# Recommender Systems for Yelp Users

Srikanth Mandru
mandru.s@husky.neu.edu

## Problem Definition

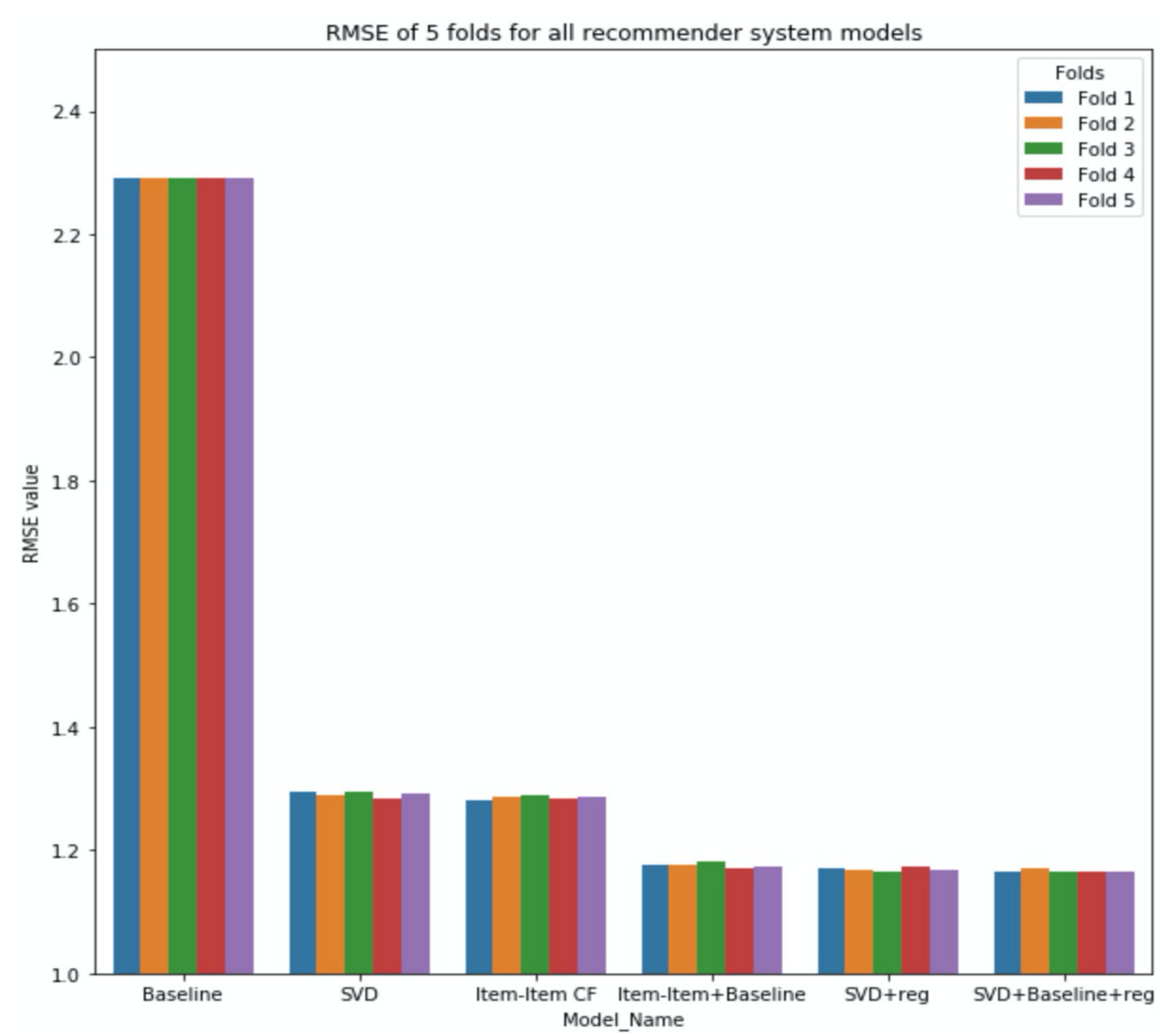
- Given the ratings for businesses by yelp users (known ratings), predict the probable ratings (unknown ratings)
- Goal is to build a recommender system with the following properties :
  - Robust - Resilient to variances in data
  - High prediction accuracy - Estimate ratings likely to be given by the user

## Existing Methods

- Content-Based recommendation system
    - Recommendations based on item and user profiles manually selecting features
- Collaborative filtering (CF)
  - Item-Item CF
    - Extrapolate item rating based on ratings of similar items.
  - User-User CF
    - Identifying similar users and recommending what similar users like
- Latent Factor Models
  - Singular Value Decomposition (SVD)
    - Represent users and items using latent (hidden) factors

