

A wide-angle photograph of a city skyline, likely Chicago, featuring a river in the foreground and numerous skyscrapers in the background under a clear blue sky. The text is overlaid on a blue banner across the middle of the image.

Predicting GeoSpatial Affluence For Disaster Recovery With **yelp***



Problem Statement & Key Objectives

Problem Statement

- Traditional methods estimate geospatial wealth via income or unemployment rates
- These data sources may not provide sufficient granular data for the agile needs of disaster response agencies

Key Objective

- **Build a model that will take zip codes or location as input and will leverage frequently updated commercial data to estimate geospatial affluence with high accuracy**



Executive Summary

- Yelp Dollar sign (\$) data alone is not a strong predictor of geospatial affluence
 - Correlation of Yelp Dollar Sign to Median Income is 0.09
- Challenges observed: limited venue/business/service representation in low affluence or low population areas, non-residents providing yelp reviews
- Augmenting Yelp Dollar Sign data with feature engineering improved scores
- Ultimately our model is able to accept a location as input and estimate the affluence category with 65.5% accuracy
- We do not recommend using Yelp price level data alone to predict affluence



Process & Model Overview

1

Data Gathering

- Yelp API
- ASC data

3

Analysis & Feature Engineering

- Affluence Score Creation
- Yelp features

2

Data Cleaning/Wrangling

- Aggregation by zip code
- Identifying & addressing null/missing data

4

Conclusions & Recommendations

Gathering & Cleaning Data



tacos, cheap dinner, Max's

Down to

Restaurants

Home Services

Auto Services



Oleana Restaurant Claimed



1500 reviews

[Details](#)

\$\$\$ · Mediterranean

[Edit](#)

