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?

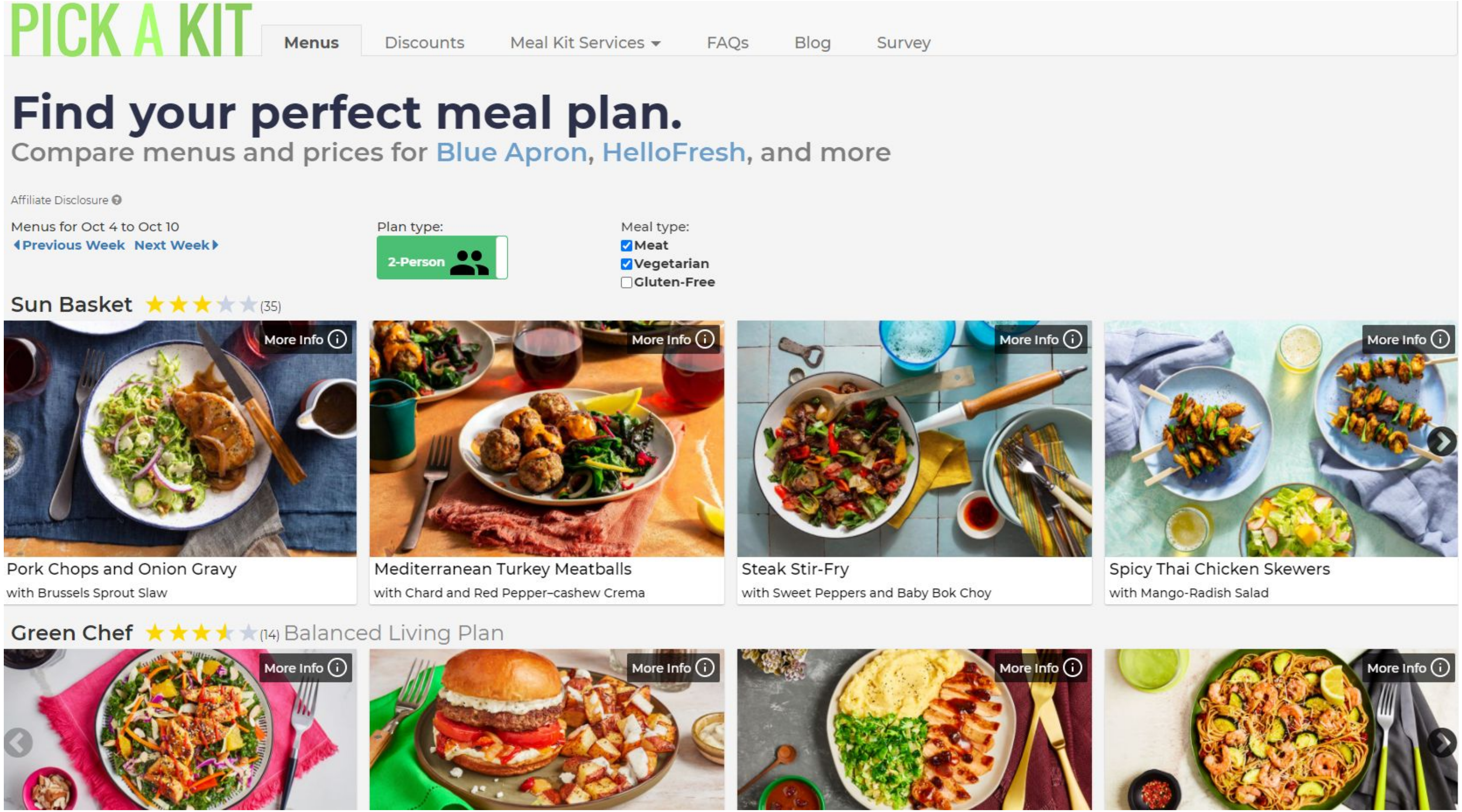
Meal kits are boxes containing premeasured and packaged ingredients for one or more recipes that are delivered to a buyer's address, oftentimes on a subscription basis.

PICK A KIT =

yelp

+

Pickakit.com

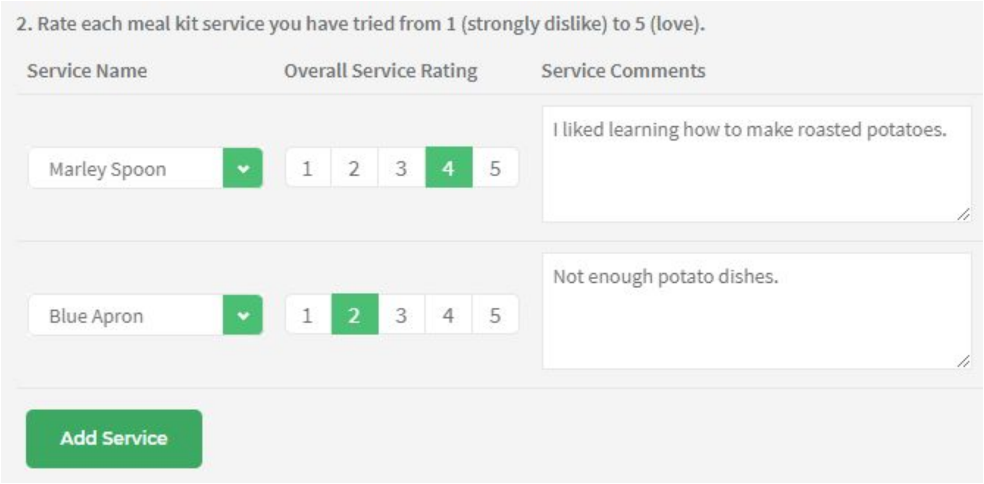


Problem Statement

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

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

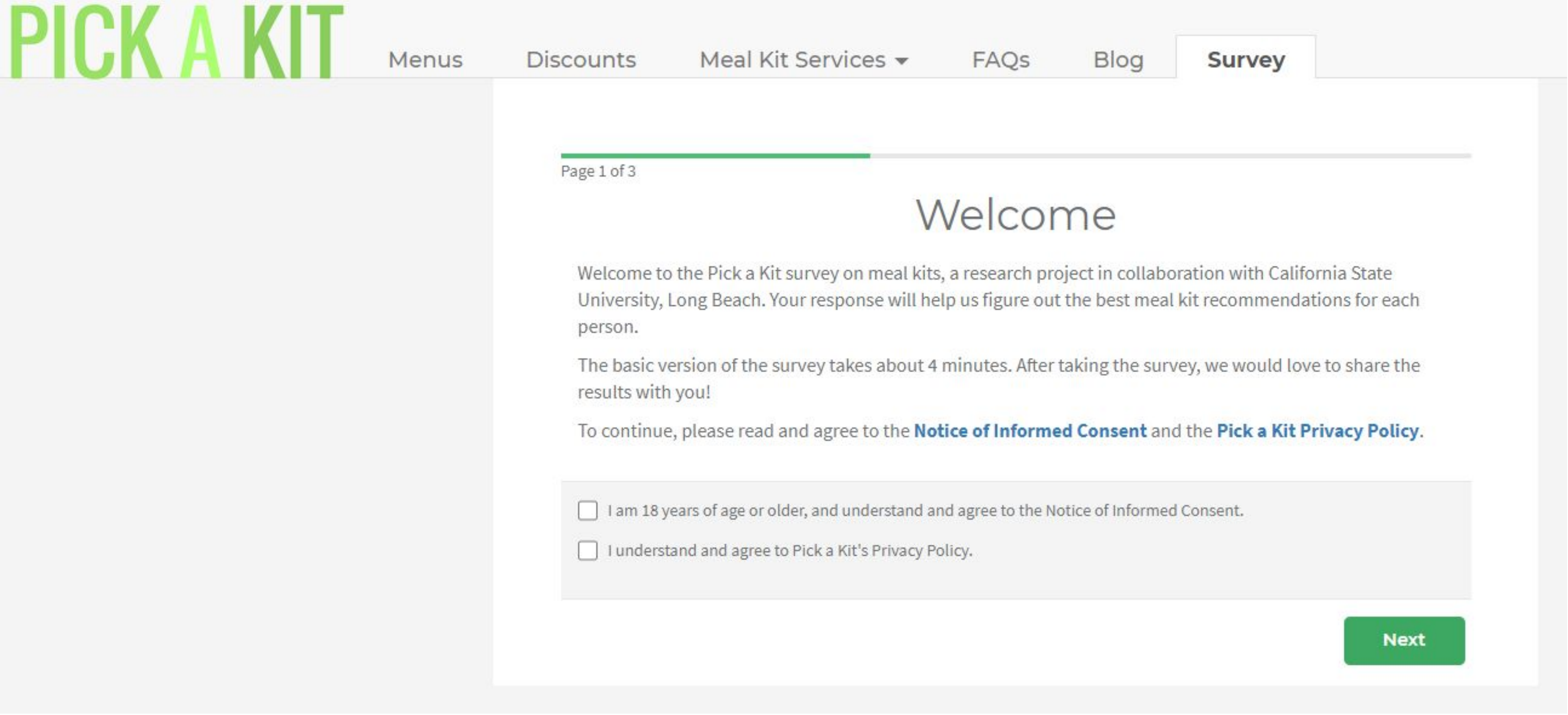
Data stored on MongoDB NoSQL databases

JSON format

Survey.js

Hosted on Pick a Kit

Pickakit.com



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:

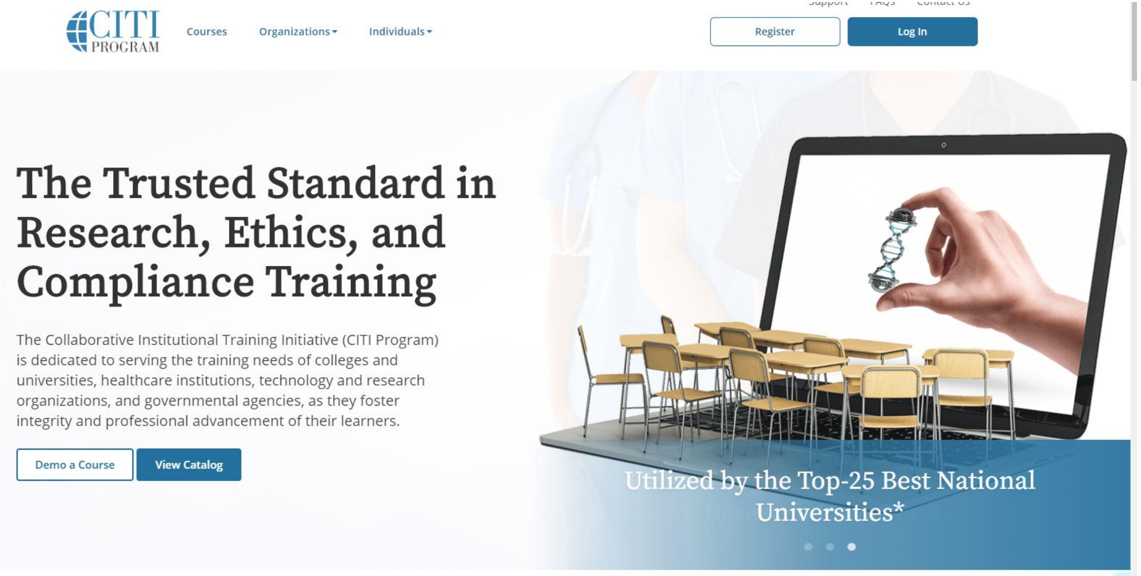
https://www.csulb.edu/office-of-research-and-sponsored-programs/institutional-review-board-irb

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 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, there-by 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

CITI PROGRAM

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 (Stage)

Under requirements set by:

