**Predicting Housing Price Fluctuation in Urban Neighborhoods Using Yelp and Zillow Data**

Christian Conroy and Lawrence Doppelt

Yelp has created an opportunity for research-minded academics to study and develop unique insights into the relationships between local restaurants and the greater fabric of the neighborhoods around them. Our goal is to use geocoded Yelp data to develop a model that can predict changes in housing prices based on Yelp rankings and text reviews. As part of the Yelp Challenge, some of the data behind the online and app-based platform has been publicly released.

While many have developed models using Yelp data to improve the efficiency or profitability of businesses, few have examined the relationship between business reviews and rankings and the broader economic development within specific neighborhoods. Our hypothesis is that an increase in average restaurant ratings, the advent of key positive words in reviews, and an increase in check-ins will lead to a detectable rise in housing prices and rents. The logic behind this is that more and better restaurants in a neighborhood will prompt other businesses to move into the neighborhood, resulting in higher property values due to increased access to amenities and overall development of the neighborhood.

There are several interrelated research questions we will explore in testing this hypothesis. What words signal an improvement in a restaurant’s overall rating? What kind of words in reviews for early entrants into a community signal increases in restaurant openings and overall investment into an area? What kind of words in reviews indicate that property values and rents will increase in a neighborhood? What is the relationship between the number of check-ins in restaurants and the property values in that neighborhood?

To develop the predictive model, we plan to merge some datasets to get a complete picture of neighborhood effects, property values, and businesses. The data from Yelp includes unique business and user ID tags, text of reviews, review dates, the star ratings provided in a review, check-ins for various businesses, and user reactions to reviews (“cool,” “useful,” “funny”). The data from Zillow includes zip-codes and geographic coordinates, median listing prices for different housing types, median rental prices for different housing types, and other supplementary metrics.

Regardless of which approaches we employ, we will use a train-test model to check the initial accuracy of our predictions, such as whether a review’s text accurately predicts the number of stars received or if the number of check-ins to local businesses could predict changes in area housing prices over time. A well validated model could ultimately be used by real estate investors seeking to value investments using a classic “buy low, sell high” strategy. At the same time, policymakers can use the model to detect if and when development might stimulate inequality often associated with gentrifying neighborhoods, pricing residents out with rising property values. By better understanding the timing of gentrification, policymakers can target inclusive zoning, low-income housing and development credits, and other policies aimed at mitigating the negative inequality effects of urban gentrification.