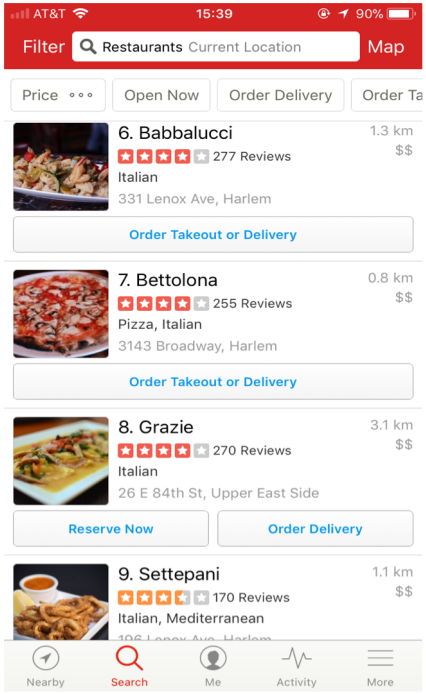
# Yelp Rating Prediction


| | |
|---|---|
| Minmin Zhu | mz2656@columbia.edu |
| Yaxin Wang | yw3042@columbia.edu |
| Tianze Yue | ty2369@columbia.edu |
| Zonghao Li | zl2613@columbia.edu |
| Jiaxi Wu | jw3588@columbia.edu |

# 1 Executive Summary

Presented by: Minmin Zhu (mz2656)

# Thesis




**Apply machine learning method to predict user ratings for each restaurant**

Recommend restaurants with high rating prediction

# Data Overview

## Data Source

- Yelp Dataset Challenge

## Data Process

- Select Las Vegas as target city ，focus on restaurants located in the city
- Select core users with more than 500 reviews

## Data Structure

- Predict rating of “?” in data matrix

| Rating | Restaurant 001 | Restaurant 002 | Restaurant 003 | Restaurant 004 |
|---|---|---|---|---|
| User 001 | 4 | ? | ? | 5 |
| User 002 | 4.5 | 3.5 | 4.5 | ? |
| User 003 | ? | 2.5 | ? | ? |
| User 004 | ? | 4 | ? | 4 |
| User 005 | 3.5 | 3 | ? | 4.5 |
| User 006 | ? | ? | 4 | 5 |

# Model Overview

## Baseline

- $\frac{1}{2}(\text{average user rating} + \text{average restaurant rating})$

## Machine Learning Model

- Collaborative Filtering Model

  Cosine similarity

- Text Based Regression Model

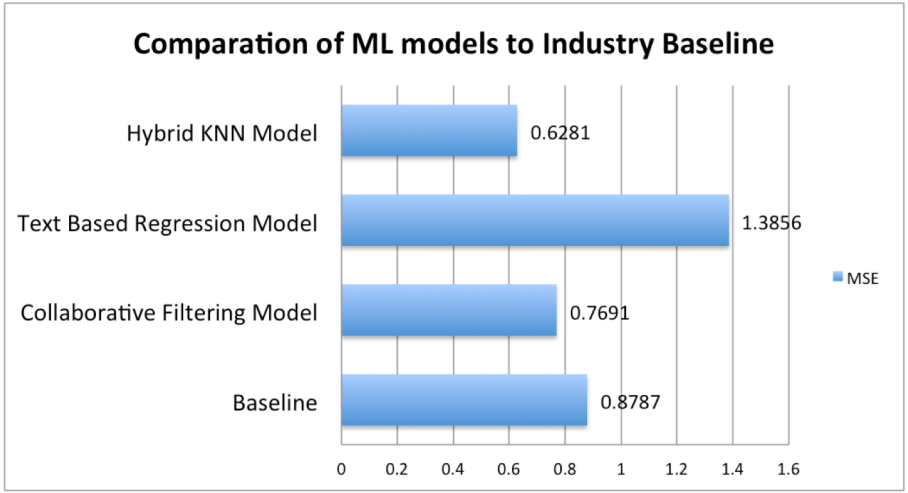
  TF-IDF(term frequency–inverse document frequency) + Cosine similarity + OLS