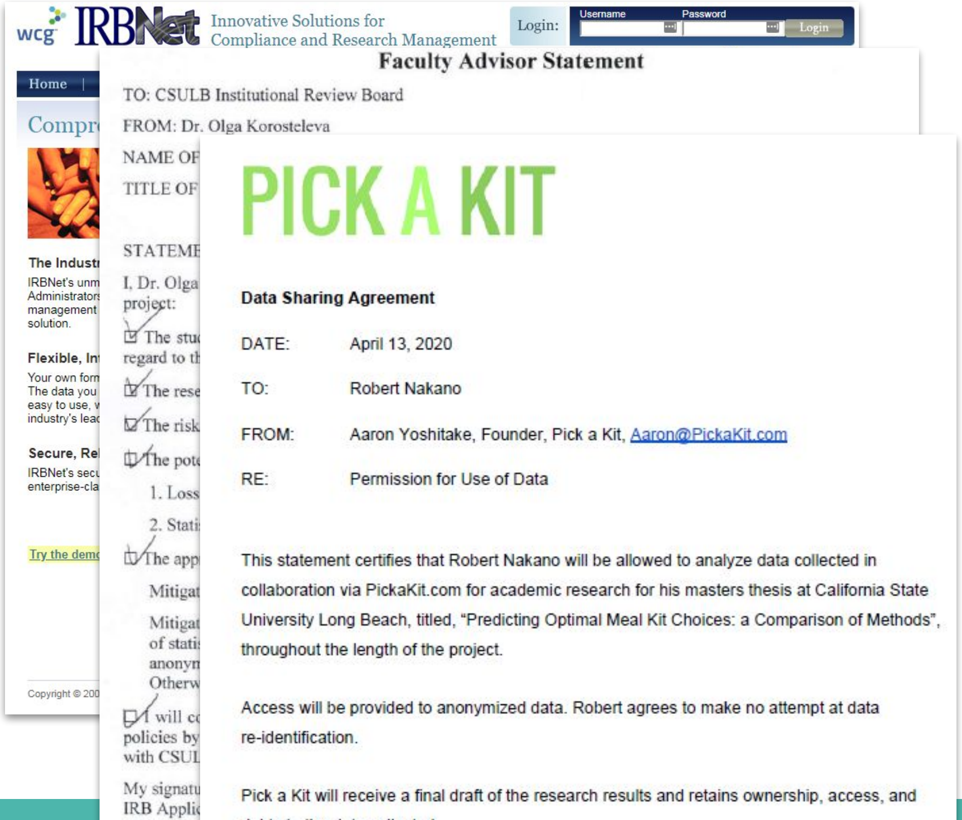
California State University, Long Beach

CITI
Collaborative Institutional Training Initiative

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



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)	☐ Yes ☐ No ☐ Not Sure
Department	Click or tap here to enter text.
Affiliation	☐ Student* ☐ Staff ☐ Faculty ☐ 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.
☐ 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

IRB Application for Existing and Secondary Data

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.

1. Basic Information

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:	☐ Student* ☐ Faculty ☐ Staff ☐ External PI

Step 5: Final Approval

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

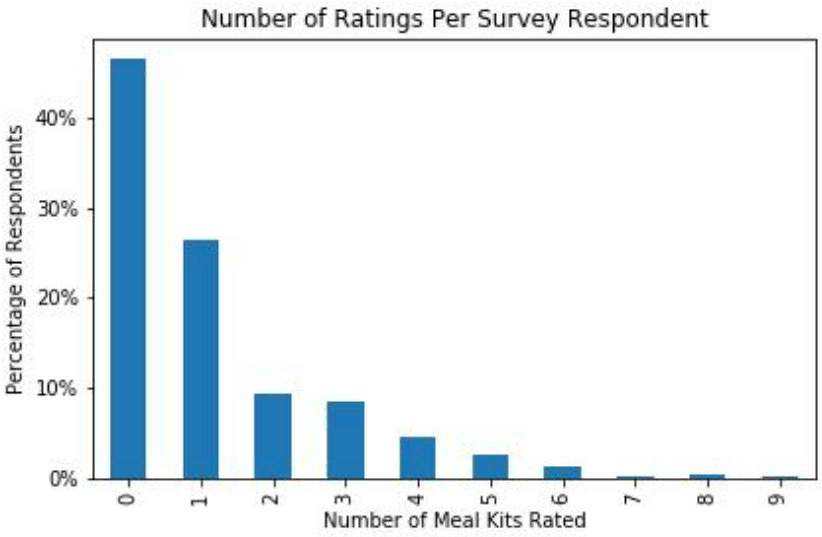
Survey Results
Descriptive Statistics

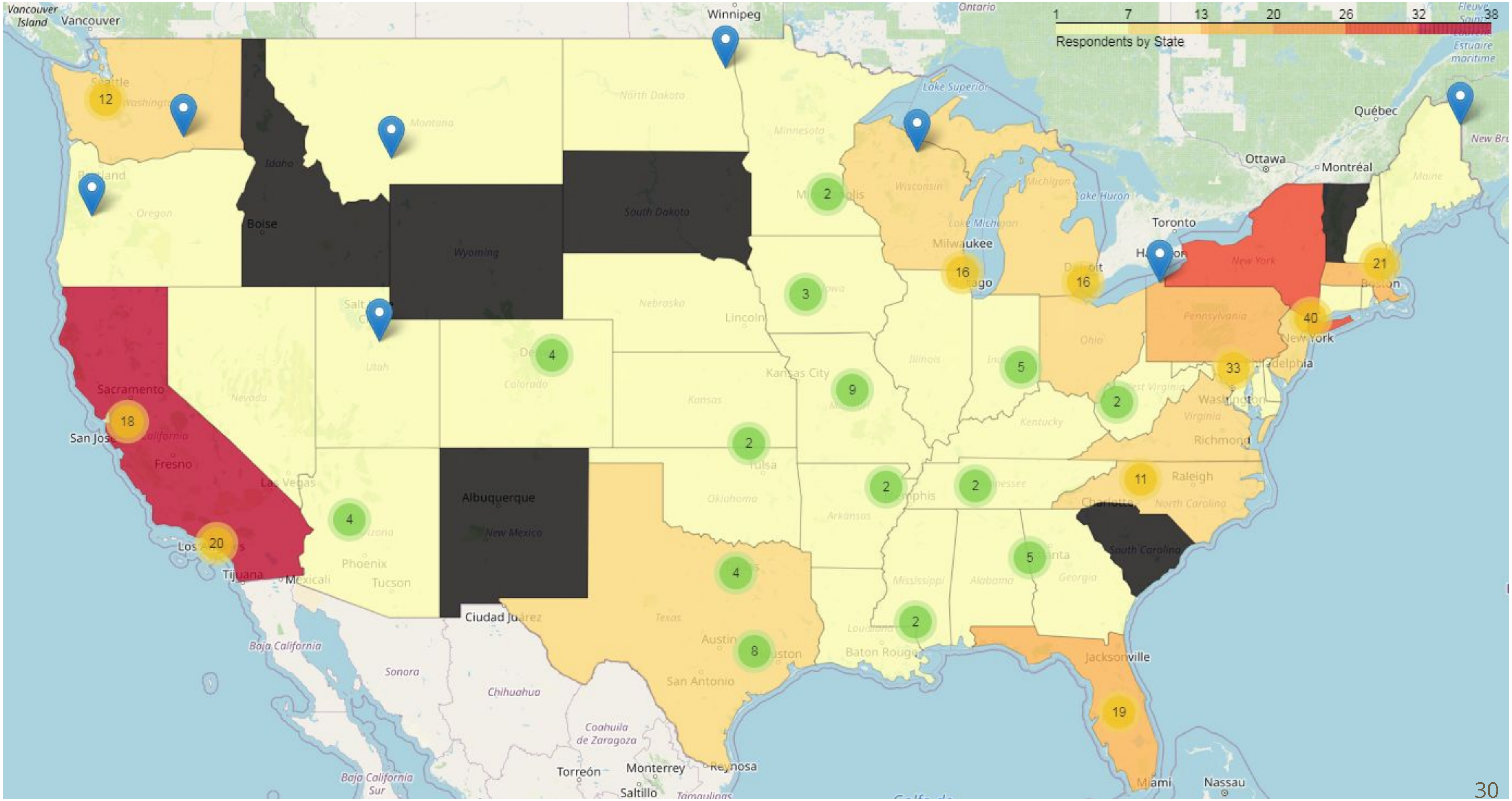
Survey Respondents

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.

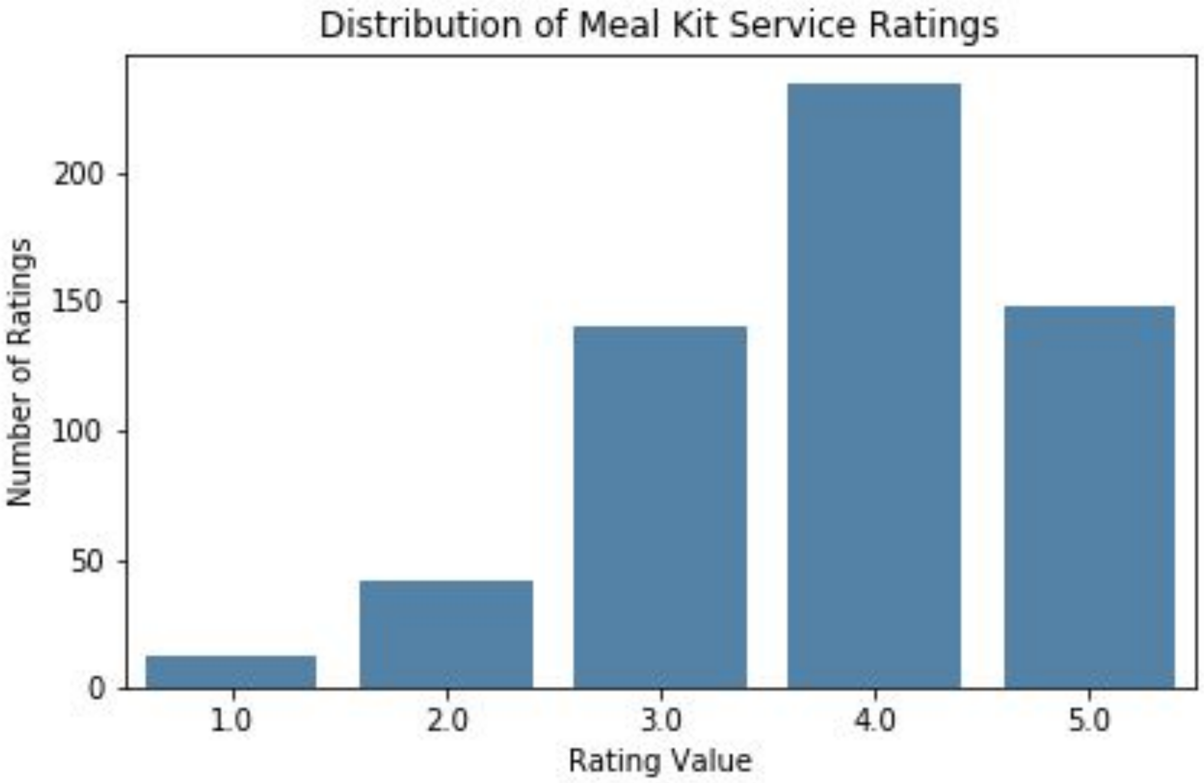



1
7
13
20
26
32
38
Respondents by State

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.



