Rent Estimations in Indianapolis, IN

This project aimed to predict rental prices in Indianapolis, IN to help renters and landlords determine fair prices.

Data Sources

The data was obtained via web scraping Apartments.com using Beautiful Soup. Apartments.com is one of the leading listing websites of rental properties. We obtained a list of approximately 700 unique listings, though some of these pages contained listings for multiple rental units. The individual listing pages were imported and key property features (e.g. the number of bedrooms, bathrooms, and square footage) were extracted as well as more specific details such as the pet policy, walking score, availability of public transportation, and neighborhood information.

Data Cleaning and EDA

After data scraping, we checked the data for missing values, duplicated columns, columns not in pep8 style and multi-unit listings. The first key step was splitting multi-unit listing items on the montly_rent, bedrooms, bathrooms and square_foot. After handling multiindex listing columns, we replaced unnecessary things such as $, whitespace, comma and words on the numerical columns. After that, we filled NaN values and dropped unnecessary columns. For the EDA, we plotted distribution of the features vs monthly_rent and saw the monthly rent prices more correlated with the number of bedrooms and bathrooms as well as neighborhood location.

Process

We looked at four models: CAT boost, gradient boost, linear regression, and random forest regression. All models shared the key features of the apartment: number of bedrooms, bathrooms, square footage, and neighborhood.

We employed a standard gradient boost and a Cat boost as improvements on linear regression. Gradient boost considered bedrooms, bathrooms, square feet, transit score, and walk score, and the Cat Boost was a slight improvement on it. The Cat Boost was chosen for its robust handling of categorical terms. For this model, we also provided the total number of rooms (instead of counting them individually), the average area for rooms, the ZIP code, walk/transit score, and GPS location.

We use Random Forest Regression since we do not need to scale the data. We considered ZIP code, bathrooms, bedrooms, square_feet and top 10 neighborhoods for modeling. We also used dummy columns for neighborhoods since it is the only categorical feature . After modeling, the r2_score is 0.69. We ran multiple linear regression on subsets of ten features. The MLR model with the lowest MSE included an additional interaction term for the number of beds/bath and square footage/bath as well as the walk score.

Results and Conclusions

Currently, we can tell if a rent is approximately fair within 10%. All models have room for improvement, as errors have larger impacts for less expensive apartments. There are some shortcomings in our data. We have a relatively small number of listings. We can improve this by including listings on additional websites (Craigslist, Zillow, etc.). We may also be viewing listings at a time where they may be fluctuating due to rental specials or other real estate trends. Our minimal feature set would also benefit from expansion and more robust treatment of listing pages with multiple rental units.