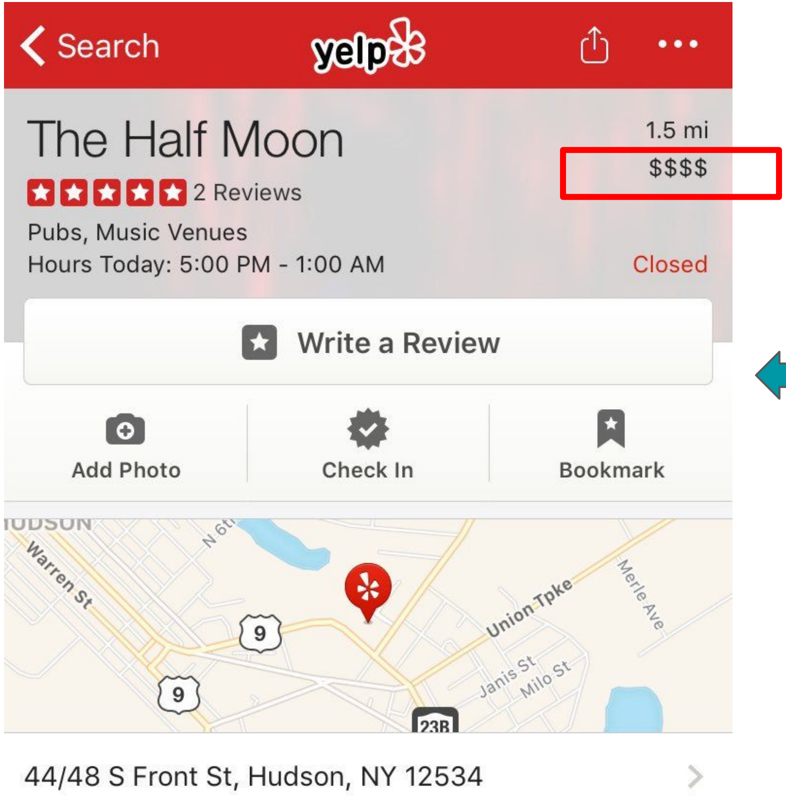
Yelp Expenditure Level & Neighborhood Affluency