- Hybrid KNN Model

  TF-IDF + Cosine similarity + KNN(K-nearest neighbors) + Cross-validation

# ML Statistical Value-add

- Collaborative Filtering Model and Hybrid KNN Model outperform industry baseline
- Hybrid KNN Model has lowest MSE
- Make more accurate user ratings prediction



1

# Business Understanding

# Business Goal

**Business Need**

Find the Restaurants

User Preference

**Predict Restaurants' Ratings**

**Recommender System**

Best match restaurants with users

**Machine Learning**

Support vector machines

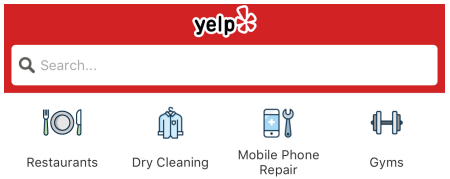
**Models**

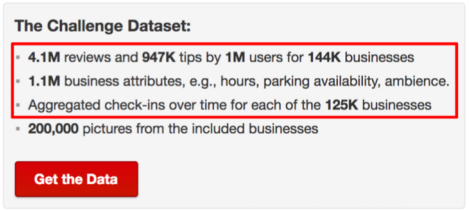
- Text based
- Collaborative Filtering
- Hybrid

Yelpers' Data

**Big Data**

Yelp Dataset



# Business reference baseline

## WHAT INDUSTRY DOES

For the baseline recommendation system, we first compute the average rating of each user and each restaurant.

$\bar{r}_u^U$: the average rating user has given

$\bar{r}_b^B$: the average rating the restaurants has received

**Prediction**: Average over the average rating of u and b, $\hat{r}_{ub} = (\bar{r}_u^U + \bar{r}_b^B)\ /2$

## WHAT WE DO

Machine Learning Techniques

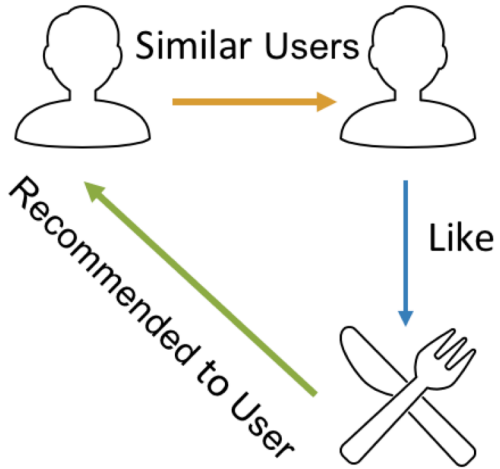
- Text-based Filtering
- Collaborative Filtering
- Hybrid

# ML statistical value-add

## Collaborative Filtering

Predict using User to User Similarity

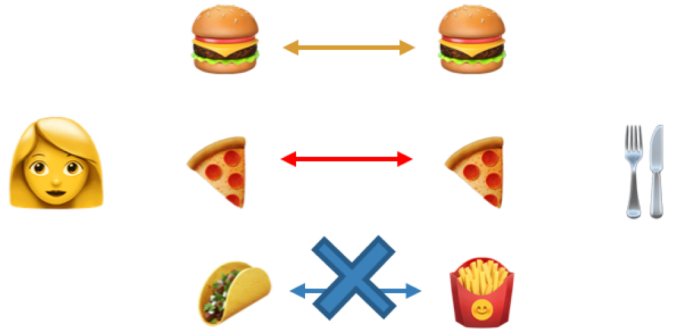
**Assumption**: Similar Users have similar preference for restaurants.




## Text Based Filtering

Match features of User with features of Restaurant

- Extract features from reviews written by a specific user
- Get how often these features were mentioned from the reviews for a specific business.
- By comparing the frequency of these words, we could match up customers and businesses by similarities.



# ML business value-add

**Baseline**

Yelp tends to recommend restaurants to users based on Distance, Prices and other single factors.

**User-based Prediction Method**

- Predict the rating of a restaurant given by a specific user based on User's Preference.
- Build models using Machine Learning Techniques.
- Optimize models to make accurate prediction.

