Overview

Challenge: We want to better understand the people who live in cities based on the businesses that surround them.

Context: We focus on New York City (NYC), the largest American city (by population) with a high density of businesses and people.

Approach:
1. Construct novel dataset of census tracts → demographic (e.g. income)
2. Engineer features based on the businesses located within each census tract
3. Train and validate models for different feature sets for income levels bucketed by quartile (e.g. top 25%)

Dataset / Pre-Processing

NYC Check-in Dataset [1]
- 227,428 check-ins collected
- Over 10 months (12 April 2012 to 16 February 2013)

Sample Check-In Data

<table>
<thead>
<tr>
<th>userId</th>
<th>venueId</th>
<th>venueCategory</th>
<th>latitude</th>
<th>longitude</th>
<th>timestamp</th>
</tr>
</thead>
<tbody>
<tr>
<td>470</td>
<td>49bbd6c</td>
<td>Arts &amp; Crafts Store</td>
<td>40.719</td>
<td>-74.002</td>
<td>Tue Apr 03...</td>
</tr>
</tbody>
</table>

NYC Census Data [2]
- Demographics for 2166 census tracts from the American Community Survey’s 5-year estimates (2015)

Sample Census Data

<table>
<thead>
<tr>
<th>BlockCode</th>
<th>TotalPop</th>
<th>White</th>
<th>Hispanic</th>
<th>...</th>
<th>Income</th>
</tr>
</thead>
<tbody>
<tr>
<td>36005000200</td>
<td>5403</td>
<td>2.3%</td>
<td>75.8%</td>
<td>...</td>
<td>72034.0</td>
</tr>
</tbody>
</table>

Preprocessing
- For all 227k check-ins, processed corresponding census tract using FCC Census Block Conversions API [3]
- ~22 hours to construct due to API rate-limits

Feature Engineering

Predicting: Income

Feature Heatmaps

Features: We evaluate the following set of features for each census tract using our baseline NaiveBayes Classifier with K-fold (k=5) cross validation

- num_total_checkins
- num_checkins_per_category (e.g. mexican_restuarant, medical_center)
- num_checkins_per_weekday (e.g. Mon, Tue)
- num_checkins_per_time_bucket (e.g. )
- num_checkins_per_category_per_time_bucket
- num_checkins_per_category_per_weekday
- num_checkins_per_category_per_time_bucket

Features Included

<table>
<thead>
<tr>
<th>Dimensions (n)</th>
<th>cv^*_{train}</th>
<th>cv^*_{dev}</th>
</tr>
</thead>
<tbody>
<tr>
<td>{a}</td>
<td>251</td>
<td>0.4284</td>
</tr>
<tr>
<td>{a,b}</td>
<td>258</td>
<td>0.4284</td>
</tr>
<tr>
<td>{a,b,c}</td>
<td>264</td>
<td>0.4521</td>
</tr>
<tr>
<td>{a,b,c,d}</td>
<td>1615</td>
<td>0.5287</td>
</tr>
<tr>
<td>{a,b,c,d,e}</td>
<td>1657</td>
<td>0.5390</td>
</tr>
<tr>
<td>{a,b,c,d,e,f}</td>
<td>9028</td>
<td>0.5283</td>
</tr>
</tbody>
</table>

Results

Comparing several models to a NaiveBayes baseline classifier:

<table>
<thead>
<tr>
<th>Model</th>
<th>cv^*_{train}</th>
<th>cv^*_{dev}</th>
<th>test</th>
</tr>
</thead>
<tbody>
<tr>
<td>Baseline: Naive Bayes</td>
<td>0.452143</td>
<td>0.3491</td>
<td>0.2707</td>
</tr>
<tr>
<td>Gradient Boosting Trees</td>
<td>0.7094</td>
<td>0.3749</td>
<td>0.3932</td>
</tr>
<tr>
<td>AdaBoost</td>
<td>0.4559</td>
<td>0.3651</td>
<td>0.3390</td>
</tr>
<tr>
<td>Quadratic Discr. Analysis</td>
<td>0.6234</td>
<td>0.3646</td>
<td>0.2735</td>
</tr>
<tr>
<td>Logistic Regression</td>
<td>0.4264</td>
<td>0.3103</td>
<td>0.2906</td>
</tr>
</tbody>
</table>

Discussion

Naive Bayes: NB failed to account for the fact that there does not exist conditional independence between the features separated by time.

Gradient Boosting Trees (GBT): GBT relies on weak learners capable of discerning relationships between lower-level features. GBT is robust against outliers.

AdaBoost: Similar to GBT, AdaBoost leverages decision trees as a robust method demographic prediction.

Quadratic Discriminant Analysis: Performed the worst out of all our models. Our data is not a quadratic surface.

Logistic Regression: Logistic regression fails to capture complex relationships between features...

Confusion Matrix
Our confusion matrix shows that the model struggles to predict the lowest income bracket, presumably because we have less signal for lower income brackets. This makes intuitive sense.

References