 10

[◀ Previous Week](#) [Next Week ▶](#)

Plan type:

[2-Person](#)

Meal type:

- Meat
- Vegetarian
- Gluten-Free

Sun Basket (35)

Pork Chops and Onion Gravy
with Brussels Sprout Slaw



Mediterranean Turkey Meatballs
with Chard and Red Pepper-cashew Crema



Steak Stir-Fry
with Sweet Peppers and Baby Bok Choy



Spicy Thai Chicken Skewers
with Mango-Radish Salad

Green Chef (14) [Balanced Living Plan](#)

Problem Statement

Which meal kit service is best for each individual?

Approach Overview

1. Run a survey to collect ratings (1-5) for meal kit services and other data
2. Use data to test ratings prediction algorithms

Evaluation Metrics

1. Accuracy- RMSE, and MAE
2. Prediction Coverage
3. Computation Time

2. Rate each meal kit service you have tried from 1 (strongly dislike) to 5 (love).

Service Name	Overall Service Rating	Service Comments
Marley Spoon	<input type="button" value="1"/> <input type="button" value="2"/> <input type="button" value="3"/> <input type="button" value="4"/> <input checked="" type="button" value="5"/>	I liked learning how to make roasted potatoes.
Blue Apron	<input type="button" value="1"/> <input checked="" type="button" value="2"/> <input type="button" value="3"/> <input type="button" value="4"/> <input type="button" value="5"/>	Not enough potato dishes.

Add Service

Survey

Survey Design

Nonprobability Survey

Recruitment over a 24 week period in early
2020

Recruitment from various channels:

- Personal networks
- Survey sharing groups
- Interest groups
- Facebook
- Twitter
- Reddit
- Pinterest
- Pick a Kit

Capture preferences on users for meal kit services (i.e. ratings 1-5)

Learn about other user preferences that may influence meal kit decisions

Provide options for further research

Survey Design- Architecture

Data stored on MongoDB NoSQL databases

JSON format

Survey.js

Hosted on Pick a Kit

Page 1 of 3

Welcome

Welcome to the Pick a Kit survey on meal kits, a research project in collaboration with California State University, Long Beach. Your response will help us figure out the best meal kit recommendations for each person.

The basic version of the survey takes about 4 minutes. After taking the survey, we would love to share the results with you!

To continue, please read and agree to the [Notice of Informed Consent](#) and the [Pick a Kit Privacy Policy](#).

I am 18 years of age or older, and understand and agree to the Notice of Informed Consent.

I understand and agree to Pick a Kit's Privacy Policy.

13

Next

The IRB Process

“The Institutional Review Board (IRB) is an administrative body established to protect the rights and welfare of human research subjects recruited to participate in research activities conducted under the auspices of the institution with which it is affiliated.”

When do you need to submit to IRB?

Human Subject + Research Activity

Project is considered research activity when:

- collecting information through interaction with individuals
- analyzing identifiable private information (individuals can directly or indirectly be identified)
- not business related

Step 0: Figure out your research project

- Research goals
 - Interests?
 - Target Population?
 - Access and Recruitment?
- Resources
 - \$
 - Time
- Team
 - Advisor
 - Committee
 - Other Researchers
 - Industry Counterparts

Fields of Study

Recommender Systems
Food Sciences and Nutrition

Goal:

Design and administer a survey
on meal kit preferences

Investigate methods to predict
optimal meal kit choices

Nonprobability Survey

Step 1: Visit CSULB IRB Website

- Instructions:

<https://www.csulb.edu/office-of-research-and-sponsored-programs/institutional-review-board-irb>

3 Types of IRB Applications

- Submission to the *IRB* is required. Which IRB Application should I submit?
- IRB Application for [Existing and Secondary Data \(DOC\)](#)
 - **For Projects involving:**
 - Secondary analysis of identifiable data
 - Retrospective and/or prospective secondary data analysis
- IRB Application for [Administrative and Limited Preview \(DOC\)](#)
 - **For Projects involving:**
 - Surveys, interview and focus groups (release of data will not place subjects at harm)
 - Benign behavioral interventions with adults
 - No children or other vulnerable populations
- IRB Application for [Expedited and Standard Review \(DOC\)](#)
 - **For Projects involving:**
 - Interventions and assessments (minimal and greater than minimal risk)
 - Behavioral interventions
 - Inclusion of children or other vulnerable populations

Step 2: Complete Citi Training

- Social & Behavioral Basic/Refresher Course
- 10 hours
- Free for CSULB students

The screenshot shows the CITI Program website. At the top, there is a navigation bar with links for Support, FAQ, Contact Us, Register, and Log In. The main header features the CITI PROGRAM logo and three dropdown menus: Courses, Organizations, and Individuals. Below the header, a large banner with a background image of two people in lab coats displays the text: "The Trusted Standard in Research, Ethics, and Compliance Training". Underneath this, a paragraph describes the CITI Program's mission to serve training needs for various institutions and organizations. At the bottom of the banner, there are two buttons: "Demo a Course" and "View Catalog". To the right of the banner, a large graphic of a hand holding a tablet is shown, with a classroom scene displayed on the screen. Below this graphic, a blue banner states: "Utilized by the Top-25 Best National Universities*". The page also includes a footer with three small circular icons.

THE BELMONT REPORT

Office of the Secretary

Ethical Principles and Guidelines for the Protection of Human
Subjects of Research

The National Commission for the Protection of Human Subjects of
Biomedical and Behavioral Research

April 18, 1979

AGENCY: Department of Health, Education, and Welfare.

ACTION: Notice of Report for Public Comment.

SUMMARY: On July 12, 1974, the National Research Act (Pub. L. 93-348) was signed into law, thereby creating the National Commission for the Protection of Human Subjects of Biomedical and Behavioral Research. One of the charges to the Commission was to identify the basic ethical principles that should underlie the conduct of biomedical and behavioral research involving human subjects and to develop guidelines which should be followed to assure that such research is conducted in accordance with those principles. In carrying out the above, the Commission was directed to consider: (I) the boundaries between biomedical and behavioral research and the accepted and routine practice of medicine, (II) the role of assessment of risk-benefit criteria in the determination of the appropriateness of research involving human subjects, (III) appropriate guidelines for the selection of human subjects for participation in such research and (IV) the nature and definition of informed consent in various research settings.

The Belmont Report attempts to summarize the basic ethical principles identified by the Commission in the course of its deliberations. It is the outgrowth of an intensive four-day period of discussions that were held in February 1976 at the Smithsonian Institution's Belmont Conference Center supplemented by the monthly deliberations of the



Completion Date 01-Dec-2019
Expiration Date 30-Nov-2022
Record ID 34321339

This is to certify that:

Robert Nakano

Has completed the following CITI Program course:

Social & Behavioral Research - Basic/Refresher (Curriculum Group)

Social & Behavioral Research - Basic/Refresher (Course Learner Group)

1 - Basic Course

CITI
(Stage)

Collaborative Institutional Training Initiative

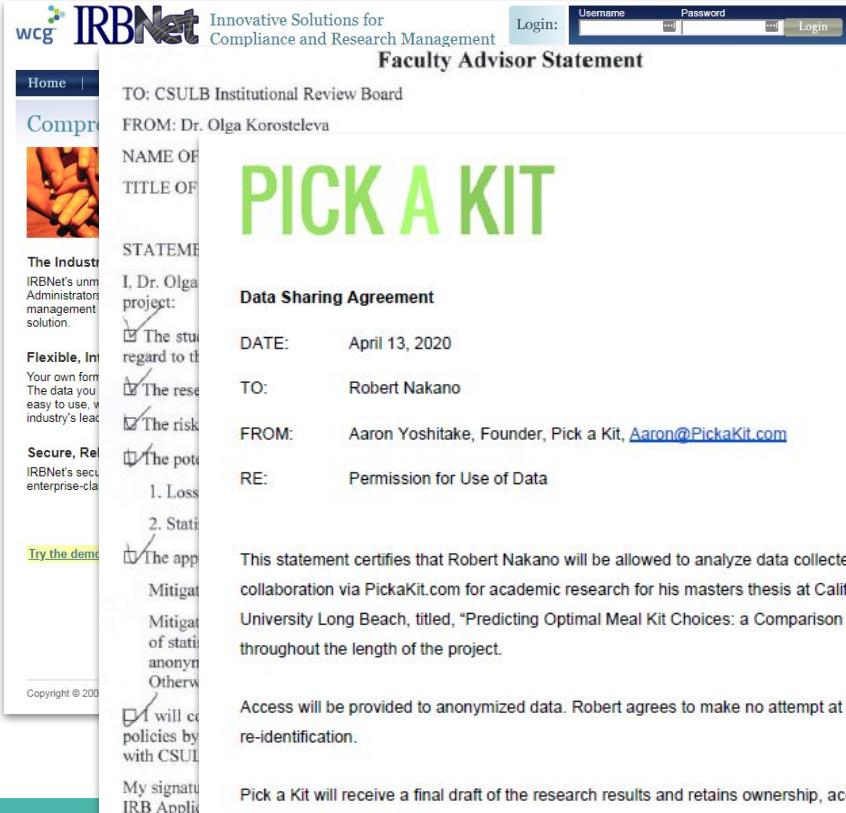
Under requirements set by:

California State University, Long Beach

Verify at www.citiprogram.org/verify/?wa5af59f5-5ba5-4780-b0fe-8ab669f9091e-34321339

Step 3: Submit Required Documents

1. Citi Training Certificate
2. Permission Letters
3. Faculty Advisor Letter
4. Online Survey
5. Consent Notice
6. Recruitment Material



The screenshot shows a software interface for 'IRBNet' from 'wgc'. The top navigation bar includes 'Home', 'About', 'Contact', 'Log In', and 'Logout'. The main title is 'Faculty Advisor Statement'. The form fields are:

TO: CSULB Institutional Review Board
FROM: Dr. Olga Korosteleva

NAME OF
TITLE OF
STATEMENT
I, Dr. Olga project:
 The stu regard to th
 The res
 The risk
 The pot
1. Loss
2. Statistic
 The app
Mitigat
Mitigat
of stati
anonym
Otherw
 I will co
policies by
with CSUL
My signatu
IRB Appli

PICK A KIT

Data Sharing Agreement

DATE: April 13, 2020
TO: Robert Nakano
FROM: Aaron Yoshitake, Founder, Pick a Kit, Aaron@PickaKit.com
RE: Permission for Use of Data

This statement certifies that Robert Nakano will be allowed to analyze data collected in collaboration via PickaKit.com for academic research for his masters thesis at California State University Long Beach, titled, "Predicting Optimal Meal Kit Choices: a Comparison of Methods", throughout the length of the project.

Access will be provided to anonymized data. Robert agrees to make no attempt at data re-identification.

Pick a Kit will receive a final draft of the research results and retains ownership, access, and rights to the data collected.

IRB Application for Administrative & Limited Review

Projects involving less than minimal risk

Instructions: Complete all questions regarding the proposed project. Use as much space as necessary and be specific. Refer to the end of the document for term definitions. Check boxes can be filled in by clicking inside the box once.

IMPORTANT: NO ACTIVITY MAY BEGIN ON THIS PROJECT UNTIL THE PRINCIPAL INVESTIGATOR HAS RECEIVED FORMAL NOTIFICATION FROM THE CSULB IRB THAT THE PROJECT HAS BEEN ACKNOWLEDGED AS A QUALITY ASSESSMENT/QUALITY IMPROVEMENT PROJECT UNDER ADMINISTRATIVE REVIEW.

1. BASIC INFORMATION

PI's Name (Last, First, Degree)	Click or tap here to enter text.
Telephone Number	Click or tap here to enter text.
Email	Click or tap here to enter text.
CITI Member ID #	Click or tap here to enter text.
Completion of CITI Social & Behavioral Basic/Refresher Course (Check one)	<input type="checkbox"/> Yes <input type="checkbox"/> No <input type="checkbox"/> Not Sure
Department	Click or tap here to enter text.
Affiliation	<input type="checkbox"/> Student* <input type="checkbox"/> Staff <input type="checkbox"/> Faculty <input type="checkbox"/> Other
*If you are a student, please complete the information below for your Faculty Advisor:	
Faculty Advisor Name	Click or tap here to enter text.
Email	Click or tap here to enter text.
Telephone Number	Click or tap here to enter text.
<input type="checkbox"/> I have attached a signed and completed letter from my faculty advisor as an appendix to this project in IRBNet.	

2. PROJECT SUMMARY

Title of Project
Click or tap here to enter text.
Describe the purpose of the project. Provide context to the importance of the research and explain

1. Basic Information
2. Project Summary
3. Risks and Mitigations
4. Data Access
5. Funding
6. Results
7. Additional Personnel
8. Investigator Assurance

- 7 page template
- Attach relevant documents

Step 4: Make Necessary Modifications

There may be mandatory changes based on ethics and compliance.

