Exploring Machine Learning's Application in Online Real Estate Marketplace

Zixuan Zhao

University of California, San Diego

ziz057@ucsd.edu

Abstract. Due to the COVID-19 pandemic and the aggressive fiscal policy by the government, the US real estate market witnessed one of the biggest price jumps in history. A significant increase in both direct and indirect economic activities happened as a result. Our interest is to explore the opportunity of applying machine learning models in the real estate industry. Several data-driven products are available in the US market, such as the Zestimate feature on Zillow.com. However, the core algorithm of these products is largely unknown to the user, and more innovative use cases could be created. We want to demystify the mechanisms behind these products and develop a proof-of-concept of machine-learning-driven model application in this area.

In this paper, we leveraged a public dataset, which compiled listing records from Zillow.com, the largest online real estate listing platform. We explored supervised (implemented with XGBoost) and unsupervised machine learning (implemented with sklearn) methods and achieved promising results in both cases. We also proposed the design of potential commercial use cases and recommendations for improvements.

Keywords: Machine-Learning-Driven Solution, Supervised Learning, Unsupervised Learning, Dimensionality Reduction, Real Estate Market

1 Introduction

During the COVID-19 pandemic, housing prices surged. The supply of homes fell to historically low levels, the demand for housing increased as people were forced to spend most of their time at home, and the interest rate decreased. In addition, remote working allowed more freedom in choosing the property's location. For example, Zillow found that nearly two million renters who could not afford homes in metro areas could now afford to buy further out because they no longer had to commute to work. [1]

However, two and half years after the pandemic's start, the once-hot housing market is starting to cool down. According to Redfin, for the first time in over 17 months, during the four-week period that ended Aug.28the average price for homes sold was below their average listing price. The supply of houses also rose nearly 27% compared to a year ago. [2] Homes no longer sell at the price they were months ago.

V. Escudero et al. (Eds.): BADS 2022, AHCS 11, pp. 1383–1391, 2023.
https://doi.org/10.2991/978-94-6463-102-9_144

Price is always the central focus of any property transaction. One of the main challenges sellers face is asking for the correct price. If listed too high, it could turn potential buyers away, but if listed too low, the seller could miss out on significant earnings. In the real estate market, most homeowners' current solution is to hire a real estate agent who is experienced and understands the local market. The agent will help the seller price his property according to the market condition and expertise. He will also utilize his professional network to find more potential buyers.

At the same time, buyers also typically hire a real estate agent to negotiate for them, but it does have an intrinsic conflict of interest. Since the buyer's agent also charges a commission fee based on the sold price, they gain more as the buyer pays more, which does not align with the buyer's best interest.

There are definitely some opportunities for technical innovation to reimagine the experience. Companies like OpenDoor takes advantage of the opportunity. For the sellers, OpenDoor completely eliminates the hassle of selling a home by sending an all-cash offer in less than 24 hours, providing a free home inspection, and taking care of the renovation. The buyers can buy directly from OpenDoor without a commission fee.[3] It is a much more seamless process for both parties.

Such applications inspire this paper. We want to explore the opportunity to utilize machine learning solutions to address these pain points. It should be noted that from a technical perspective, the problem seems to be very straightforward, however, we also need to consider user experiences in our design. This aspect remains one of the most challenging areas in applied machine learning.

2 Methodology and Basic Data Structure

This research employs a set of published data from Zillow, the leading real estate marketplace. The dataset covers different aspects of the listing, including the source of the seller, price, location, Zestimate (Zillow's estimate of the property's value), and house features like the number of beds and baths and total area.[4]

The first part of this research involves supervised learning. We train our model with the Zillow dataset, and our goal is to accurately predict the price of a never-before-seen house with its features.

We preprocess the data before training our model by deleting the unnecessary columns, unifying data types, and dealing with null values. We want to prevent overfitting, which is why we use train test split. We use 75% of the data set to train and the rest 25% to fine-tune the model.

