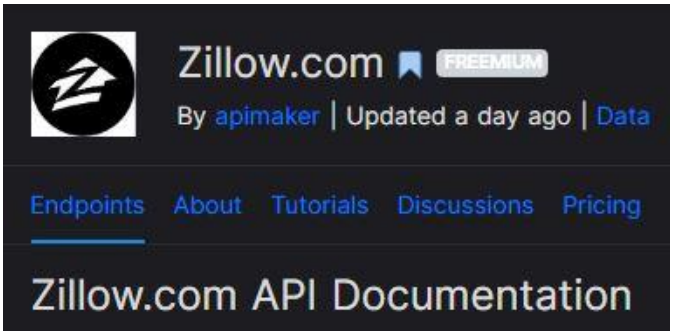
Name: Khoa Truong
Date: 09/30/2022
Abstract, Motivation, Problem:

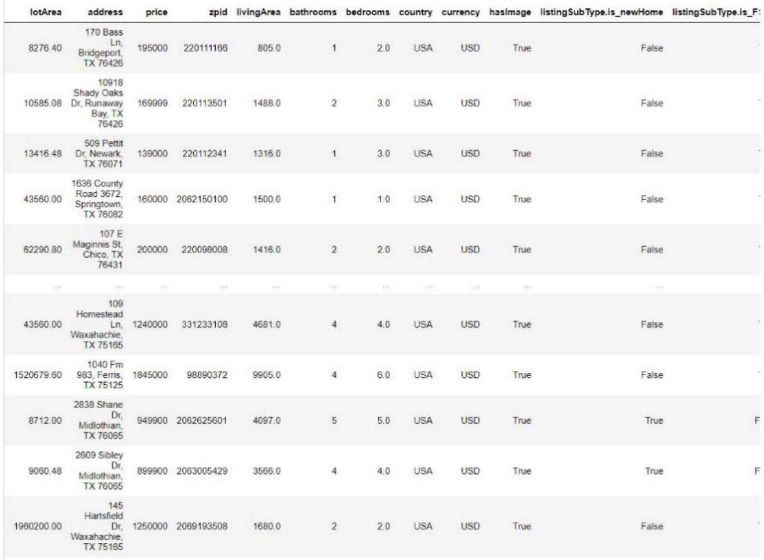
Financial market has been a hot topic since January 2021 due to massive movements in stocks' prices. For example, stocks such as SPY and AMD rallied more than 30% upward and then dropped from their peaks significantly and still couldn't reach that peak around 2022. This shows that the market has been volatile as ever and this can also be seen in the housing market. For instance, based on Zillow, a single-family house can be $450k in June and then dropped to $399k in August. Moreover, in the 2020 pandemic, some houses' prices dropped 7-10% within a month. Therefore, even though it's not as volatile as stocks, housing prices have moved significantly over these years. This ends up with people don't know when they invest in houses or simply buy one for living. Hence, these scenarios pique my interest and I want to create a ML model that predicts housing prices monthly as best as possible. I have to admit that there are many articles have tried their best to predict house prices using multiple different techniques, from linear regression to artificial neural networks (**House Price Prediction: Hedonic Price Model vs. Artificial Neural Network, Limsombunchai**, **V.**). Most people also use features of the houses such as bathrooms and bedrooms to improve the predictions (**House Price Forecasting Using Machine Learning, Alisha Kuvalekar**). Moreover, there are researchers that use historical house prices to predict future prices for a specific area with time series models. Although the data, specifically features that are used in these papers, are necessary components, they are not enough to predict house prices. In an economy, there are many external factors that can heavily affect housing prices, especially the Federal Funds Rate. The amount of interest that banks charge while lending can also affect housing prices. That's why in this capstone project, I will try my best to incorporate major economic indicators into the housing dataset. Then, we can begin to look into different ML methods of predicting prices. Just a head up, we will only look at single-family houses since these are commonly traded real estate properties. Therefore, prices are usually up to date with the state of the economy.

Datasets (Types, basic statistics, quality, etc.):

To get real and up-to-date house data, we have to get information from current single-family house sales on Zillow. To do so, we are pulling Zillow data with a website API called 'RapidAPI' to get the latest house prices and properties at the end of every single month around 28th or 29th. The reason is that by getting house prices at the end of the month, we manage to capture all of the price movements of that one month.