Write a Review

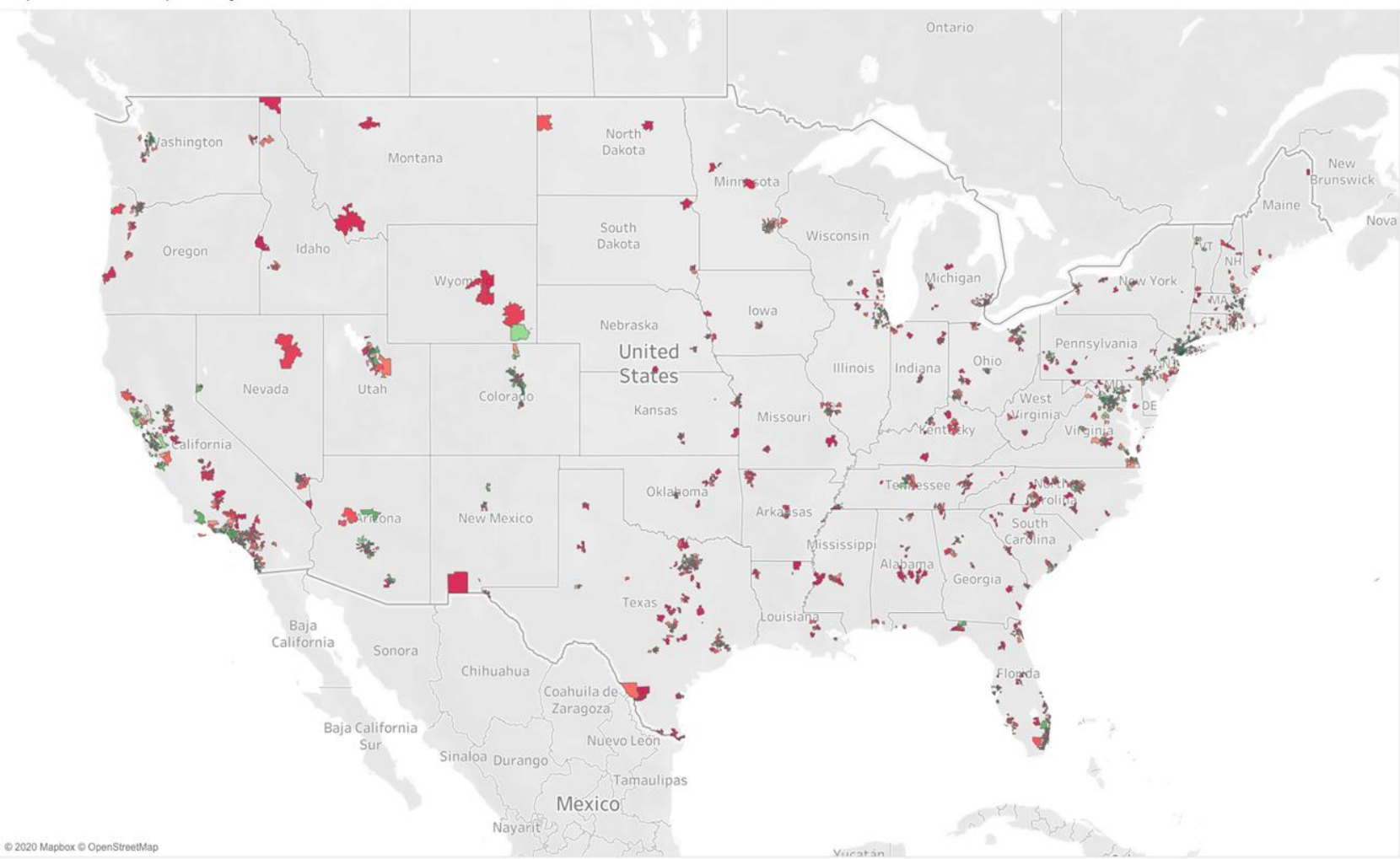
Add Photo

Share

Save

- Yelp records with \$ values were extracted for over **300** cities:
 - **Richest** cities in every U.S. state (median household income)
 - **Poorest** cities in every U.S. state
 - **200** most populous U.S. cities
- Over **100K** unique records were obtained, providing representation for the entire nation
- Data Extraction Tools and Approach:
 - **Web scraping** with BeautifulSoup for City Lists
 - Custom Python code leveraging **YelpAPI** used to pull Yelp JSON files spanning multiple hours
 - Python code used to parse, clean and merge datasets
- Data from American Fact Service (census data) downloaded for **income**, **median home value**, etc.

Zipcodes Scraped by Affluence Score





Key Challenges



- Correlation between price and income: 0.09
- There are more \$\$ establishments than the rest of the price categories combined
- Tried smaller dataframe where each zip code had at least 250 observations, the results did not improve. Lack of data is not the problem.



Data Cleaning and Wrangling

Cleaning

- Aggregating data by zip code
 - ~3400 unique zip codes
 - Mean and sum of particular features
- Removing the overpopulous zip codes
 - Consistent amounts of NaNs
 - Lack of reviews and restaurants per zip
- Binning appropriate features

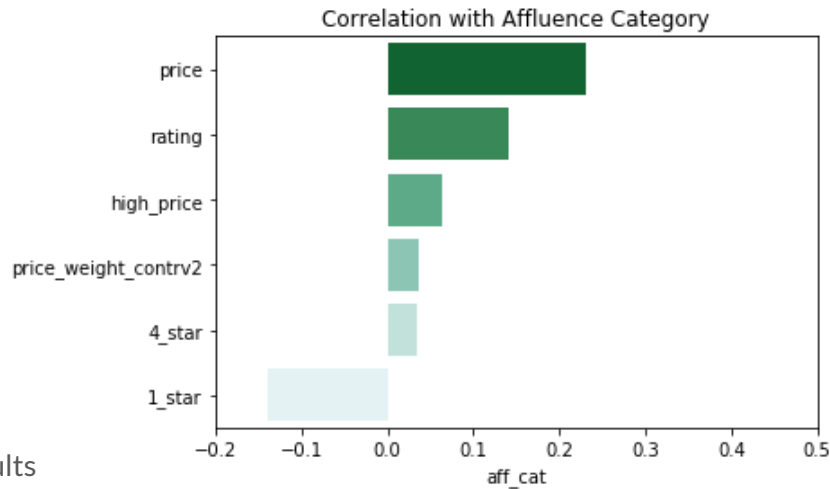
Wrangling

- Merging data from several data sources
 - Yelp scrape
 - Census data
- Normalization of Yelp Prices

