Network Recommendation for Small Businesses

Presented by: Zhenru Wang, Catharine Wu, Royce Yap
Supervised by: Will Claybaugh, Micah Lanier
Video Walk-through: https://youtu.be/8lYnGJo8iiY
Introduction

Alignable
The Small Business Referral Network
Problem Statement and Motivation

Who should connect with whom?
Outline

Data
- Unification
- Missing Values
- Transform Variables
- Train - Test Split
- Model Preparation

Model
- Baseline Model
- Predictive Models
- Collaborative Filtering

Result
- Evaluation
- Interpretation
Data sources

Business-level attributes (nodes)
- Businesses
- Industries
- Roles
- Communities
- Users
- Discussion Posts

Relationships between businesses (edges)
- Recommendations
- Connections
- Business Requests
Data points for Alignable businesses

**Community**
- Longitude
- State
- Latitude
- Recommendations

**Industry / Role**
- Customer Type
  - (B2B / B2C)
- Customer Location
  - (Local/Regional/National)
- Added logo
- Date of joining
- Most recent sign-in
- Added banner
- Discussion Posts

**Existing connections**
- Business Requests
Data Exploration - Geographical Distribution

Users across the Platform

Example: Connections to Los Angeles Users
Data Exploration - Connections

Number of connections per business

Connections across industries
Data Exploration - Industry / Role

Distribution across 29 industries

Distribution across 332 roles (sub-categories)
Data Wrangling / Preprocessing

- Unification
- Missing Values
- Transforming Variables
E.g. Industry/Role Affinity Scores

Industry Affinity Score
## Train-test split

**Connections for a given business**

<table>
<thead>
<tr>
<th>connected businesses</th>
<th>A</th>
<th>B</th>
<th>C</th>
<th>D</th>
<th>E</th>
<th>F</th>
<th>G</th>
<th>H</th>
<th>I</th>
<th>J</th>
</tr>
</thead>
<tbody>
<tr>
<td><em>created_at</em> (ascending)</td>
<td>0.12</td>
<td>0.2</td>
<td>0.27</td>
<td>0.29</td>
<td>0.32</td>
<td>0.48</td>
<td>0.57</td>
<td>0.68</td>
<td>0.79</td>
<td>0.98</td>
</tr>
</tbody>
</table>

### Train

<table>
<thead>
<tr>
<th>A</th>
<th>B</th>
<th>C</th>
<th>D</th>
<th>E</th>
<th>F</th>
<th>G</th>
<th>H</th>
<th>I</th>
<th>J</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.12</td>
<td>0.2</td>
<td>0.27</td>
<td>0.29</td>
<td>0.32</td>
<td>0.48</td>
<td>0.57</td>
<td>0.68</td>
<td>0.79</td>
<td>0.98</td>
</tr>
</tbody>
</table>

### Test
Modeling Approach

Baseline
Randomly predict whether connected

Predictive Models
Linear & Logistic Regression, Decision Tree, Random Forest, AdaBoost, Neural Network

Collaborative Filtering
Singular Value Decomposition (SVD), Alternating Least Squares (ALS)
Modeling Approach

Baseline
Randomly predict whether connected

Predictive Models
Linear & Logistic Regression, Decision Tree, Random Forest, AdaBoost, Neural Network

Collaborative Filtering
Singular Value Decomposition (SVD), Alternating Least Squares (ALS)
# Predictive Models

<table>
<thead>
<tr>
<th>source_business_id</th>
<th>target_business_id</th>
<th>source_business_profile</th>
<th>target_business_profile</th>
<th>distance</th>
<th># of mutual connections</th>
<th>connected</th>
</tr>
</thead>
<tbody>
<tr>
<td>2422</td>
<td>7257010</td>
<td>...</td>
<td>...</td>
<td>14.072</td>
<td>2</td>
<td>1</td>
</tr>
<tr>
<td>2422</td>
<td>8550114</td>
<td>...</td>
<td>...</td>
<td>1.8673</td>
<td>8</td>
<td>0</td>
</tr>
</tbody>
</table>

**Business Profile**
- type, scale, location, industry, role, profile completeness, time joined, latest activity
Modeling Approach

Baseline
Randomly predict whether connected

Predictive Models
Linear & Logistic Regression, Decision Tree, Random Forest, AdaBoost, Neural Network

Collaborative Filtering
Singular Value Decomposition (SVD), Alternating Least Squares (ALS)
Collaborative Filtering

<table>
<thead>
<tr>
<th>Source Businesses</th>
<th>Target Businesses</th>
</tr>
</thead>
<tbody>
<tr>
<td>ID 1</td>
<td>Target 1</td>
</tr>
<tr>
<td>ID 2</td>
<td>0</td>
</tr>
<tr>
<td>ID 3791</td>
<td>0</td>
</tr>
</tbody>
</table>
PROBLEM 1

Data Sparsity

SOLUTION

- Community Clustering
- Data Balancing
PROBLEM 2

Expensive Computation

SOLUTION

Business Clustering
## Model Comparison

- **Metric**: Normalized Discounted Cumulative Gain (NDCG)

<table>
<thead>
<tr>
<th>City</th>
<th>Baseline</th>
<th>Random Forest</th>
<th>SVD</th>
</tr>
</thead>
<tbody>
<tr>
<td>LA</td>
<td>0.0002</td>
<td>0.0031</td>
<td>0.0030</td>
</tr>
<tr>
<td>NYC</td>
<td>0.0003</td>
<td>0.0086</td>
<td>0.0073</td>
</tr>
<tr>
<td>Boston</td>
<td>0.0023</td>
<td>0.0322</td>
<td>0.0270</td>
</tr>
</tbody>
</table>
Averaged Overall Accuracy among all communities by Random Forest

83.44%
Evaluation

Accuracy distribution

- Blue: baseline
- Orange: random forest
Model performs better on larger communities
Feature Importance

![Feature Importance Top 10 Chart]

- joined_neighborhood_at_2
- distance
- last_sign_in_at_2
- last_sign_in_at_1
- joined_neighborhood_at_1
- num_mutual
- has_added_logo_2
- has_added_logo_1
- has_added_banner_2
- same_community
Deliverables

A **csv** file containing all Alignable’s businesses and 5 ranked recommendations for each

A **python** file to produce the csv

Modeling approaches and analysis


Review

Data
- Unification
- Missing Values
- Transform Variables
- Train - Test Split
- Model Preparation

Model
- Baseline Model
- Predictive Models
- Collaborative Filtering

Result
- Evaluation
- Interpretation