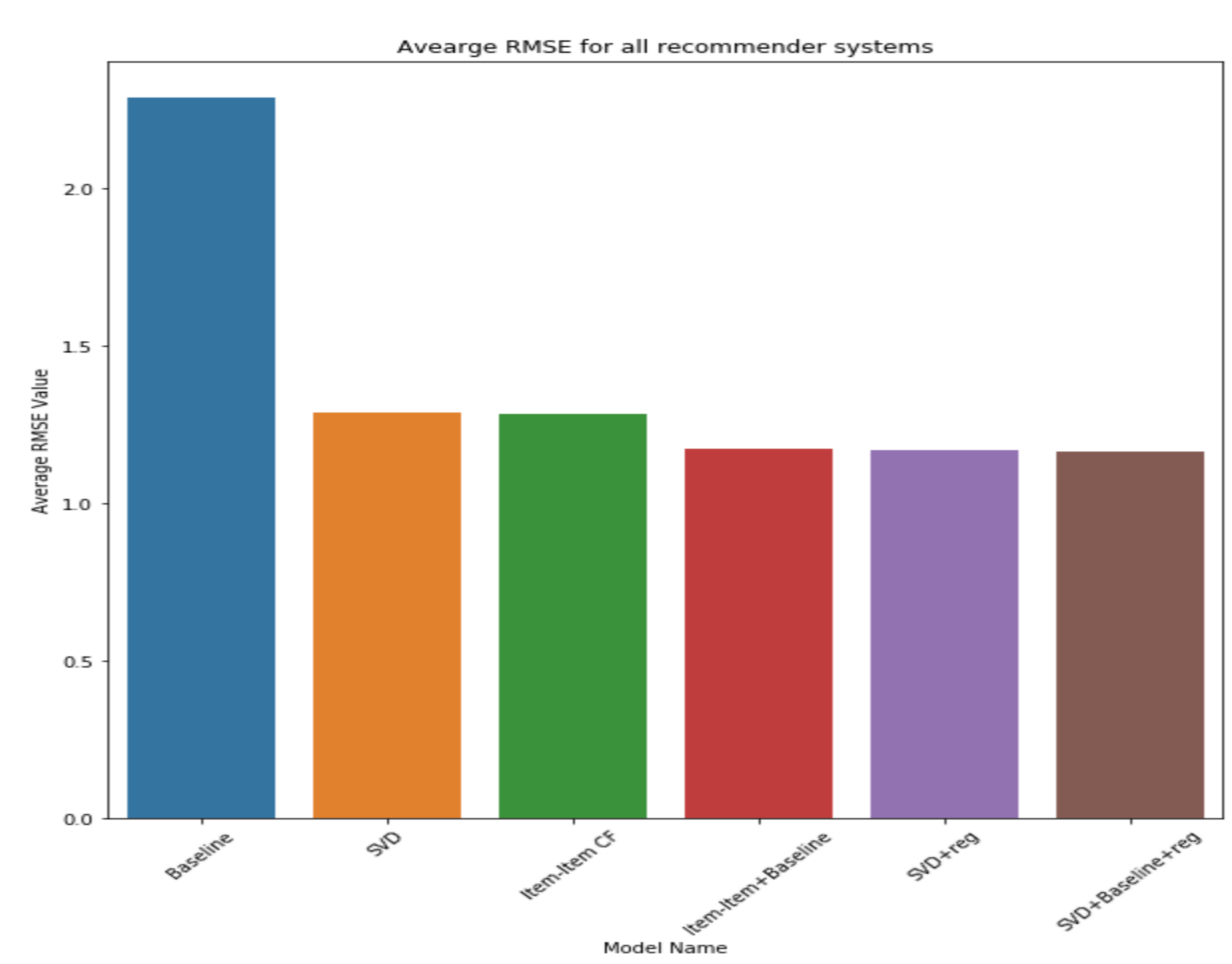
## Proposed Method

- Build Recommender systems using the different existing methods (stated below) and compare the models based on the average of RMSE values computed through k-fold cross-validation
- Methods implemented:
  - Baseline approach
  - Item–Item collaborative filtering
    - Similarity metric used – Pearson correlation coefficient
  - SVD
  - SVD with Regularization
  - Hybrid methods:
    - Baseline and Item Item collaborative filtering
    - Baseline and SVD with Regularization



