Understanding New York City Demographics from Local Business Activity

Vincent Chen [vschen@stanford.edu], Dan Yu [dxyu@stanford.edu]

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

userId	venueId	venueCategory	latitude	longitude	timestamp
470	49bbd6c0f96...	Arts & Crafts Store	40.719	-74.002	Tue Apr 03...

NYC Census Data [2]

- Demographics for 2166 census tracts from the American Community Survey's 5-year estimates (2015)

Sample Census Data

BlockCode	TotalPop	White	Hispanic	...	Income
36005000200	5403	2.3%	75.8%	...	72034.0

Preprocessing

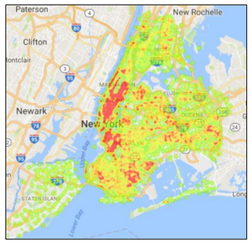
- For all 227k check-ins, processed corresponding census tract using FCC Census Block Conversions API [3]
- ~22 hours to construct due to API rate-limits

Feature Engineering

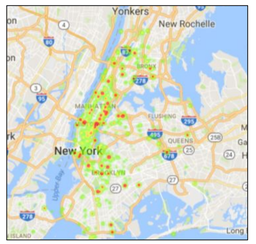
Feature Heatmaps

Predicting: Income

Num. Total Checkins



Num. Residential Condos



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

a. num_total_checkins
b. num_checkins_per_category (e.g. mexican_restuarant, medical_center)
c. num_checkins_per_weekday (e.g. Mon, Tue)
d. num_checkins_per_time_bucket (e.g.)
e. num_checkins_per_category_per_time_bucket
f. num_checkins_per_category_per_weekday
g. num_checkins_per_category_per_time_bucket

*cross validation

Features Included	Dimensions (n)	cv*_train	cv*_dev
{a}	251	0.4284	0.3182
{a,b}	258	0.4284	0.3251
{a,b,c}	**264**	**0.4521**	**0.3491**
{a,b,c,d}	1615	0.5287	0.3057
{a,b,c,d, e}	1657	0.5390	0.3229
{a,b,c,d,e,f}	9028	0.5283	0.2857

underfit ↕ overfit

Results

Comparing several models to a NaiveBayes baseline classifier:

Model	cv*_train	cv*_dev	test
Baseline: Naive Bayes	0.452143	0.3491	0.2707
Gradient Boosting Trees	**0.7094**	**0.3749**	**0.3932**
AdaBoost	0.4559	0.3651	0.3390
Quadratic Discrim. Analysis	0.6234	0.3646	0.2735
Logistic Regression	0.4264	0.3103	0.2906

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.

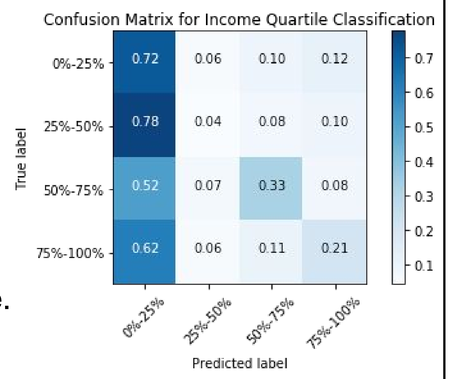
AdaBoost: Similar to GBT, AdaBoost leverages decision trees as a robust method demographic prediction.

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

[1] Daqing Zhang, Vincent W. Zheng, Zhiyong Yu. Modeling User Activity Preference by Leveraging User Spatial Temporal Characteristics in LBSNs. IEEE Trans. on Systems, Man, and Cybernetics: Systems, (TSMC), 45(1), 129-142, 2015.
[2] American Community Survey, MuonNeutrino (Kaggle). (2015) New York City Census Data. Retrieved from https://www.kaggle.com/muonneutrino/new-york-city-census-data.
[3] Federal Communications Commission. "Census Block Conversions API." (2017) Retrieved From https://www.fcc.gov/general/census-block-conversions-api.