CSULB IRB Application for Existing and Secondary Data	Version: 01/02/2018
<h2>IRB Application for Existing and Secondary Data</h2>	
<p>Instructions: Please confirm that the research activities meet the definition of research with human subjects (the data has identifiers or links to identifiers). Fill out the form completely. Any incomplete forms will be returned. Check boxes can be filled by clicking once inside the box. Please include all applicable supporting documents for this submission such as permission letters and faculty supervisor letter.</p>	
<h3>1. Basic Information</h3>	
Principal Investigator:	Click or tap here to enter text.
CITI Member ID Number:	Click or tap here to enter text.
Department:	Click or tap here to enter text.
Telephone Number:	Click or tap here to enter text.
Email:	Click or tap here to enter text.
Affiliation:	<input type="checkbox"/> Student* <input type="checkbox"/> Faculty <input type="checkbox"/> Staff <input type="checkbox"/> External PI

Step 5: Final Approval

- An email notice is sent updating your status
- Updates to the research require updates to the IRB application

Survey Results

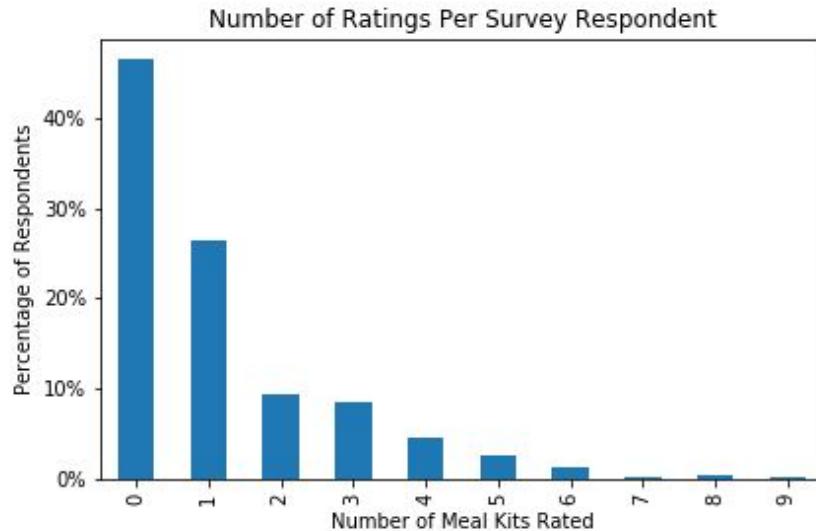
Descriptive Statistics

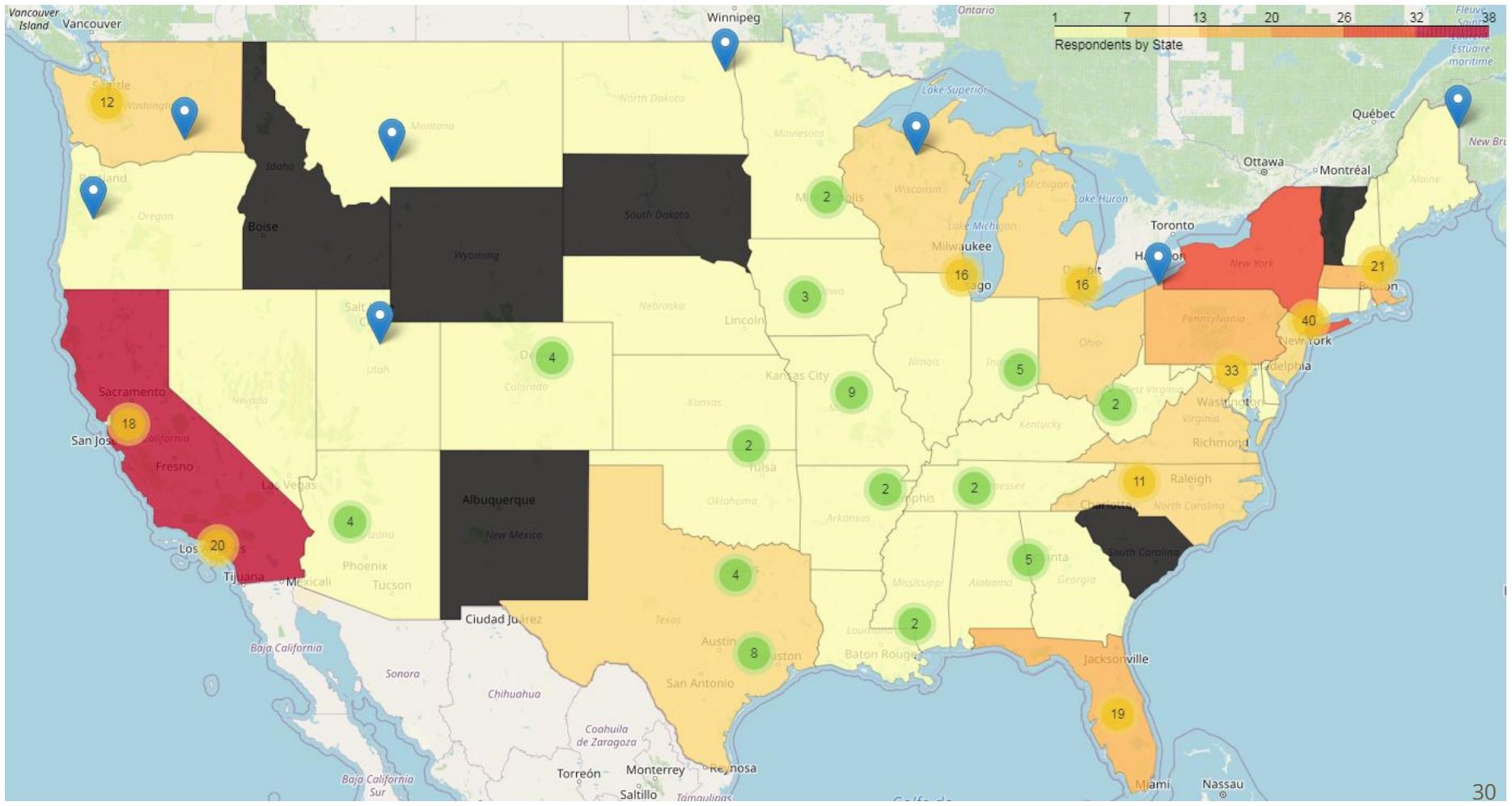
Survey Respondents

499 survey respondents over 24 weeks

267 respondents rated meal kits

After data cleaning, the resulting user rating matrix contains 577 ratings, 360 features, and 1 target variable.



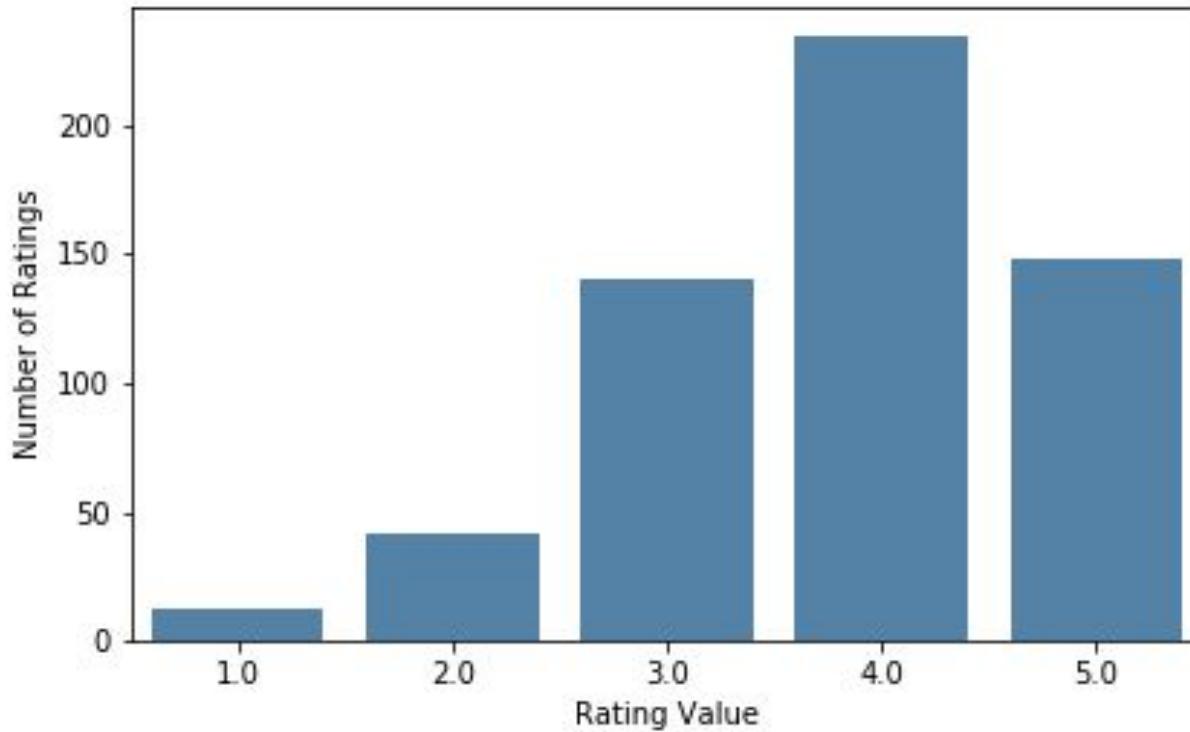


Top 10 Number of Meal Kit Survey Respondents in United States per Capita

Rank	State	Respondents per Million of Population
1	Maine	2.98
2	Massachusetts	2.47
3	New Hampshire	2.21
4	Washington	1.71
5	New York	1.54
6	District of Columbia	1.42
7	Wisconsin	1.37
8	North Dakota	1.31
9	Pennsylvania	1.25
10	Missouri	1.14

- Northeastern United States shows higher participation per capita.

Distribution of Meal Kit Service Ratings



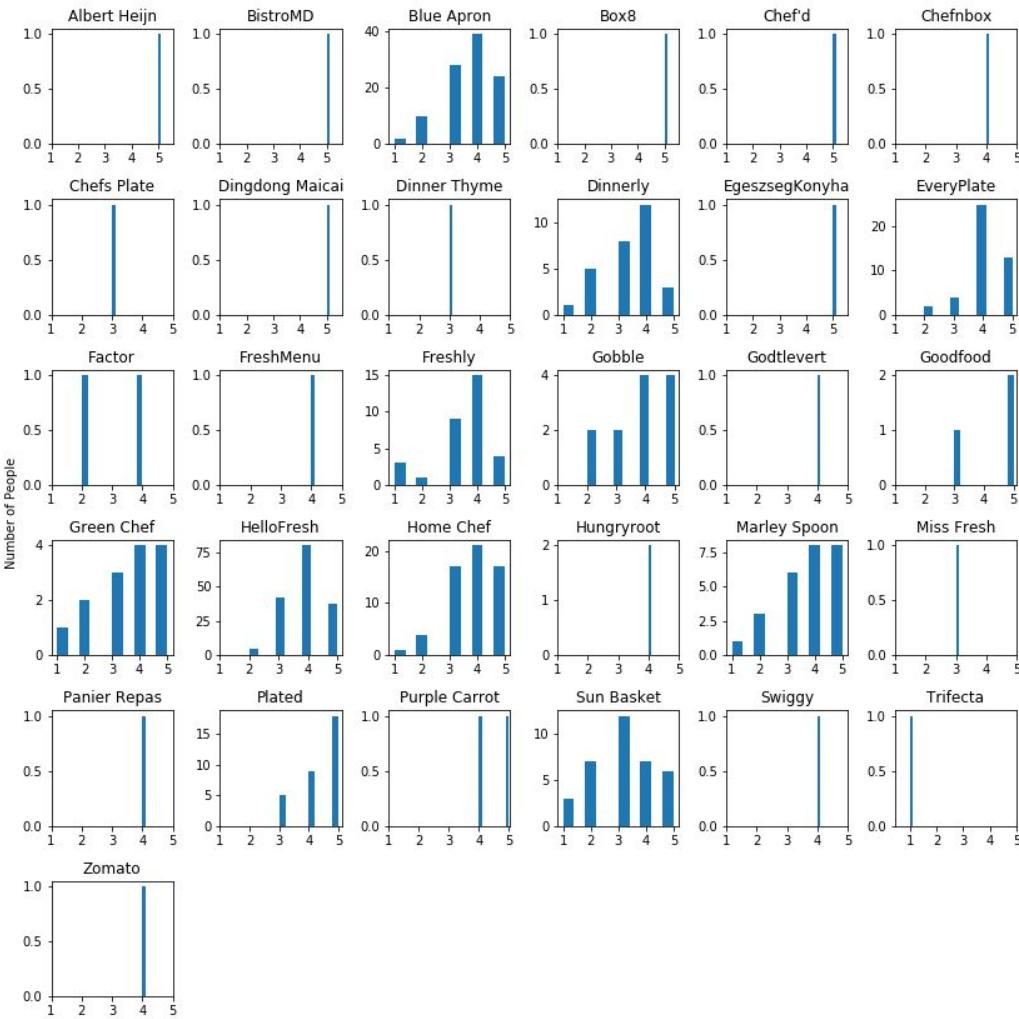
499 survey respondents

276 respondents included meal kit service ratings

Meal Kit Service	Rating Mean	Rating Count	Rating Std dev.
Plated	4.41	32	0.756
Goodfood	4.33	3	1.155
EveryPlate	4.11	44	0.754
HelloFresh	3.92	165	0.776
Gobble	3.83	12	1.115
Home Chef	3.82	60	0.983
Marley Spoon	3.73	26	1.151
Blue Apron	3.71	103	0.996
Green Chef	3.57	14	1.284
Freshly	3.50	32	1.078
Dinnerly	3.38	29	1.015
Sun Basket	3.17	35	1.200

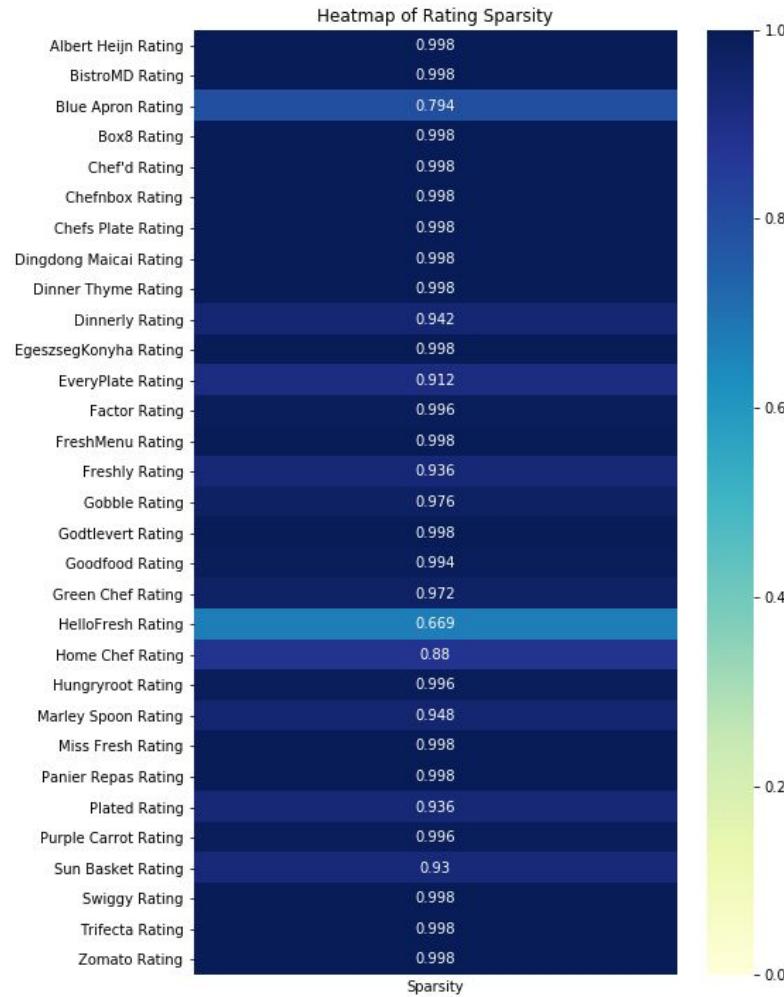
Average Ratings for Meal Kit Services with 3 or More Ratings

