# Advanced Machine Learning Project Proposal

Piyush Bhargava, Sakshi Bhargava, Chhavi Choudhary

## 1 Team Members

Piyush Bhargava
Sakshi Bhargava
Chhavi Choudhury

## 2 Project title

Yelp Business Rating Prediction

## 3 Background and motivation

With the explosion of social network websites and online user-generated content platforms, there is an ever increasing desire of users for personalization. Personalization also presents a strategic opportunity for businesses to expand and enhance their offerings. Hence, developing and improving recommendation systems to provide reasonable predictions about people's preferences is becoming a constant priority for businesses. A significant progress has already been made in application of new techniques in recommender systems and there is still focus on further improvements. We are motivated by the widespread application of recommendation systems in various industries. We would like to leverage this opportunity to gain hands-on experience in this rapidly evolving field. Also, data from a popular website such as Yelp is extensive and presents a challenging recommendation problem.

## 4 Project Objectives

Our project objective is to apply recommendation systems such as collaborative filtering methods to a real-life dataset - dataset from 2013 Yelp Business Rating prediction competition. We will create a model to predict the rating that a Yelp user would assign to a business. We wish to re-enact the competition and learn from the best algorithms in the competition. This will provide us an exposure to a variety of efficient and accurate recommendation techniques that can be customized for other businesses and also prepare us to participate in live competitions on recommendation systems. We will also strive to incorporate other learnings from this course such as Apache Spark in the project, thus making the recommender system scalable.

## 5 What data?

This recommender system challenge is organised by Yelp and hosted by Kaggle. The data available on Kaggle is a detailed dump of Yelp reviews, businesses, users, and checkins for the Phoenix, AZ metropolitan area. The data is in json format consisting of three json objects:

Business - Contains metadata about the business. The information mainly includes location, type of restaurant and count of reviews for the restaurant

Review - The reviews by a user id for a business id is detailed around the star ratings, text review and votes

User - The user object provides average star ratings, count of reviews and counts of votes for user Some user profiles are omitted from the data because they have elected not to have public profiles

Checking - Provides checking information of users at various hours with the business.

Since this is a Kaggle Competition, the sufficiency of data is ensured to model something interesting!

## 6 Techniques Overview

We will be exploring various popular Recommendation System algorithms. Ensembling of various models will be performed to improve the predictions of our algorithm. Some of the proposed algorithm we will be exploring are:

**Neighborhood-based**

- User-User or User-Item based similarity-based
- Content Based recommendations

**Latent-factor models**

- Low-Rank matrix factorization via ALS, SVD

**Classifier-based**

- Machine Learning algorithms like logistic regression,
- Clustering based on content data

Further, we intend to create a flask based front end to display the demo or our work.
In future, we can Gather more data, if possible, to test the effectiveness of scale-able algorithm

## 7 Optional outcomes

We would like to try these some specialized algorithms, however we are not sure at this point if they can be done in a scaleable way.

- Restricted Boltzmann Machines
- LambdaMart

## 8 Evaluation

The metric to assess the performance of the recommender system provided by Kaggle is root mean square error (RMSE)

## 9 Schedule, timeline, and team responsibilities

We plan to have the following weekly milestones for our project:

- Week 1 - Exploratory data analysis and Missing value treatment. Research and apply feature engineering techniques on data. All the team members to work on different approaches.
- Week 2 - Research various recommendation systems techniques and algorithms including the some of the best algorithms from the competition. Choose a subset of algorithms and also build our own algorithms. All team members to focus on separate techniques.
- Week 3 - Build the models (using the chosen algorithms) on the training set using Python/R and validate the models on test set. All team members to focus on coding different models.
- Week 4 - Finalize the best performing models (lowest RMSE) and also attempt a blended approach to get better predictions. Write the paper detailing the modeling approaches and the conclusions. All team members to focus on different sections of the report.
- Week 5 - Prepare the presentation.

We plan to have multiple weekly checkpoints among the team members.