Presented by Daniel Stern, Rex Chang, Julian Sweet and Jane Liang

01/18/2018 GA_DSI_LA

Expenditure Level & Affluency



The Half Moon

1.5 mi

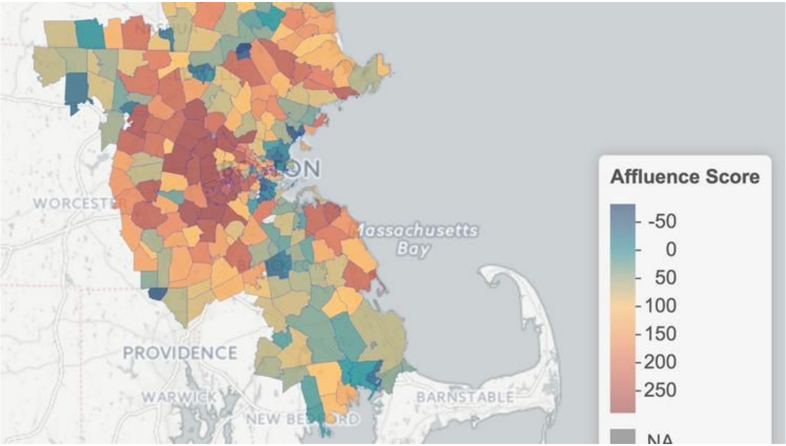
$$$$

★★★★★ 2 Reviews

Pubs, Music Venues

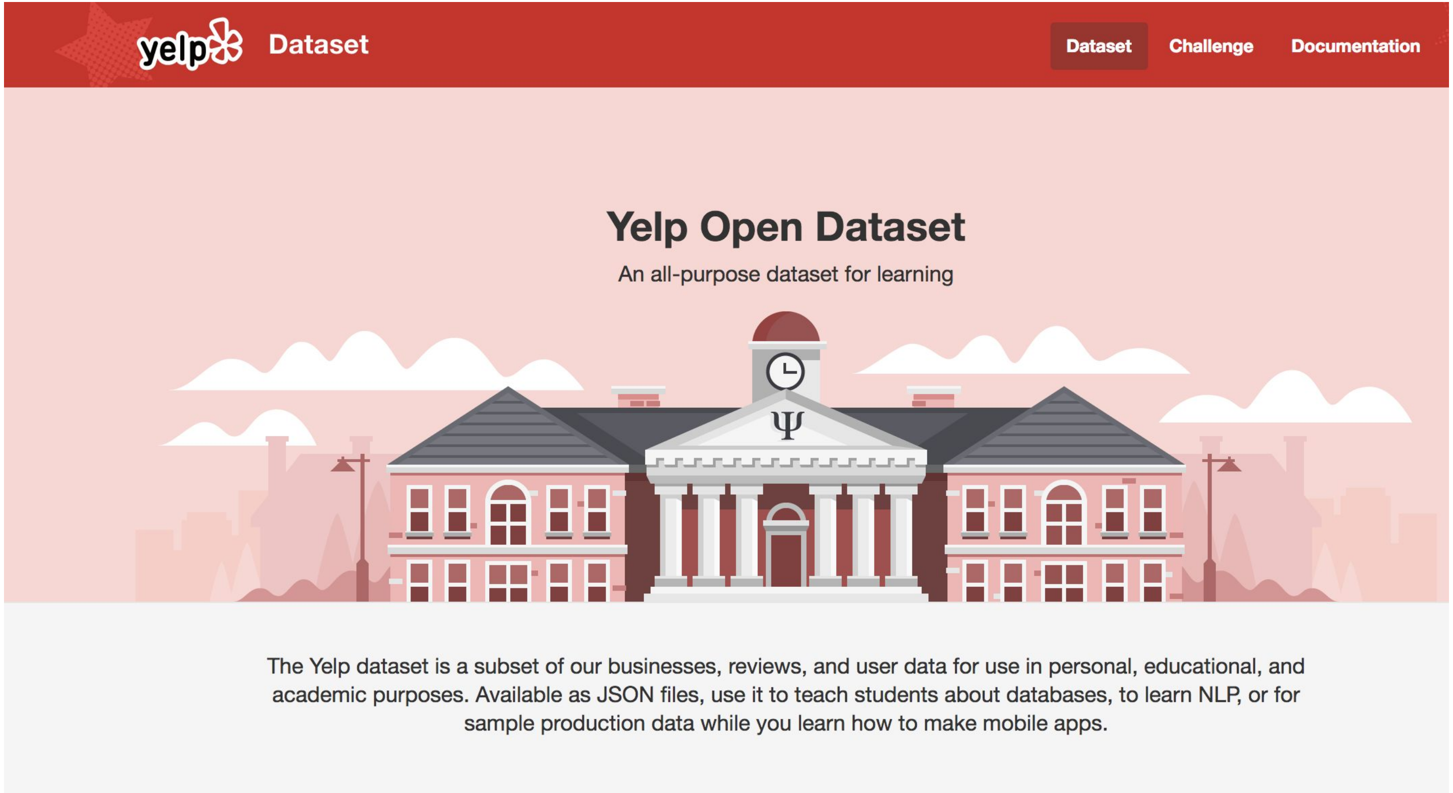
Hours Today: 5:00 PM - 1:00 AM Closed

44/48 S Front St, Hudson, NY 12534



Yelp Open Dataset

https://www.yelp.com/dataset (174,000 business attributes)

