Network Recommendation for Small Businesses
Harvard Computational Science and Engineering Capstone Project, Fall 2020

Overview
Alignable aims to help business owners and operators grow their networks by finding peers with related professional interests. To build valuable relationships through referrals and recommendations, the key question here is “How do we best recommend potential connections?”. Our solution is to build a recommendation system that would generate ranked lists of businesses so that Alignable can send out referral emails accordingly.

Key Question
Who should connect with whom?

Modeling Approaches
Our strategy is to fit models on a community level. We started building our models with the biggest community in our dataset, LA (42,468 businesses). After the best model is selected, we run it on all available data.

Data
To start the project, Alignable provided a thoroughly anonymized dataset. It contains basic information about around 3 million businesses, including geographical location, business type and scale, and their activities on the Alignable platform.

Baseline
Our baseline model randomly predict whether two businesses are connected.

Predictive Models
We attempted multiple models used for binary classifications - regressions (linear, logistic), SVM, KNN, and trees (decision tree, random forest, boosted trees). The metric for predictive models is the overall accuracy, and the results are as follows. Random Forest performed the best.

Collaborative Filtering
We focused on the user-based collaborative filtering approach, namely, recommending items based on preferences of similar users. What is unique in this case is that users (those given recommendations) and items (those being recommended) are both businesses within the Alignable network.

Evaluation and Interpretation
We used the unordered Normalized Discounted Cumulative Gain (NDCG) as a metric and fit the best performing models, Random Forest and SVD, on three communities, LA (~40,000 businesses), Midtown East in New York (~25,000), and Downtown Boston (~5,000). In the table below, Random Forest out-performed SVD.

Finally, we ran Random Forest on all communities. To mitigate runtime difficulties, we clustered target businesses using k-means, and only recommended those that are in the same cluster as the source business. We achieved an 83.4% overall accuracy on average among all communities. By plotting the feature importance of our model, we identified significant effects from attributes such as the time a business joined Alignable, when it was most recently active, both businesses’ distance and their mutual connections.

By exploring the relationship between accuracy and community size, we see that our model performed better on larger communities.

Final Deliverables
1. A CSV file containing 5 ranked recommendations for each business.
2. A single python file that produces this CSV, such that the client can reproduce the result or update it when new data comes in.
3. Our modeling approaches and analysis.