CX 4242 Report
United States Real Estate Visualization
Team 4: The Underdogs

Intro and Motivation (Heilmeier Q's 2, 4, 5)

Housing is a huge driving factor of the economy. In 2011, the total value of residential properties in the US was $16 trillion [13]. In 2020, it was worth $36.2 trillion, meaning that it more than doubled over the past 9 years. Any American who owns a home would be interested in how the market will change in their area. Furthermore, investors looking for the hottest place to buy real estate want to know where they should grow their business and where they should avoid.

When an individual wants to sell a house, much of this decision making is done by the real estate agent that sellers hire. The agent will generally conduct a Comparative Market Analysis (CMA) which is based largely on subjective features of the house, leading to inconsistencies in valuation across agents. In practice, buyers and sellers often are forced to speculate on property prices with little knowledge of how property prices are changing in their area.

Our project will make an impact on the stakeholders by increasing their knowledge of the market and thus increasing their confidence when they make decisions. We can quantitatively measure this impact with user surveys. For example, survey a pool of home sellers and ask how confident they are in their list price; then give the seller our tool to use for five minutes and ask how confident they are again. This is one procedure that we could use to quantify the impact of our tool.

Problem Definition

How can we assist and easily inform different stakeholders (buyers, sellers, investors) when they make economic decisions related to real estate and future housing plans by creating a user-friendly web-app that users can leverage in under 60 seconds?

Literature Survey

As a housing price prediction model is so important, there has been a significant amount of research done, with a rapidly growing focus on using Machine Learning.

Non machine learning prediction

[8] uses a macroeconomic dynamic stochastic general equilibrium model (DSGE) based on behavioral economics. The model relies on a multitude of variables, and, as such, relies on many assumptions about the relationships between those variables. [13] is a review of many other papers that attempt to predict price using variables such as rent-to-price ratio, vacancy rates, and monetary policy and concludes that although these may be accurate for capturing the current state of the market, it struggles to predict future prices.

Machine learning prediction

Robey et. al. found that a multiple linear regression was able to predict housing prices with an r^2 of over 90% [1]. Also, certain regression methods such as Support Vector Regression[15] or Fuzzy Least-Squares Regression[16] are found to generate accurate and efficient real estate price prediction models that do not use excessive amounts of variables. By comparing MSE (mean squared error), we can choose the best model[11]. With various regression models, it's effective to take the weighted mean of different models to get more accurate results[10]. Multiple other papers such as [7] - which studied housing prices in Beijing using a number of physical attributes about the house - and [3] - which used macroeconomic variables to determine house price in London - found that a Random Forest Model most

accurately predicted the price in their holdout set, however it may be prone to overfitting. [9] used 27 different variables to run multiple machine learning algorithms to determine closing price of houses, to determine that RIPPER was the best classifier to determine if the closing price was above or below the listed price. Papers in this section provide us with several different machine learning methods from general, like random forest and XGB[14], to advanced, like merged models and weighted combinations or RIPPER which is designed for predicting house price. These papers gave us ideas for which models can accurately predict real estate data but fall short because we will do time-series analysis on price alone.

Visualization

One of the important aspects of our project will be the visualization of the predictions from the model. Bao et. al[2] and M.Li[12] created a visualization tool called *HomeSeeker* for users to interactively search for properties based on user-specified requirements. [4] is about balance between information loss and complexity of the graph when visualizing a large data set. Some papers introduce mainstream tools such as excel and tableau[6]. Other papers go further in detail, like interactive user interface and cluster analysis[5]. There are some powerful visualization tools available that specialize in geographic visualization, such as ArcGis that could provide effective visualization for housing price comparison using representations such as Choropleth map [17]. One powerful visualization method is NDS, an interactive web-based system built with D3 and Mapbox GL JS[18]. Those resources helped us perform a logical and concise visualization for real estate. However, those papers were not designed for time series regression visualization, so some ideas may not be the most appropriate for this project.

Project Summary (Heilmeier Q's 1, 3, 6, 7, 8, 9)

Our team collected real estate data to perform time-series modeling with ARIMA(autoregressive integrated moving average) in order to predict future changes in property values across the United States, and created a visual tool to illustrate our findings.

This project created a tool that people can use to easily access how property prices have changed in the past and get an estimate for how much they will change in the future with forecasts created by a machine learning algorithm. By leveraging data visualization, consumers will have a revolutionary way of getting a "feel" for the real estate market in their state over time, with the ability to refine their results based on the type of property that they are concerned with. Ultimately, the tool is unique in that our users should be able to draw conclusions from our data in less than 60 seconds and come out with a better understanding of the real estate market.