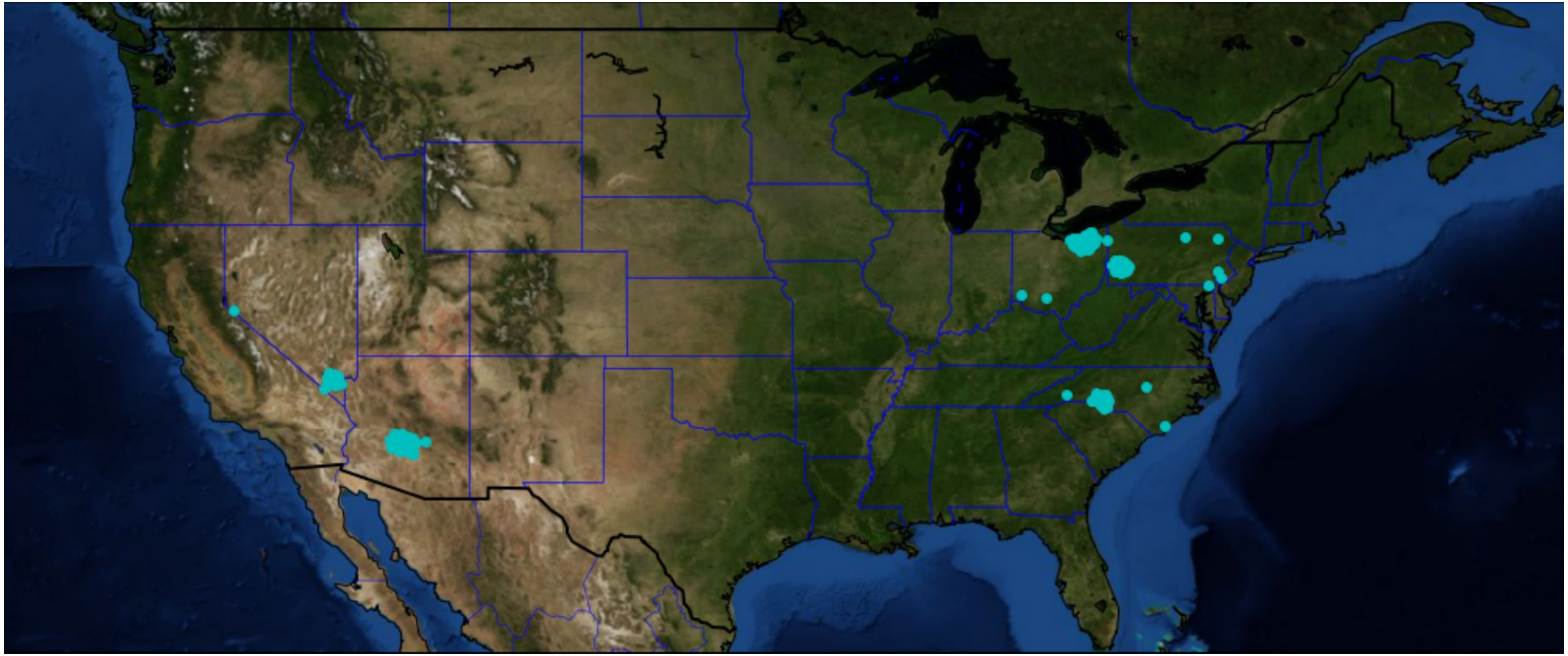


Yelp Open Dataset

An all-purpose dataset for learning

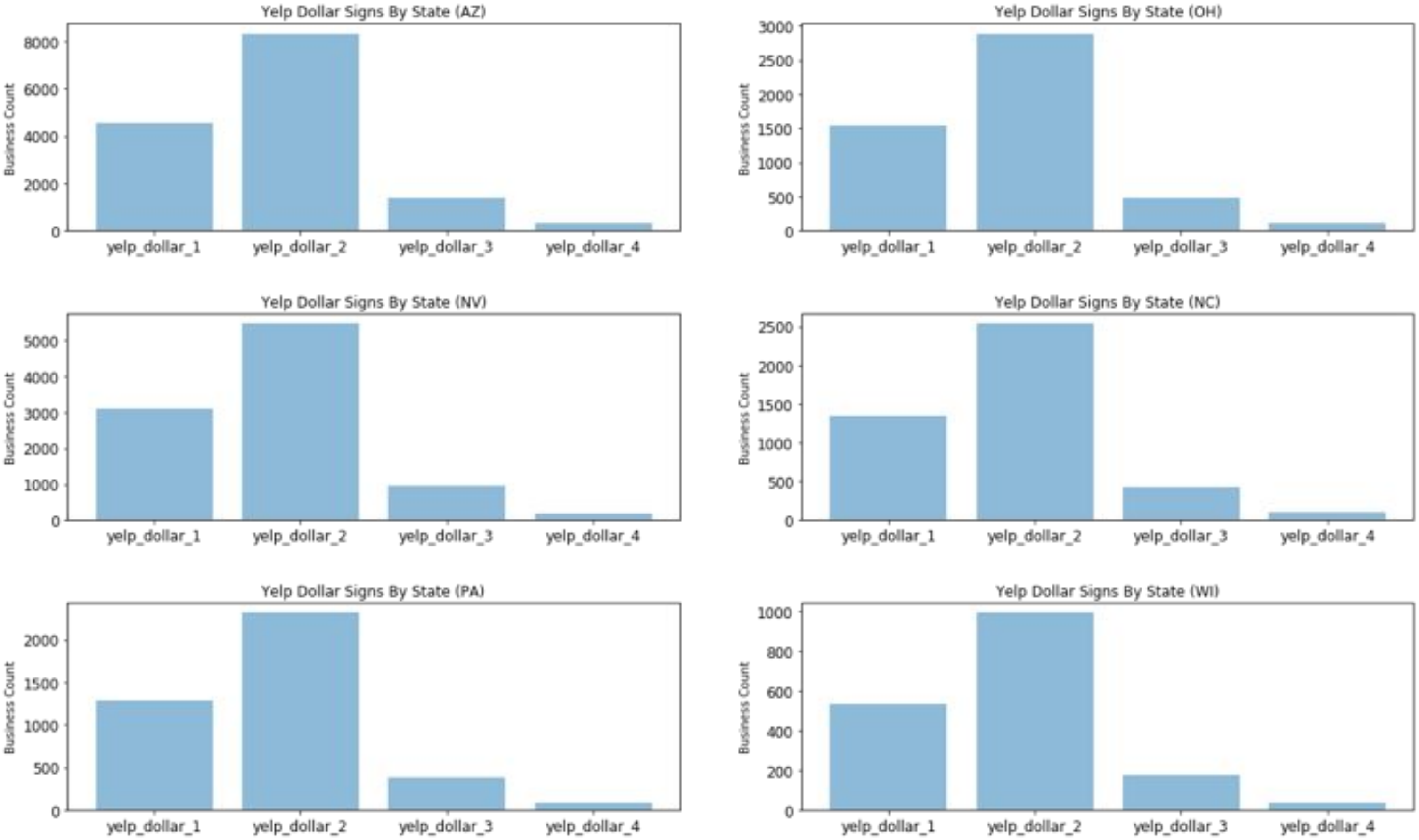
The Yelp dataset is a subset of our businesses, reviews, and user data for use in personal, educational, and academic purposes. Available as JSON files, use it to teach students about databases, to learn NLP, or for sample production data while you learn how to make mobile apps.

Metropolitan Zip Code Clusters



Five Major Metros

Distribution of Yelp Prices By Metro Area in Dataset



Best way to engineer income target feature?