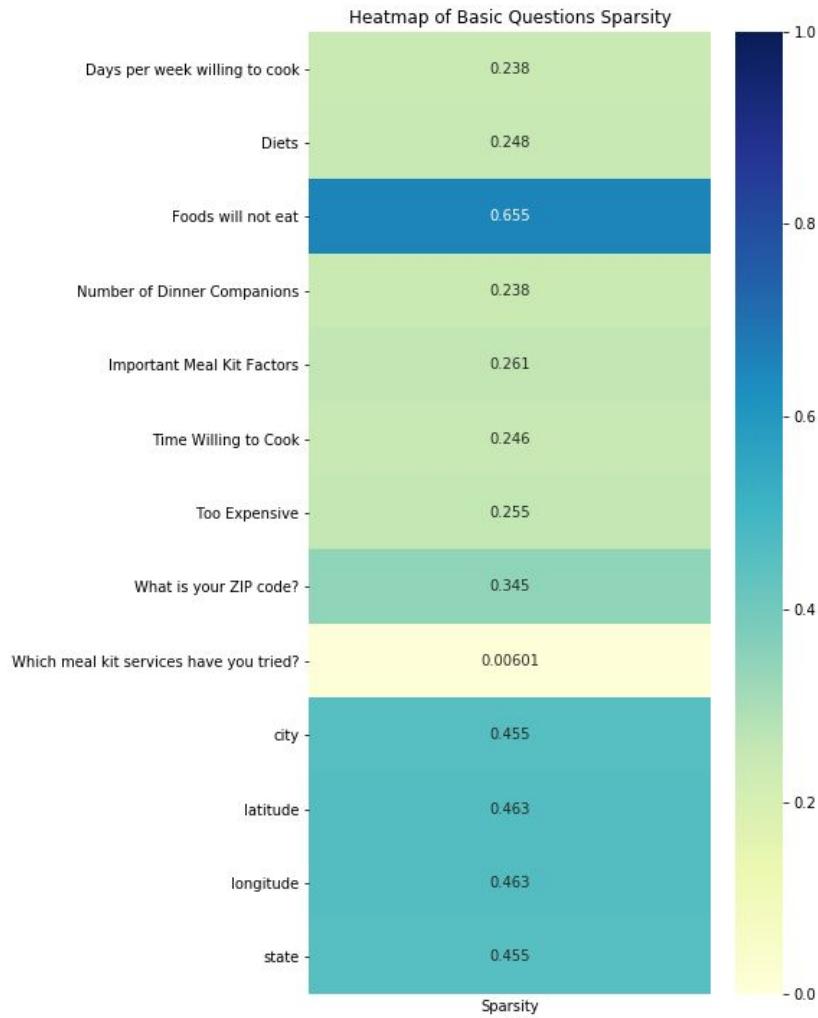
Meal Kit Service Ratings Histograms



Data Sparsity

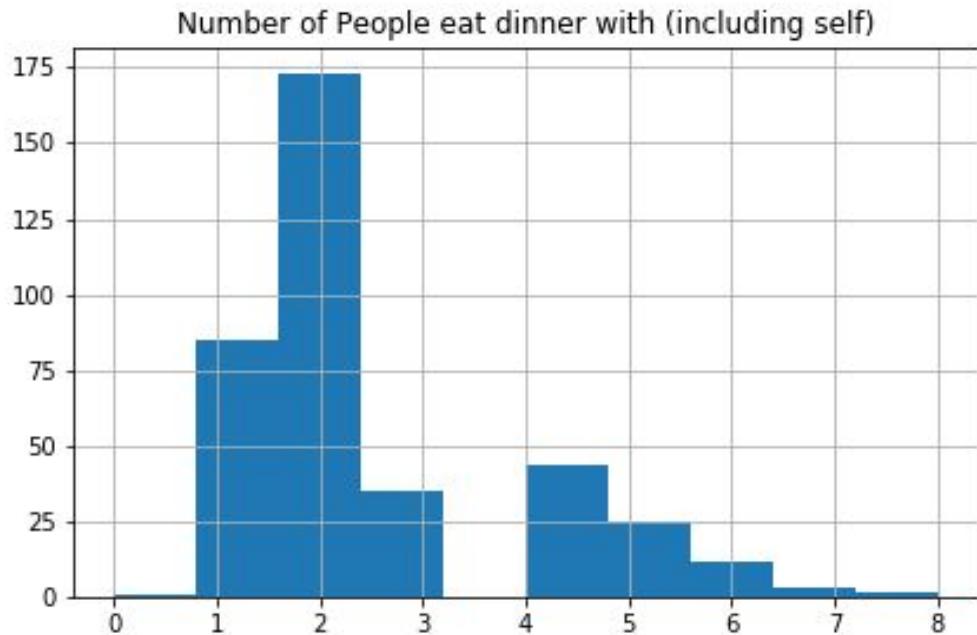
$$= \frac{\text{Total Missing Data Points}}{\text{Total Possible Data Points}}$$



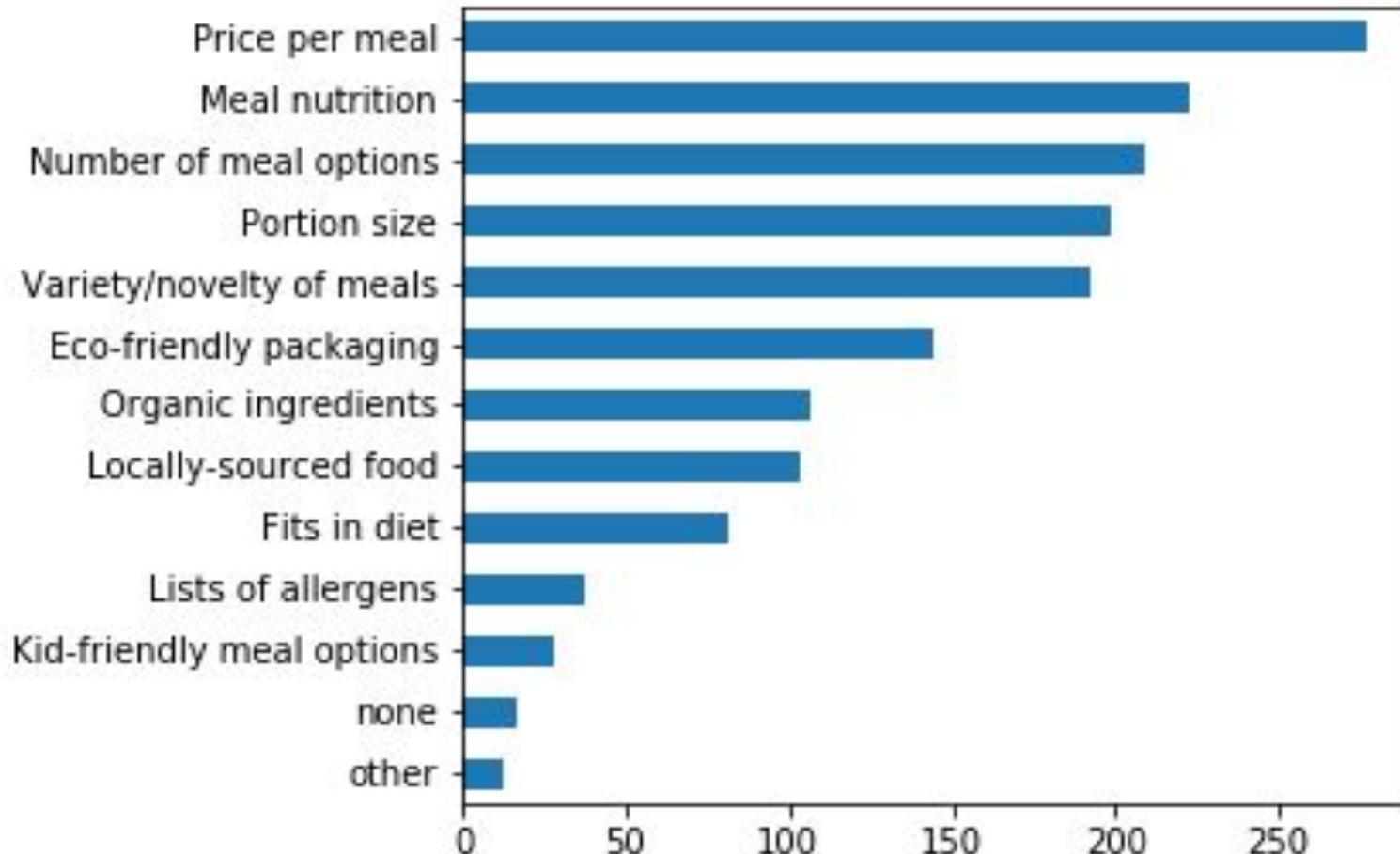


Dinner Companions

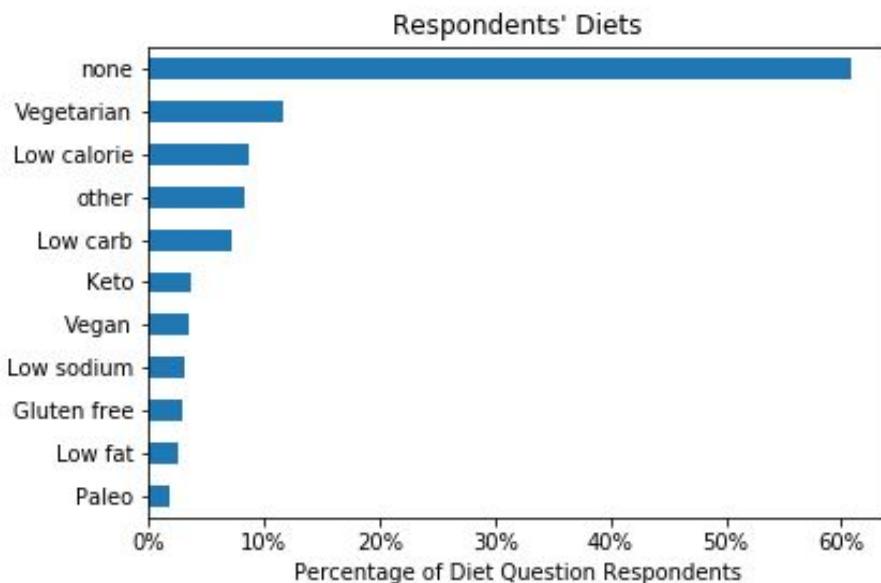
Many respondents eat dinner with 1 other person



Important Meal Kit Factors



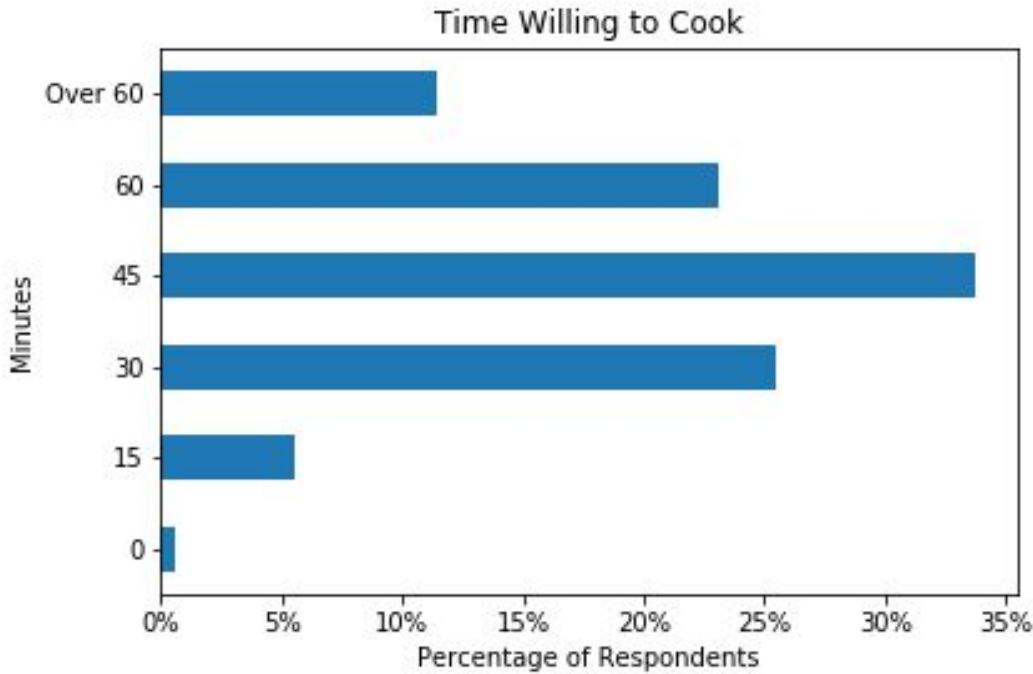
Diets of Respondents



Diet	Count	Percentage
No Diet	228	60.8%
Vegetarian	44	11.7%
Low Calorie	33	8.8%
Other	31	8.3%
Low Carb	27	7.2%
Keto	14	3.7%
Vegan	13	3.5%
Low Sodium	12	3.2%
Gluten Free	11	2.9%
Low Fat	10	2.7%
Paleo	7	1.9%

Cooking Time

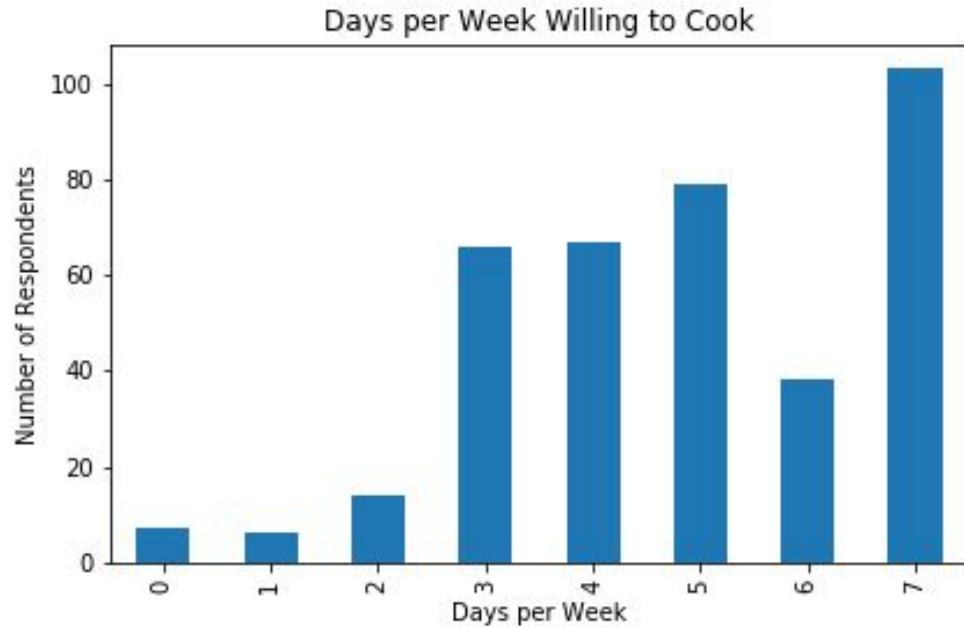
45 minutes was the most common time willing to cook



