LE Lee, Nicholas Perry and Andrew Mouat
D590 - Time Series Analysis
06-MAR-2022

Analysis and Forecasting of Indiana Residential Real Estate Pricing

Project Summary

The problem that we aim to solve is the lack of tools available to potential home buyers for accurate and timely reporting of price signals. It is important to have readily available analysis and forecasting on real estate pricing because this will allow users to buy the right properties at the right time in order to maximize their investment. The web page we create will be able to assist buyers in maximizing their return on investment by allowing users to input relevant parameters to better estimate and forecast home values.

Project Description

Objectives:

The objective of this project is to create a web-based application that allows users to generate price predictions for the current value of a home and value forecasts for residential properties within an input, Indiana-based zip code. Furthermore, users will have the ability to enter various parameters, such as number of bedrooms, number of bathrooms, square footage, age of house, condition of house, etc. to refine the tools accuracy. In addition to historical pricing data (see “Dataset” section below) and user input, our model will incorporate other external, influencing factors including, but not limited to, mortgage interest rates, construction spending and total home inventory. In order to determine the prediction for the current value of the home, machine learning methods such as K-Nearest Neighbors and Random Forest,

amongst others, will be explored. Then, price forecasts will be determined using a dynamic regression time series model.

This tool will allow prospective buyers/investors the opportunity to independently compare the current listed value of a property to the predicted and forecasted value from our model to make informed decisions. Additionally, the tool can be used by current home owners and real estate agents to determine a listing price for their property that is reasonable and competitive with the rest of the local market.

Usefulness:

The housing market is noticeably susceptible to changes over time. These fluctuations have a vast impact on local economies, personal finances, and general public perceptions of the economy as a whole. Forecasting real estate prices can help stakeholders (buyers, sellers, investors, real estate agents, etc.) in the decision-making process of purchasing and selling properties. Our web tool will show the viability of investments for future profit. It will also help to identify undervalued homes, which could be exploited for profit through real estate arbitrage. Finally, it will also assist users by properly valuing their current primary residence and their potential next home. While Zillow has an estimated home value, or "Zestimate," to help users value properties, it does not have them for every property. The estimates are also tied to individual properties and so they do not allow a user to get an idea of the accepted price range for homes in an area with certain parameters. Zillow also does not provide future forecasts of prices based on time series principles (seasonality, trends, etc.).

Dataset:

In order to implement a tool that meets all of the objectives outlined for our project (see "Objectives" section above), our team will require data that falls into two over-arching categories: Home Price Data and Influencing Factors Data. Home Price Data includes data associated with historical prices of homes in addition to the key metadata our team is interested

in (number of bedrooms, number of bathrooms, square footage, etc.). Influencing Factors Data is data that will be used as inputs to the dynamic regression model to account for explanatory factors (interest rates, home inventory, etc.).

Home Price Data

Initially, our team will utilize research data provided by Zillow (https://www.zillow.com/research/data/) to build the overall, generalized framework for our project. However, the ready-to-use data provided by Zillow only provides pricing data at the regional level and does not provide the granularity required by our team to allow for implementation of the additional user inputs (number of bedrooms, number of bathrooms, square footage, etc.). Therefore, our team will develop a custom script that will perform web-scraping of Zillow home listings to generate a custom dataset including desired factors.

Influencing Factors Data

As we progress through the project, our team may elect to investigate additional factors . However, as a start, the three primary influencing factors that our team will investigate are: mortgage interest rates, construction spending and total home inventory. Below, an initial data source is provided for each of these factors. As needed, additional data sources will be integrated or custom data sources generated to fit the requirements of our project.

- Mortgage Interest Rates - FreddieMac, a government-sponsored enterprise that purchases, guarantees, and securitizes home loans, provides historical data (weekly) on mortgage interest rates (https://www.freddiemac.com/pmms).
- Construction Spending - The U.S. Census Bureau puts out a monthly report on new domestic construction spending activity, by dollar value, in the country (https://www.census.gov/construction/c30/c30index.html). The report gives a

breakdown by residential and nonresidential spending, as well as by private and public spending.

- Total Home Inventory - The Federal Reserve Economic Data (FRED) website (https://fred.stlouisfed.org/categories/28041) provides county-by-county total number of listings data for Indiana.

Functionalities:

As described above, the primary functionality of our tool will be to provide a web-based means for a user to input key information about a property: location information, number of bedrooms, etc. and then output a price prediction based on historical data. In addition, our tool will utilize a time series model to provide future price forecasts (with various confidence levels), taking into account a variety of external factors such as mortgage interest rates, inventory, etc. A stretch goal for our team will be to allow the user to put in multiple entries at once to allow for side by side comparison ("What-If" analysis). The tool will output the results of the forecasting analysis in an easy to interpret, aesthetically pleasing data visualization that includes the historical data used and the forecasted horizon. A stretch goal would be for the visualization to be interactive.

In addition to the functionalities described above, a stretch goal for our data processing workflow would be the functionality to engineer custom features from the input data sources and determine if they positively impact the model accuracy. An example of a feature that may be custom engineered could be"Proximity to Amenities". This could be accomplished using Natural Language Processing to scrape pertinent key words (park, restaurants, etc.) from the Zillow listing or by using geographic coordinates of the property. This could be compared to the Zillow Walk Score for validation.

Task Divisions:

In an effort to ensure each member of our team has the opportunity to develop the broad range of skills that will be required to implement this project, we have elected not to explicitly assign individuals to certain aspects of the project. Instead, tasks required to be completed by each team member will be discussed and assigned on a weekly basis based on a weekly check-in meeting (see "Communication and Sharing" section below).

Communication and Sharing:

For Part 1, our team communicated through a group email chain and will continue to do so throughout the duration of the project. However, in addition, our team has established a WhatsApp group chat for instances where email is not the appropriate channel of communication.

In addition to communications via email and WhatsApp, our team will meet via Zoom once per week (Sunday evenings), at a minimum. This weekly meeting will be utilized to discuss barriers during project implementation, assign tasks based on upcoming milestones (see "Milestones" section below) and have working sessions as a team on project deliverables.

Project deliverables and working files are to be stored in two collaboration environments that all team members have access to.

1. GoogleDrive: Files associated with administrative files (reports, presentations, references, etc.) will be managed via a shared GoogleDrive, allowing for simultaneous collaboration amongst team members.
2. GitHub Repository: Files associated with code development will be managed via a GitHub repository. This will provide a suitable environment for version control and collaboration amongst team members. Our team intends on utilizing R and Python, leveraging the extensive, existing libraries that are available for data

import and export, data pre-processing, time series modeling and forecasting, and data visualization.

Milestones:

Our plan is to use Week 9 for data scraping, gathering, and cleansing. Week 10 will be used for exploratory data analysis and preliminary visualizations of the data. Week 11 will include initial model designs and evaluation, as well as submission of Part 2. Week 12 will be spent refining and further evaluating the model from Week 11. Week 13 will be used to mock up the web design and for back-end development, as well as submission of Part 3. Week 14's work will include further back-end development, implementation of user interface, and model finalization. Week 15 will be used for full deployment and the submission of a working demonstration.