“Restaurant Pandora”: A Restaurant Recommendation System Based on Yelp Information

Linxuan Shi, Peiying Yu, Yuan Ye

Project Goals:

- Implement user-based collaborative filtering (KNN algorithm) to build a restaurant recommendation system

- Find restaurant attributes that have impact on restaurant’s Yelp rating

Dataset:

- Limitation on Yelp's own API (only provides the first 3 reviews, location information and overall rating)
- Build a Yelp info scraper in Python with Beautifulsoup
- Scraped 36 attributes for 2,500 NYC restaurants
- Scraped Attributes: Hygiene score, Noise Level, Parking,Wi-Fi, Price Range and so on
- Scraped all the reviews for the 90 Morningside heights restaurants (user name and user's rating)

Restaurant Dataset:

```
In [11]:  1 df_final.columns
```

```
Out[11]: Index(['Accepts Bitcoin', 'Accepts Credit Cards', 'Alcohol', 'Ambience',
       'Attire', 'Bike Parking', 'Category', 'Caters', 'Delivery',
       'Dogs Allowed', 'Gender Neutral Restrooms', 'Good For',
       'Good for Groups', 'Good for Kids', 'Happy Hour',
       'Has Dairy-free Options', 'Has Gluten-free Options',
       'Has Soy-free Options', 'Has TV', 'Hygiene_score', 'Liked by Vegans',
       'Liked by Vegetarians', 'Noise Level', 'Outdoor Seating', 'Parking',
       'Source', 'Take-out', 'Takes Reservations', 'Waiter Service',
       'Wheelchair Accessible', 'Wi-Fi', 'price_range', 'restaurant_name',
       'restaurant_neighborhood', 'restaurant_rating',
       'restaurant_reviewcount', 'restaurant_zipcode', 'retaurant_address'],
      dtype='object')
```

Review Dataset:

- Implement a join function in R to merge all scraped review info for restaurants in the morningside heights

- Only keep top 100 individuals who rate the largest number of restaurants

Prediction

- Accept.Credit.Cards: Yes, No
- Hygiene_score: A, B, C
- Noise Level: Average, Loud, Quite
- Parking: Garage, No, Private lot, Street
- Category: American, Chinese, French, Indian, Italian, Japanese, Korean, Mexico
- Waiter Service: Yes, No
- WiFi: Yes, No, Paid
- Price Range: under $10, $11-$30, $31-$60, above $60
- Review Count

Prediction: Simple Regression

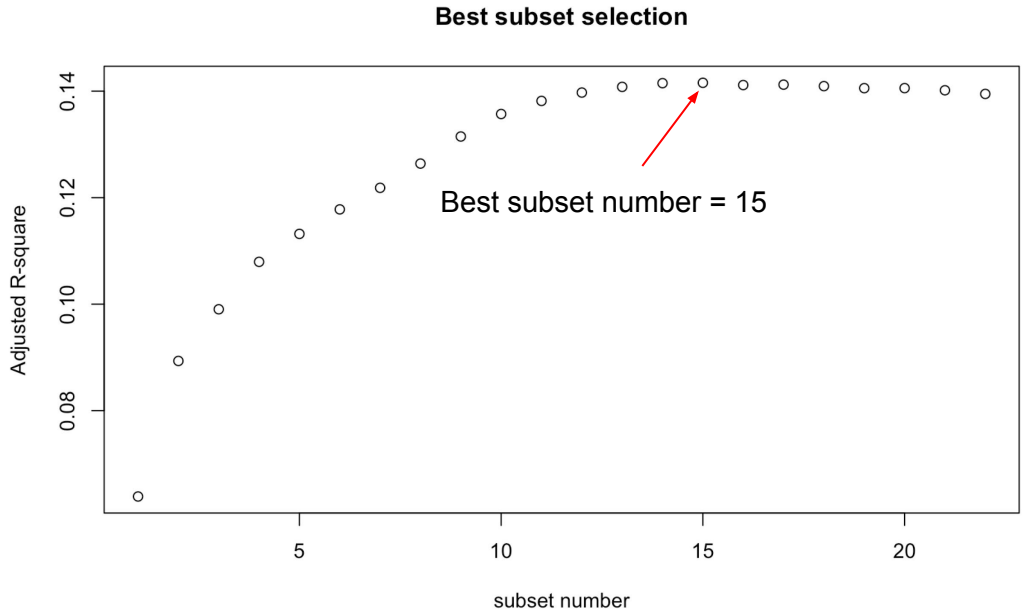
Only the intercept explains something!