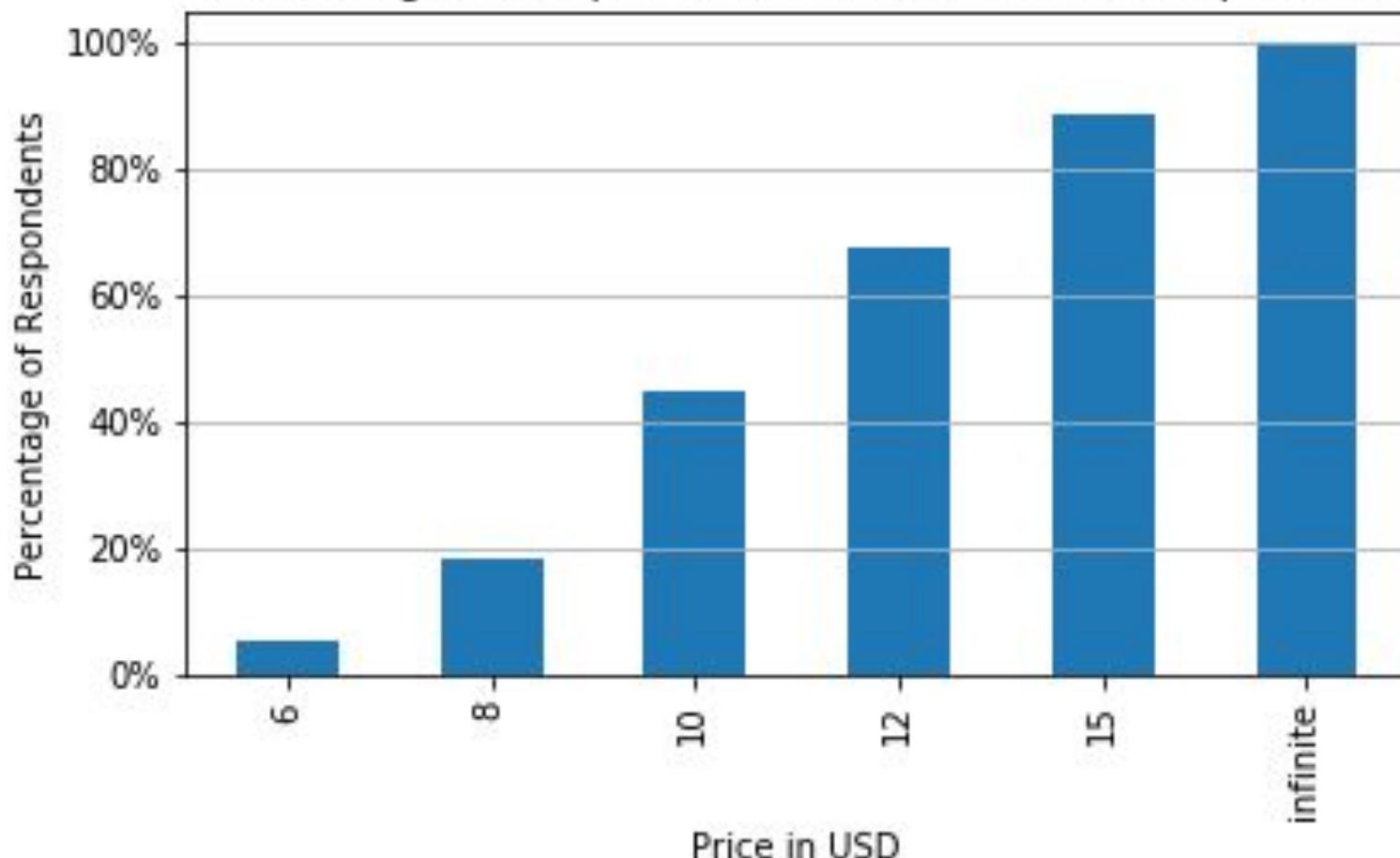
Cooking Frequency

Many respondent reported willingness to cook every day of the week

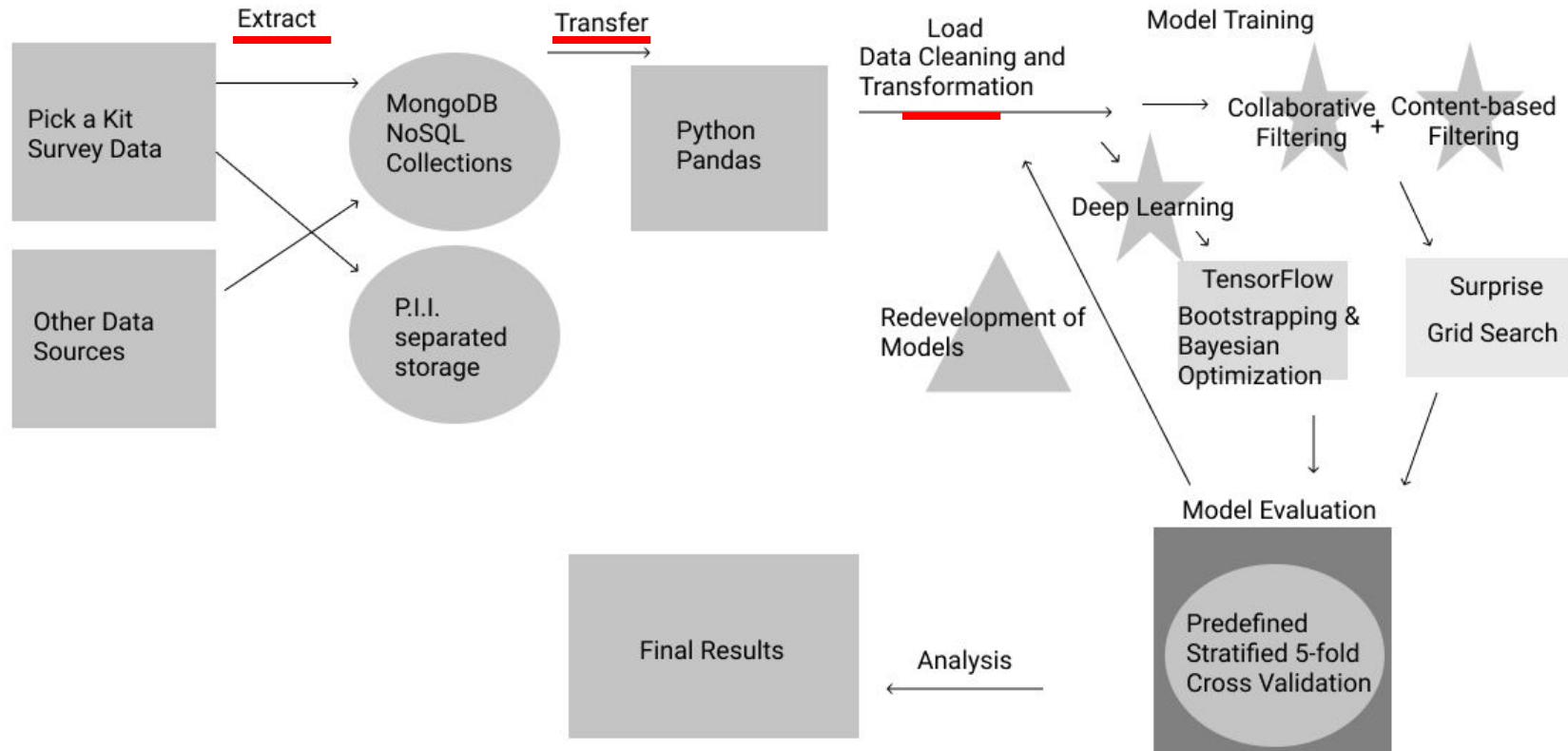
Cooking between 3-5 days per week was also a common response



Percentage of Respondents Think Price is Too Expensive



Methodology



Preprocessing

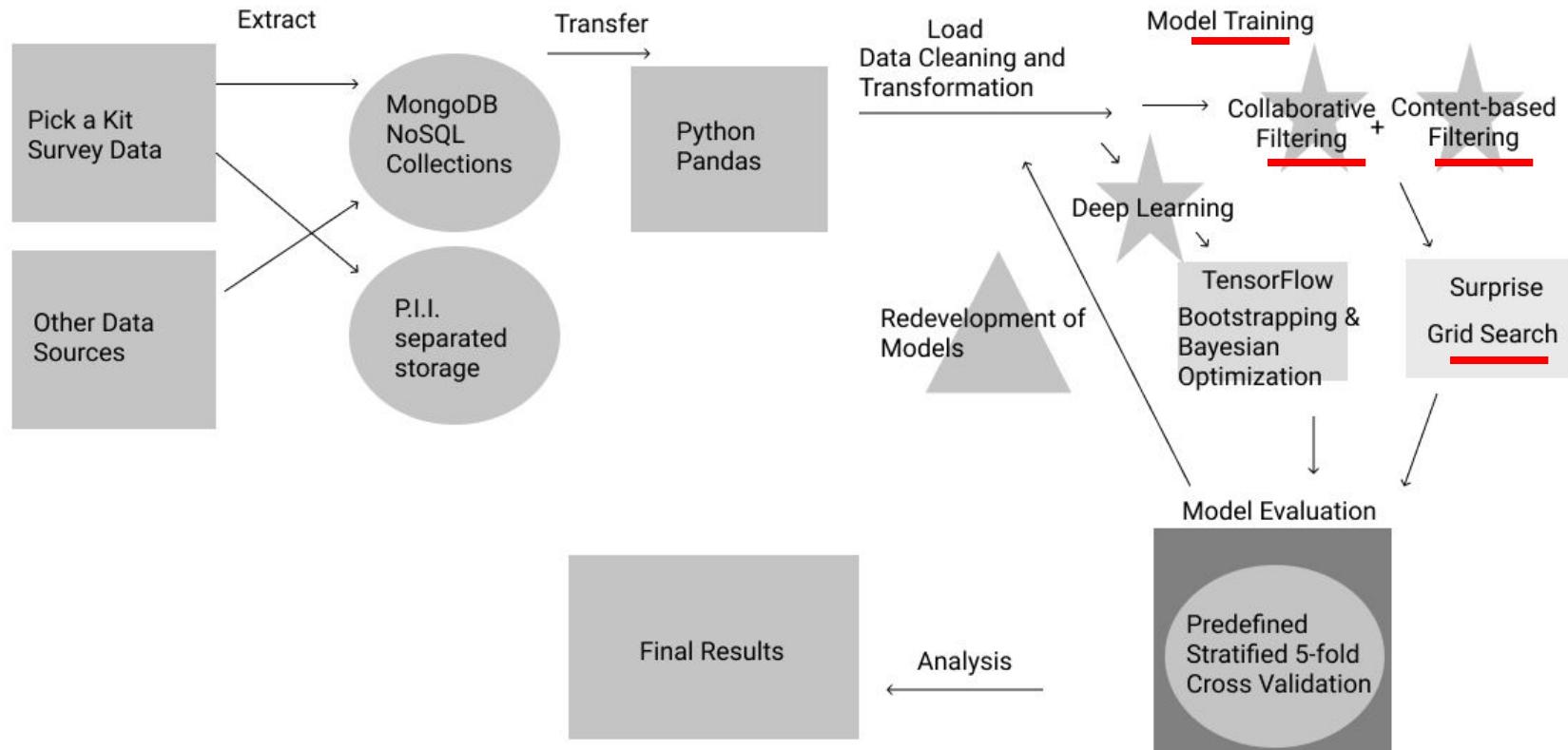
- Empty Responses Dropped
- Ordinal Encoding of User and Item Variables
- Median Imputation
- Binary Encoding of categorical variables
- Zipcodes-> Latitude and Longitude (Numeric)
- 5 stratified predefined folds for Cross-Validation
- Python
- Google Colab
- Pandas
- Numpy
- Sci-Kit Learn

Prediction Matrix Data Frame

			How many people do you eat dinner with on a typical night? Include yourself.	Days per week	Time willing to cook	Too Expensive	latitude	longitude	Which meal kit services have you tried?	_Blue Apron	Which meal kit services have you tried?	_Dinnerly	Which meal kit services have you tried?	_EveryPlate	Which meal kit services have you tried?	_Freshly	Which meal kit services have you tried?	_Gobble	...		
179		Albert Heijn	5.0	2.0	7.0	45.0	8.0	0.00	0.00	0	0	0	0	0	0	0	0	0	0		
607		BistroMD	5.0	1.0	6.0	15.0	12.0	0.00	0.00	0	0	0	1	0	0	0	0	0	0		
612		Blue Apron	4.0	2.0	5.0	45.0	12.0	0.00	0.00	0	0	0	0	0	0	0	0	0	0		
614		Blue Apron	3.0	4.0	4.0	45.0	15.0	34.08	-118.14	1	0	0	0	0	0	0	0	0	0		
617		Blue Apron	5.0	2.0	5.0	45.0	12.0	0.00	0.00	1	0	0	0	0	0	0	0	0	1		
618		Blue Apron	4.0	3.0	2.0	60.0	10.0	40.56	-105.13	1	0	0	0	0	0	0	0	0	1		
620		Blue Apron	4.0	1.0	5.0	45.0	12.0	38.90	-92.40	1	1	0	0	1	0	1	0	0	0		
621		Blue Apron	4.0	2.0	4.0	30.0	15.0	37.32	-121.93	1	0	0	0	0	0	0	0	0	0		
626		Blue Apron	3.0	2.0	3.0	60.0	10.0	33.74	-117.81	1	0	0	0	0	0	0	0	0	0		
630		Blue Apron	3.0	2.0	4.0	30.0	10.0	33.68	-117.83	1	1	0	0	0	0	0	0	0	0		

Algorithm Groups

Collaborative Filtering, Content-based Filtering, and Deep Learning

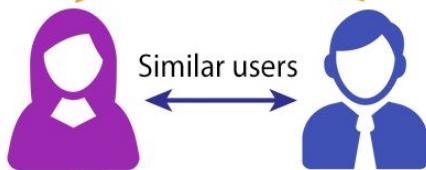


COLLABORATIVE FILTERING

Read by both users



Similar users



CONTENT-BASED FILTERING

Read by user



Similar articles



Recommended
to user

Meal Kit Service Examples

Collaborative Filtering

User	Hello Fresh	Blue Apron	Gobble
Alan			2
Olga	5		
Kagba	2	3	5
Yale	2		

Content-based Filtering

Item Attributes	Hello Fresh	Blue Apron	Gobble
Price	\$8.99	\$8.99	\$12.99
Avg. Calories	740	800	1000

Predicts Yale will like Gobble

Predicts Olga will like Blue Apron

Normal predictor

Algorithm Summary

Algorithm predicting a random rating based on the distribution of the training set, which is assumed to be normal. The prediction is generated from a normal distribution, estimated from the training data using Maximum Likelihood Estimation.