The biggest risk is that there is no 100% accurate method of forecasting housing prices. Users should understand that the forecasting only takes historical data into account and it doesn't account for unexpected future events. However, the payoff is a potentially huge user-base who are interested in our tool and our predictions. Our data sources are free to use and the prototype should be free to deploy for beta testing.

Proposed Method

Housing price forecast was conducted using the ARIMA model, or Autoregressive Integrated Moving Average. We explored different machine learning methods as well as non-machine learning methods. One of the biggest candidate methods was using a linear regression model on our time series property price data. However, knowing that many property price forecasts based on linear regression were already available in the real estate market such as Zillow who creates a Zestimate for each house, we wanted to use a different approach.

Compared to other candidate models, ARIMA model sets an emphasis on capturing the autocorrelation present in the data. Autocorrelation is the measure of how each observation is related to

the observation of its recent past. For the housing data, it is reasonable to say that tomorrow's property price is very likely to be related to today's property price. Thus, we decided to use the ARIMA model among other models that could be less accurate as they do not capture autocorrelation. Parameter selection on the ARIMA model will be discussed later in the Experiments/Evaluation section.

Visualization is created using D3.js. We created a choropleth map giving a visual of real estate prices in each on a color gradient with features like a tooltip and parameter selection. This approach is unique in that we combine macro real estate data visualization on gradient map with advanced machine learning techniques, allowing users to quickly gain insight from historical and forecasted trends in the real estate market.

Although there are several real estate marketplaces out there such as Zillow, Trulia, and realtor.com that provide visualization for historical price of the property, our approach is intuitive and innovative for several reasons. The state of the art techniques either report prices on an individual or local level but miss out on national trends or they are able to report national trends but fail to visualize the data to be easily accessible to non-expert users such as homeowners. Also, current real estate marketplaces try to avoid providing forecasted data as they are at a higher risk of losing their reputation and potential customers when they fail to forecast the property prices correctly. Therefore, we are unique in that we provide statewide and national trends of property price changes as well as their forecasted data for next 5 years, combined with a visualization friendly enough to every user as they are accessible and interactive.

Following is description of our approach:

1. Download data from Kaggle

We will use this data set from Paul Mooney created a year ago:
https://www.kaggle.com/paultimothymooney/zillow-house-price-data
To be specific, only 15 files will be used, which are 5 different bedroom size data of City_zhvi, State_Zhvi, and State_MedianRentalPrice. The data set after merging all 15 files has about 105,000 rows and size is approximately 200 MB.

2. Analyze the data

291 columns of the data set is the price observed monthly since January of 1996 upto March of 2020. The remaining columns are information about the region, such as region id, type, and its belonging state as well as its rank in terms of population.

3. Build a forecasting model

The model we used is the ARIMA model, or auto regressive integrated moving average as explained above. ARIMA is composed of three parts: auto regressive(AR), integrated(I), and moving average(MA). Auto regressive model is a time series model that uses past forecasts to predict the value at the next time step. Moving average model is a time series model that uses past forecast errors to predict the next time step. Integrated(I) is a tool to smooth the time series.

4. Train model

We used the ARIMA package from the statsmodel python library to create the ARIMA model. This model would forecast the housing price of every housing data by city and the average housing price of each state. After training the model with Zillow historical time series housing data as an input, the model is trained and ready to be used for prediction. Then, we used future years as an input for a trained ARIMA model to forecast future housing prices. The python script writes future housing prices for each city and average housing price for each state into csv files.

5. Build a user interactive web app using d3.js.

A user-interactive visualization map was created using HTML, Javascript and d3.js. We created a choropleth map of the U.S at the state level by using GeoJSON. The user interactive web app has several features for users to apply filters and interact with the map. The first feature is a dropdown list for users to select the house type by number of bedrooms: 1 bedroom, 2 bedrooms, 3 bedrooms, 4 bedrooms, and 5 bedrooms + and a year. The second feature is a chart visualization and a tooltip when a user hovers over a state in a choropleth map. The tooltip includes median rental price and median house price for a selected state. Two charts pop up at the bottom when a user hovers over a state and the first chart on the left shows historical and predicted housing prices for a selected state. The second chart on the right shows only historical rental prices for a selected state. The third feature is a color scale of each state in a choropleth map by median housing price to make a map more intuitive for users. By adding various features, the map is more user-engaging and it conveys information to users in various ways: number, chart, color, and a map.