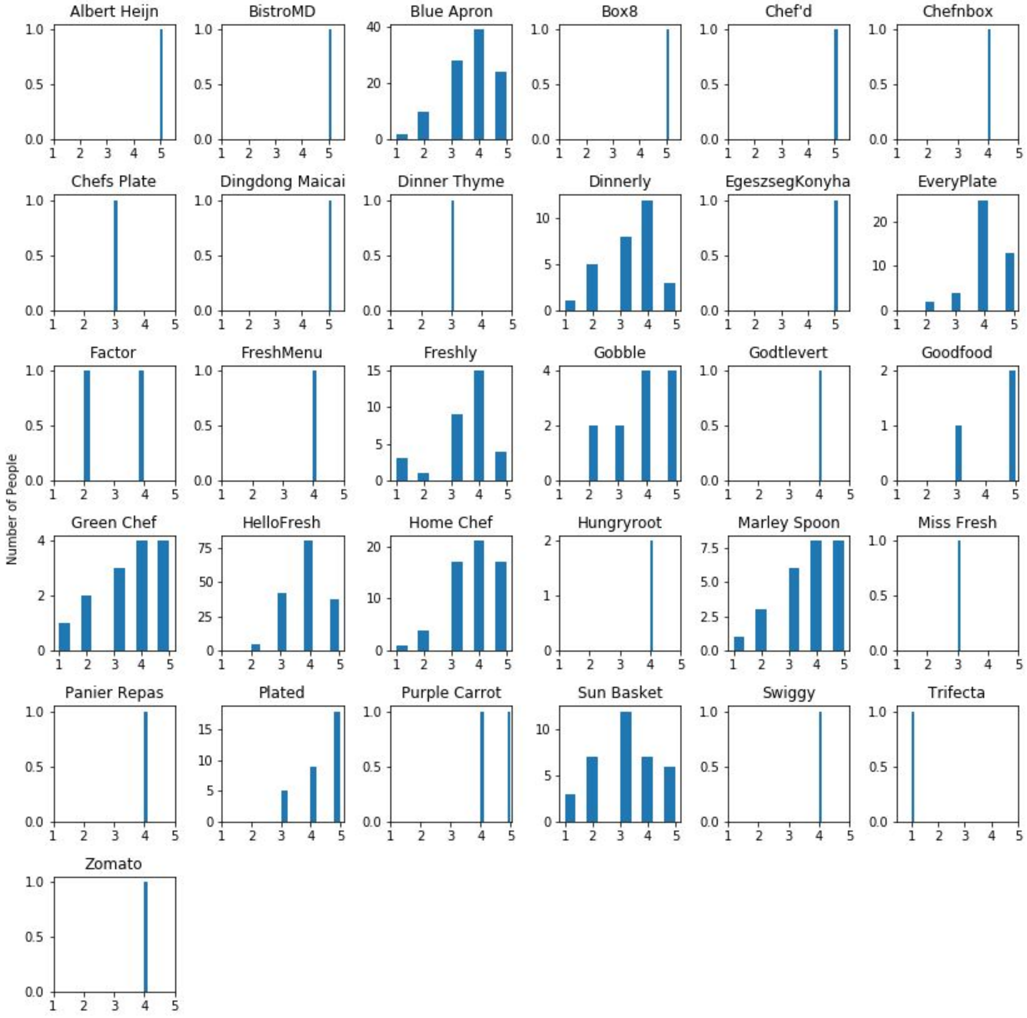
499 survey respondents

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

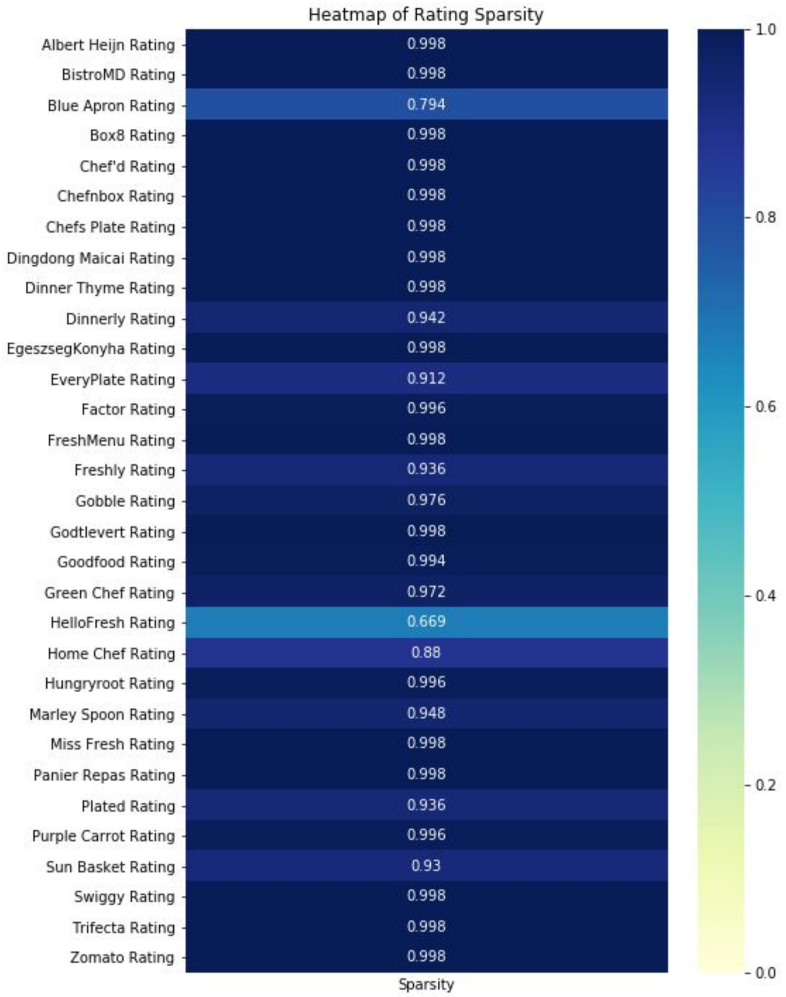
Average Ratings for Meal Kit Services with 3 or More Ratings

Meal Kit Service Ratings Histograms

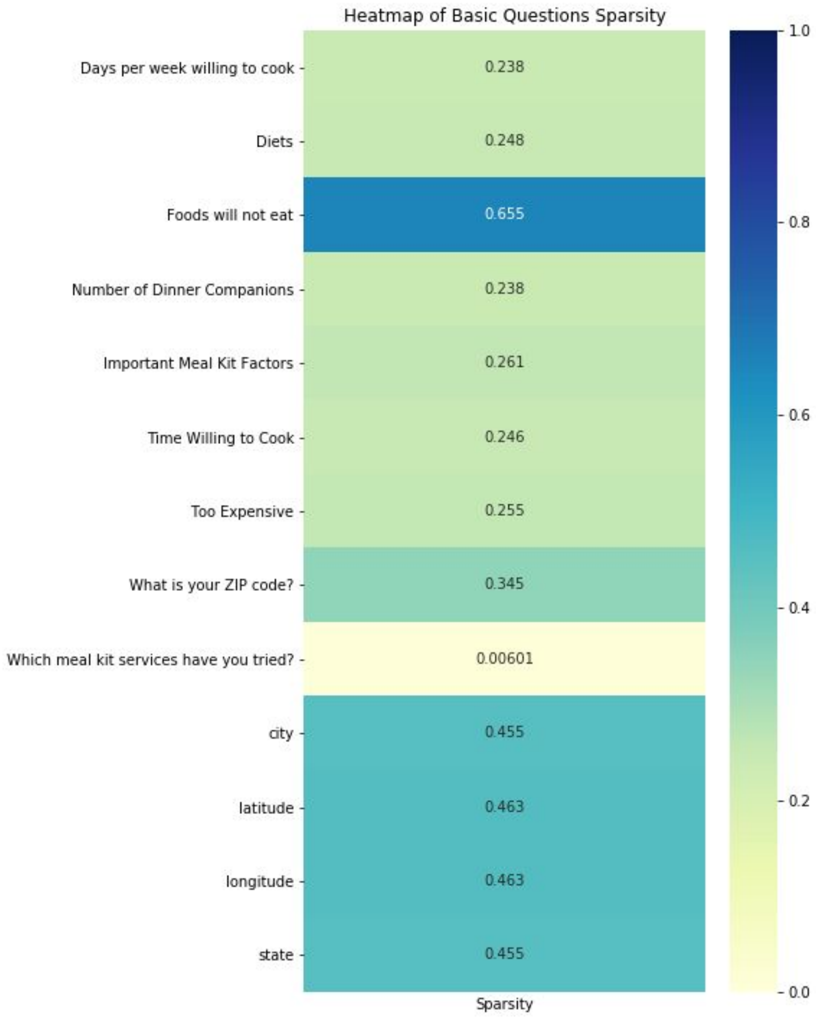


Data Sparsity

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



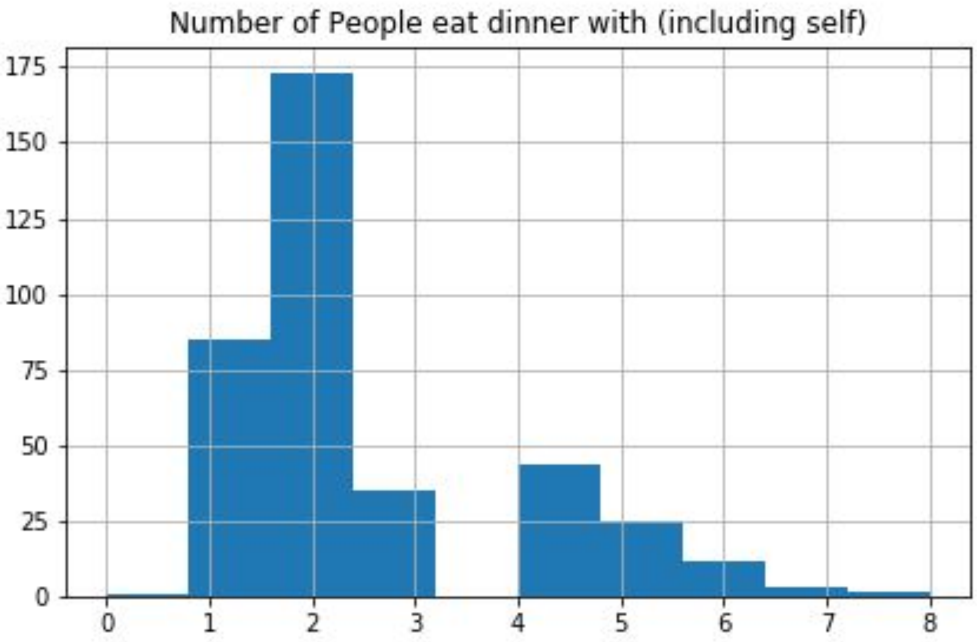
Heatmap of Rating Sparsity
Albert Heijn Rating
BistroMD Rating
Blue Apron Rating
Box8 Rating
Chef'd Rating
Chefnbox Rating
Chefs Plate Rating
Dingdong Maicai Rating
Dinner Thyme Rating
Dinnerly Rating
EgeszsegKonyha Rating
EveryPlate Rating
Factor Rating
FreshMenu Rating
Freshly Rating
Gobble Rating
Godtlevert Rating
Goodfood Rating
Green Chef Rating
HelloFresh Rating
Home Chef Rating
Hungryroot Rating
Marley Spoon Rating
Miss Fresh Rating
Panier Repas Rating
Plated Rating
Purple Carrot Rating
Sun Basket Rating
Swiggy Rating
Trifecta Rating
Zomato Rating
0.998
0.998
0.794
0.998
0.998
0.998
0.998
0.998
0.998
0.942
0.998
0.912
0.996
0.998
0.936
0.976
0.998
0.994
0.972
0.669
0.88
0.996
0.948
0.998
0.998
0.936
0.996
0.93
0.998
0.998
0.998
Sparsity
1.0
0.8
0.6
0.4
0.2
0.0