All the dummy variables are meaningless!

```
Coefficients:
                                  Estimate Std. Error t value Pr(>|t|)
(Intercept)                      4.183e+00  1.065e-01  39.293  < 2e-16 ***
train_Xrestaurant_reviewcount    6.797e-05  3.782e-05   1.797 0.072534 .
train_XAccepts.Credit.Cards_Yes -1.894e-01  5.446e-02  -3.477 0.000526 ***
train_XHygiene_score_B          -2.513e-01  5.824e-02  -4.316 1.73e-05 ***
train_XHygiene_score_C          -1.104e-01  1.116e-01  -0.989 0.323083
train_XNoise.Level_Loud         -3.842e-01  6.614e-02  -5.809 8.10e-09 ***
train_XNoise.Level_Quite         3.338e-02  5.078e-02   0.657 0.511039
train_XParking_No                9.492e-02  7.119e-02   1.333 0.182711
train_XParking_Private_lot       1.280e-01  1.025e-01   1.249 0.212034
train_XParking_Street            7.896e-02  6.513e-02   1.212 0.225627
train_XChinese                  -3.012e-01  5.675e-02  -5.308 1.33e-07 ***
train_XFrench                   -6.303e-02  6.212e-02  -1.015 0.310529
train_XIndian                    1.945e-02  6.800e-02   0.286 0.774884
train_XItalian                   2.671e-01  1.344e-01   1.987 0.047144 *
train_XJapanese                 -9.112e-02  8.246e-02  -1.105 0.269348
train_XKorean                   -6.866e-02  6.449e-02  -1.065 0.287241
train_XMexico                   -1.934e-01  8.729e-02  -2.216 0.026878 *
train_XWaiter.Service_Yes       -1.836e-01  5.492e-02  -3.343 0.000854 ***
train_XWifi_No                  -7.124e-02  3.515e-02  -2.027 0.042903 *
train_XWifi_Paid                -1.116e+00  3.779e-01  -2.953 0.003212 **
train_X`11-30dollar`             1.551e-01  4.664e-02   3.325 0.000911 ***
train_X`31-60dollar`             2.342e-01  6.765e-02   3.461 0.000557 ***
train_Xabove60dollar             5.089e-01  1.067e-01   4.771 2.07e-06 ***
```

Prediction: Best Subset Selection

Although 15 features subset shows the highest adjusted R-square, it's still only 0.14.



Prediction: Cross Validation Regression

- N-folds: 3
- Train-Validation Data: Restaurants in Manhattan except Morningside Heights
- Test Data: 90 Restaurants in Morningside Heights

```
> mean( (test_y - predict(cv.out, newx = test_X ) )^2 )
[1] 0.2064494
>
```