6. Connect a webapp to historical and predicted data.

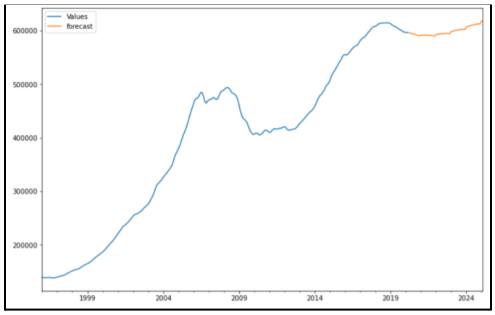
After testing visualization with city and state level housing price data, we chose to use state-level housing and rental price for optimized user-experience. Representing city-level house prices on the choropleth map seemed too specific and wasn't visually appealing. However, city-level forecasted house price data is stored in a dataset folder for a reference.

7. UI / UX revision based on user feedbacks

After conducting a user survey, we made UI/UX revisions based on user feedback each iteration.

Experiments

First Experiment: Proof of Concept



↑ Sample Arima Forecast (New York, NY); blue - historical data, orange - forecast

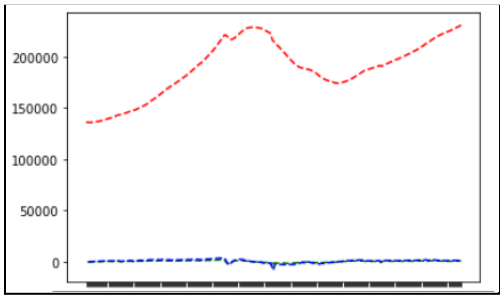
Questions:

1. Is the ARIMA model a good option to pursue for the project?
2. How hard is it to train the ARIMA model to our data?

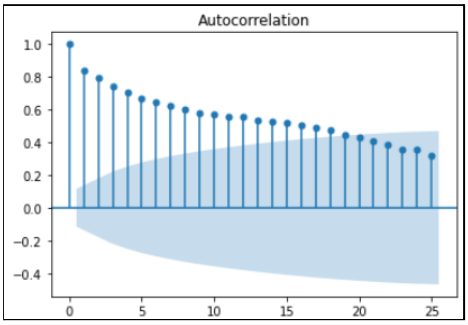
In our first experiment, we wanted to see if the Autoregressive Integrated Moving Average (ARIMA) would actually be able to create a reliable forecast for the next 5 years, which is March of 2020 to February of 2025. The graph above is an example of ARIMA model forecasting performed on one sample data observation, which is New York City of the state of New York. As we can observe, the forecast successfully captured the ongoing trend of decreasing slightly as well as the overall trend of increasing property prices. We generated numerous plots based on different cities, and all group members agreed that the forecast output of the ARIMA model looks valid and the resulting forecast would be reliable.

Also, we realized that it is not difficult to train the ARIMA model to our data. We were able to create code on jupyter notebook to filter the necessary columns and rows and use the functions from statsmodel library to create a 60 months long forecast.

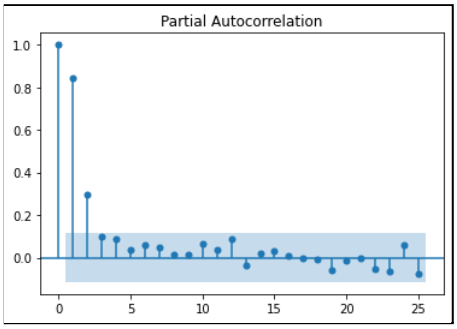
Second Experiment: Parameter Selection and Seasonality



← Differencing plot; Red - 1st order differential, blue - 2nd order differential



← ACF plot



← PACF plot

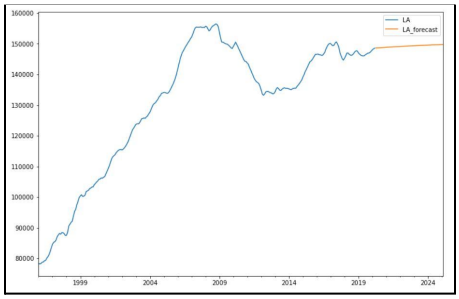
Questions:

1. What parameters should we select for the ARIMA model?
2. Does adding seasonal parameters improve the forecasting?

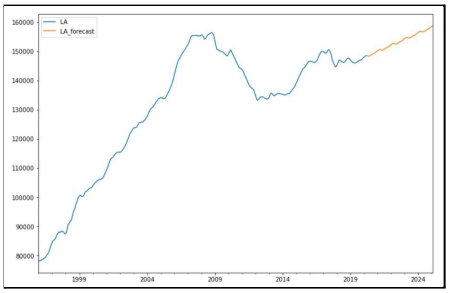
Firstly, we needed to decide what parameters we should select for the ARIMA model. The ARIMA model has three parameters, which are (p, d, q). p represents the number of lag observations in the AR model. d represents the number of differentiation. q represents the number of lag observations in the MA model. Firstly, we selected d. From the differencing plot above, we observed a significant improvement from the original data to the first order differential, while we observed nearly no difference between the first and second order differential. Therefore, we selected d = 1. Then, we selected p and q. General rule of thumb is p + q < 2 and p * q = 0. Since ACF plot shows a steady decrease of value while the PACF plot shows rapid decrease after lag 2, we concluded that our time series data resembles the feature of AR model and thus selected p = 0 and q = 1.

Also, we figured that the seasonality exists within the data after some testing. As you can compare from the two graphs below, which are the sample from the state of Louisiana, the seasonal ARIMA model generates a more convincing forecast than a non-seasonal ARIMA model. So we decided to add a seasonal parameter to create a forecast. We set our seasonal parameter to be (1,1,0,12) since the ACF is positive at the first significant lag, the series has a stable seasonal pattern over time, and observations were made 12 times a year.

In conclusion, we set our parameter to be (p,d,q) (P,D,Q,m) = (0,1,1) (1,1,0,12).



← No seasonality



← Seasonality

Third Experiment: User Impact

Questions:

1. How much value do users derive from our tool?
2. What are users' feedback from our visualization?
3. How much do users learn by using our tool?

In order to measure the impact of our product, our group conducted a user survey with students at Georgia Tech. The procedure was as follows: a new user will be asked on a scale from 1 to 10 how confident they are in their understanding of the national real estate market; then the user will be given our tool to use however they like for 10 minutes; then they will be asked to rate their confidence again on the same scale. This will measure the impact of our tool on the user and serve as a proof-of-concept for our product.

User Impact Survey Results:

Our group performed the user impact survey as described above. We surveyed 22 participants and measured the difference in their knowledge of the U.S. real estate market before and after using the web tool. On average, users reported their knowledge of the market increasing by 3.72 points on the 1-10 scale. The majority of participants reported their knowledge increased by 2 or 3 out of 10 points. Before using the tool, participants had an average knowledge of 2.7 points out of 10 and after using the tool participants reported an average knowledge of 6.4 points out of 10.

This user impact survey proves that our tool provides valuable insights into the United States real estate market. Every user reported an increase of their knowledge in this market which is proof that this tool could be valuable to a wide range of potential users.

Distribution of team member effort: All team members have contributed a similar amount of effort.

Conclusion

Overall, this project was challenging but successful. Our team had a lofty goal in mind, which was to build a visualization that leverages machine learning to provide insights in the real estate market. We are proud of the final product and the feedback we have received from user testing has been positive. We created an interactive web tool that allows users to explore historical and forecasted trends in the U.S. real estate market on a macro level, accomplishing most of our vision in doing so.

Our team overcame many obstacles in completing the project. We learned about developing code in a medium-sized team, which was challenging at first. We faced tough decisions when it came to dividing and conquering this project. Our team learned quickly that we needed to leverage each of our individual strengths to tackle each phase of the project efficiently.

So what’s next with this project? We believe that real estate data visualization has a strong demand and not many companies are supplying these types of insights. The fact that our team was able to create such a powerful and insightful tool in just one semester shows how much potential there is for data scientists to create products in the real estate segment.

This project gave our team a great opportunity to apply the tools we learned in CX 4242. It was a great way to wrap up the class we’re proud of the final product.

- # References

[1] S. Robey, M. McKnight, M. Price and R. Coleman, "Considerations for a Regression-Based Real Estate Valuation and Appraisal Model: A Pilot Study", *Accounting and Finance Research*, vol. 8, no. 2, p. 99, 2019. Available: 10.5430/afr.v8n2p99 [Accessed 12 October 2021].