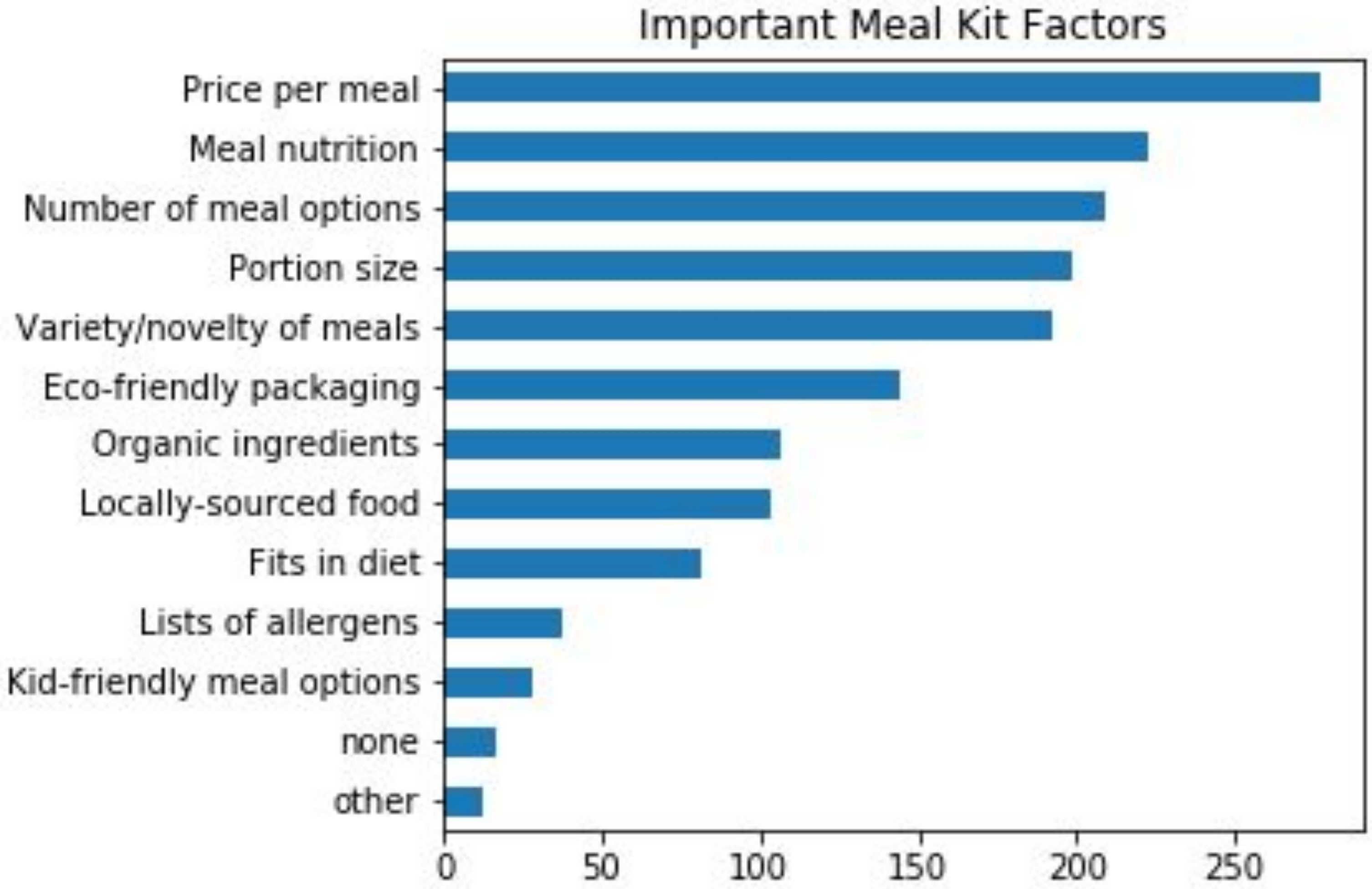


Heatmap of Basic Questions Sparsity
Days per week willing to cook
Diets
Foods will not eat
Number of Dinner Companions
Important Meal Kit Factors
Time Willing to Cook
Too Expensive
What is your ZIP code?
Which meal kit services have you tried?
city
latitude
longitude
state
0.238
0.248
0.655
0.238
0.261
0.246
0.255
0.345
0.00601
0.455
0.463
0.463
0.455
Sparsity
1.0
0.8
0.6
0.4
0.2
0.0

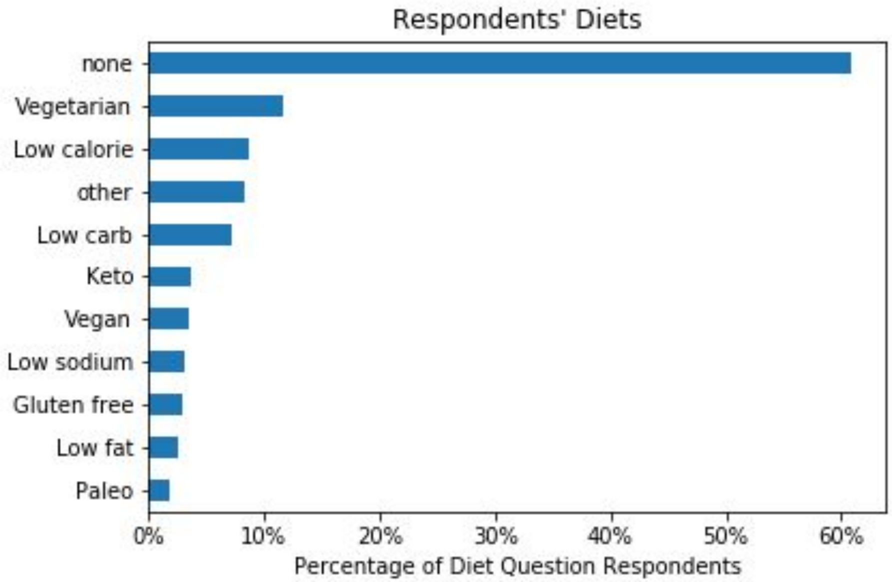
Dinner Companions

Many respondents eat dinner with 1 other person




Important Meal Kit Factors
Price per meal
Meal nutrition
Number of meal options
Portion size
Variety/novelty of meals
Eco-friendly packaging
Organic ingredients
Locally-sourced food
Fits in diet
Lists of allergens
Kid-friendly meal options
none
other
0
50
100
150
200
250

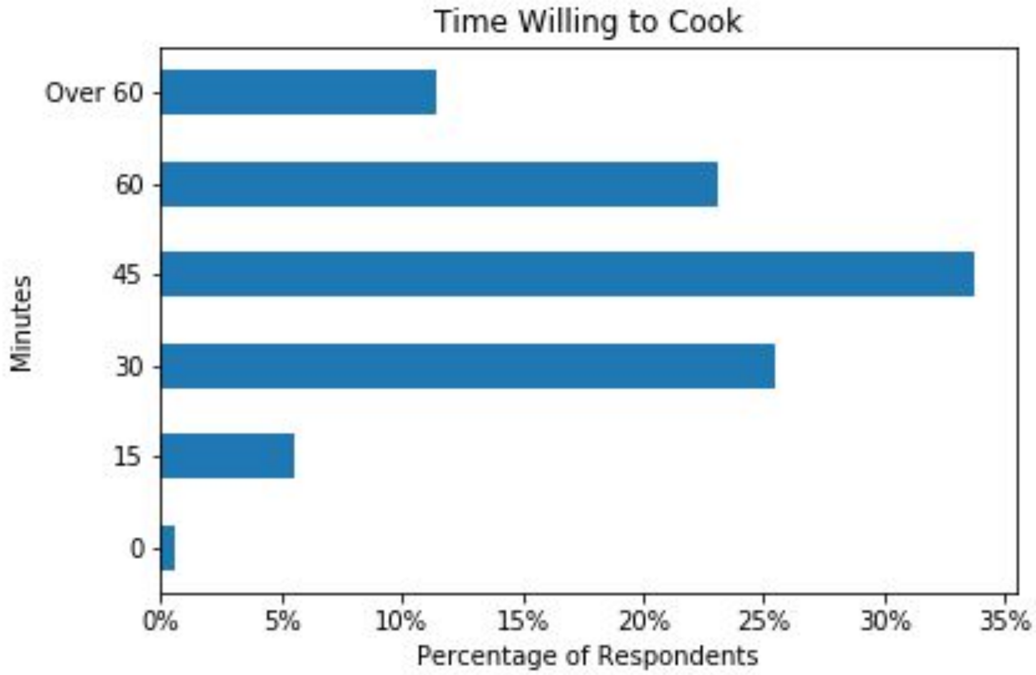
Diets of Respondents



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

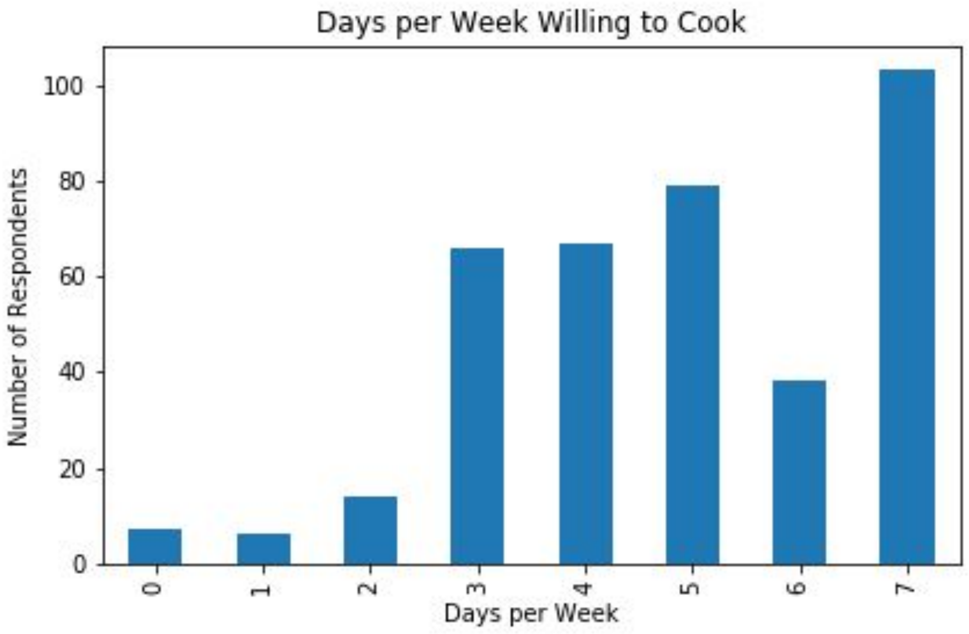
45 minutes was the most common time willing to cook



Cooking Frequency

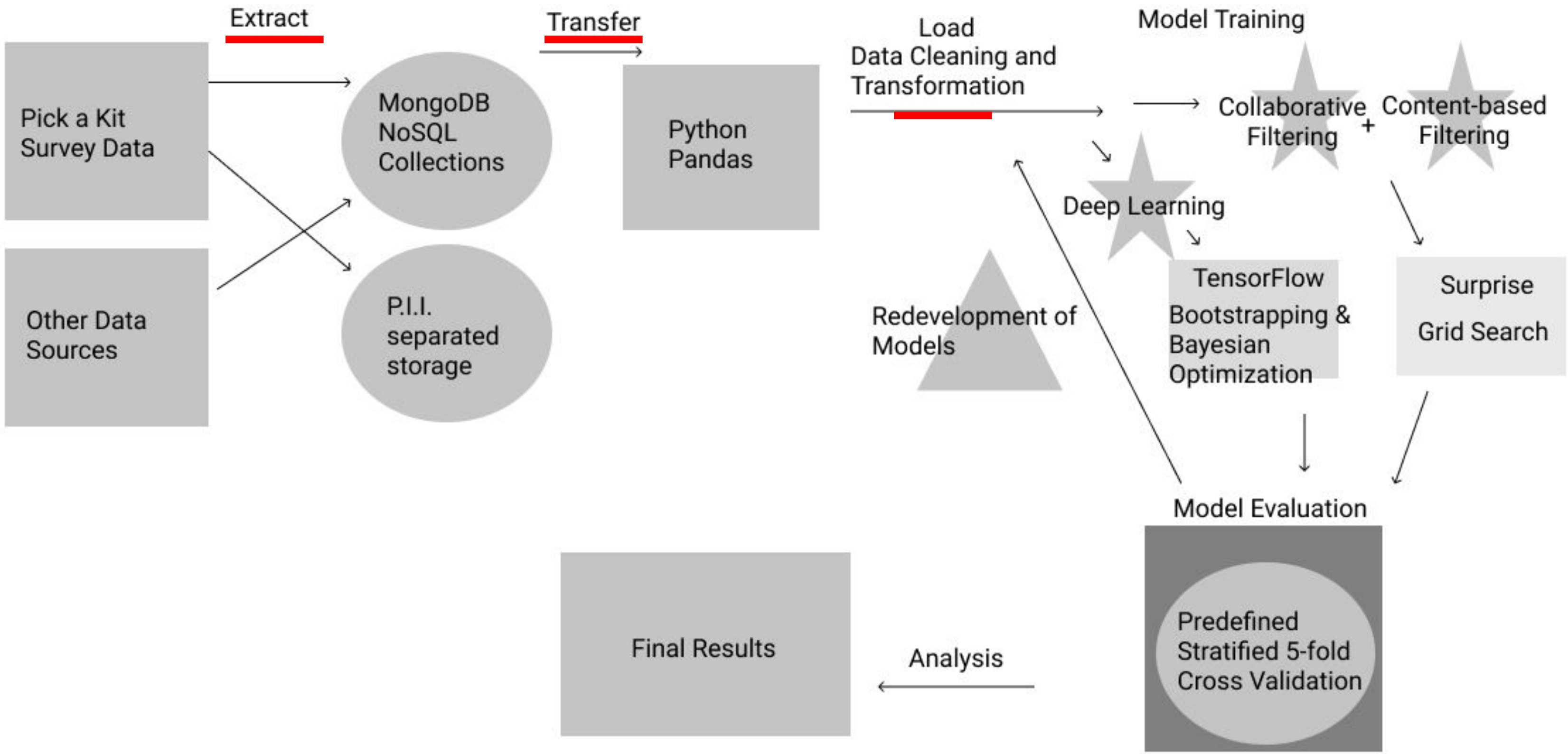
Many respondent reported willingness to cook every day of the week