**More Customized Recommendation**

- Sort restaurants by ratings given by user customized rating criteria.
- Recommend restaurants with highest ratings to user.

3

# Data Understanding

# Independent, Dependent and Control Variables

5,200,000 user reviews, information on 174,000 businesses

- Dependent Variable: New rating (rating of business that user never been to)
- Independent Variables: Existing rating and review text
- Filtering Variables: city, business_review_count, user_review_count, categories

Why do not contains user features age and gender:

- Not enough user features in dataset
- Not predictive for different business

# Independent and Dependent Variables

| | A | B | C | D | E | F | G | H |
|---|---|---|---|---|---|---|---|---|
| 1 | user_id | business_id | stars | city | business_review_count | categories | user_review_count | text |
| 2 | uEvusDwoSymbJJ0auR3muQ | pH0BLkL4cbxKzu471VZnuA | 3 | Ahwatuke | 22 | Dentists;General De | 23 | I live very near |
| 3 | Dgx185t0xrXT8nboVCNBqA | U9aA5H13y7t9xWnoQslV0Q | 4 | McMurray | 11 | Hair Stylists;Hair Sal | 20 | We've been wanting to |
| 4 | WPl2gULxrrh_GoypMzHF-A | yfxDa8RFOvJPQh0rNtakHA | 3 | Phoenix | 18 | Departments of Mo | 44 | I've eaten here a couple |
| 5 | ugk0g6vyv2XavpoTaJSy6A | aT_SsfZ6GQgJGyulv1Hapw | 5 | Tempe | 9 | Sporting Goods;Sho | 2 | Some of my girlfriends |
| 6 | Kp5KHBSmV-Htc9NeSjJfPA | 44kd3YdkhXj5XiSPs5XNjQ | 4 | Cuyahoga | 116 | American (New);Nig | 31 | I work here at the hotel |
| 7 | awQhT121Pe0R33sukfTyuA | ahSFUPojs9X3-1jP-QPb-w | 5 | Stuttgart | 5 | Italian;Restaurants | 15 | I have only three words f |
| 8 | YE54kKTuqJJPNYWlKlpOEQ | 364hhL5st0LV16UcBHRJ3A | 3 | Las Vegas | 5 | Real Estate Services; | 48 | This place is an OK |
| 9 | 8ZryN_S-n48g6rsa3W3QtQ | xVEtGucSRLk5pxxN0t4i6g | 3 | Las Vegas | 9 | Shopping;Sporting ( | 12 | I am not a seafood fan. |
| 10 | ZmWLeLU_bGrNiqBVAGo-eg | rioQ_p2pILNbJ4Xp5jW6-Q | 5 | Wexford | 15 | Coffee & Tea;Ice Cre | 6 | Great view of the strip fr |
| 11 | GF-UBlwA0gEcUbAkve6s3A | MqYYYNA-ZYvV-1w5qcmMoA | 4 | Henderso | 7 | Automotive;Auto De | 41 | On paper this is a |
| 12 | 2nE0zU6y_F7gkwHi3yL6cQ | cYwJA2A6l12KNkm2rtXd5g | 4 | Houston | 3 | Breakfast & Brunch;( | 20 | The food was great. Fro |
| 13 | YdFcEQaOCef-ojBAJ7Mxjg | QJR4qBUHegWEozSQrGmBPw | 5 | Chandler | 23 | Local Services;Self St | 47 | The food is good. It is be |
| 14 | ZIlZhI7xT9gztKRIWk8Uvg | YnmtUJGqQlQL9tz1MRyfqA | 4 | Markham | 38 | | | |
| 15 | bLbSNkLggFnqwNNzzq-Ijw | Fi-2ruy5x600SX4avnrFuA | 3 | Homestea | 5 | | | |
| 16 | 8Aq_UdlsrjhBwGWb_U-xRA | LtXy1VinKWfuLFslVarKRw | 2 | Charlotte | 7 | | | |
| 17 | 4ljzfbO7XKFNWWVufHNhww | eoHdUeQDNgQ6WYEnP2aiRw | 2 | Toronto | 12 | | | |
| 18 | YMgZqBUAddmFErxLtCfK_w | GHS1rVjO-RMcRB6WJLpCDQ | 4 | Peoria | 20 | | | |
| 19 | 0zDHiDbj79uBOy3OAcE25A | g8OnV26ywJlZpezdBnOWUQ | 3 | Sun Prairi | 9 | | | |
| 20 | 25NDlYy0wzjq7mh3liCp0Q | k8BqCejnaMlw7aoVthvqyw | 5 | Goodyear | 65 | | | |

