An Interactive Visualization Interface and Time Series Forecasting Model for U.S. Real Estate Data

Alex Ford Josh Janda John Morris Kris Parker Phillip Smith Kevin West

Summary

Team “Binary Refinery” has developed an interactive visual interface that predicts home prices and trends across the country, using the publicly available Zillow dataset with additional regional data provided from the U.S. Census and Department of Housing and Urban Development (HUD). Our customized model improves on the current Zillow ”Zestimate” and identifies local housing trends through our interactive dashboard.

Problem & Motivation

Home price projections are often based on fragmented data available through regional or national stakeholders, or through paid services accessible to real estate professionals.

Methodology is often opaque or based on qualitative factors (e.g., surveys or opinion polling) versus listing or sales data.

A wide range of models have been used to track and forecast real estate prices . Our unique innovations are:

1. The integration of macro housing indicators (e.g., inventory, percent of homes changing in value) to improve upon Zillow estimates for some zip codes
2. The addition of a user-friendly visual interface to inform decision making

Dataset

Our primary data was provided by Zillow and available for download on Kaggle.com. The dataset contains 4,393,885 observations and 76 columns, with observations beginning in April 1996 and ending in December 2017.

Missing data presented the largest challenge to an accurate model forecast. The risk was mitigated through 2 data cleaning steps:

1. Merging Zillow and HUD data for improved coverage
2. Imputing missing values by zip code, county, and/or state

Average % Missing – Original Zillow Dataset	81.2%
Average % Missing – Post-2010 Zillow Dataset	72.7%
Average % Missing – Post Imputation Dataset	3.6%

Model

The goal of the predictive model was to use macro housing market indicators to predict listing price trends in zip codes throughout the US.

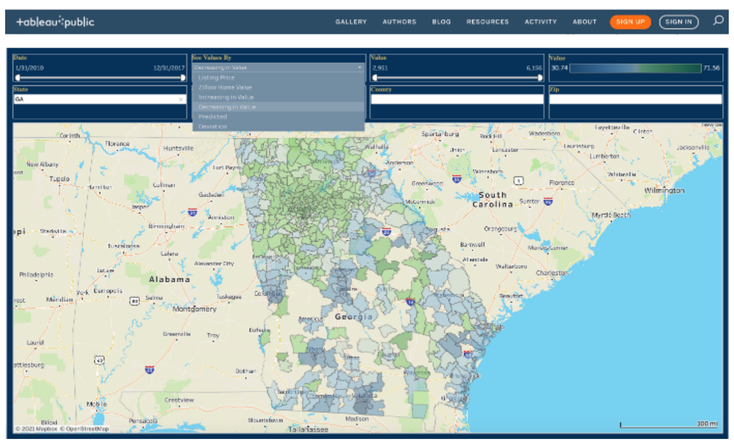
Three variable (total home inventory, percent of homes increasing/decreasing in value, and median Zillow home value index) we’re proven to have a correlation to listing price. A 6-month lag effect was built into the model to account for the slow timeline of home sales.

Several regression models were tested – including ridge regression, elastic net regression and OLS regression. OLS regression was found to produce the most promising results on average across the widest range of zip codes.

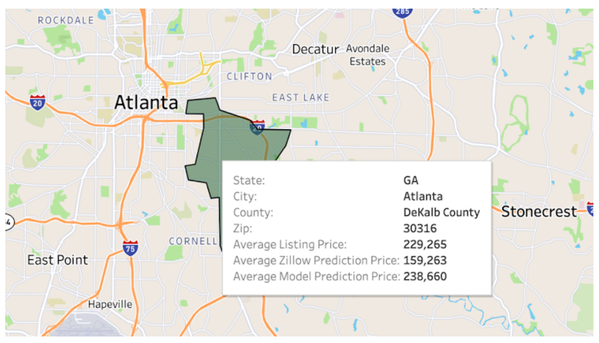
Visualization

The Tableau developed interface allows the user to choose from state, city, county, and zip code level of granularity for evaluation. The sliding scale allow the user to visualize pricing data or model predictions by date, sorting by multiple factors including the selected model, with scaling by a user-determined range of values.

https://public.tableau.com/profile/kevin.west7084#!/vizhome/GTCSE6242Spring2021BinaryRefinery/HomeValues



The Tooltip included in the visualization shows the predicted values versus average listing price and ZHVI for for each zip code.



Evaluation

While the model does not appear to be able to accurately predict actual listing values (right), it does appear to improve upon the Zillow Home Value Index in many zip codes as a predictive tool for general housing market trends.

Throughout the test period of 2016-2017, the model outperformed the Zillow model estimate (ZHVI) values in ~77% of zip codes based on mean squared error (MSE).

For more details or to continue our exploration, please visit: https://github.com/joshjanda1/CSE6242_Project