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




Percentage of Respondents Think Price is Too Expensive
Percentage of Respondents
100%
80%
60%
40%
20%
0%
6
8
10
12
15
infinite
Price in USD

Methodology



Extract
Transfer
Load
Data Cleaning and Transformation
Model Training
Pick a Kit Survey Data
Other Data Sources
MongoDB NoSQL Collections
P.I.I. separated storage
Python Pandas
Collaborative Filtering
+
Content-based Filtering
Deep Learning
TensorFlow
Bootstrapping & Bayesian Optimization
Surprise
Grid Search
Redevelopment of Models
Model Evaluation
Predefined Stratified 5-fold Cross Validation
Analysis
Final Results

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

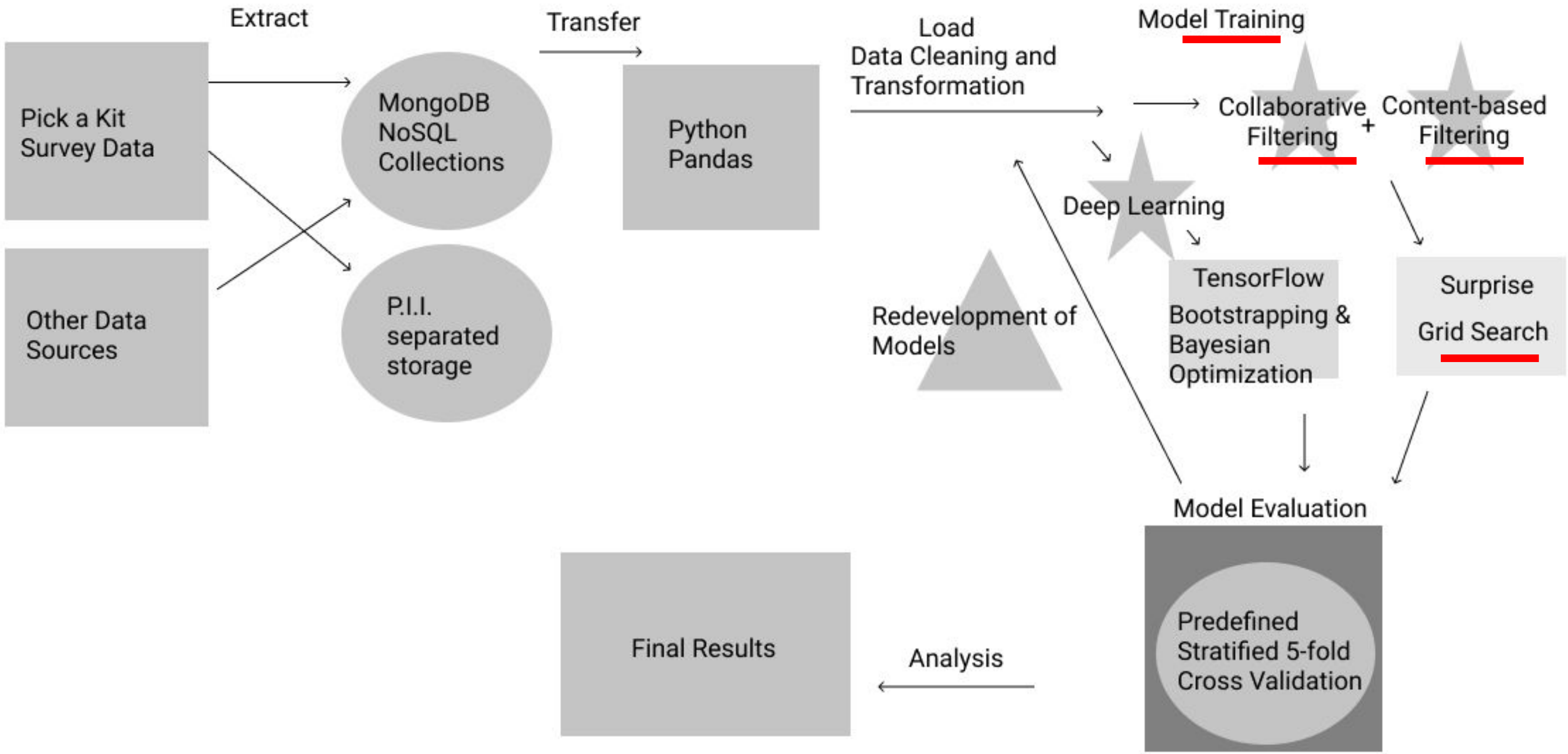
- Python
- Google Colab
- Pandas
- Numpy
- Sci-Kit Learn

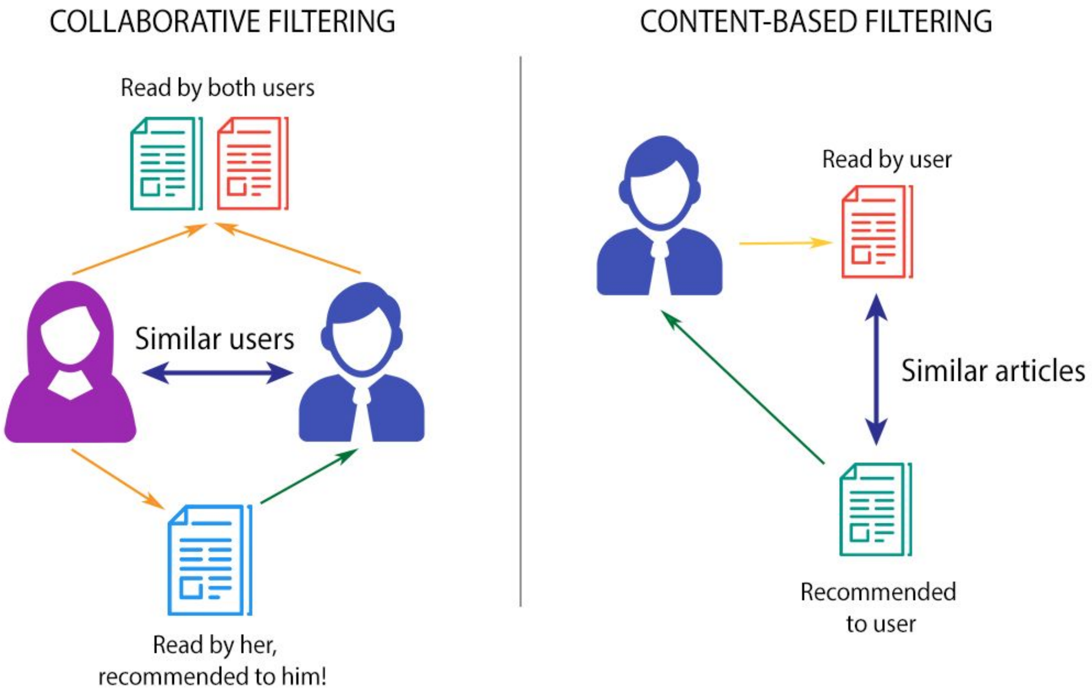
Prediction Matrix Data Frame

	_id	service	rating	How many people do you eat dinner with on a typical night? Include yourself.	Days per week willing to cook	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	[redacted]	Albert Heijn	5.0	2.0	7.0	45.0	8.0	0.00	0.00	0	0	0	0	0	
607	[redacted]	BistroMD	5.0	1.0	6.0	15.0	12.0	0.00	0.00	0	0	1	0	0	
612	[redacted]	Blue Apron	4.0	2.0	5.0	45.0	12.0	0.00	0.00	0	0	0	0	0	
614	[redacted]	Blue Apron	3.0	4.0	4.0	45.0	15.0	34.08	-118.14	1	0	0	0	0	
617	[redacted]	Blue Apron	5.0	2.0	5.0	45.0	12.0	0.00	0.00	1	0	0	0	1	
618	[redacted]	Blue Apron	4.0	3.0	2.0	60.0	10.0	40.56	-105.13	1	0	0	0	1	
620	[redacted]	Blue Apron	4.0	1.0	5.0	45.0	12.0	38.90	-92.40	1	1	0	1	0	
621	[redacted]	Blue Apron	4.0	2.0	4.0	30.0	15.0	37.32	-121.93	1	0	0	0	0	
626	[redacted]	Blue Apron	3.0	2.0	3.0	60.0	10.0	33.74	-117.81	1	0	0	0	0	
630	[redacted]	Blue Apron	3.0	2.0	4.0	30.0	10.0	33.68	-117.83	1	1	0	0	0	

Algorithm Groups

Collaborative Filtering, Content-based Filtering, and Deep Learning


Extract
Transfer
Load
Data Cleaning and
Transformation
Model Training
Pick a Kit
Survey Data
Other Data
Sources
MongoDB
NoSQL
Collections
P.I.I.
separated
storage
Python
Pandas
Collaborative
Filtering
+
Content-based
Filtering
Deep Learning
TensorFlow
Bootstrapping &
Bayesian
Optimization
Surprise
Grid Search
Redevelopment of
Models
Model Evaluation
Predefined
Stratified 5-fold
Cross Validation
Analysis
Final Results



https://towardsdatascience.com/brief-on-recommender-systems-b86a1068a4dd

Meal Kit Service Examples

Collaborative Filtering

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

Predicts Yale will like Gobble

Content-based Filtering

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

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.

Equation

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

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

Best Results:

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

Baseline Algorithm

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

Equation

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

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

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}

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.

Equation

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

Best Results:

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

{'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

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

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)	(carbohy
service_name										
Albert Heijn	1.74	9.00	4.69	1.40	1.0	NaN	NaN	NaN	NaN	
BistroMD	9.50	13.00	11.15	1.27	5.0	NaN	NaN	NaN	NaN	
Blue Apron	7.49	9.99	9.20	0.99	3.0	18.769231	191.0	26.0	178.0	
Box8	0.13	16.90	2.19	2.33	1.0	NaN	NaN	NaN	NaN	
Chef'd	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	
Chefnbox	13.56	13.56	13.56	0.00	1.0	NaN	NaN	NaN	NaN	
Chefs Plate	8.99	9.99	9.49	0.50	2.0	NaN	NaN	NaN	NaN	
Dingdong Maicai	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	
Dinnerly	4.29	4.99	4.72	0.23	2.0	19.076923	205.0	1.0	160.0	
EgeszsegKonyha	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	
Factor	11.00	15.00	12.54	1.39	5.0	NaN	NaN	NaN	NaN	
Freshly	7.99	11.50	9.37	1.30	4.0	38.307692	56.0	16.0	68.0	
FreshMenu	1.73	10.71	3.35	3.91	1.0	NaN	NaN	NaN	NaN	
Gobble	11.99	11.99	11.99	0.00	3.0	NaN	NaN	NaN	NaN	
Godtlevert	6.09	15.68	9.89	3.61	1.0	NaN	NaN	NaN	NaN	
Goodfood	7.48	9.60	8.47	0.87	5.0	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

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

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

Deep Learning Approaches

Deep Neural Networks (DNN)