IRS Income Data

Number of residents in each category by zip code

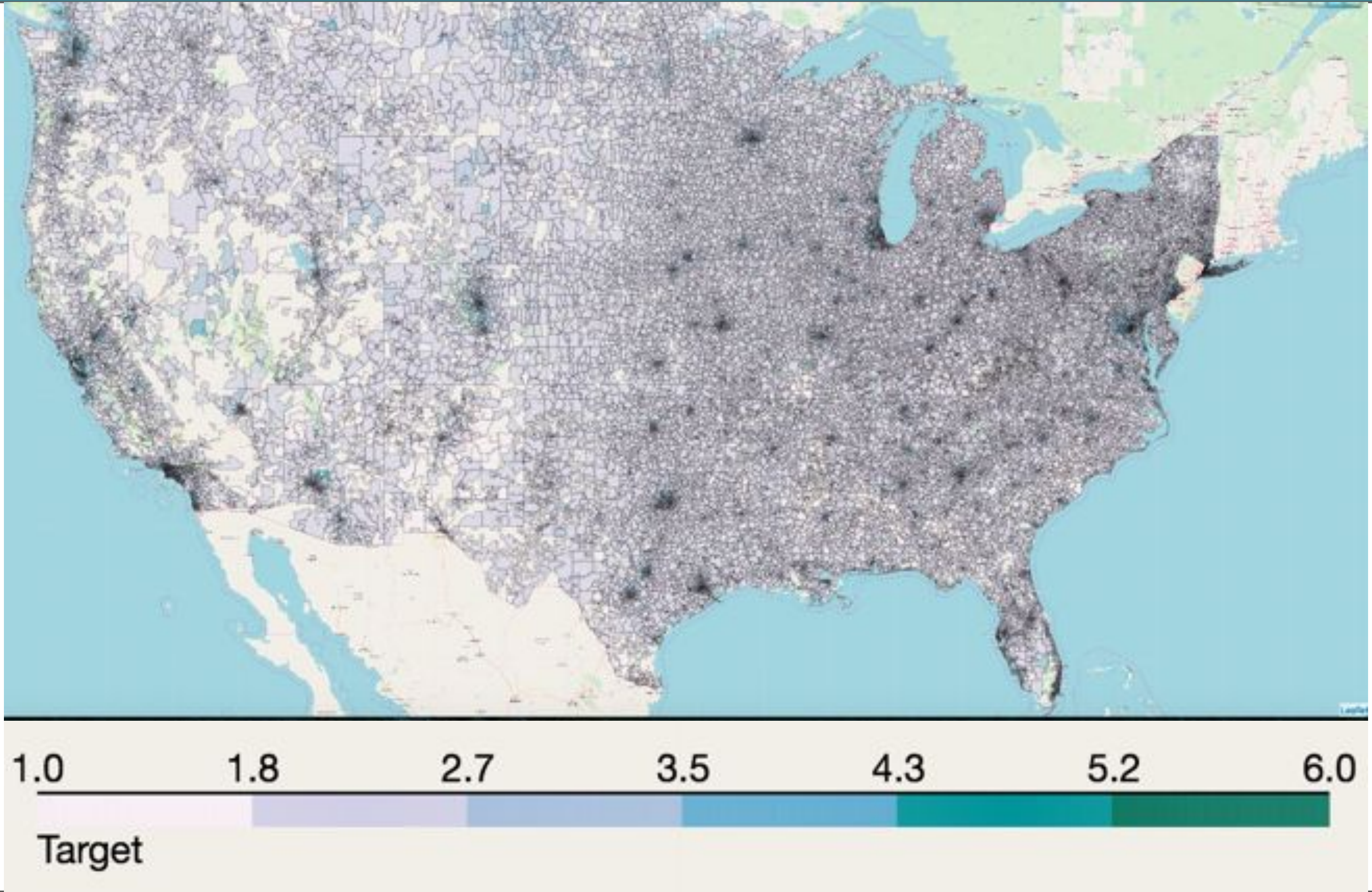
- 1 = $1 under $25,000
- 2 = $25,000 under $50,000
- 3 = $50,000 under $75,000
- 4 = $75,000 under $100,000
- 5 = $100,000 under $200,000
- 6 = $200,000 or more