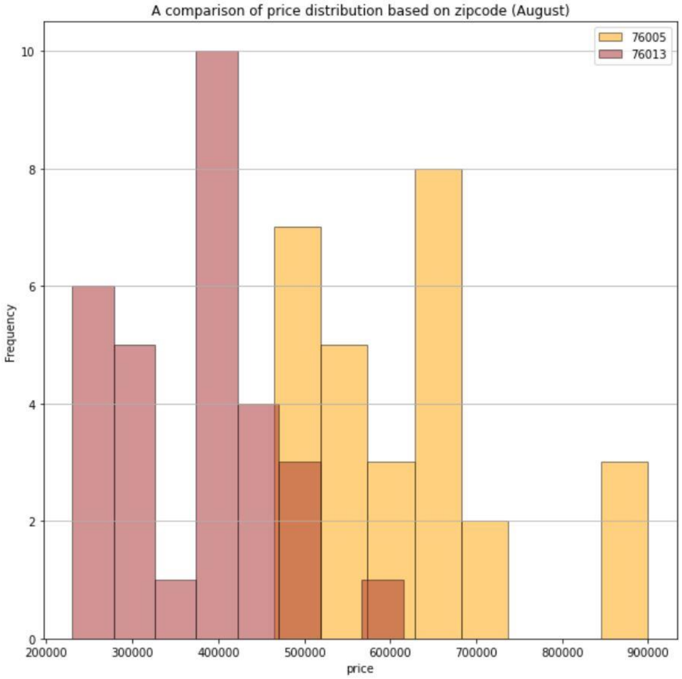

Zillow.com
FREEMIUM
By apimaker | Updated a day ago | Data
Endpoints
About
Tutorials
Discussions
Pricing
Zillow.com API Documentation

Moreover, this first pull only gets us the fundamental properties of houses such as the living Area, number of bedrooms, bathrooms, etc. So, we have to clean the data to ensure that we get the datapoints with non-null essential features. For instances, every single house has a specific id called 'zpid', a Zillow id that represents a house. A row without that should not be included in the dataset. This idea can also apply to rows with null or zero lot area. We also have to remove row with address equals to ‘Available Soon’ since these are listings only represent house blueprints. Consequently, we end up with a dataset contains only the necessary houses that we want to analyze. Here is a subset of the first dataset:



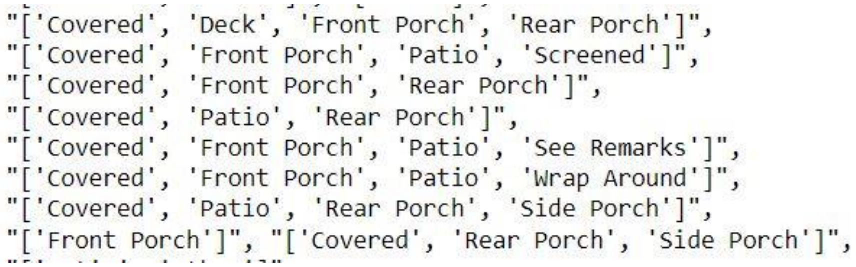
lotArea	address	price	zpid	livingArea	bathrooms	bedrooms	country	currency	hasImage	listingSubType.is_newHome	listingSubType.is_F:
8276.40	170 Bass Ln, Bridgeport, TX 76426	195000	220111166	805.0	1	2.0	USA	USD	True	False	
10585.08	10918 Shady Oaks Dr, Runaway Bay, TX 76426	169999	220113501	1488.0	2	3.0	USA	USD	True	False	
13416.48	509 Pettit Dr, Newark, TX 76071	139000	220112341	1316.0	1	3.0	USA	USD	True	False	
43560.00	1636 County Road 3672, Springtown, TX 76082	160000	2062150100	1500.0	1	1.0	USA	USD	True	False	
62290.80	107 E Maginnis St, Chico, TX 76431	200000	220098008	1416.0	2	2.0	USA	USD	True	False	
...	...	...	...	...	...	...	...	...	...	...	
43560.00	109 Homestead Ln, Waxahachie, TX 75165	1240000	331233108	4681.0	4	4.0	USA	USD	True	False	
1520679.60	1040 Fm 983, Ferris, TX 75125	1845000	98890372	9905.0	4	6.0	USA	USD	True	False	
8712.00	2838 Shane Dr, Midlothian, TX 76065	949900	2062625601	4097.0	5	5.0	USA	USD	True	True	F
9060.48	2609 Sibley Dr, Midlothian, TX 76065	899900	2063005429	3566.0	4	4.0	USA	USD	True	True	F
1960200.00	145 Hartsfield Dr, Waxahachie, TX 75165	1250000	2069193508	1680.0	2	2.0	USA	USD	True	False	