Two layered fully connected neural networks

Experts

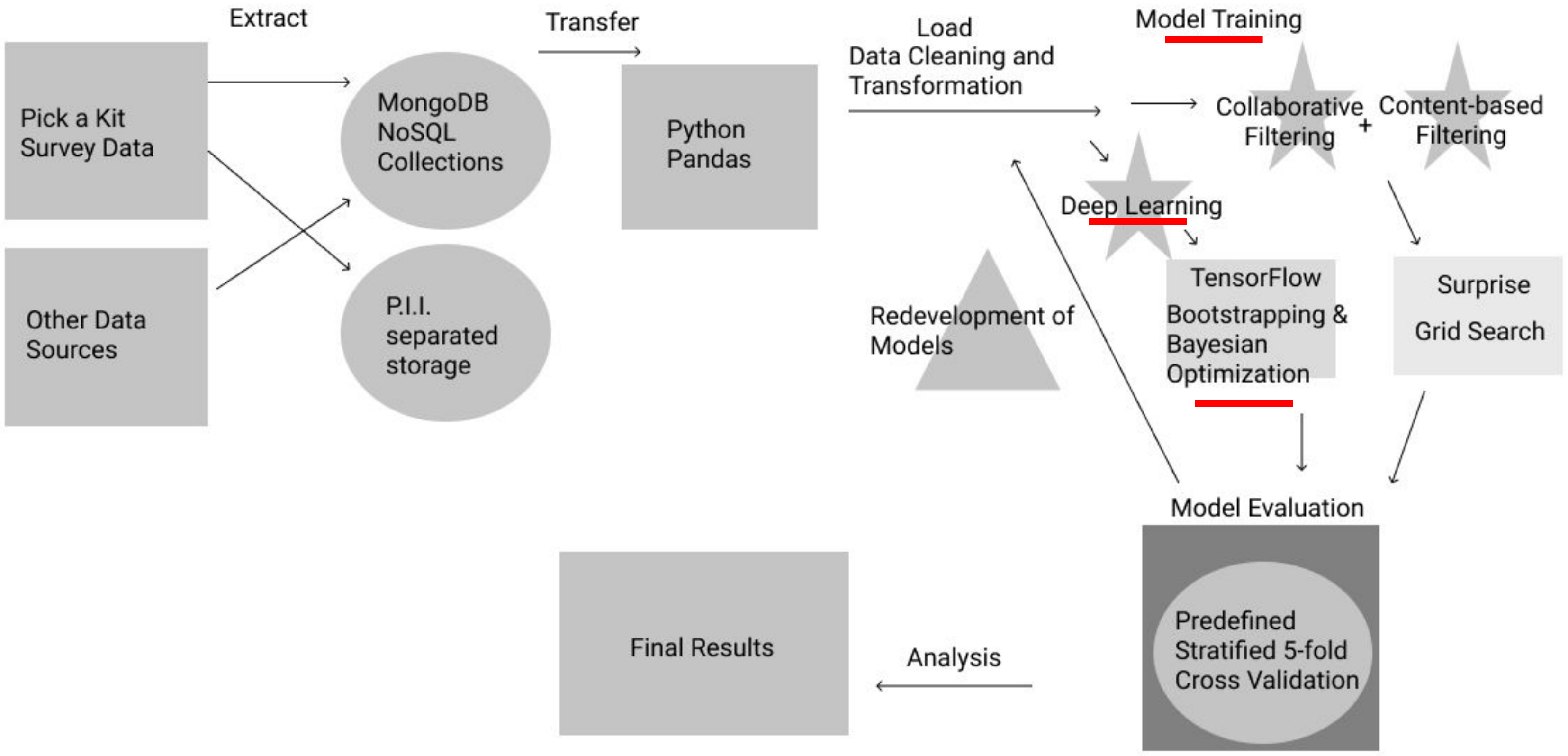
TensorFlow Keras version 2.3.0

Adam Optimization

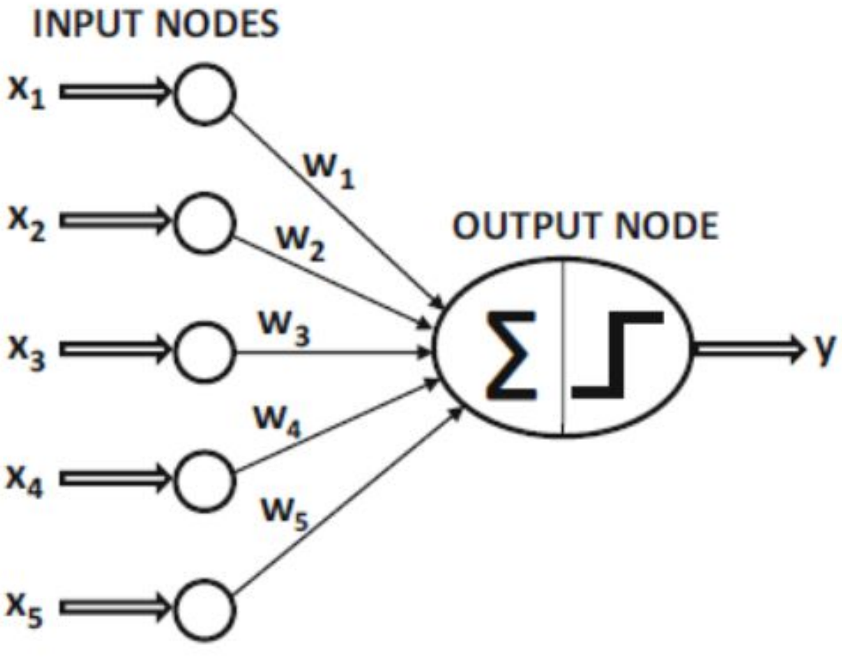
PReLU and ReLU Activation Functions

Early Stopping

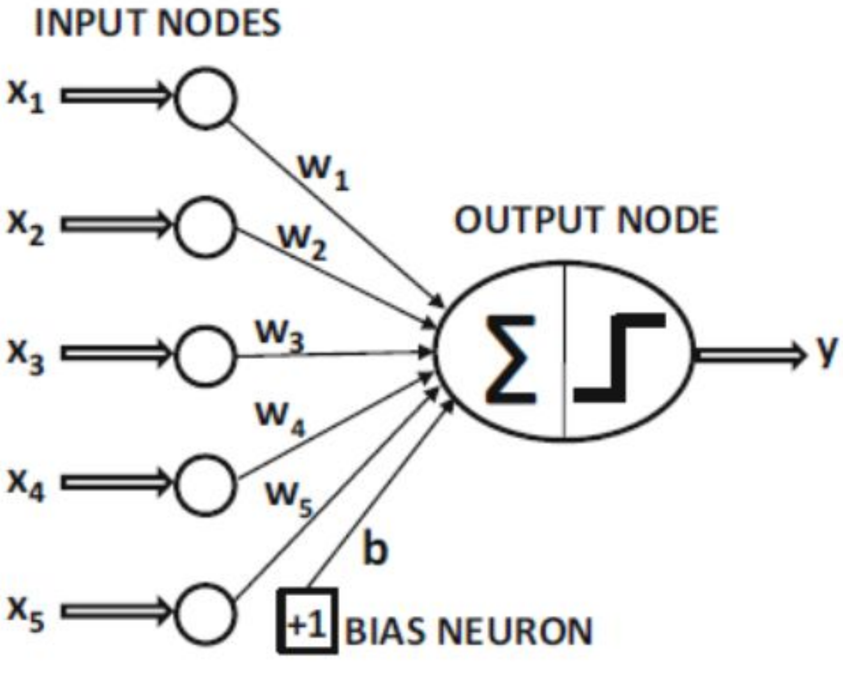
Bootstrapping 10x


Extract
Pick a Kit
Survey Data
Other Data
Sources
MongoDB
NoSQL
Collections
P.I.I.
separated
storage
Transfer
Python
Pandas
Load
Data Cleaning and
Transformation
Model Training
Collaborative
Filtering
+
Content-based
Filtering
Deep Learning
TensorFlow
Bootstrapping &
Bayesian
Optimization
Surprise
Grid Search
Redevelopment of
Models
Model Evaluation
Predefined
Stratified 5-fold
Cross Validation
Analysis
Final Results

The Basic Architecture of the Perceptron



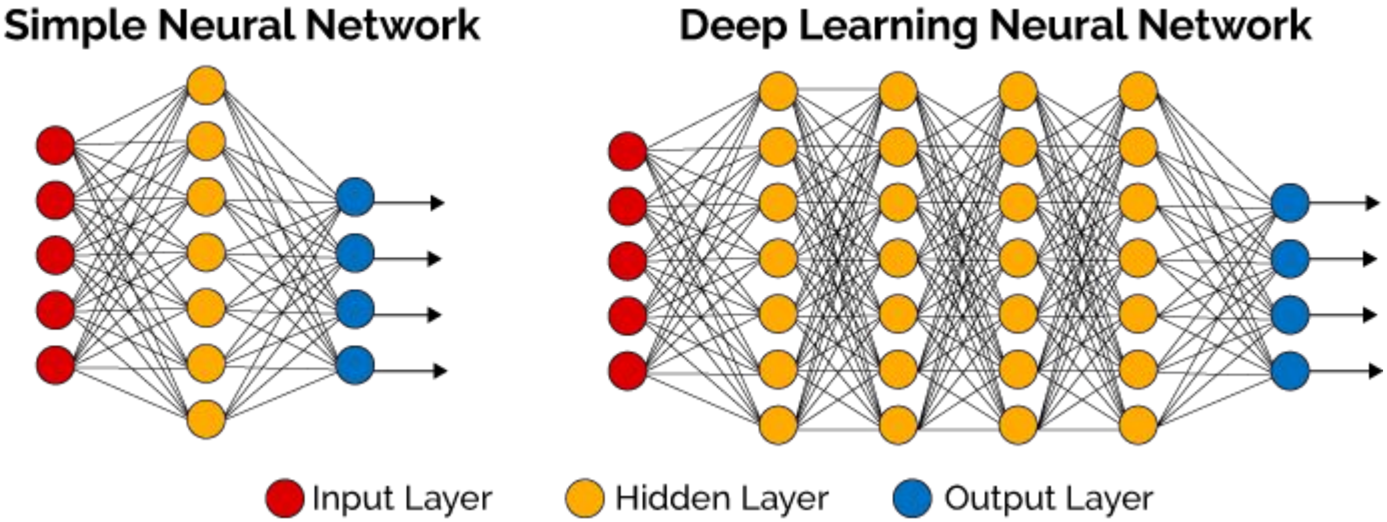
(a) Perceptron without bias



(b) Perceptron with bias

Charu Aggarwal 2018, Neural Networks and Deep Learning

Neural Networks Nodes and Layers



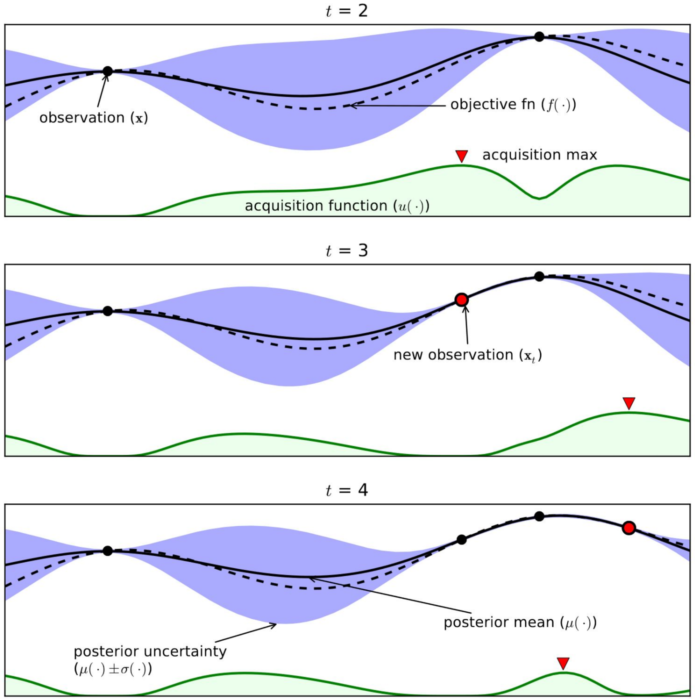
https://www.datadriveninvestor.com/deep-learning-explained/