## Metadata

| Variable | Description | DataType |
|---|---|---|
| user_id | user's ID | String |
| business_id | business's ID | String |
| ratings | star of business given by user | Numeric |
| city | city of business | String |
| business_review_count | total number of reviews of business | Numeric |
| categories | category tags of business | String |
| user_review_count | total number of reviews of user | Numeric |
| text | review text of business given by user | String |

## Data Preparation Techniques

1. Select the business with three conditions: its city label is “Las Vegas”, its category contains several key words such as “restaurant”, “food” and its review count is larger than 500
2. Select the user whose review count is larger than 500
3. Extract the review and rating data of by business and user lists

The first three steps greatly reduce sparsity

4. Construct a data matrix with the business ID list as column name and user ID list as row names.
5. It contains 420 columns and 3730 rows.

# Dataset Quality

## Matrix Sparsity: 98% to 72%

If not, we need to use 2% data to predict 98%, which will lead to poor prediction result. In this way, we can reduce around 25% MSE.

| user_ID | business_ID | rating |
|---|---|---|
| User 001 | Restaurant 001 | 4 |
| User 001 | Restaurant 004 | 5 |
| User 002 | Restaurant 001 | 4.5 |
| User 002 | Restaurant 002 | 3.5 |
| User 002 | Restaurant 003 | 4.5 |
| User 003 | Restaurant 002 | 2 |
| User 003 | Restaurant 005 | 1 |

| Rating | Restaurant 001 | Restaurant 002 | Restaurant 003 | Restaurant 004 | Restaurant 005 |
|---|---|---|---|---|---|
| User 001 | 4 | NA | NA | 5 | NA |
| User 002 | 4.5 | 3.5 | 4.5 | NA | 2 |
| User 003 | NA | 2.5 | NA | NA | 1 |
| User 004 | NA | 4 | NA | 4 | NA |
| User 005 | 3.5 | 3 | NA | 4.5 | NA |
| User 006 | NA | NA | 4 | 5 | 3 |

## Sampling / Other Key Tasks

- Divide the review data set into 25% hold-out set and 75% training set.
- Text data processing: Feature extraction

Converting text data into a matrix of TF-IDF (term frequency–inverse document frequency) features. These features make it possible to implement new machine learning techniques.

| TF-IDF | hot | seafood | ramen | cheap | pasta |
|---|---|---|---|---|---|
| User 001 | 22.67 | 7.83 | 13.03 | 26.72 | 2.95 |
| User 002 | 35.80 | 20.73 | 30.25 | 44.00 | 49.39 |
| User 003 | 44.16 | 3.69 | 29.03 | 22.82 | 30.02 |
| User 004 | 26.37 | 39.76 | 29.59 | 0.95 | 44.94 |
| User 005 | 42.66 | 19.41 | 40.48 | 34.01 | 9.26 |

| TF-IDF | hot | seafood | ramen | cheap | pasta |
|---|---|---|---|---|---|
| Restaurant 001 | 47.66 | 45.37 | 29.94 | 10.07 | 3.97 |
| Restaurant 002 | 10.59 | 5.67 | 31.92 | 23.22 | 4.37 |
| Restaurant 003 | 29.00 | 44.80 | 5.52 | 27.95 | 37.73 |
| Restaurant 004 | 16.08 | 3.16 | 15.78 | 14.79 | 25.84 |
| Restaurant 005 | 40.82 | 8.59 | 36.74 | 38.26 | 4.55 |