1 to 6 – Categorical, but also ordinal

Weighted average? No

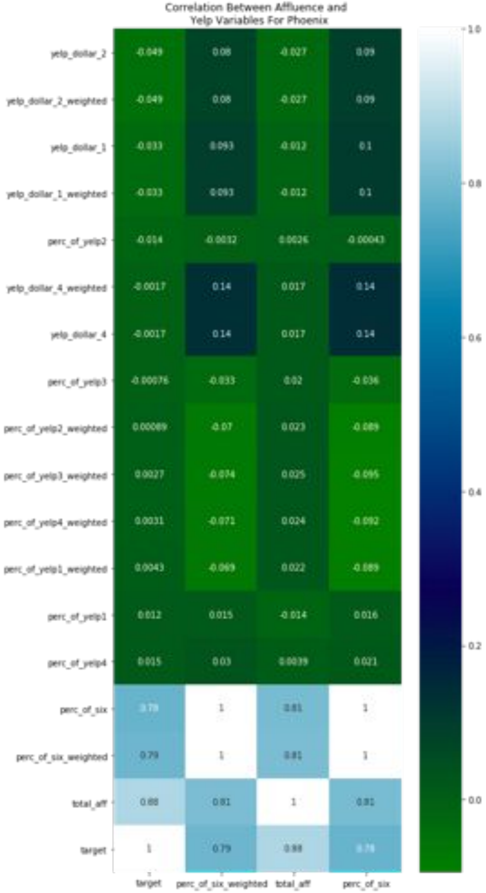
Instead, median category #

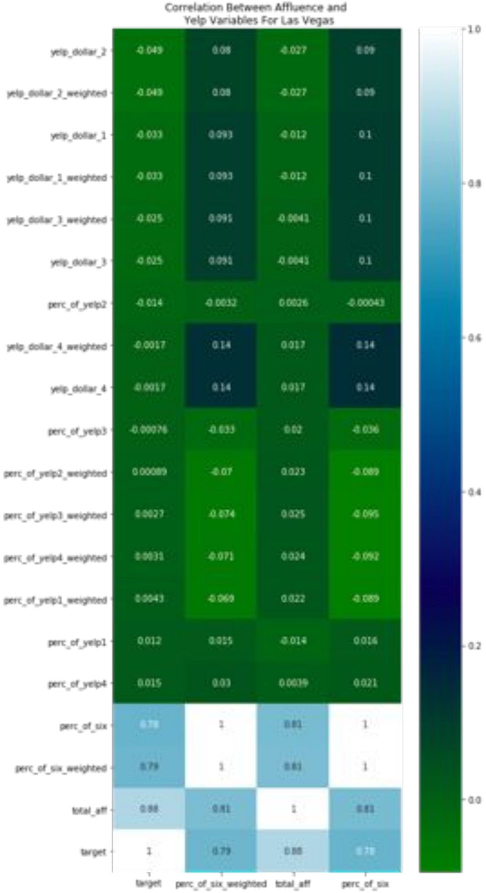
All US Zipcodes By Income Class



Sample Heatmap of Yelp + Affluency Data

- Very low correlation between Yelp & Affluence-related data.
- Correlation seems to be similar for each distinctive metro area.





Engineered Feature – Metro Distance

- For 5 largest metros, Metro centroid was assigned to city hall for the respective metro. Latitude and longitude were determined by lookup.

- Specialized GeoPy library used to to determine distance in miles from city hall to each zip code.

- This served as an engineered feature to determine if proximity, or lack thereof to center of the metro correlated with affluence.

Machine Learning: Logistic Regression Model

Baseline Accuracy

63.19%

Machine Learning: Logistic Regression Model

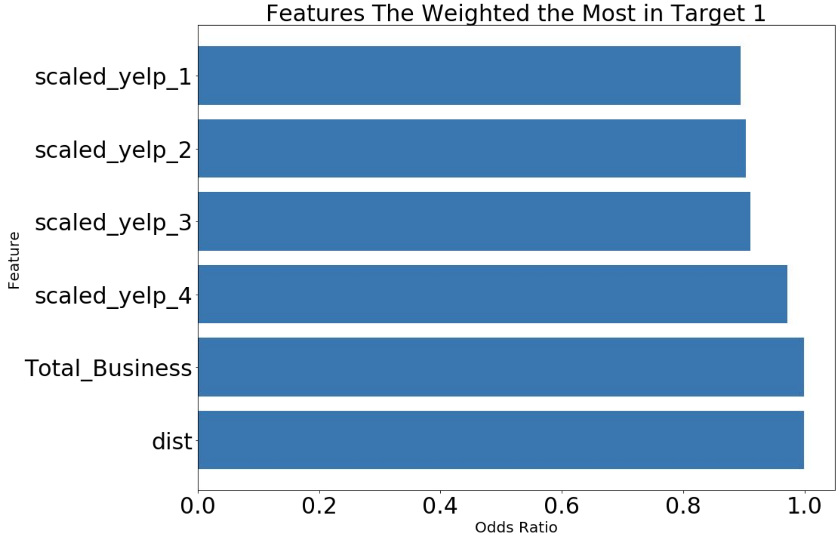
Modeling Accuracy :