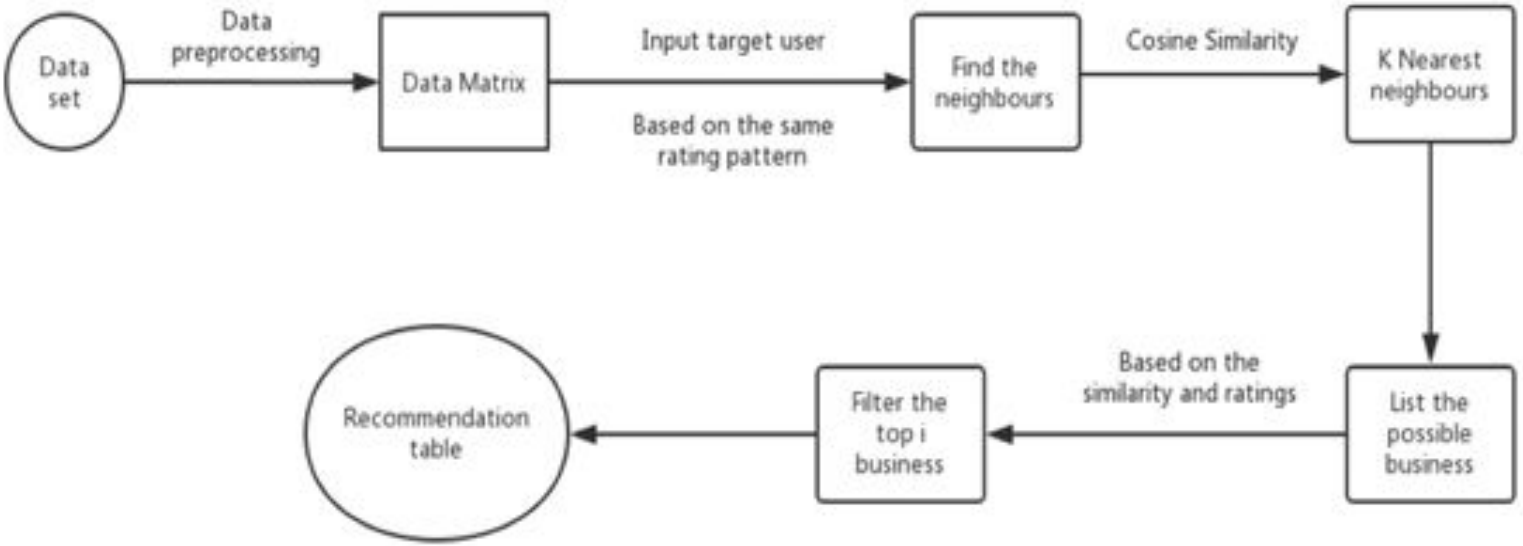
Prediction: Cross Validation Regression

Restaurant	Predicted	Actual	Restaurant	Predicted	Actual	Restaurant	Predicted	Actual
Amelie	3.914973	4.5	Doaba Deli	3.914973	4.5	Kikoo Sushi	3.914973	4.5
Amy Ruth's	3.914973	4	Dun Huang	3.785996	4	Kingston	3.914973	4
Antojitos Mexicano El Taco Taco	3.914973	4	e's BAR	3.914973	4	Kitchenette Uptown	3.914973	3.5
Artopolis	3.914973	3	El Puerto Seafood	3.914973	4	Koko Wings	3.914973	3.5
Atlas Kitchen	3.785996	4	Flat Top	3.914973	4	Koronet Pizza	3.914973	3.5
Awash Ethiopian Restaurant	3.914973	4	Flor de Mayo	3.785996	4	La Diagonal	3.914973	4
Babbalucci	3.914973	4	Friedman's	3.914973	3.5	La Piccola Cucina	3.914973	4
Bar314	3.914973	5	Go! Go! Curry!	3.914973	3.5	La Savane	3.914973	4.5
Belle Harlem	3.914973	5	Grain House	3.785996	4	Le Monde	3.914973	3
Bettolona	3.914973	4	Greek Taverna	3.914973	3.5	Lido	3.914973	4
BLVD Bistro	3.914973	4	Harlem Ale House	3.914973	5	Lolo's Seafood Shack	3.914973	4
Broadway Restaurant	3.914973	4	Harlem Taco & Bowl	3.914973	4.5	Loui Loui	3.785996	4
Buceo 95	3.870583	4	Hex & Company	3.914973	4	Malaysia Grill	3.914973	4
Cantina Taqueria & Tequila Bar	3.914973	4	Hula Poke	3.914973	4	MAMA's TOO!	3.914973	4
Carmine's Italian Restaurant	3.870583	4	Infamous Chicken	3.914973	4	Manna Korean Food	3.914973	5
Casa Mexicana	3.914973	4	Isola on Columbus	3.914973	4	Marlow Bistro	3.914973	4
Chapati House	3.914973	4	Jin Ramen	3.914973	4	Massawa	3.914973	4
Community Food & Juice	3.914973	3.5	JJ's Place	3.870583	5	Max Caffe	3.914973	3.5
Dig Inn	3.914973	4	Junzi Kitchen	3.785996	4	Max Soha	3.870583	4
Dive 106	3.914973	4.5	KALBI	3.914973	4.5	Mel's Burger Bar	3.914973	3.5