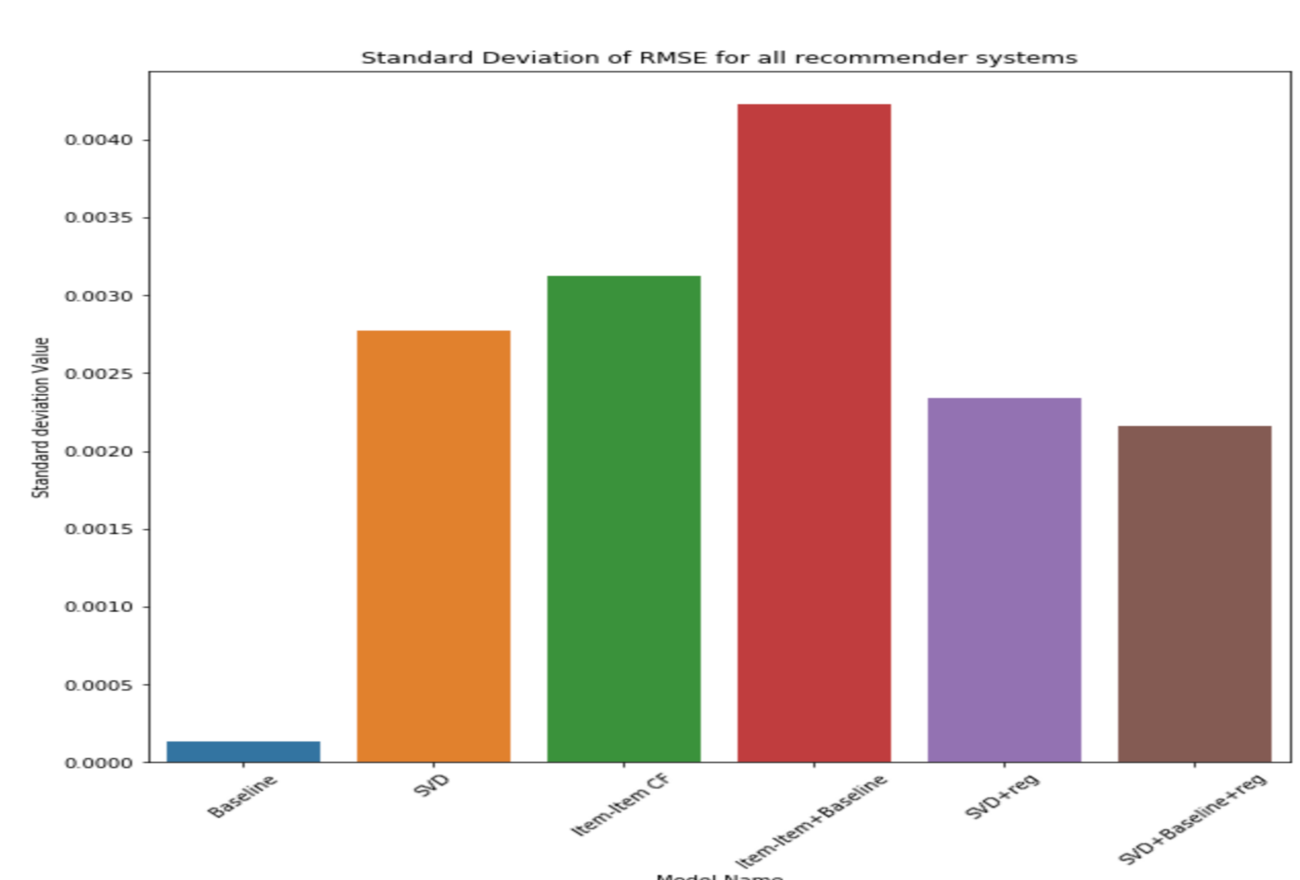

## Data Description & Experimental Setup

- Data (after filtering) used for building models comprises of 200,000 ratings given by 31,824 users to 2061 food-related businesses
- **Handling Cold Start problem**:
  - Filtered data with constraints :
    - Business rated by minimum 20 users
    - Users with a minimum of 2 ratings
- **Experimental setup to obtain RMSE**:
  1. Divide data into k-folds
  2. For each model do
     i. Repeat k times
        - Run model on k-1 folds and compute RMSE on hold out fold
  3. Report the RMSE of each fold for all models




## Results and Discussion

- **Model efficiency**: Across all models,
  - Non-hybrid models do not have enough parameters to capture the patterns in data which resulted in high error
  - Hybrid methods performed well showing low errors on all folds
- **Robustness** :
  - Models with regularization gave better results with less deviations in RMSE values




## Takeaway Points & Future Work

- Use combinations of different methods
- Regularization prevents overfitting
- **Future Work**: Extend to incorporate the seasonal biases