# Yelp Recommendation System


Jason Ting, Swaroop Indra Ramaswamy


## Background and Motivation

### The Problem

Yelp is a web/mobile application that publishes crowd-sourced reviews about local businesses and restaurants. The rise of Yelp's popularity created an influx of data on people's personal preferences as modern customers to the businesses that they go to. Through this project we utilized Yelp's data to make personalized business recommendations for Yelp users by making a model to predict the number of review stars that a user would assign to a business.

### Data

The dataset comes from the Yelp recommendation Kaggle competition. This information contains actual business, user, and users' review data from the greater Phoenix, AZ metropolitan area. By using and combining various data fields, we can aggregate similar users to create models to predict how users will rate businesses they have not been to.

### Evaluation Metric

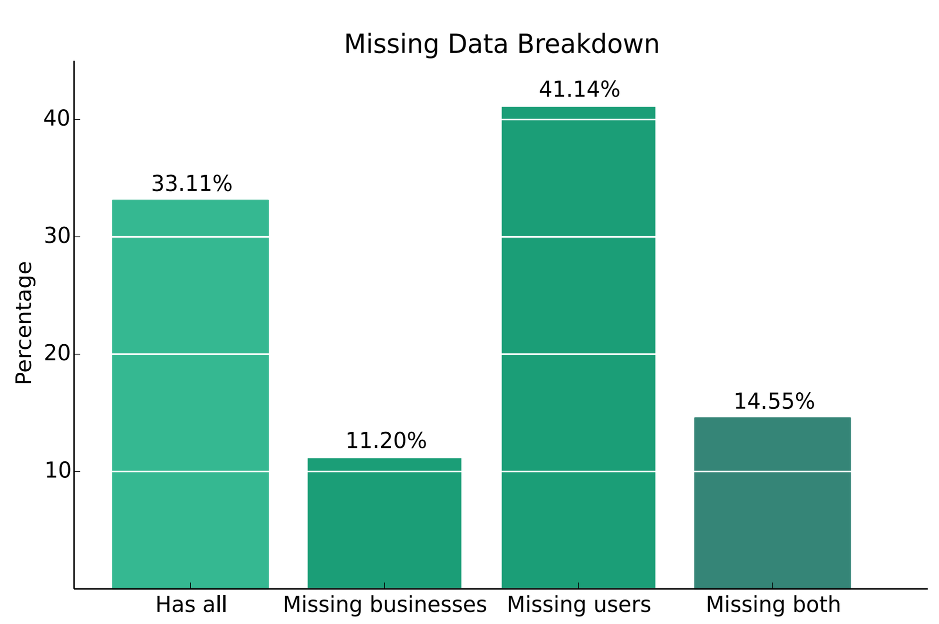
We chose to evaluate our model through the root mean squared error (RMSE) to measure the accuracy, where n is the total number of reviews to predict, p is the predicted rating, and a is the actual rating.

$$RMSE = \sqrt{\frac{\sum_{i=1}^{n}(p_i - a_i)^2}{n}}$$

### Missing Data

A significant portion of the data is missing in the test set, so we used simple imputation and replaced the missing data with the following:

- The mean of the training set.
- A random sample from the training set.
- Predicted regression values from using the other features.




## Machine Learning

### Features

Business open, business stars, business review count, user review count, user average, number of cool votes, number of useful votes, number of funny votes. Computed features: gender, category averages, franchise averages, and interaction terms

### Models

- Linear Regression- fits a linear model by solving the following optimization problem to find the estimated parameters.

$$\min_{\theta} ||X\theta - y||_2^2$$

- Ridge Regression- a linear model with regularization with the l2 norm and a tuning parameter that was chosen with leave one out cross validation which solves the following problem.

$$\min_{\theta} ||X\theta - y||_2^2 + \lambda\,||\theta||_2^2$$

- The Lasso- a linear model with regularization with the l1 norm and a tuning parameter that was chosen with leave one out cross validation which solves the following problem.

$$\min_{\theta} ||X\theta - y||_2^2 + \lambda\,||\theta||_1$$

- Elastic Net- regression model trained with L1 and L2 prior as regularizer, which minimizes the following problem.