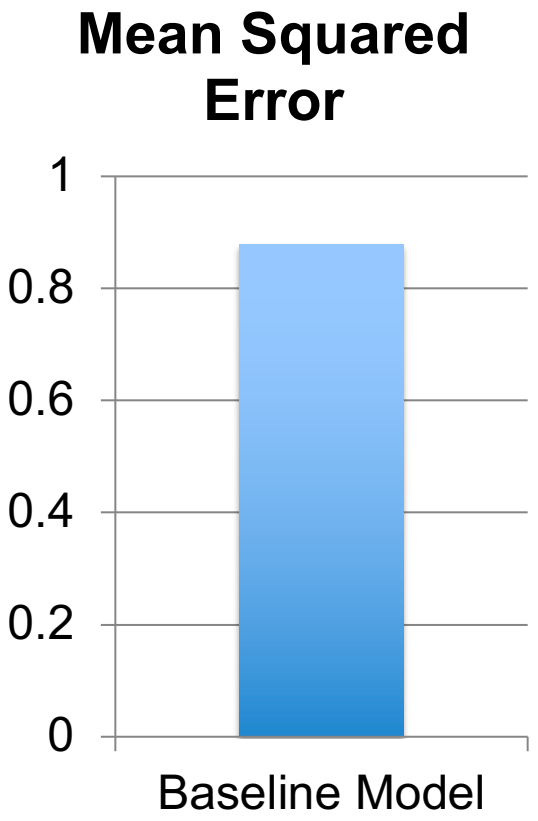
4

# Data Modeling

# Data Modeling

**Baseline Model: Average of user rating and business rating**

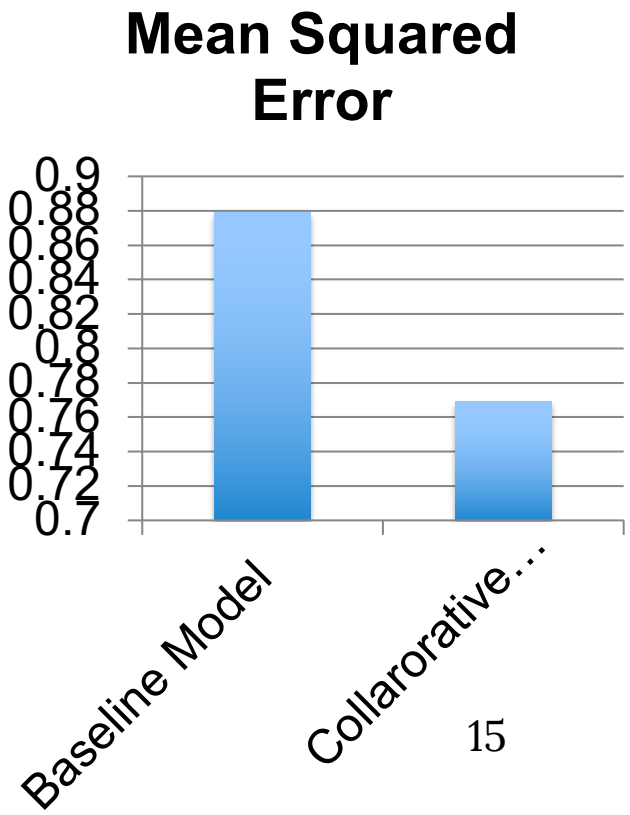
| Rating | Restaurant 001 | Restaurant 002 | Restaurant 003 | Restaurant 004 |
|---|---|---|---|---|
| User 001 | 4 | 5 | X | 3 |
| User 002 | 4 | 5 | 2 | |
| User 003 | 3 | 3.5 | | |
| User 004 | 2 | | 4 | 4 |