Best Results:

RMSE	MAE	Prediction Coverage	Test Time
1.35427	1.05111	1	5.20E-03

Equation

$$\hat{\mu} = \frac{1}{|R_{train}|} \sum_{r_{ui} \in R_{train}} r_{ui}$$

$$\hat{\sigma} = \sqrt{\sum_{r_{ui} \in R_{train}} \frac{(r_{ui} - \hat{\mu})^2}{|R_{train}|}}$$

Baseline Algorithm

Algorithm Summary:

Computes baseline estimates for users and items using stochastic gradient descent or alternating least squares.

Equation

$$\hat{r}_{ui} = b_{ui} = \mu + b_u + b_i$$

Best Results:

	RMSE	MAE	Prediction Coverage	Test Time
ALS	0.95308	0.75626	1	8.00E-04
SGD	0.95489	0.76127	1	8.00E-04

```
{'bsl_options': {'method': 'als', 'reg': 0.001}, 'verbose': False}  
{'bsl_options': {'method': 'sgd', 'reg': 0.03}, 'verbose': False}
```

$$\sum_{r_{ui} \in R_{train}} (r_{ui} - (\mu + b_u + b_i))^2 + \lambda (b_u^2 + b_i^2)$$

Memory based Collaborative Filtering Algorithms

- KNN with Means
- KNN with ZScore
- KNN Baseline

Uses similarity metrics on dataset
to make predictions

KNN with Means

Algorithm Summary:

A basic collaborative filtering algorithm, taking into account the mean ratings of each user.

Best Results:

RMSE	MAE	Prediction Coverage	Test Time
1.02592	0.80950	0.53507	1.40E-03

Equation

$$\hat{r}_{ui} = \mu_u + \frac{\sum_{v \in N_i^k(u)} sim(u,v) \cdot (r_{vi} - u_v)}{\sum_{v \in N_i^k(u)} sim(u,v)}$$

```
{"bsl_options": {"method": "sgd", "reg": 1}, "learning_rate": 0.5, "k": 50, "sim_options": {"name": "pearson_baseline", "min_support": 5, "user_based": False}, "verbose": False}
```

KNN with Z-Score

Algorithm Summary:

Mean centered and standardized nearest neighbor ratings

Best Results:

RMSE	MAE	Prediction Coverage	Test Time
1.02191	0.80624	0.53507	1.40E-03

Equation

$$\hat{r}_{ui} = \mu_u + \sigma_u \frac{\sum_{v \in N_i^k(u)} sim(u, v) \cdot (r_{vi} - \mu_v) / \sigma_v}{\sum_{v \in N_i^k(u)} sim(u, v)}$$

$$\hat{r}_{ui} = \mu_u + \sigma_u \frac{\sum_{j \in N_i^k(i)} sim(i, j) \cdot (r_{uj} - \mu_j) / \sigma_j}{\sum_{j \in N_i^k(i)} sim(i, j)}$$

```
{'bsl_options': {'method': 'sgd', 'reg': 1}, 'learning_rate': 0.001, 'k': 3, 'sim_options': {'name': 'pearson_baseline', 'min_support': 5, 'user_based': False}, 'verbose': False}
```

KNN Baseline

Algorithm Summary:

User and item baselines adjusted to KNN algorithm

Best Results:

RMSE	MAE	Prediction Coverage	Test Time
0.95080	0.75474	1	4.40E-03

Equation

$$b_{ui} = \mu + b_u + b_i$$

$$\hat{r}_{ui} = b_{ui} + \frac{\sum_{v \in N_i^k(u)} sim(u, v) \cdot (r_{vi} - b_{vi})}{\sum_{v \in N_i^k(u)} sim(u, v)}$$

$$\hat{r}_{ui} = b_{ui} + \frac{\sum_{j \in N_u^k(i)} sim(i, j) \cdot (r_{uj} - b_{uj})}{\sum_{j \in N_u^k(i)} sim(i, j)}$$

```
{'bsl_options': {'method': 'als', 'reg': 2}, 'learning_rate': 0.1, 'k': 3, 'sim_options': {'name': 'pearson', 'min_support': 6, 'user_based': True}, 'verbose': False}
```

Model based Collaborative Filtering Algorithms

SVD, SVD++, NMF, Slope One, and Co-Clustering

Develops models to make predictions

Singular Value Decomposition (SVD)

Algorithm Summary:

Matrix factorization technique that uncovers latent factors in ratings utility matrix

Best Results:

RMSE	MAE	Prediction Coverage	Test Time
0.94919	0.75616	1	1.20E-03

{'n_factors': 160, 'n_epochs': 20, 'biased': True, 'lr_all': 0.005, 'reg_all': 0.1}

Equation

$$\hat{r}_{ui} = \mu + b_u + b_i + q_i^T p_u$$

$$\sum_{r_{ui} \in R_{train}} (r_{ui} - \hat{r}_{ui})^2 + \lambda(b_i^2 + b_u^2 + \|q_i\|^2 + \|p_u\|^2)$$

$$b_u \leftarrow b_u + \gamma(e_{ui} - \lambda b_u)$$

$$b_i \leftarrow b_i + \gamma(e_{ui} - \lambda b_i)$$

$$p_u \leftarrow p_u + \gamma(e_{ui} \cdot q_i - \lambda p_u)$$

$$q_i \leftarrow q_i + \gamma(e_{ui} \cdot p_u - \lambda q_i)$$

$$\text{where } e_{ui} = r_{ui} - \hat{r}_{ui}$$

SVD++

Algorithm Summary:

SVD algorithm with the inclusion of implicit ratings preferences

Equation

$$\hat{r}_{ui} = \mu + b_u + b_i + q_i^T(p_u + |I_u|^{-\frac{1}{2}} \sum_{j \in I_u} y_j)$$

Best Results:

RMSE	MAE	Prediction Coverage	Test Time
0.94992	0.76033	1	2.40E-03

```
{'n_factors': 25, 'n_epochs': 10, 'lr_all': 0.01, 'reg_all': 0.1}
```

Nonnegative Matrix Factorization (NMF)

Algorithm Summary:

Matrix Factorization technique similar to SVD, where factored matrices are composed of only positive user and item features

Best Results:

RMSE	MAE	Prediction Coverage	Test Time
1.02749	0.83511	1	1.20E-03

{'n_factors': 4, 'n_epochs': 4, 'biased': True}

Equation

$$\hat{r}_{ui} = q_i^T p_u$$

$$p_{uf} \leftarrow p_{uf} \cdot \frac{\sum_{i \in I_u} q_{if} \cdot r_{ui}}{\sum_{i \in I_u} q_{if} \cdot \hat{r}_{ui} + \lambda_u |I_u| p_{uf}}$$

$$q_{if} \leftarrow q_{if} \cdot \frac{\sum_{u \in U_i} p_{uf} \cdot r_{ui}}{\sum_{u \in U_i} p_{uf} \cdot \hat{r}_{ui} + \lambda_i |U_i| q_{if}}$$

$$\hat{r}_{ui} = \mu + b_u + b_i + q_i^T p_u$$

Slope One

Algorithm Summary:

Uses $f(x) = x+b$ model without a coefficient (i.e. slope = 1) for simplified popularity adjusted ratings

Best Results:

RMSE	MAE	Prediction Coverage	Test Time
1.13234	0.89565	0.53507	1.40E-03

Equation

$$\hat{r}_{ui} = \mu_u + \frac{1}{|R_i(u)|} \sum_{j \in R_i(u)} dev(i, j)$$

$$dev(i, j) = \frac{1}{|U_{ij}|} \sum_{w \in U_{ij}} r_{ui} - r_{uj}$$

Co-Clustering

Algorithm Summary:

Assigns users and items to clusters using a k-means like optimization method. If the item is unknown, the prediction is set to the user average. If both the user and the item are unknown, the prediction is set to the global average.

Equation

$$\hat{r}_{ui} = \overline{C}_{ui} + (\mu_u - \overline{C}_u) + (u_i - \overline{C}_i)$$

Best Results:

RMSE	MAE	Prediction Coverage	Test Time
1.12311	0.88992	1	6.00E-04

{'n_cltr_u': 2, 'n_cltr_i': 2, 'n_epochs': 5}

Content-based Filtering

Content Based Basic

- Uses underlying item attributes to make predictions
- Does not use ratings data of other users

Meal Kit Service Profiles

	Price_min	Price_max	Price_average	Price_std_dev	Plan_Count	agg_meals	unique_meals	(carbohydrate_grams, min)	(carbohydrate_grams, max)	(carbohydrate_grams, avg)	...
service_name											
Albert Heijn	1.74	9.00	4.69	1.40	1.0	NaN	NaN	NaN	NaN	NaN	
BistroMD	9.50	13.00	11.15	1.27	5.0	NaN	NaN	NaN	NaN	NaN	
Blue Apron	7.49	9.99	9.20	0.99	3.0	18.769231	191.0	26.0	178.0	102.0	
Box8	0.13	16.90	2.19	2.33	1.0	NaN	NaN	NaN	NaN	NaN	
Chefd	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	
Chefnbox	13.56	13.56	13.56	0.00	1.0	NaN	NaN	NaN	NaN	NaN	
Chefs Plate	8.99	9.99	9.49	0.50	2.0	NaN	NaN	NaN	NaN	NaN	
Dingdong Maicai	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	
Dinner Thyme	2.00	15.00	11.68	1.66	1.0	NaN	NaN	NaN	NaN	NaN	
Dinnerly	4.29	4.99	4.72	0.23	2.0	19.076923	205.0	1.0	160.0	100.0	
EgeszsegKonyha	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	
EveryPlate	4.99	4.99	4.99	0.00	1.0	16.307692	162.0	36.0	110.0	80.0	
Factor	11.00	15.00	12.54	1.39	5.0	NaN	NaN	NaN	NaN	NaN	
Freshly	7.99	11.50	9.37	1.30	4.0	38.307692	56.0	16.0	68.0	40.0	
FreshMenu	1.73	10.71	3.35	3.91	1.0	NaN	NaN	NaN	NaN	NaN	
Gobble	11.99	11.99	11.99	0.00	3.0	NaN	NaN	NaN	NaN	NaN	
Godtlevvert	6.09	15.68	9.89	3.61	1.0	NaN	NaN	NaN	NaN	NaN	
Goodfood	7.48	9.60	8.47	0.87	5.0	NaN	NaN	NaN	NaN	NaN	

Content Based Basic

Algorithm Summary:

A nearest neighbors approach to content-based filtering. Calculates the cosine similarity of item attributes with an option for unweighted similarity.

Best Results:

RMSE	MAE	Prediction Coverage	Test Time
0.99645	0.71229	0.03859	2.42E-01

{'weights': 'cosine', 'k': 1}

Equation

$$\hat{r}_{ux} = \frac{\sum_{\substack{y \in N \\ u(y)}} \text{cosine}(\bar{X}, \bar{Y}) \cdot r_{uy}}{\sum_{\substack{y \in N \\ u(y)}} \text{cosine}(\bar{X}, \bar{Y})}$$

$$\text{Cosine}(\bar{X}, \bar{Y}) = \frac{\sum_{i=1}^d x_i y_i}{\sqrt{\sum_{i=1}^d x_i^2} \sqrt{\sum_{i=1}^d y_i^2}}$$

Deep Learning Approaches

Deep Neural Networks (DNN)