After collecting the necessary data points, we begin to do multiple API pulls from the same API. The number of times we call the API is equal to the length of the first curated dataset. Each pull represents a house with more than 50 features. However, only some features affect the housing prices. For example, ‘listing provider’ column contains more than 3000 names, and each monthly pull ends up with more different names than these 3000 names. Therefore, we can’t do categorical encoding on this column, and we have to drop it. ‘Zipcode’ column is too important to the dataset since house prices always depend on the location. For example, this is the graph of price distribution between 2 zip codes: 76005 and 76013.



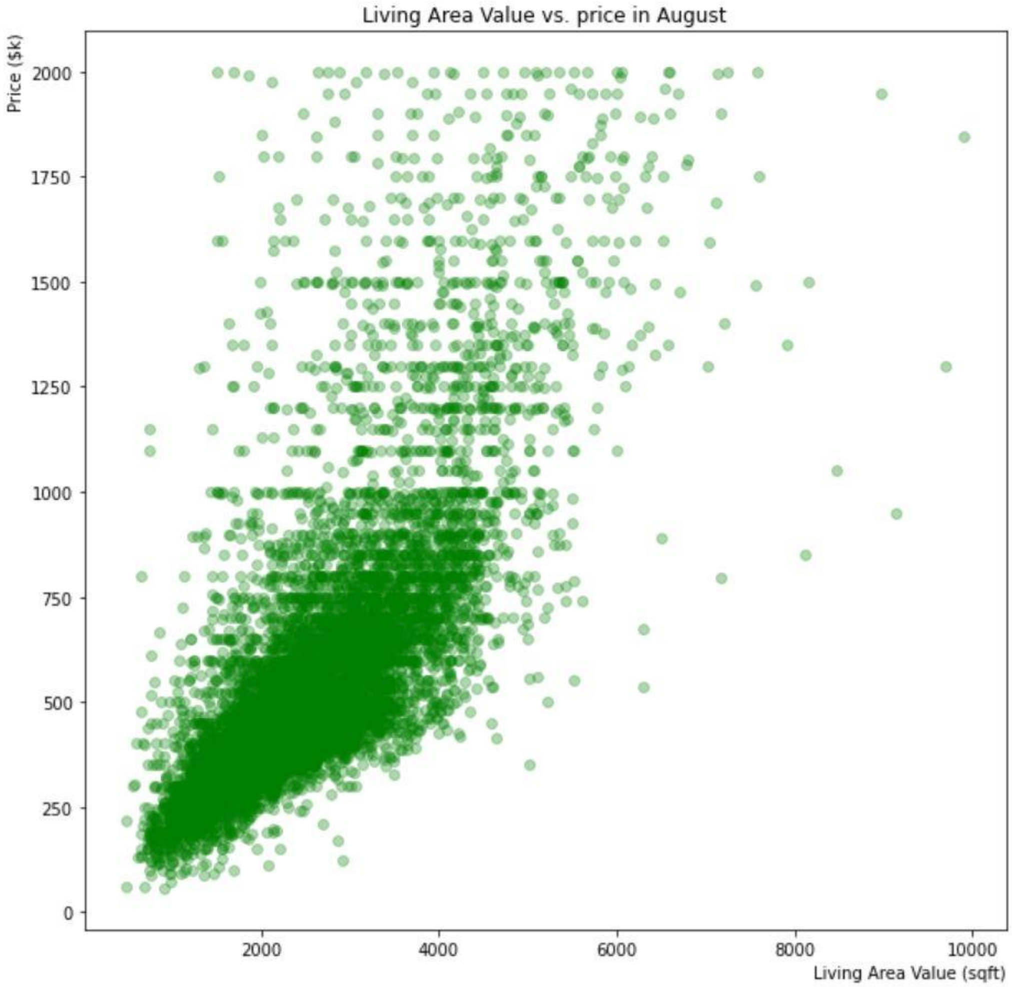
```
The mean and median of houses's prices in 76005 (August): 618637 | 587373
The mean and median of houses's prices in 76013 (August): 374713 | 387000
```