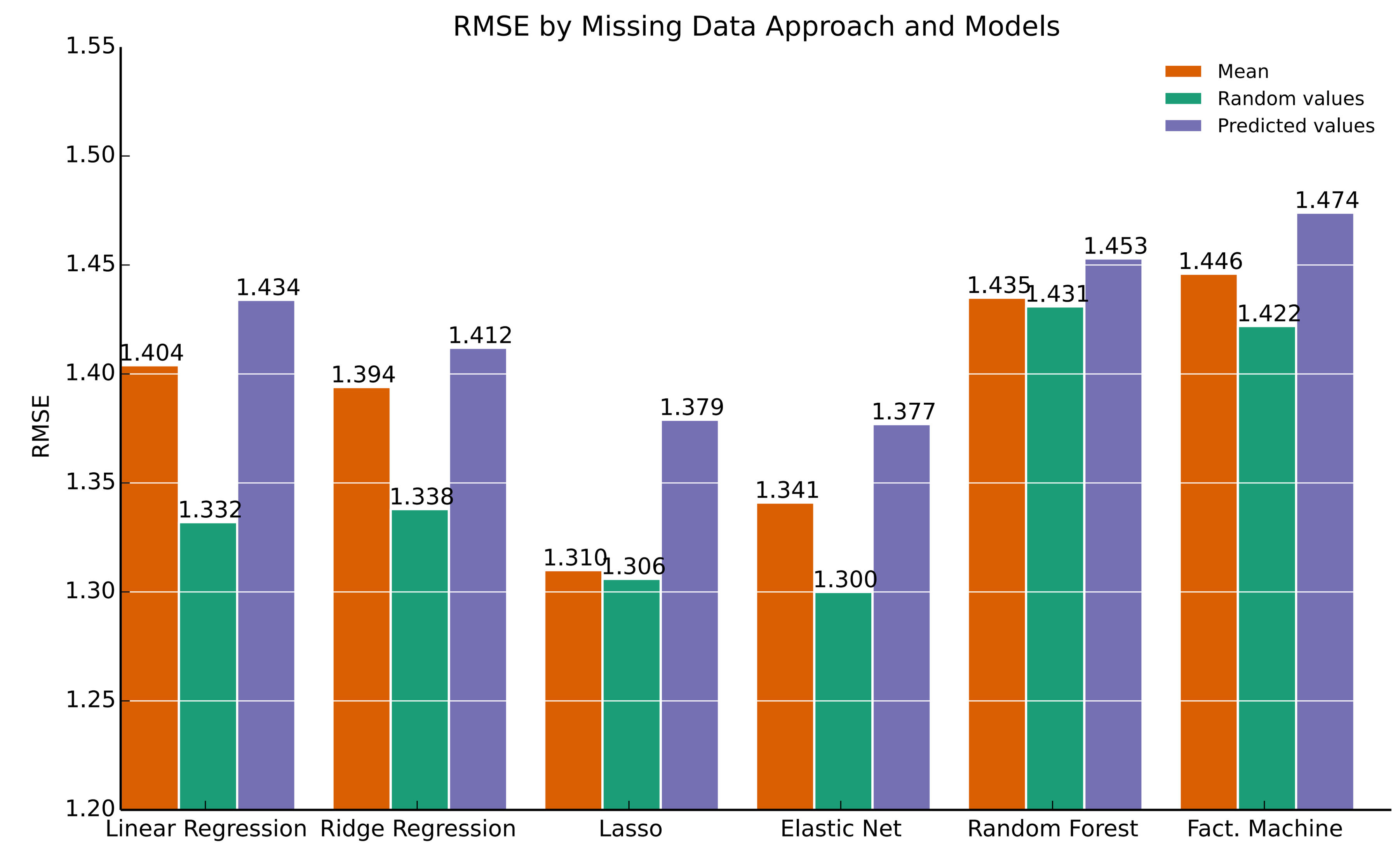
$$\min_{\theta} \frac{1}{2n} ||X\theta - y||_2^2 + \lambda\, p ||\theta||_1 + \frac{\lambda(1-p)}{2} ||\theta||_2^2$$