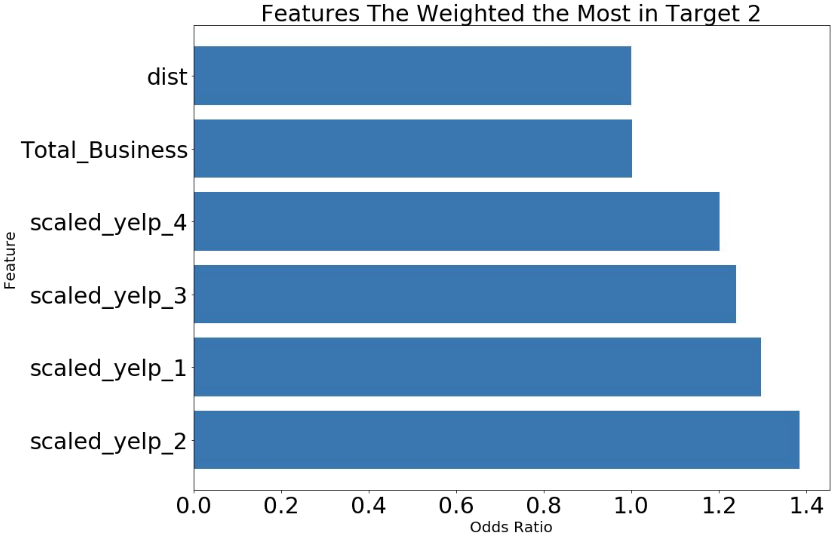
Train : **64.5%**

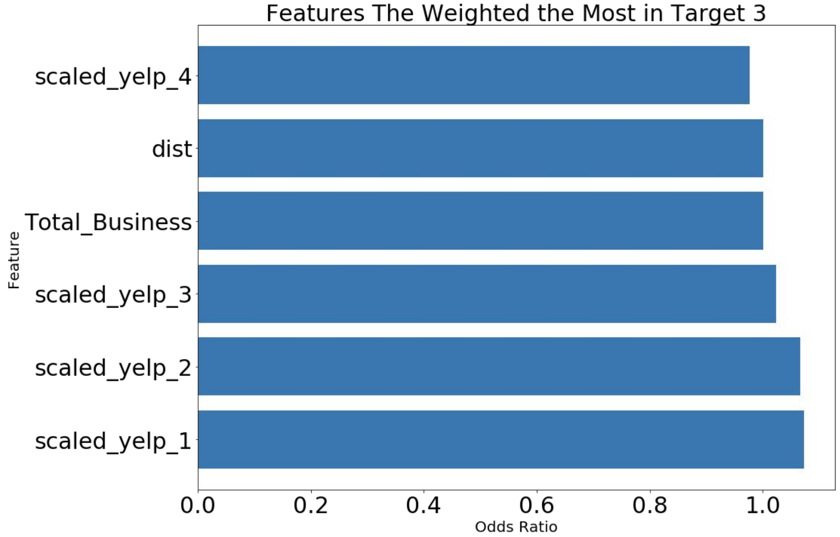
Test : **59.3%**

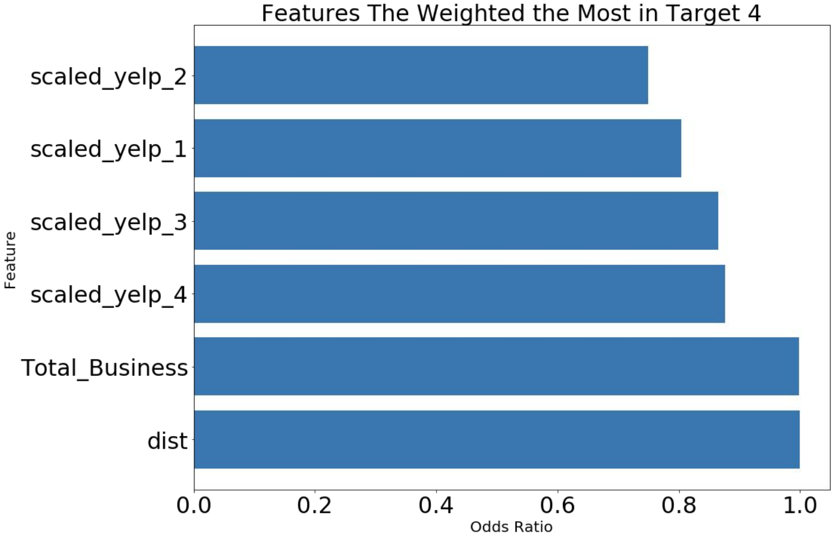
Intercepts :

Target 1	Target 2	Target 3	Target 4
-0.378	**1.080**	**0.134**	**-0.836**


Features The Weighted the Most in Target 1


Features The Weighted the Most in Target 2


Features The Weighted the Most in Target 3


Features The Weighted the Most in Target 4

Conclusions and Recommendations

Poor Model performance due to :

1. **Limited Scope:** currently not representative of the overall U.S. population.
2. **Limited Features:** aggregate expenditure level and distance from downtown for any given zip code. This is not sufficient when building our model.
3. **Other ML models**.