Two layered fully connected neural networks

Experts

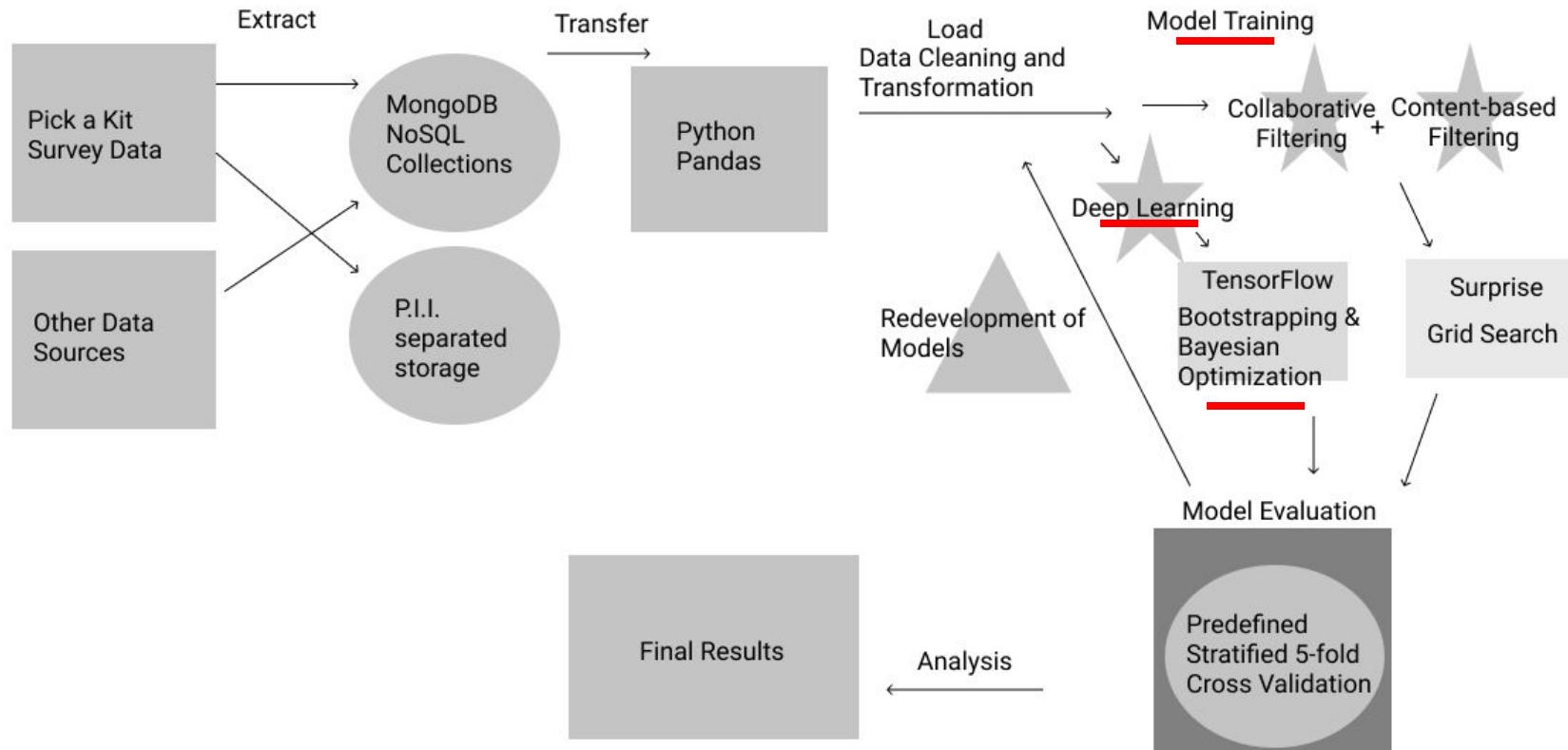
TensorFlow Keras version 2.3.0

Adam Optimization

PReLU and ReLU Activation Functions

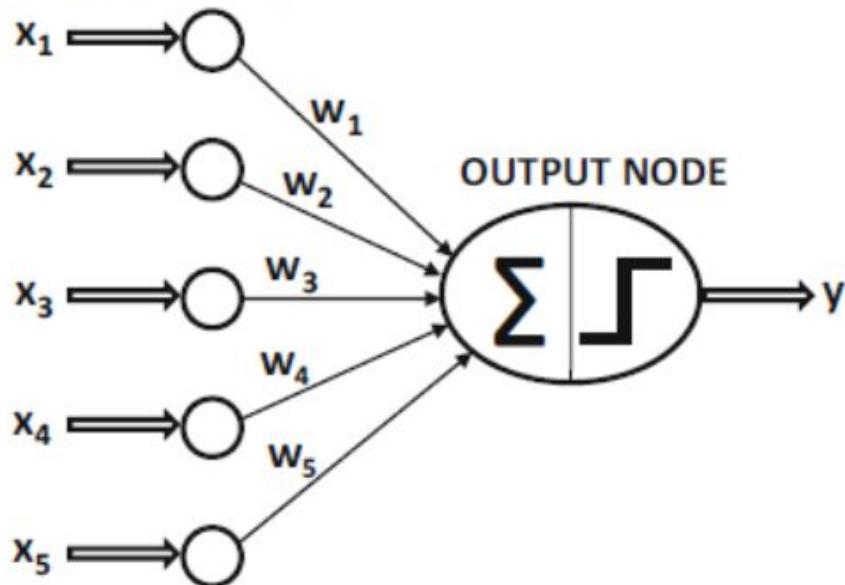
Early Stopping

Bootstrapping 10x



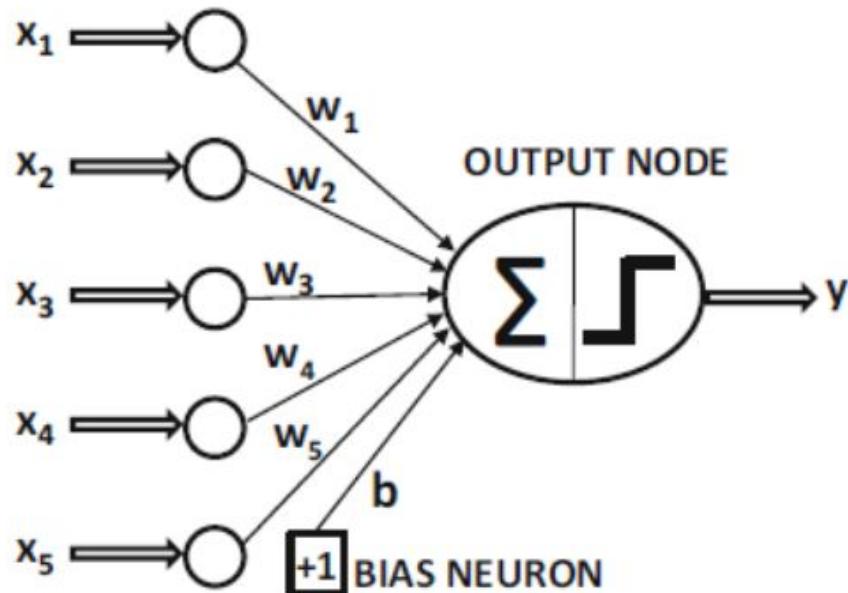
The Basic Architecture of the Perceptron

INPUT NODES



(a) Perceptron without bias

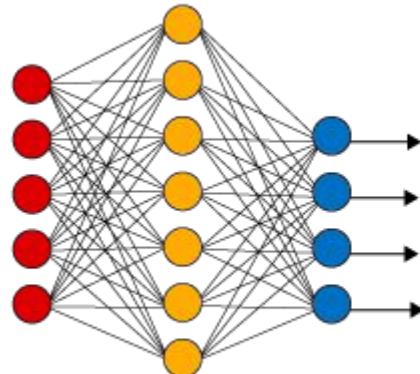
INPUT NODES



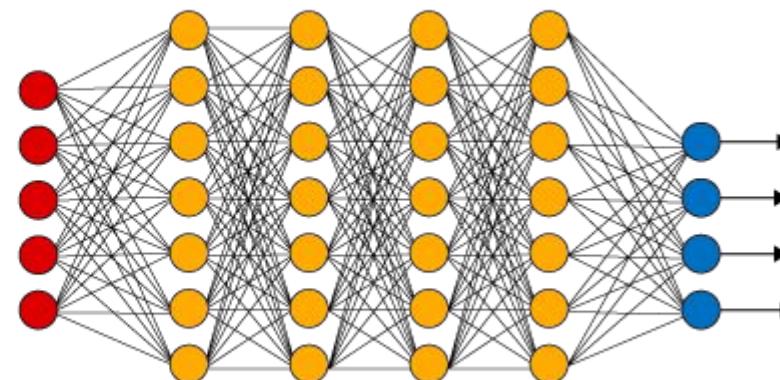
(b) Perceptron with bias

Neural Networks Nodes and Layers

Simple Neural Network



Deep Learning Neural Network



● Input Layer

● Hidden Layer

● Output Layer

Bayesian Optimization

Objective Function: Average RMSE of 10 Bootstrap Iterations

Hyperparameters:

Data Groups: User item ratings matrix, Item Profiles, User Profiles (DNN Accommodates CF, CBF, and Hybrid methods)

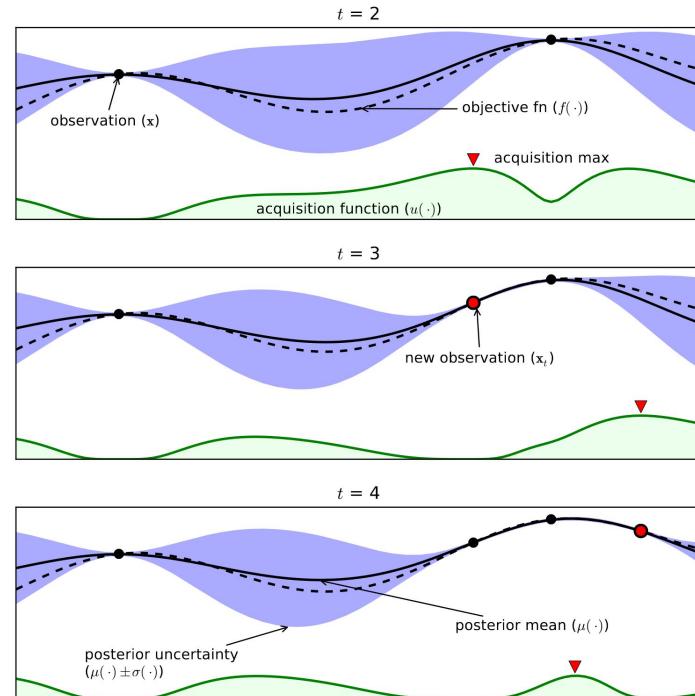
Neuron Percentage

Neuron Shrink

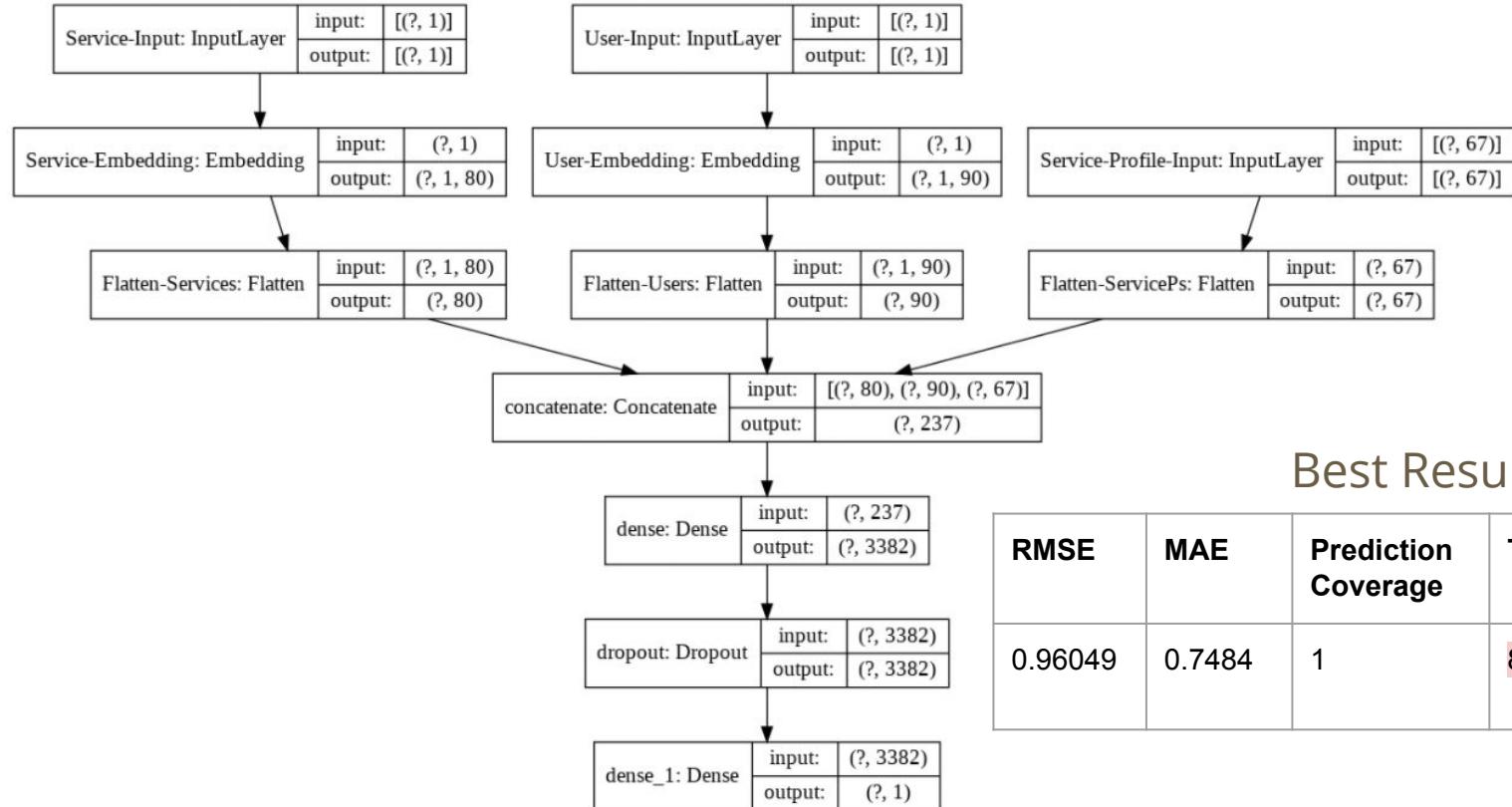
Number of Layers

Learning Rate

Embeddings Dimensions

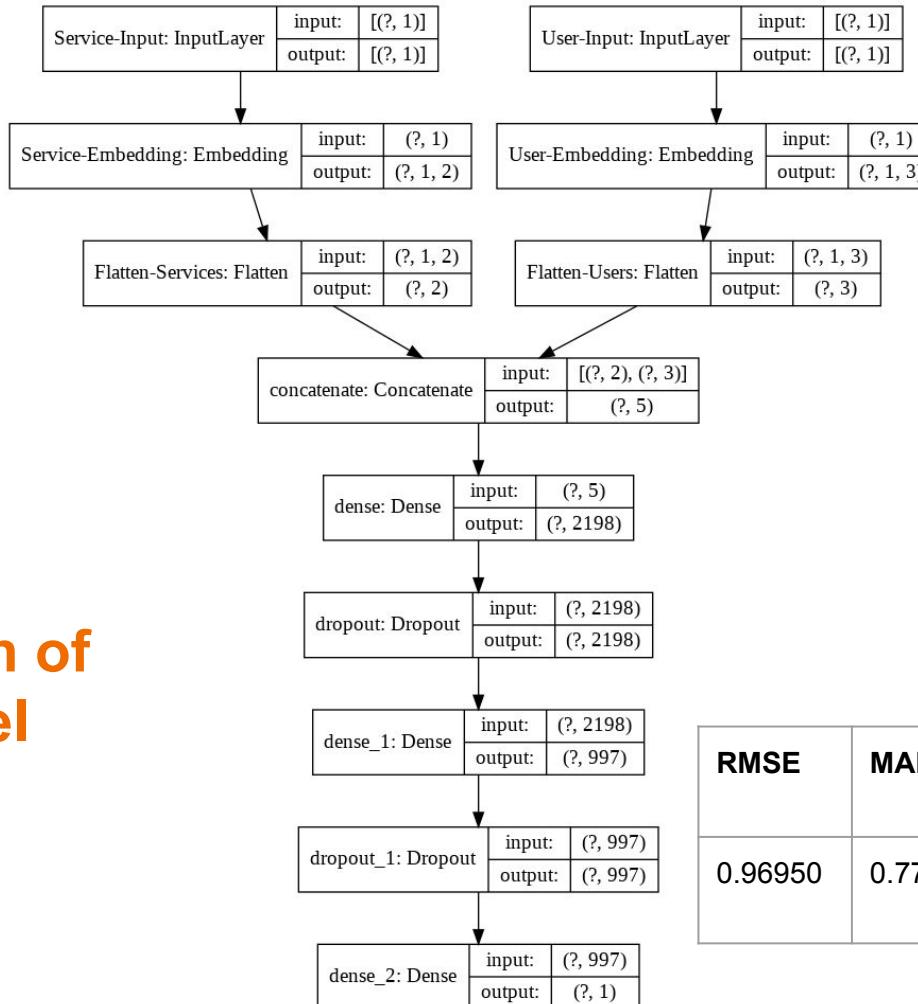


DNN Network Graph Example



Best Results:

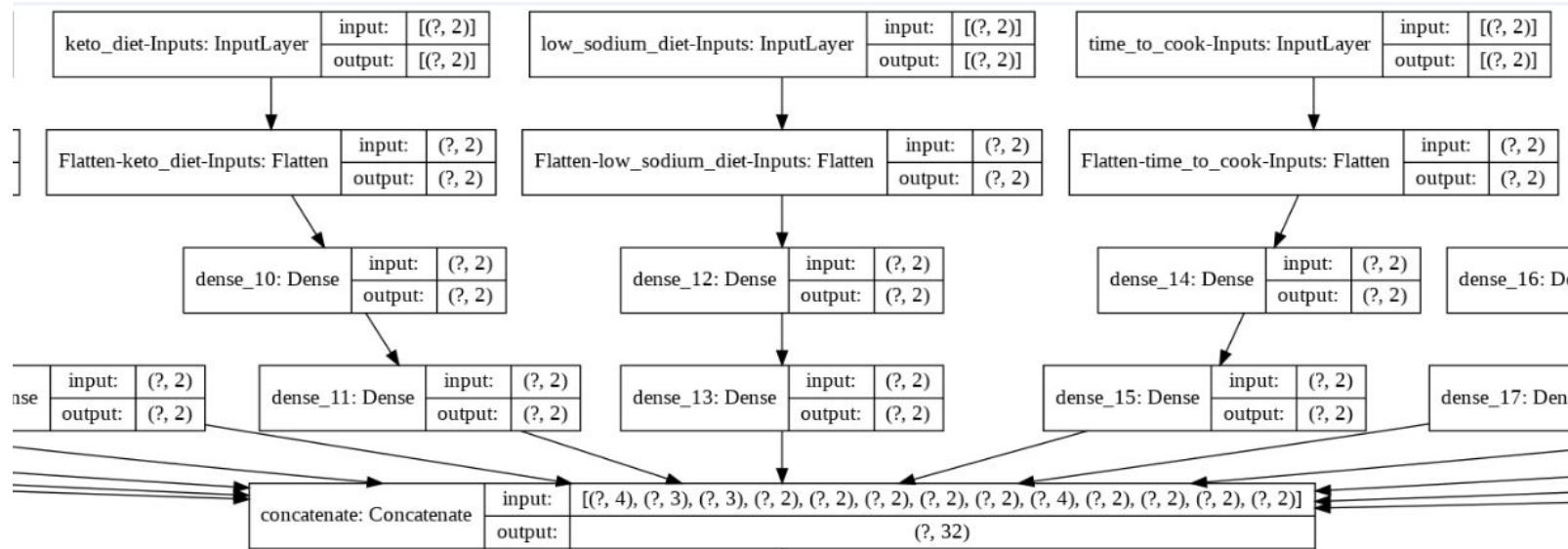
RMSE	MAE	Prediction Coverage	Test Time
0.96049	0.7484	1	8.42E+01



Network Graph of Two FCP Model Example

Best Results:

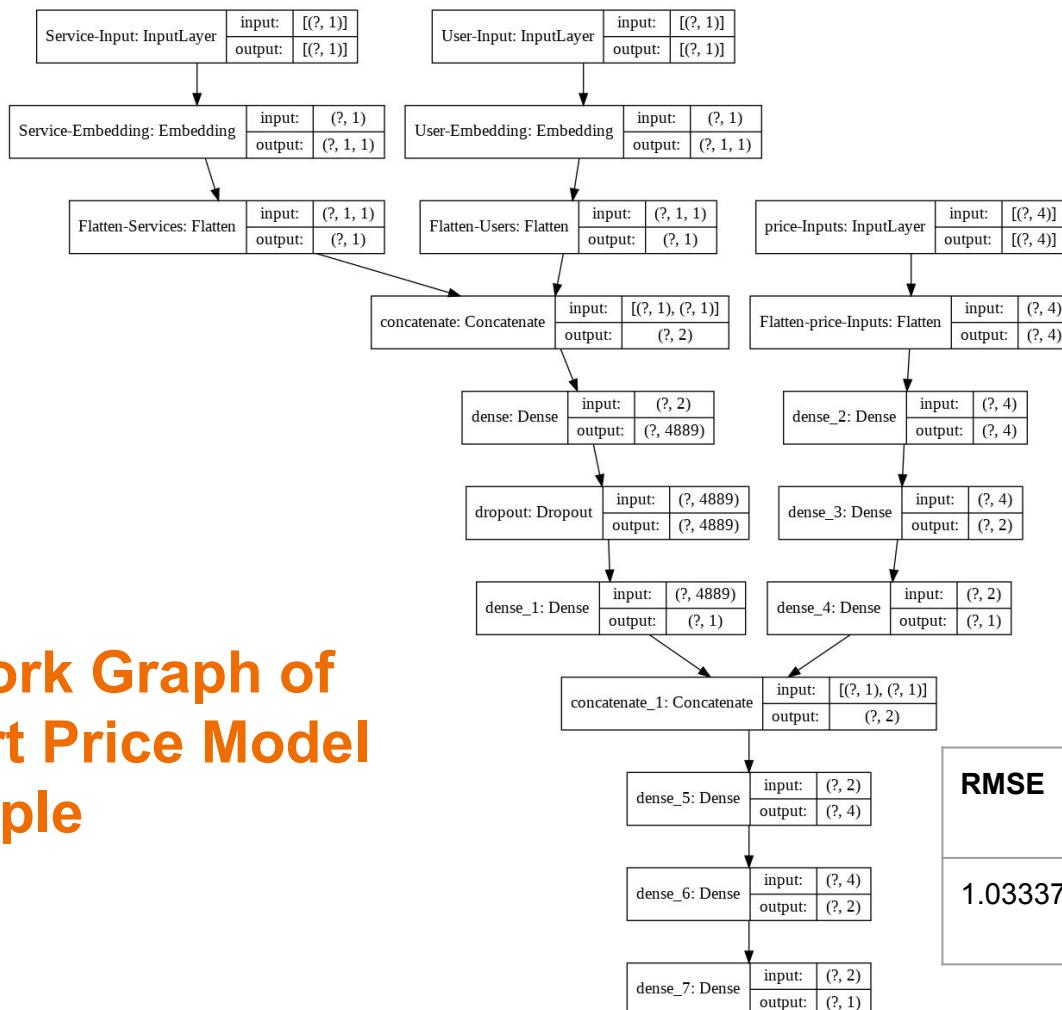
RMSE	MAE	Prediction Coverage	Test Time
0.96950	0.77799	1	1.68E+00



Network Graph of Experts Model Example

Best Results:

RMSE	MAE	Prediction Coverage	Test Time
0.98388	0.78141	1	7.49E+00



Network Graph of Expert Price Model Example

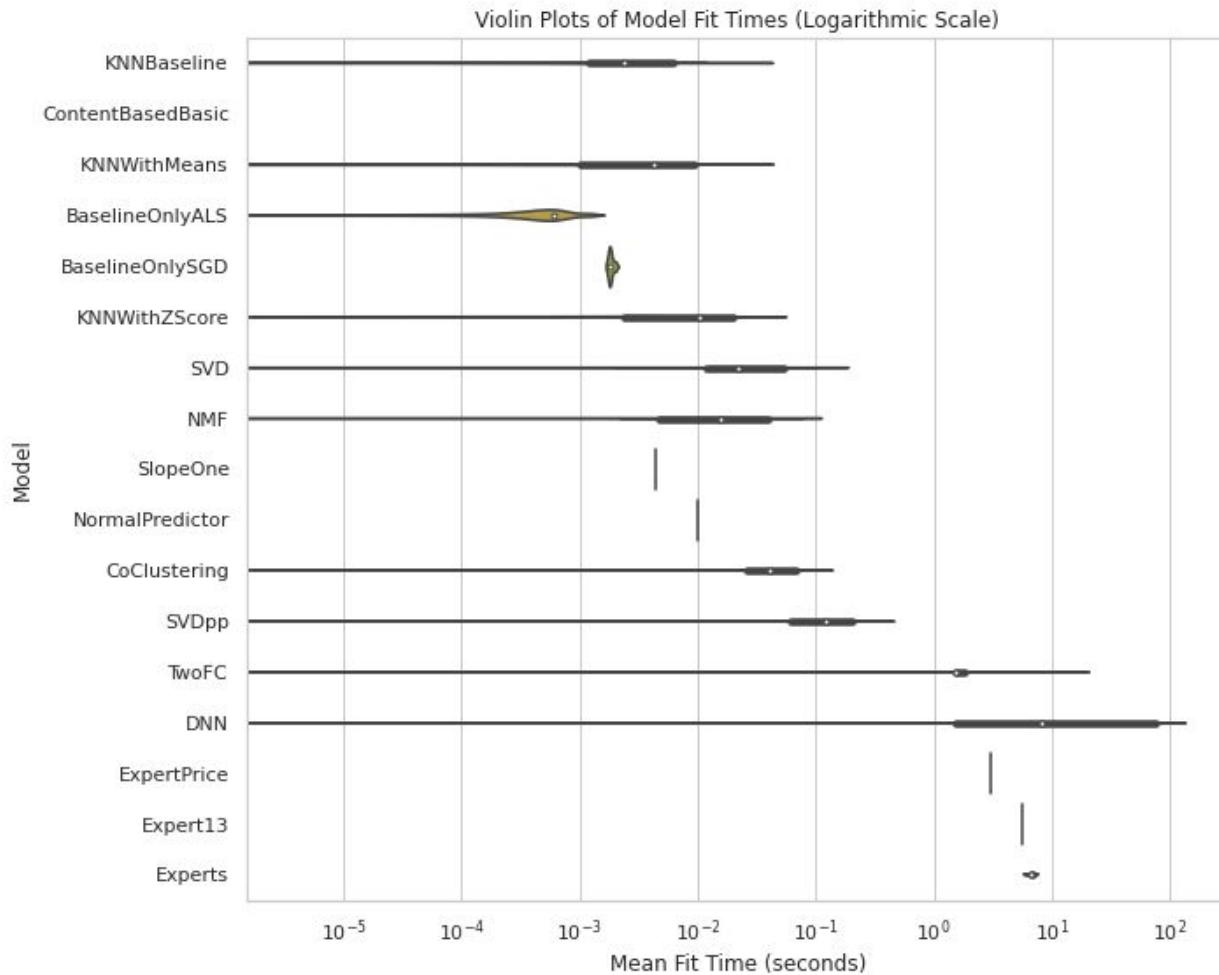
Best Results:

RMSE	MAE	Prediction Coverage	Test Time
1.03337	0.84518	1	3.26E+00

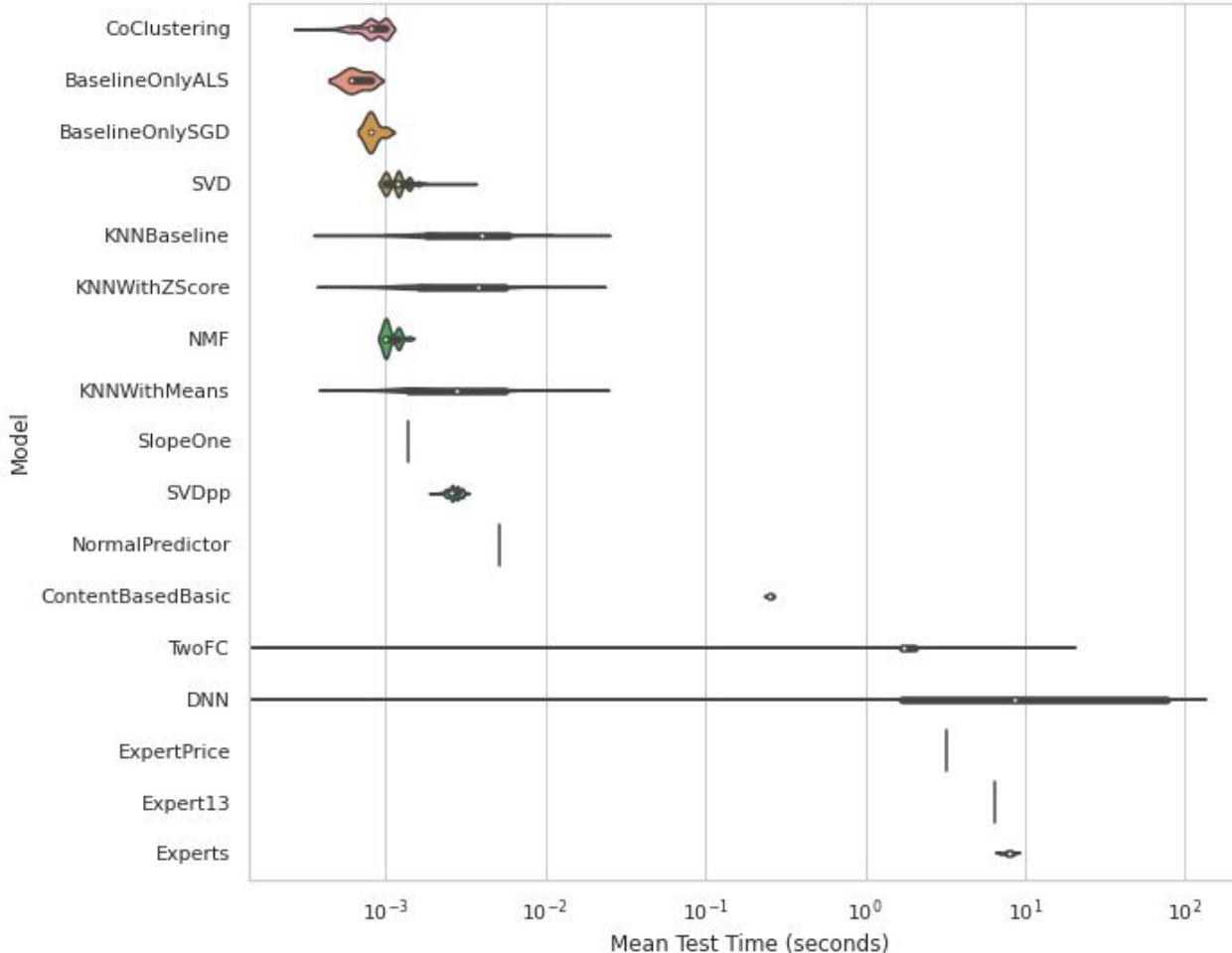
Results

Best Model Results by Mean Test RMSE

Model	Mean Test RMSE	Mean Test MAE	Mean Test Time (sec)	Prediction Coverage	Standard Deviation RMSE	Standard Deviation MAE
SVD	0.9492	0.7562	0.0012	1.000	0.0189	0.0297
SVDpp	0.9499	0.7603	0.0024	1.000	0.0214	0.0304
KNNBaseline	0.9508	0.7547	0.0044	1.000	0.0200	0.0262
BaselineOnlyALS	0.9531	0.7563	0.0008	1.000	0.0197	0.0257
BaselineOnlySGD	0.9549	0.7613	0.0008	1.000	0.0251	0.0337
DNN	0.9605	0.7485	84.2458	1.000	0.0110	0.0235
TwoFC	0.9695	0.7780	1.6770	1.000	0.0328	0.0327
Experts	0.9839	0.7814	7.4900	1.000	0.0130	0.0150
Expert13	0.9920	0.7904	6.3966	1.000	0.0276	0.0330
ContentBasedBasic	0.9965	0.7123	0.2420	0.039	0.0097	0.0041
KNNWithZScore	1.0219	0.8062	0.0014	0.535	0.0536	0.0500
KNNWithMeans	1.0259	0.8095	0.0016	0.535	0.0578	0.0535
NMF	1.0275	0.8351	0.0012	1.000	0.0297	0.0280
ExpertPrice	1.0334	0.8452	3.2612	1.000	0.0657	0.0544
CoClustering	1.1231	0.8899	0.0006	1.000	0.0474	0.0528
SlopeOne	1.1323	0.8957	0.0014	0.535	0.0414	0.0528
NormalPredictor	1.3543	1.0511	0.0052	1.000	0.0680	0.0554