Prediction Result

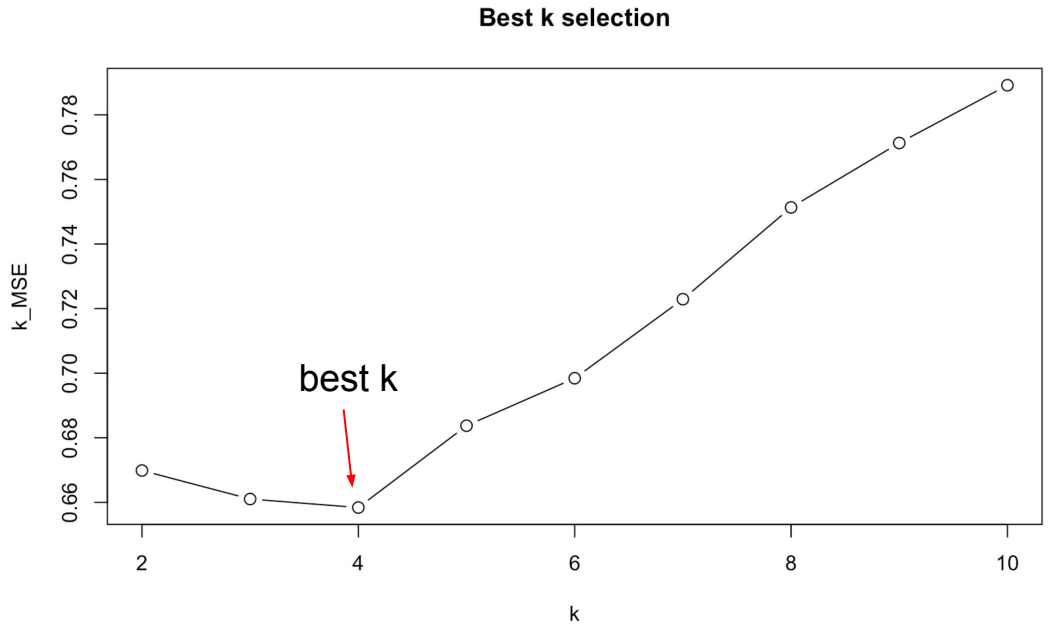
- No matter what regression method we choose, the variables “Accept.Credit.Cards”, “Hygiene_score”, “Noise Level”, “Parking”, “Category”, “Waiter Service”, “WiFi”, “Price Range”, “Review Count” cannot explain a majority of restaurants rating.
- What we missed?
- The quality of food!

Recommendation Methodology:



Recommendation System:

- Model selection for KNN (choose best k):
  we tried k = 2:10 and found k = 4 gives the smallest MSE (0.658)
- Use 4-NN as our recommendation system



Recommendation Result:

```
> head(sort(preds,decreasing = TRUE),10)
                Belle.Harlem                     Dive.106
                        5.00                         5.00
          harlem.ale.house                  hex_company
                        5.00                         5.00
                kikoo.sushi                   La.Savane
                        5.00                         4.75
             Marlow.Bistro                     Melba.s
                        4.75                         4.75
          Milano.Market Nous.Espresso.Grad.School.Cafe
                        4.75                         4.75
```

Potential Application:

- Expand to entire Manhattan
- Other major metropolitan areas
- May not function in less populated areas

Q&A