# Data Modeling

**Collaborative Filtering Model:**
**Calculate similarities between users and**

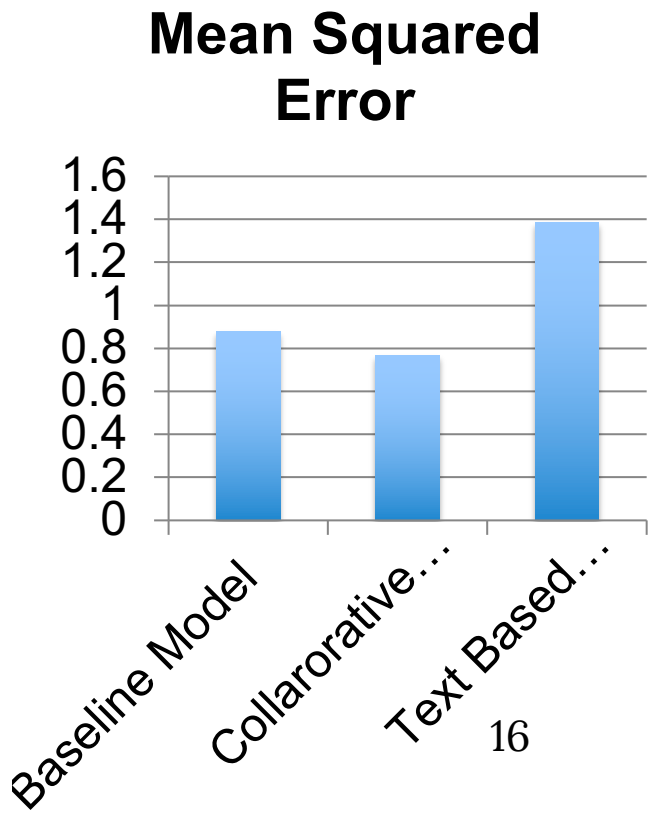
| Rating | Restaurant 001 | Restaurant 002 | Restaurant 003 | Restaurant 004 |
|---|---|---|---|---|
| User 001 | 4 | 5 | X | 3 |
| User 002 | 4 | 5 | 2 | |
| User 003 | 3 | 3.5 | | |
| User 004 | 2 | | 4 | 4 |



# Data Modeling

**Text Based Regression Model: Extract features from reviews and calculate similarity**

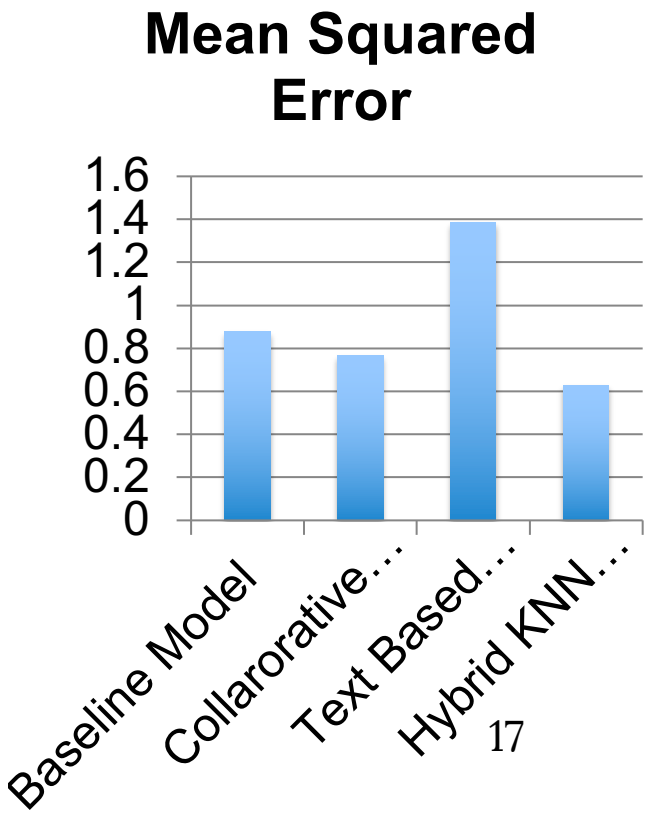
| TF-IDF | hot | seafood | ramen | cheap | pasta |
|---|---|---|---|---|---|
| User 001 | 22.67 | 7.83 | 13.03 | 26.72 | 2.95 |
| User 002 | 35.80 | 20.73 | 30.25 | 44.00 | 49.39 |
| User 003 | 44.16 | 3.69 | 29.03 | 22.82 | 30.02 |
| User 004 | 26.37 | 39.76 | 29.59 | 0.95 | 44.94 |
| User 005 | 42.66 | 19.41 | 40.48 | 34.01 | 9.26 |



