Motivation

• Transactions of real estate properties involve huge amount of money
• It is important for the stakeholders to analyze the performance of market in order to avoid risk of financial loss
• Analysis would benefit previously depressed areas starting to see growth in real estate prices, and wanting to take advantage of upward trends

Objective

• To analyze historical records of property price estimates
• Find correlation between factors influencing price variations
• Predict price estimates for the near future

Prediction Model

• Linear regression model was trained on instances corresponding to years 2010 till 2016
• Testing and validation of model was done over the instances corresponding to the year 2017

Correlation Analysis

• Population estimates since 2010 till 2016 for one of the counties in western New York region were collected from the U.S. Census Bureau
• As seen in Figure 4, population estimates are observed to be inversely proportional to the average house prices
• Figure 4 explains a positive difference in people moving out and former residents returning which is a potential reason for the growth of average property prices

Data Collection, Cleaning and Analysis

• Data was collected from Zillow [1], using web crawlers implemented in Python with Beautiful Soup library
• Dataset consists of 7,204 non-commercial properties spanning across 25 zip codes
• Another dataset consisting of average property price estimates grouped by zip code is obtained
• Monthly price estimates since August 2010 till December 2017 were derived using the above two datasets. Final dataset contains 648,360 instances

Architecture

Data collection

Row Data

Pre-processing

Attributes

Zip code
Latitude
Longitude
Property type
Number of beds
Number of baths
Lot size in sq ft
Finished sq ft
Year built
Tax assessment year
Tax assessment value
Neighborhood
Current price estimate
Year-month
Zip code average price
Property price estimate

Initial training model

Training with k-fold cross validation

Trained model

Testing

Final model

Figure 1. Architecture of proposed solution (left) and list of attributes in the final dataset (right)

Figure 2. Zip code heat map for average current price estimates of properties

Figure 3. Visualization for types of non-commercial properties in the dataset

Figure 4. Population estimate versus average property prices within Cattaraugus County, NY

Results

<table>
<thead>
<tr>
<th>Consideration</th>
<th>Correctly predicted instances</th>
<th>Incorrectly predicted instances</th>
<th>Accuracy</th>
</tr>
</thead>
<tbody>
<tr>
<td>Exact price prediction</td>
<td>39,012</td>
<td>47,436</td>
<td>45.12%</td>
</tr>
<tr>
<td>Price prediction with confidence band (+/− $1000)</td>
<td>67,949</td>
<td>18,499</td>
<td>78.6%</td>
</tr>
</tbody>
</table>

Conclusion

• Model performed well when the confidence band was set to +/− $1000
• Input attributes showing higher correlation with the output variable play a major role in improving accuracy

Future Work

• Non-linear regression models can be tried out to compare the accuracy of prediction
• Other environmental factors such as population and economic indices can be considered for inputs to the regression model

References