[2] M. Li, Z. Bao, T. Sellis and S. Yan, "Visualization-Aided Exploration of the Real Estate Data", *Lecture Notes in Computer Science*, pp. 435-439, 2016. Available: 10.1007/978-3-319-46922-5_34 [Accessed 12 October 2021].

[3] S. Levantesi and G. Piscopo, "The Importance of Economic Variables on London Real Estate Market: A Random Forest Approach", *Risks*, vol. 8, no. 4, p. 112, 2020. Available: 10.3390/risks8040112 [Accessed 12 October 2021].

[4] E. Gorodov and V. Gubarev, "Analytical Review of Data Visualization Methods in Application to Big Data", *Journal of Electrical and Computer Engineering*, vol. 2013, pp. 1-7, 2013. Available: 10.1155/2013/969458 [Accessed 12 October 2021].

[5] M. Khotilin and A. Blagov, "Visualization and cluster analysis of social networks", 2021. .

[6] S. Ali, N. Gupta, G. Nayak and R. Lenka, "Big data visualization: Tools and challenges", *2016 2nd International Conference on Contemporary Computing and Informatics (IC3I)*, 2016. Available: 10.1109/ic3i.2016.7918044 [Accessed 12 October 2021].

[7] Q. Truong, M. Nguyen, H. Dang and B. Mei, "Housing Price Prediction via Improved Machine Learning Techniques", *Procedia Computer Science*, vol. 174, pp. 433-442, 2020. Available: 10.1016/j.procs.2020.06.111 [Accessed 12 October 2021].

[8] Y. He and F. Xia, "Heterogeneous traders, house prices and healthy urban housing market: A DSGE model based on behavioral economics", *Habitat International*, vol. 96, p. 102085, 2020. Available: 10.1016/j.habitatint.2019.102085 [Accessed 12 October 2021].

[9] B. Park and J. Bae, "Using machine learning algorithms for housing price prediction: The case of Fairfax County, Virginia housing data", *Expert Systems with Applications*, vol. 42, no. 6, pp. 2928-2934, 2015. Available: 10.1016/j.eswa.2014.11.040 [Accessed 12 October 2021].

[10] A. Varma, A. Sarma, S. Doshi and R. Nair, "House Price Prediction Using Machine Learning and Neural Networks", *2018 Second International Conference on Inventive Communication and Computational Technologies (ICICCT)*, 2018. Available: 10.1109/icicct.2018.8473231 [Accessed 12 October 2021].

[11] G. Milunovich, "Forecasting Australia's real house price index: A comparison of time series and machine learning methods", *Journal of Forecasting*, vol. 39, no. 7, pp. 1098-1118, 2020. Available: 10.1002/for.2678 [Accessed 12 October 2021].

[12] M. Li, Z. Bao, T. Sellis, S. Yan and R. Zhang, "HomeSeeker: A visual analytics system of real estate data", *Journal of Visual Languages & Computing*, vol. 45, pp. 1-16, 2018. Available: 10.1016/j.jvlc.2018.02.001 [Accessed 12 October 2021].

[13] E. Ghysels, A. Plazzi, R. Valkanov and W. Torous, "Forecasting Real Estate Prices", *Handbook of Economic Forecasting*, pp. 509-580, 2013. Available: 10.1016/b978-0-444-53683-9.00009-8 [Accessed 12 October 2021].

[14] J. Manasa, R. Gupta and N. Narahari, "Machine Learning based Predicting House Prices using Regression Techniques", *2020 2nd International Conference on Innovative Mechanisms for Industry Applications (ICIMIA)*, 2020. Available: 10.1109/icimia48430.2020.9074952 [Accessed 12 October 2021].

[15] J. Y. Wu, “Housing price prediction using support vector regression,” *San Jose State University*, May 2017. Available: 10.31979/etd.vpub-6bgs [Accessed 12 October 2021].

[16] A. G. Sarip, M. B. Hafez, and M. N. Daud, “Application of fuzzy regression model for real estate price prediction,” *Malaysian Journal of Computer Science*, vol. 29, no. 1, pp. 15–27, 2016. [Accessed 12 October 2021].

[17] M. Nöllenburg, “Geographic visualization,” *Human-Centered Visualization Environments*, pp. 257–294. [Accessed 12 October 2021].

[18] Y. Lan, E. Delmelle, and E. Delmelle, “NDS: An interactive, web-based system to visualize urban neighborhood dynamics in the United States,” *Journal of Maps*, vol. 17, no. 1, pp. 62–70, 2021. [Accessed 12 October 2021].