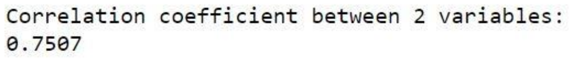
A simple comparison of price distribution of 2 zip codes shows that just having zip codes as our categorical features can significantly help us predicting prices. There are a few features that contains a list of variables that describes on feature. For example, ‘patio and porch’ feature contains a list of descriptions of what it looks like. Some houses can have ['Covered', 'Front Porch', 'Patio', 'Wrap Around'] while others have ['Covered', 'Rear Porch', 'Side Porch']:



```
"['Covered', 'Deck', 'Front Porch', 'Rear Porch']",
"['Covered', 'Front Porch', 'Patio', 'Screened']",
"['Covered', 'Front Porch', 'Rear Porch']",
"['Covered', 'Patio', 'Rear Porch']",
"['Covered', 'Front Porch', 'Patio', 'See Remarks']",
"['Covered', 'Front Porch', 'Patio', 'Wrap Around']",
"['Covered', 'Patio', 'Rear Porch', 'Side Porch']",
"['Front Porch']", "['Covered', 'Rear Porch', 'Side Porch']",
```

Hence, we have to spend some time to do label encoding with these types of variables. Overall, many categorical features in this dataset are too important to ignore. As more categorical features we get, the better the ML model performs. Next, there are only a few quantitative features within this dataset. Moreover, only one or two of them matter. For instance, this is the scatter plot of living area (square feet) vs. price in August:

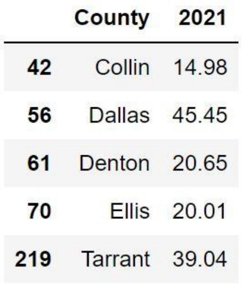




```
Correlation coefficient between 2 variables:
0.7507
```

As you can see, the correlation between these 2 variables is very high considering that we haven't done any transformations to our features. However, we noticed that there are outliers such as a house with only 2500ish living area, but the price is $2 million. Of course, there are reasons why the house is priced that way.

Next, we have to collect economy indicators such as Fed Funds Rate and CPI data over many years if we want to do a time-series analysis for houses. The reason is that these indicators can tell us if the economy is in a good or bad position (https://fred.stlouisfed.org/series/FEDFUNDS). For instance, high CPI usually represents that the economy is experiencing inflation so house prices would increase. However, the Fed will increase Fed Funds Rate to slower the economy activity, which will lower the house prices. We also have to look at crime rate within a specific county. A place with low crime rate would definitely increase house prices and vice versa. Therefore, I will use the latest crime data that I can get my hands on. This website publishes all types of crimes occurred in a Texas county 2021 (https://txucr.nibrs.com/Home/Index). Even though the rate doesn't represent the current year, the number of crimes can represent which county is riskier or safer to live in. Hence, house prices are affected by it.



	County	2021
42	Collin	14.98
56	Dallas	45.45
61	Denton	20.65
70	Ellis	20.01
219	Tarrant	39.04

(The number on the right represents number of crimes per 1000 population)

Of course, the main libraries that are used for the processing of tabular data are numpy, pandas, and requests. We might plan to push the data to a cloud database such as AWS, but the size is pretty small. Hence, this is just an optional objective if we have the time to do so. Resources such as high-performance GPUs or big storage won't be necessary.

Goals (Algorithm, metrics, etc.):

For metrics, we use R2-score from the library sklearn to measure the performance of ML regression model. It shows that how well our features explain the housing prices. The score is in the range 0-1, where 0 is bad and 1 is perfect. R2-score can end up a negative number, but a negative R2-score just means really poor performance.

I know that there many big real-estate companies with very complex housing models, but they all keep them private. Hence, we decide to stick with models we can get our hands on without undesired consequences. There are many traditional ML regression models: Random Forest Regression, Linear Regression, Support Vector Machine, XGB, etc. As long as we pick the right features, our R2-score will be usually high with any of these models. Nonetheless, we want the highest R2 since it means that this type of ML does best with our dataset. We will also explore deep learning with this dataset since we have over 8000 houses per month. Moreover, according to Limsombunchai in House Price Prediction: Hedonic Price Model vs. Artificial Neural Network, he shows that ANN beats traditional hedonic price model with an R2 equals to 0.9.