Bayesian Optimization

Objective Function: Average RMSE of 10 Bootstrap Iterations

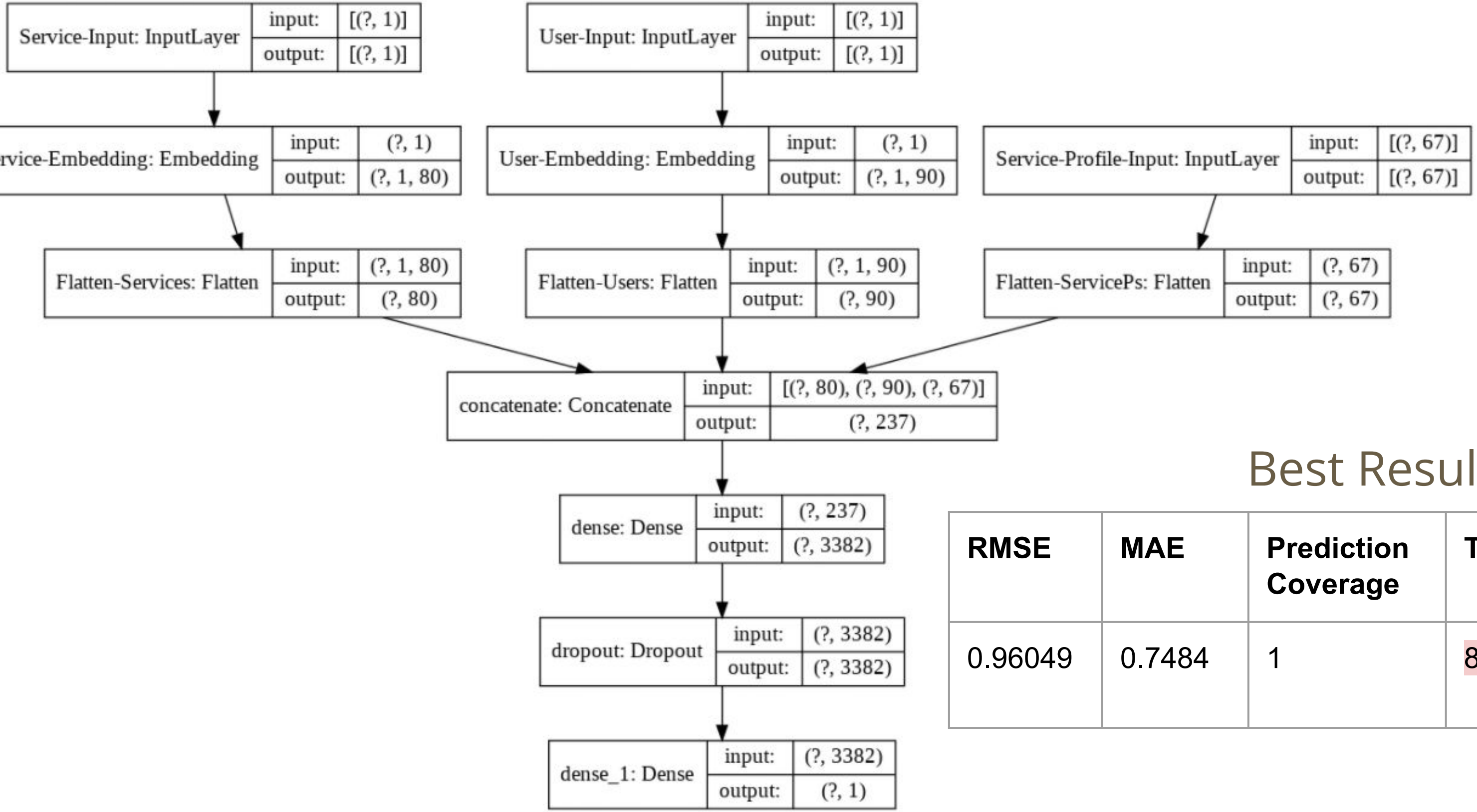
Hyperparameters:

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



https://towardsdatascience.com/shallow-understanding-on-bayesian-optimization-324b6c1f7083

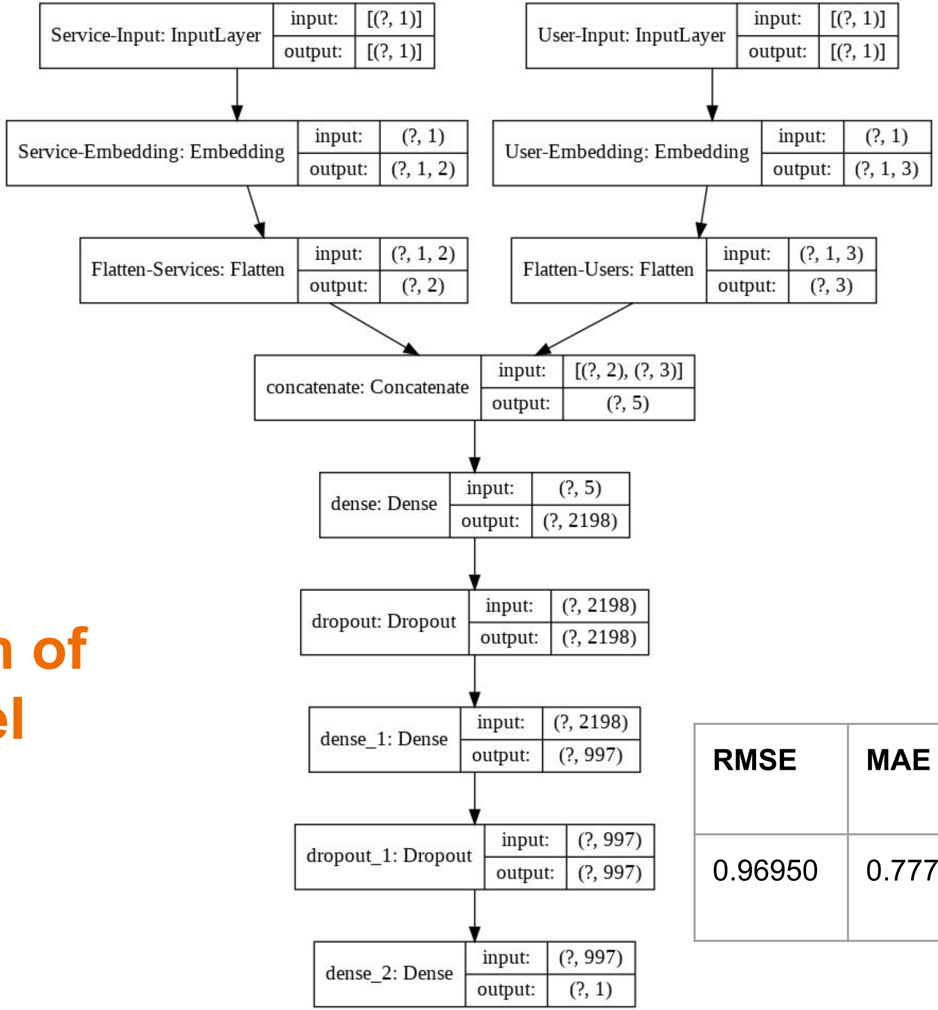
DNN Network Graph Example



Best Results:

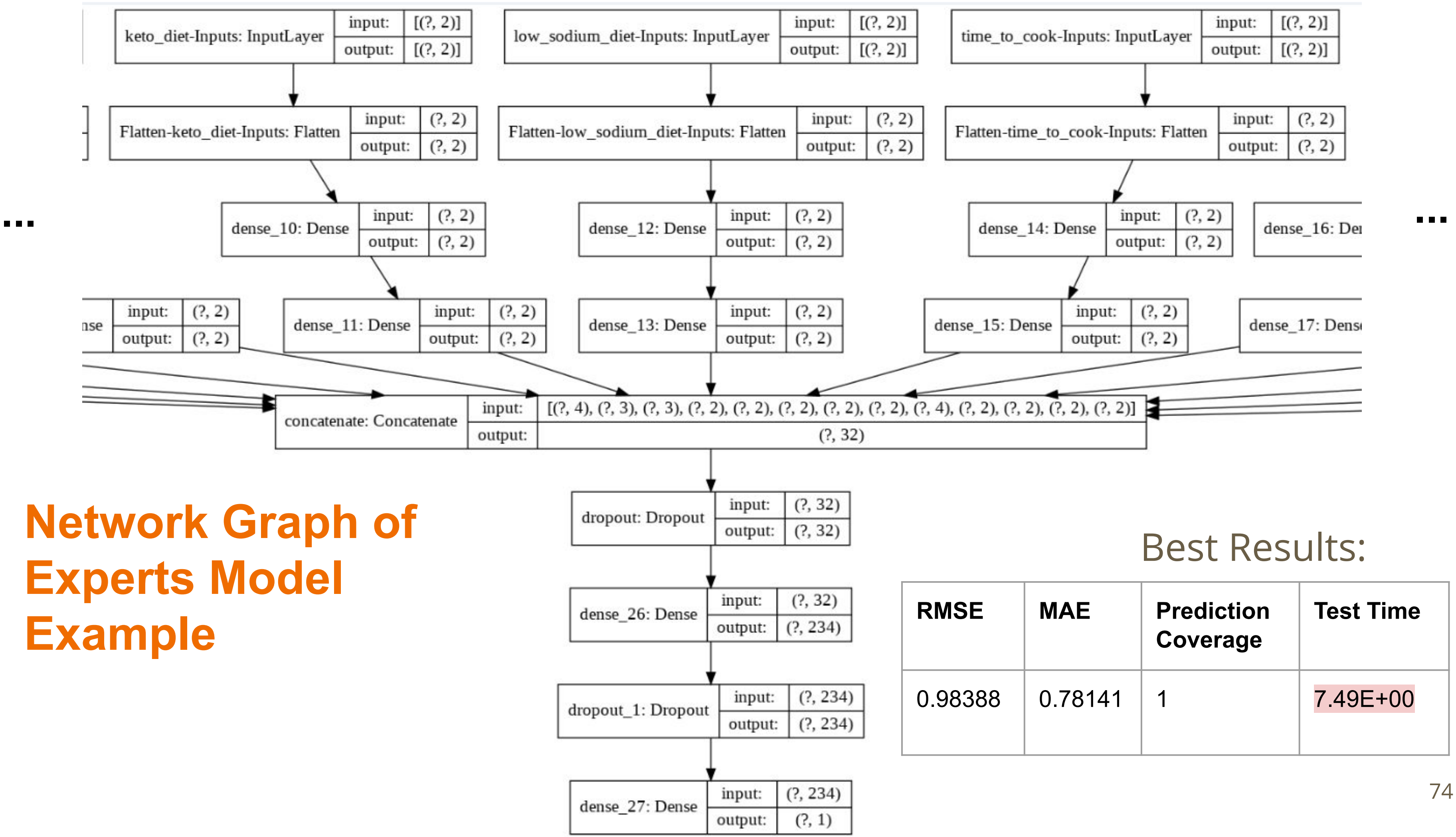
RMSE	**MAE**	**Prediction Coverage**	**Test Time**
0.96049	0.7484	1	8.42E+01

Network Graph of Two FCP Model Example



Best Results:

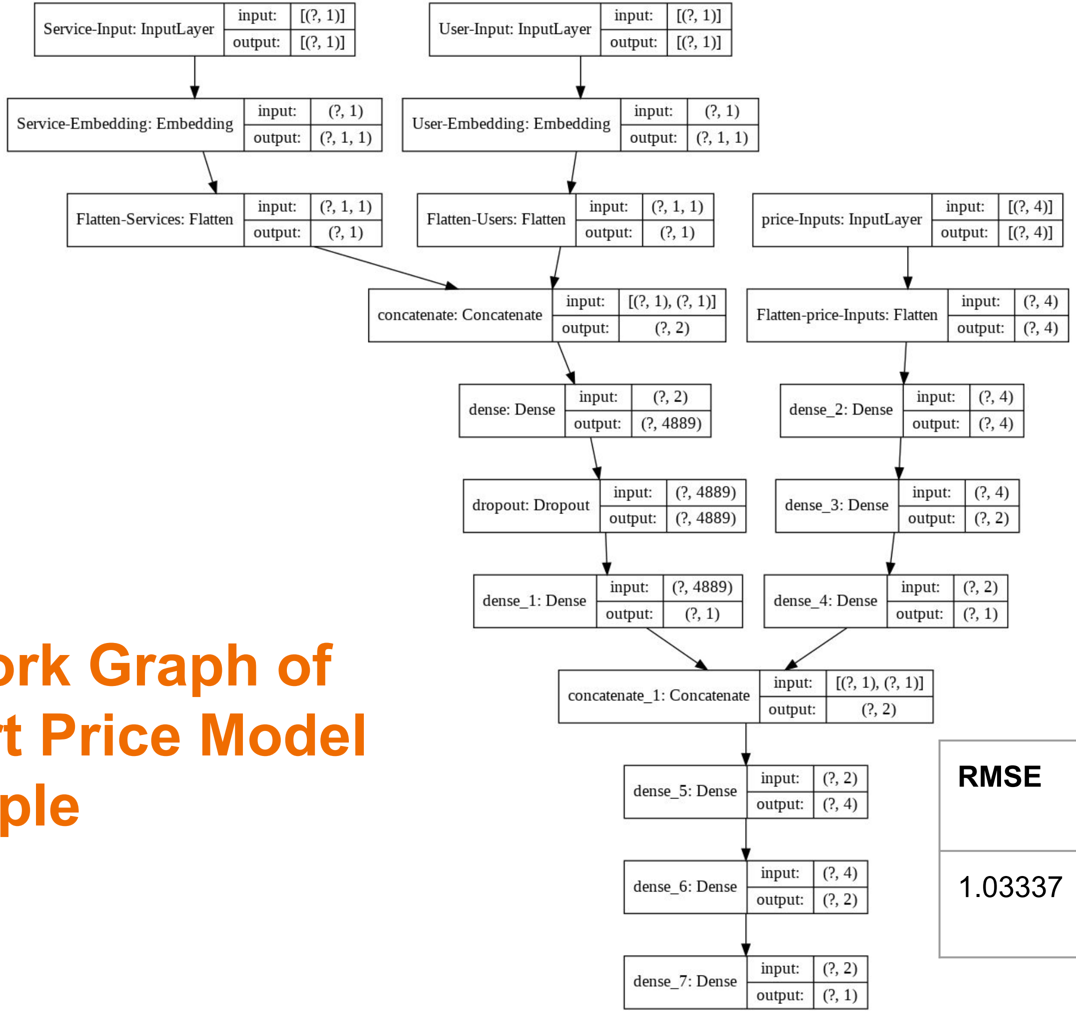
RMSE	MAE	Prediction Coverage	Test Time
0.96950	0.77799	1	1.68E+00



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

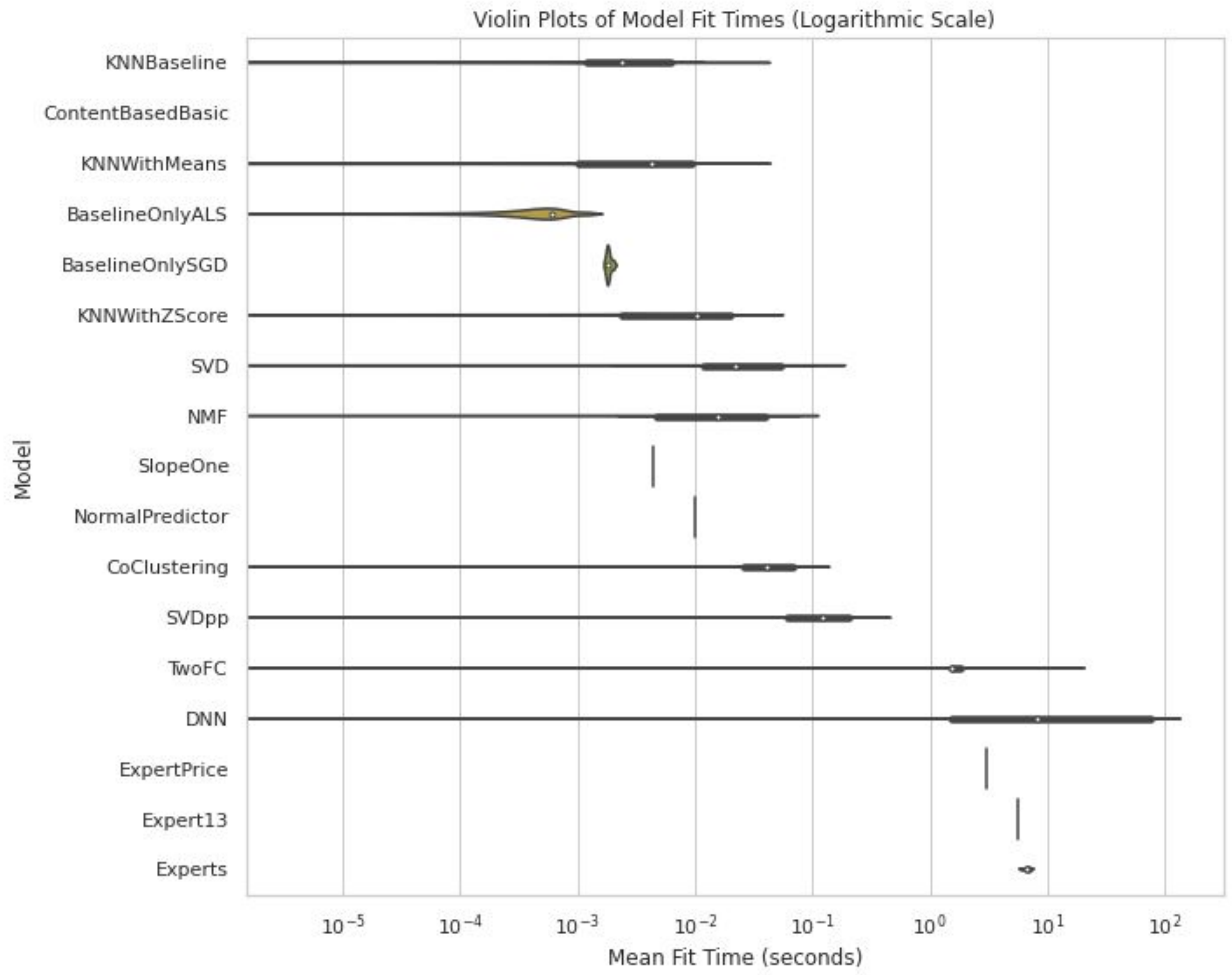
Best Results:

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

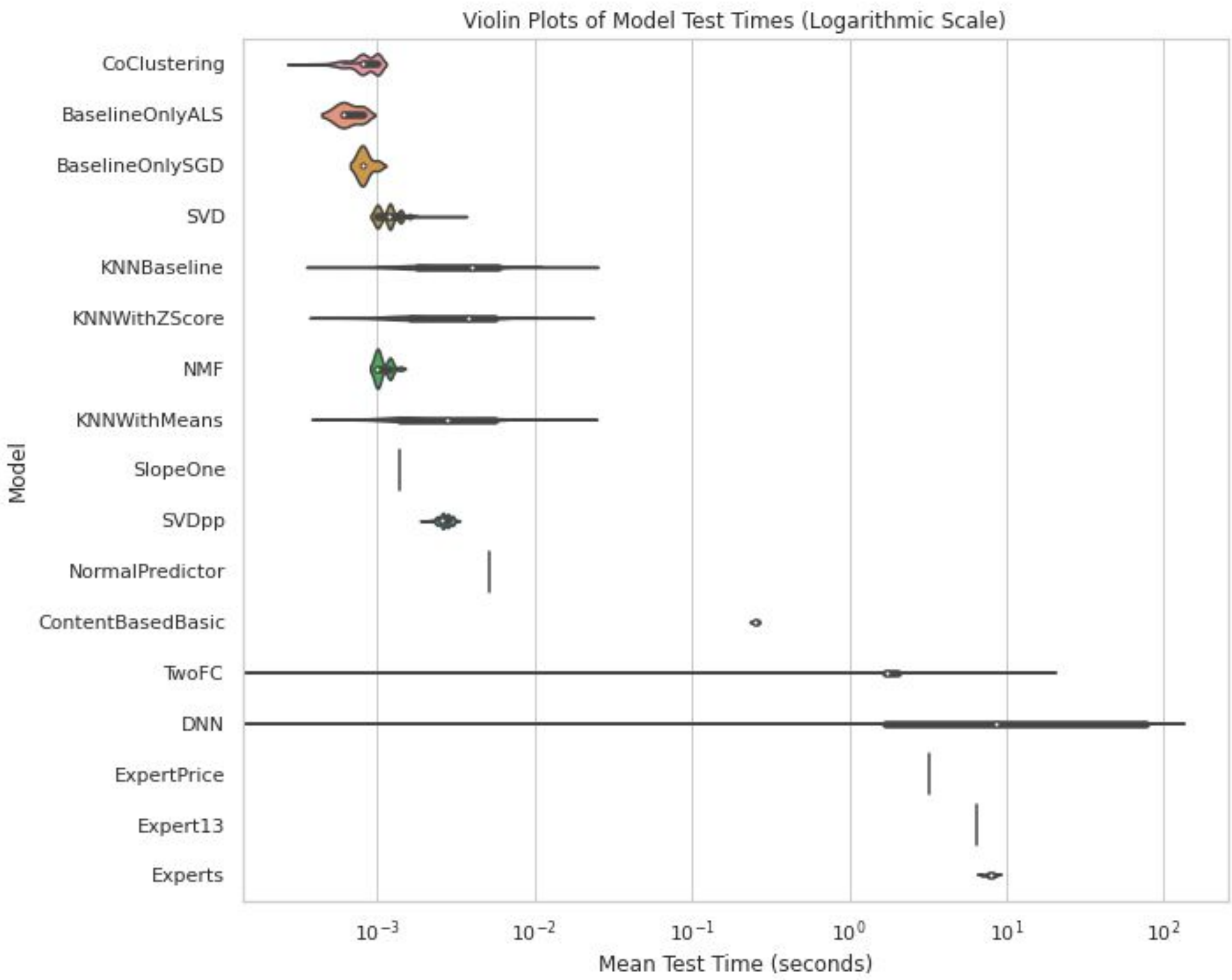
Results

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 Fit Times (Logarithmic Scale)
KNNBaseline
ContentBasedBasic
KNNWithMeans
BaselineOnlyALS
BaselineOnlySGD
KNNWithZScore
SVD
NMF
SlopeOne
NormalPredictor
CoClustering
SVDpp
TwoFC
DNN
ExpertPrice
Expert13
Experts
Model
10^{-5}
10^{-4}
10^{-3}
10^{-2}
10^{-1}
10^{0}
10^{1}
10^{2}
Mean Fit Time (seconds)



Violin Plots of Model Test Times (Logarithmic Scale)
Model
CoClustering
BaselineOnlyALS
BaselineOnlySGD
SVD
KNNBaseline
KNNWithZScore
NMF
KNNWithMeans
SlopeOne
SVDpp
NormalPredictor
ContentBasedBasic
TwoFC
DNN
ExpertPrice
Expert13
Experts
10^{-3}
10^{-2}
10^{-1}
10^{0}
10^{1}
10^{2}
Mean Test Time (seconds)

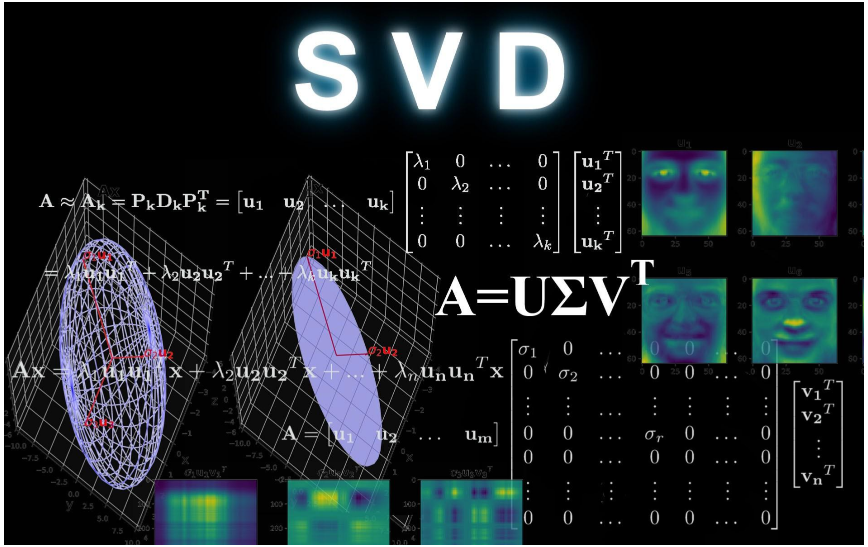
SVD model has overall best performance

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