Violin Plots of Model Test Times (Logarithmic Scale)



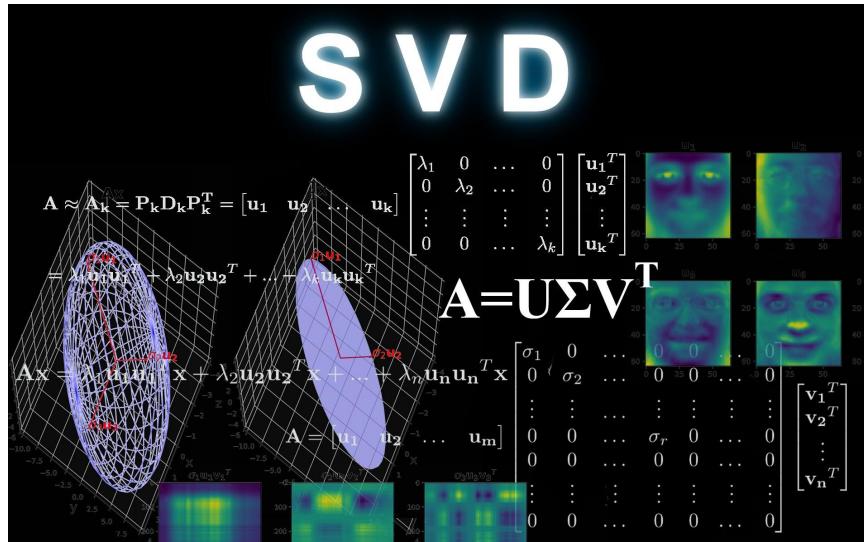
SVD model has overall best performance

SVD performed the best in RMSE

Competitive results in MAE

Full Prediction Coverage

Fast Model



<https://towardsdatascience.com/understanding-singular-value-decomposition-and-its-application-in-data-science-388a54be95d>

Future Work

1. Dataset expansion
2. Ranked list testing in online production format
3. Diversity, serendipity, and user feedback metrics
4. Expanded review of algorithms
5. Endless possibilities

Thank you

RobertNakano@gmail.com

Appendices

Similarity Metrics

$$msd_sim(u, v) = \frac{1}{msd(u, v)+1}$$

$$msd_sim(i, j) = \frac{1}{msd(i, j)+1}$$

$$\text{cosine_sim}(u, v) = \frac{\sum_{i \in I_{uv}} r_{ui} \cdot r_{vi}}{\sqrt{\sum_{i \in I_{uv}} r_{ui}^2} \cdot \sqrt{\sum_{i \in I_{uv}} r_{vi}^2}}$$

$$\text{cosine_sim}(i, j) = \frac{\sum_{u \in U_{ij}} r_{ui} \cdot r_{uj}}{\sqrt{\sum_{u \in U_{ij}} r_{ui}^2} \cdot \sqrt{\sum_{u \in U_{ij}} r_{uj}^2}}$$

$$pearson_sim(u, v) = \frac{\sum_{i \in I_{uv}} (r_{ui} - \mu_u) \cdot (r_{vi} - \mu_v)}{\sqrt{\sum_{i \in I_{uv}} (r_{ui} - \mu_u)^2} \cdot \sqrt{\sum_{i \in I_{uv}} (r_{vi} - \mu_v)^2}}$$

or

$$pearson_sim(i, j) = \frac{\sum_{u \in U_{ij}} (r_{ui} - \mu_i) \cdot (r_{uj} - \mu_j)}{\sqrt{\sum_{u \in U_{ij}} (r_{ui} - \mu_i)^2} \cdot \sqrt{\sum_{u \in U_{ij}} (r_{uj} - \mu_j)^2}}$$

$$pearson_baseline_sim(u, v) = \hat{\rho}_{uv} = \frac{\sum_{i \in I_{uv}} (r_{ui} - b_{ui}) \cdot (r_{vi} - b_{vi})}{\sqrt{\sum_{i \in I_{uv}} (r_{ui} - b_{ui})^2} \cdot \sqrt{\sum_{i \in I_{uv}} (r_{vi} - b_{vi})^2}}$$

or

$$pearson_baseline_sim(i, j) = \hat{\rho}_{ij} = \frac{\sum_{u \in U_{ij}} (r_{ui} - b_{ui}) \cdot (r_{uj} - b_{uj})}{\sqrt{\sum_{u \in U_{ij}} (r_{ui} - b_{ui})^2} \cdot \sqrt{\sum_{u \in U_{ij}} (r_{uj} - b_{uj})^2}}$$

$$\text{pearson_baseline_shrunk_sim}(u, v) = \frac{|I_{uv}|-1}{|I_{uv}|-1 + \text{shrinkage}} \cdot \hat{\rho}_{uv}$$

or

$$\text{pearson_baseline_shrunk_sim}(i, j) = \frac{|U_{ij}|-1}{|U_{ij}|-1 + \text{shrinkage}} \cdot \hat{\rho}_{ij}$$

Category

Meal Kit Service Profile Features

Price

Price_min, Price_max, Price_average, Price_std_dev

Plan

Plan_Count, agg_meals, unique_meals

Nutrition

Preparation Time

('prep_time_minutes', 'min') ('prep_time_minutes', 'max') ('prep_time_minutes', 'mean') ('prep_time_minutes', 'mad')

('prep_time_minutes', 'median') ('prep_time_minutes', 'std') ('prep_time_minutes', 'count')

Diet

(‘carbohydrate_grams’, ‘min’),	(‘sodium_milligrams’, ‘min’)
(‘carbohydrate_grams’, ‘max’),	(‘sodium_milligrams’, ‘max’)
(‘carbohydrate_grams’, ‘mean’),	(‘sodium_milligrams’, ‘mean’)
(‘carbohydrate_grams’, ‘mad’)	(‘sodium_milligrams’, ‘mad’)
(‘carbohydrate_grams’, ‘median’)	(‘sodium_milligrams’, ‘median’)
(‘carbohydrate_grams’, ‘std’)	(‘sodium_milligrams’, ‘std’)
(‘carbohydrate_grams’, ‘count’)	(‘sodium_milligrams’, ‘count’)
(‘fat_grams’, ‘min’)	
(‘fat_grams’, ‘max’)	(‘calories’, ‘min’)
(‘fat_grams’, ‘mean’)	(‘calories’, ‘max’)
(‘fat_grams’, ‘mad’)	(‘calories’, ‘mean’)
(‘fat_grams’, ‘median’)	(‘calories’, ‘mad’)
(‘fat_grams’, ‘std’)	(‘calories’, ‘median’)
(‘fat_grams’, ‘count’)	(‘calories’, ‘std’)
(‘protein_grams’, ‘min’) (‘protein_grams’, ‘max’)	(‘protein_grams’, ‘count’)
(‘protein_grams’, ‘mean’)	
(‘protein_grams’, ‘mad’)	(‘protein_grams’, ‘median’)
(‘protein_grams’, ‘median’)	(‘protein_grams’, ‘std’)
(‘protein_grams’, ‘count’)	

('hasBeef', 'sum') ('hasBeef', 'count')	('hasFish', 'sum') ('hasFish', 'count')
('hasLamb', 'sum') ('hasLamb', 'count')	('hasPork', 'sum') ('hasPork', 'count')
('hasPoultry', 'sum') ('hasPoultry', 'count')	('hasSeafood', 'sum') ('hasSeafood', 'count')
('hasShellfish', 'sum') ('hasShellfish', 'count')	('isVegan', 'sum') ('isVegan', 'count') ('isVegetarian', 'sum') ('isVegetarian', 'count')

Cover Letter

November 21, 2019

Dear Participant:

My name is Robert Nakano and I am a graduate student in applied statistics at California State University, Long Beach. For my final project, I am examining statistical and machine learning methods implemented in recommendation systems for meal kit services. Because you are a potential meal kit subscriber, I am inviting you to participate in this research study by completing the attached surveys.

The following questionnaire will require approximately 15 minutes to complete. There is no compensation for responding nor is there any known risk. In order to ensure that all information will remain confidential, please do not include your name. Copies of the project will be provided to my CSULB instructor and to the Institutional Review Board (IRB). Additionally, data will be shared with PickaKit.com to help provide users with better meal kit recommendations. If you choose to participate in this project, please answer all questions as honestly as possible and return the completed questionnaires promptly. Participation is strictly voluntary and you may refuse to participate at any time.

I sincerely appreciate your time and help in this survey. The data collected will provide useful information regarding studying applications of predictive methods and eventually help individuals select better meals. Without your experience, this project would not be possible. Thank you for taking the time to assist me in my educational endeavors. If you would like a summary copy of this study please complete and detach the Request for Information Form and return it to me in a separate envelope or via email. Completion and return of the questionnaire will indicate your willingness to participate in this study. If you require additional information or have questions, please feel free to contact me at Robert.Nakano@student.csulb.edu.

If you are not satisfied with the manner in which this study is being conducted, you may report (anonymously if you so choose) any complaints to the Department of Mathematics and Statistics, California State University, Long Beach, 1250 Bellflower Blvd., Long Beach, CA, 90840-1001

Sincerely,

Robert Nakano
Robert.Nakano@student.csulb.edu
M.S. Applied Statistics, Candidate

Dr. Olga Korosteleva
Olga.Korosteleva@csulb.edu
Professor of Statistics

Questionnaire

Meal Kit Questionnaire

Demographics

1. What gender do you identify with?
Male Female Other/Non-binary
2. What is your birthday? Day ____ Month ____ Year ____
3. In what zip code do you primarily reside in? _____
4. What is your Myers Briggs (MBTI) personality type?
If unknown, take this 4 minute quiz: <https://youtopia-project.com/youtopia-16-assessment/>
ESTP ESFP ISTP ISFP
ESFJ ESTJ ISTJ ISFJ
ENTP ENTJ INTP INTJ
ENFP ENFJ INFJ INFJ
5. How many people are in your household?
1 2 3 4 5 6 7 Other: _____

Income & Price Sensitivity

6. What is your approximate annual income? _____
7. Do you have any food allergies? _____
8. Do you have any dietary restrictions due to religion? _____
9. Do you have any other dietary restrictions? _____
10. List any foods you don't like: _____
11. Is eating organic foods an important factor to you? Yes No
12. Is eating locally sourced foods an important factor to you? Yes No
13. Are you following any particular diets?

Health

14. What is your weight? ____
15. What is your height? ____
16. What is your blood pressure at the last reading? ____ ____ Don't recall
17. Do you have any health goals? _____
18. How much time do you have to cook? ____
19. Ideally, how much time would you like to spend cooking? ____
20. Do you enjoy cooking? Often Sometimes rarely
21. Is eco friendliness an important factor to you in food packaging?

Service Ratings

Please rate the following services you have tried: 1 being terrible, 10 being perfect!

22. Blue Apron: Never Tried 1 2 3 4 5 6 7 8 9 10
23. Gobble: Never Tried 1 2 3 4 5 6 7 8 9 10
24. Green Chef: Never Tried 1 2 3 4 5 6 7 8 9 10
25. HelloFresh: Never Tried 1 2 3 4 5 6 7 8 9 10
26. Home Chef: Never Tried 1 2 3 4 5 6 7 8 9 10
27. Marley Spoon: Never Tried 1 2 3 4 5 6 7 8 9 10
28. Plated: Never Tried 1 2 3 4 5 6 7 8 9 10
29. Sun Basket: Never Tried 1 2 3 4 5 6 7 8 9 10
30. Dinnerly: Never Tried 1 2 3 4 5 6 7 8 9 10
31. Everyplate: Never Tried 1 2 3 4 5 6 7 8 9 10
32. Freshly: Never Tried 1 2 3 4 5 6 7 8 9 10

Please fill in and rate any other meal kit services you have tried.

33. _____ : 1 2 3 4 5 6 7 8 9 10
34. _____ : 1 2 3 4 5 6 7 8 9 10
35. _____ : 1 2 3 4 5 6 7 8 9 10

Do you have any comments on the meal kit services you have tried above?

IRB Application for Expedited and Standard Review

Projects involving minimal or greater than minimal risk

Instructions: The Principal Investigator (PI) is responsible for all information in this form. The PI should fill out the form completely; incomplete forms will be returned. This form should be completed in Microsoft Word; any other word processing program may compromise the formatting of this document. Use as much space as necessary to answer each question adequately. Check boxes can be filled and unfilled by clicking once inside the box you intend to select or de-select. Please upload the completed application and all relevant appendices to IRBNet (www.irbnet.org).

IMPORTANT: NO ACTIVITY MAY BEGIN ON THIS PROJECT UNTIL THE CSULB IRB ISSUES FORMAL NOTIFICATION TO THE PI REGARDING PROJECT APPROVAL VIA IRBNet. In addition, the IRB will confirm that all individuals listed on this project have successfully completed the CITI online training for the Social & Behavioral Basic/Refresher course before issuing

Investigator Information

1. Basic Information

Principal Investigator (Name)	Click or tap here to enter text.
CITI Member ID #	Click or tap here to enter text.
Telephone Number	Click or tap here to enter text.
Email	Click or tap here to enter text.
Department	Click or tap here to enter text.
Affiliation with CSULB	<input type="checkbox"/> Student* <input type="checkbox"/> Faculty <input type="checkbox"/> Staff <input type="checkbox"/> External PI <input type="checkbox"/> Other:
<small>*If you are a student, please fill out the information for your faculty advisor in the next section and attach a signed permission letter from your advisor as an appendix to your project in IRBNet.</small>	

2. PI Qualifications

Describe any training, certification, experience or expertise relevant to this research. Also, explain your relationship or affiliation with any organization, course or group involved in with the project.

Variable Descriptions

https://drive.google.com/file/d/1HYFar2VoNnqfu2o8xL-log6mRJK8_z-b/view?usp=sharing