Real Estate Price Prediction through Machine Learning

Osaze Shears
oshears@gmu.edu

Maryam Heidari
mheidari@gmu.edu

Advisor:
Houman Homayoun

Dept. of Electrical and Computer Engineering

I. Introduction

Being able to determine the value of housing assets is critical for real estate investors adjusting their property prices due to market competition and volatility. If a house is valued too low, the investor will not gain optimal returns, but if the price is too high, prospective tenants will consider other options. This project aims to explore the prediction capabilities of several machine learning algorithms for determining the rental price of several instances of houses and other properties.

II. Methods

In order to begin predicting rental prices for real estate data, sample instances were first downloaded from Zillow.com's database of properties. To construct highly accurate models for predicting the rent price of the properties, it was decided that machine learning regression models would be built at the state, city and zip code levels. The data set used at this point in the project only included instances of houses in the state of Virginia.

Data Set Division Procedure

For each level of the data set, multiple regression algorithms were trained and evaluated. At the state level one regression model was built, while at the city and zip code levels approximately 240 and 349 models were built respectively. Within the data set there were a total of 15,342 properties, however only the houses for sale were considered in this part of the project. A standardized work flow was derived and carried out on each subset of instances in order to increase prediction accuracy.

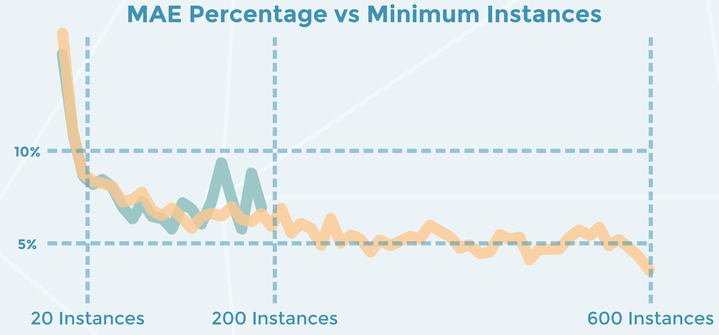
Data Set Preprocessing Work Flow

III. Results

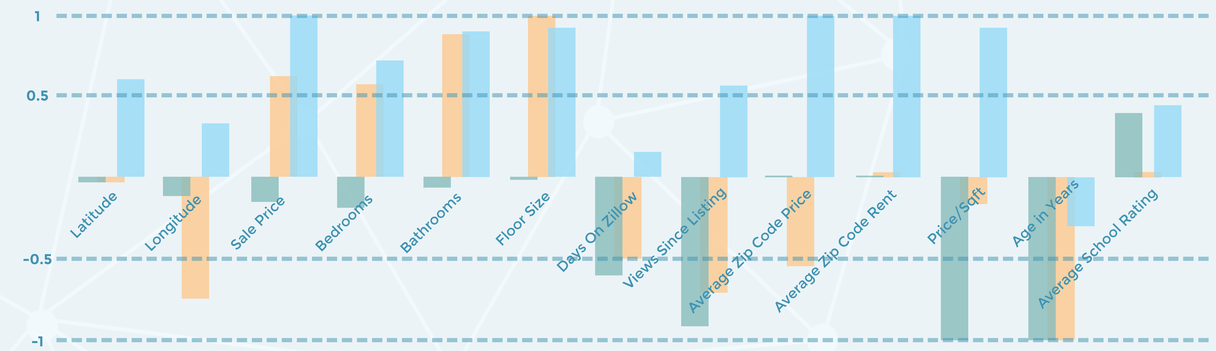
After constructing and validating the regression models using a 80%/20% Train-Test schema, it was discovered that the "Random Forest Regression" algorithm proved to yield the lowest amount of average mean absolute error (MAE) percentage in the state, city and zip code level regressions (~8.38%). When the minimum amount of instances needed for the regression model to be constructed at the city and zip code levels was increased, the average MAE percentage began to decrease and could eventually reach 5% at the city level with a minimum of 300 instances per city. The weight of each attribute per level was also evaluated during the regression processes and insights were obtained as to which attributes of the houses were the most important at each level respectively.

Average MAE Percentage per Classifier

MAE Percentage vs Minimum Instances



Relative House Attribute to Rent Price Correlation



IV. Discussion

From this research a greater understanding of the factors that influence the prices of houses in Virginia was obtained. In order to further utilize this research, the constructed regression models will be optimized for mobile, low-power processors and the algorithms will further be trained with data from other states as well as data that includes more information about non-traditional properties for rent.

V. References

[1] Zillow.com Real Estate and Mortgage Data (2017)

[2] Frank, E. et al. The WEKA Workbench. (2016)

[3] Kaufmann, M. Online Appendix for "Data Mining: Practical Machine Learning Tools and Techniques", Fourth Edition. (2016)