# Data Modeling

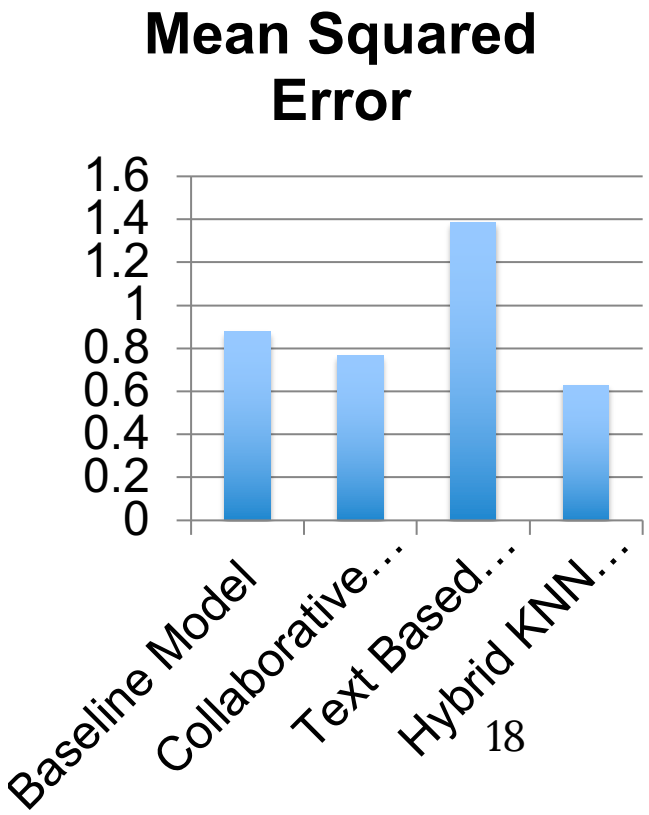
**Hybrid KNN Model: Combine features extracted from TFIDF matrix with KNN method**

- Comparing to Collaborative Filtering: similarity matrix is less sparse
- Comparing to Text Based Model: more available data for prediction



# Value Add

| | Baseline | Collaborative | Text | KNN |
|---|---|---|---|---|
| MSE | 0.8787 | 0.7691 | 1.3856 | 0.6281 |



# 5 Technical Appendix

## Variable and Parameter Selection

- Variable/Feature selection in text analysis

Delect function words such as “a, the, we, and”

Use meaningful words as features such as “seafood, hot, cheap, parking space”

- KNN parameter

Use 10 folds cross-validation

## Collaborative Filtering

- Step 1: Compute cosine similarity for user i and user j.

$$S_{ij}^{U} = \text{sim}\,(X_i^U, X_j^U) = \frac{(X_i^U)^{\mathrm{T}} X_j^U}{\|X_i^U\|_2 \|X_j^U\|_2}$$

- Step 2: Predict rating by weighted-average of the ratings from all users.

$$\hat{r}_{ub} = \bar{r}_u + \frac{\sum_i S_{ui}^U \,(X_{ib} - \bar{r}_i^U)}{\sum_i |S_{ui}^U|}$$

## Text-based Model

- Step 1: Concatenate all the reviews written by a specific user and all the reviews for a specific business.
- Step 2: Extract user and restaurant features by creating TF-IDF matrix.
- Step 3: Calculate the user-business cosine similarities.
- Step 4: Fit a simple linear regression model of similarities depending on available ratings.
- Step 5: Predict missing ratings based on similarities.

## Hybrid KNN Model

- Step 1: Extract user features by creating TF-IDF matrix.
- Step 2: Calculate the cosine similarities for user i and j and used as correlation distance
- Step 3: Find Cluster $N_K(x)$ which contains the nearest $K$ users of user $x$
- Step 4: Calculate mean rating within cluster as prediction

$$\hat{f}(x) = \frac{1}{K} \sum_{i \in N_K(x)} y_i$$