Feature Engineering

We leveraged the price level data point and rating into additional valuable features:

- One star
 - Four Star
 - High Price
 - Mean Price
 - Price-weight-control
-
- Dummifying establishment category provided no results
 - Possibly because some types of establishments are far more likely to have price level than others



Affluence Score

- Income alone does not equate to “affluence”
- Combined several measures for better representation:
 - Median income
 - Percent with bachelor’s degree
 - Median home value





Model Performance

Pre- Feature Engineering Results
Engineering Results

Model	Train/Test Scores
Logistic Regression	0.5441 / 0.5445
KNN Neighbors	0.5185 / 0.5445
Decision Tree	0.5446 / 0.548
Random Forest	0.5466/ 0.5482
AdaBoost	0.5426 / 0.5482
SVC	0.5466 / 0.5468

Feature

Model	Train/Test Scores
Logistic Regression	0.6416 / 0.6192
KNN Neighbors	0.7397 / 0.5752
Decision Tree	1.0000 / 0.5000
Random Forest	0.9722/ 0.5856
AdaBoost	0.6718 / 0.625
SVC	0.6745/ 0.617

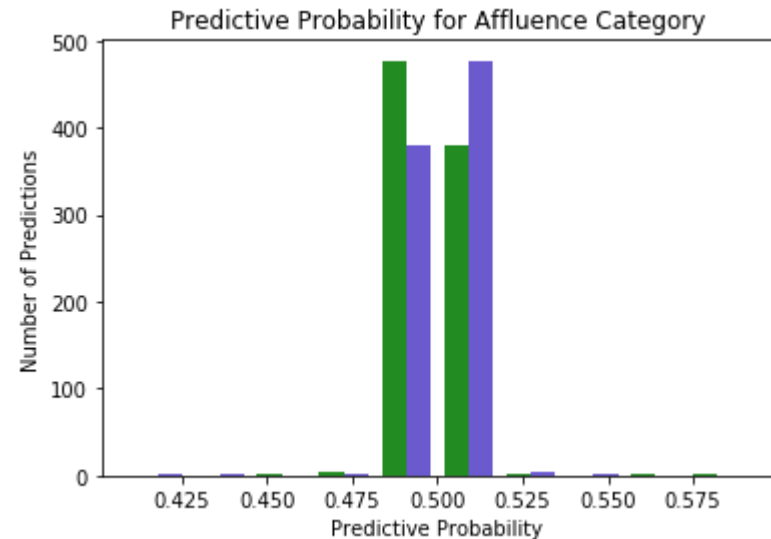


Model Results

AdaBoost with 50 Decision Trees

65.5% accuracy

- Confidence of model generally low
- Mean price for high affluence: 1.88
- Mean price for low affluence: 1.67





Conclusion

- Yelp Dollar sign (\$) data alone is not a strong predictor of geospatial affluence
 - Limited differentiation
- Turning the problem into two class classification problem made it more feasible
- Engineered features significantly improved performance
- Challenges observed: limited venue/business/service representation in low affluence or low population areas



Recommendations

- Do not heavily rely on Yelp price level data to predict affluence
- Continue building robustness of model with scheduled periodic Yelp Data updates to continue training model and improving accuracy
 - Consider using category data
- Consider purchase of Commercial Grade Yelp API access in order to streamline data extraction
- Expand functionality and improve user-experience by building front-end with Flask or similar tool