- Random Forest- ensemble learning method that is based from decision trees. Each tree in the ensemble is built from a bootstrap sample from the training set. The split that is picked is the best split among a random subset of features.
- Factorization Machine- generic approach that allows to mimic factorization models by feature engineering, which combine the generality of feature engineering with the superiority of factorization models in estimating interactions between categorical variables of large domain.
- Collaborative Filtering- Collaborative filtering is a technique that identifies patterns of user preferences towards certain items and makes recommendations. Collaborative filtering uses a sparse matrix holding the rating of users to businesses and calculates a similarity score to estimate the features.

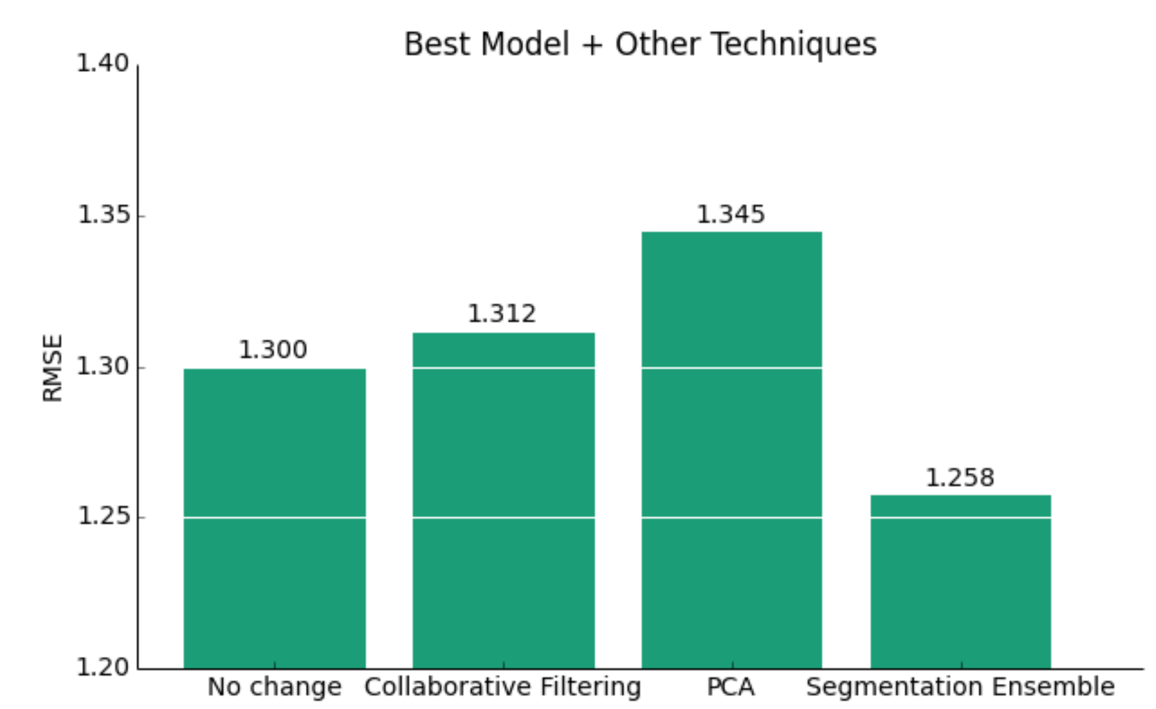
## Results

### Performance

Using a training set of 229,907 and a test set size of 36,404 we get the following result for each approach on the missing data and each model. We applied other techniques to improve all of the above models. We used filtering using PCA- dimensionality reduction, segmentation ensembles-building separate models for different portions of the data, and blending with collaborative filtering with data with nothing missing.




In general, using random values for simple imputation and Elastic Net and Lasso models with segmentation ensemble performed the best.





**Jason Ting:** jmting@stanford.edu
**Swaroop Indra Ramaswamy**: swaroopr@stanford.edu