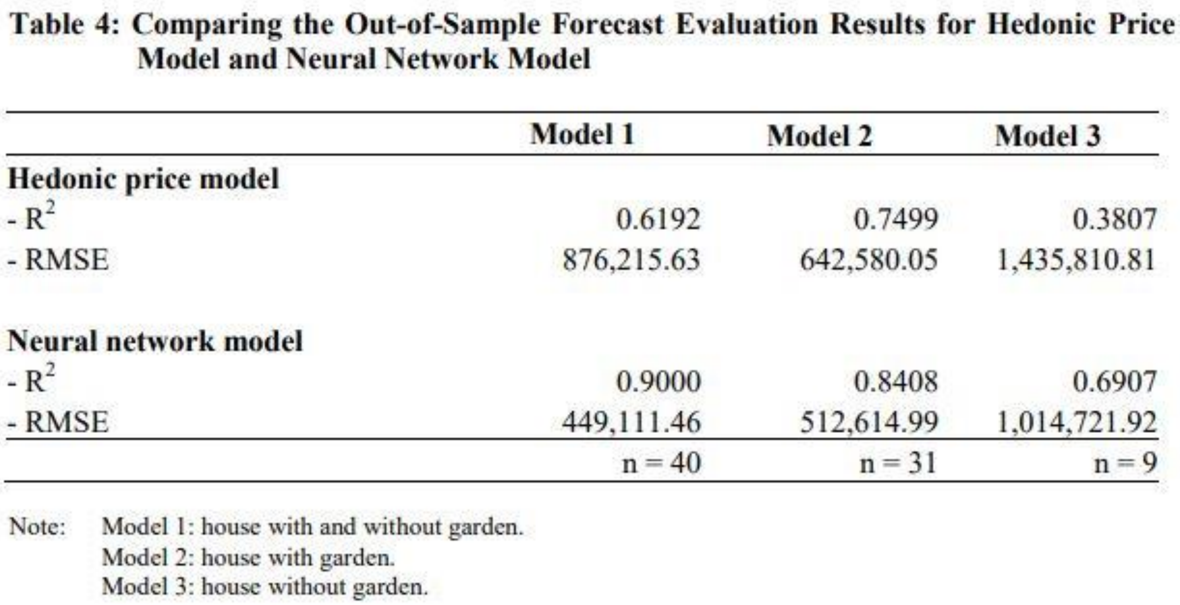


Table 4: Comparing the Out-of-Sample Forecast Evaluation Results for Hedonic Price Model and Neural Network Model

	Model 1	Model 2	Model 3
Hedonic price model			
- R^2	0.6192	0.7499	0.3807
- RMSE	876,215.63	642,580.05	1,435,810.81
Neural network model			
- R^2	0.9000	0.8408	0.6907
- RMSE	449,111.46	512,614.99	1,014,721.92
	n = 40	n = 31	n = 9

Note: Model 1: house with and without garden.
Model 2: house with garden.
Model 3: house without garden.

In my opinion, I want to use Pytorch for this particular case because it is easier to customize DL models. Specifically for DL models, we use 'Mean squared error' as our loss function to optimize our regression model. In addition, we have to split our datasets into 3 different categories: Train, Validation, and Test. The ratio would be 70:15:15.

Workplan and Github:

The github repository will hold all the codes (except the API key) I have written for this project. The reason that I will 'blur out' the API key is that I paid for it. Therefore, if someone actually gets their hands on my API key, they can actually use it and I will get charged for extra uses. For the datasets, I would like it to be on a database cloud so anyone can access it without storage limitation. Still, the datasets don't take much space in the computer, so this is an optional objective. I will also post the versions of the libraries that I use to ensure that other people can also run my codes. I have been working with Dr. Keaton on this project and we usually meet on Friday. So, the milestone for this semester is get a well-curated dataset, do data analysis, and price prediction within a given month. For instance, if we pull and curate data from October, then we will predict current house prices in October. The mid-semester report for the first semester is to show that we have well-curated datasets that are ready for analysis. Then, for the second semester, we will do time-series analysis on the dataset since we have collected several months of data at that point of time. The mid-semester report for this one is the time-series analysis on monthly house prices. The only risk that I worry about is accidentally leaking the API key.