Finally, we use package XGboost's fit method to generate our model. We choose XGboost because it is "a highly flexible and versatile tool that can work through most regression"[5] and we use RMSE to evaluate the result. This practice aims to replicate Zestimate's basic functionality and seeks improvement opportunities.

The second part of our research involves unsupervised learning and cluster segmentation, where we partition the houses into clusters based on similarities in observations. We first perform dimensionality reduction with PCA to minimize error, then use K-means to cluster the data. Then We use the silhouette score to evaluate the clustering. By applying this methodology, we propose a way of segmenting listings. Potential use cases of this model include providing personalized recommendations, more precise targeting, and a high-level overview of listing by category.

3 Preliminary analysis

We conduct exploratory data analysis before building our models. The data initially had over 60 columns, and we narrowed it down to 30 columns. We deleted unnecessary features such as detail URL and broker name and repetitive features such as address and price.

We unified all numeric data to float, filled in null values with String "NA", and replaced all boolean values with String "True" and "False". We also performed feature engineering on the Latitude Longitude column. Since it was in dictionary form that we could not directly use, we transformed it into two numeric features Latitude and Longitude.

Observing the data set, we discovered many insights. To start with, Texas, California, Florida, Arizona, and Illinois are the five states with the most property listings in that order (Figure 1). Moreover, most houses are single-family, making up 60.8% of all listings, followed by condos, making up 15.9% of all listings (Figure 2).

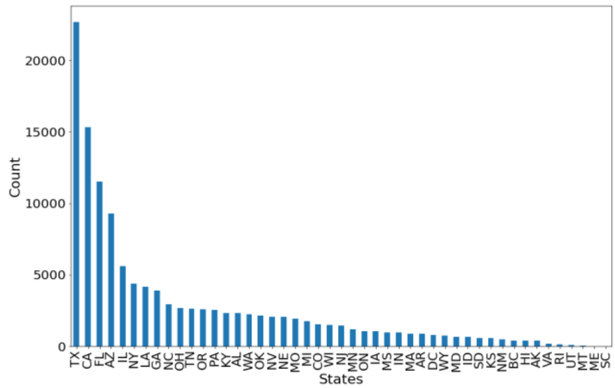

Fig. 1. Distribution of State of Property [Owner-draw]

We also see that over 85% of the houses are for sale through traditional brokers, and the portion of listings startups like OpenDoor are trying to take advantage of.

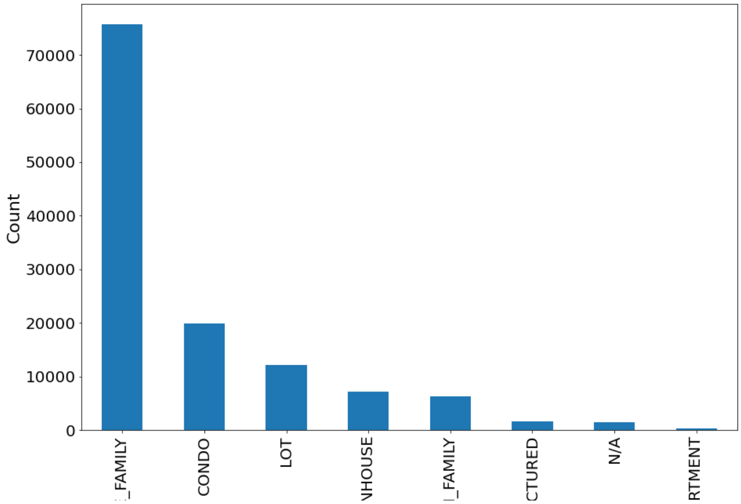

Fig. 2. Distribution of Listing Types [Owner-draw]

Finally, before building the models, we conducted outlier analysis. Figure 3 shows the distribution of the target variable, and we can see that this dataset has a very long tail on the right side. Empirically, extreme outliers will not help improve model performance, and under a business context, such listings are often ultra-high-end properties that account for only a meager fraction of all listings. As a result, we remove the outliers by only looking at the listing with unformatted prices under 1 million.

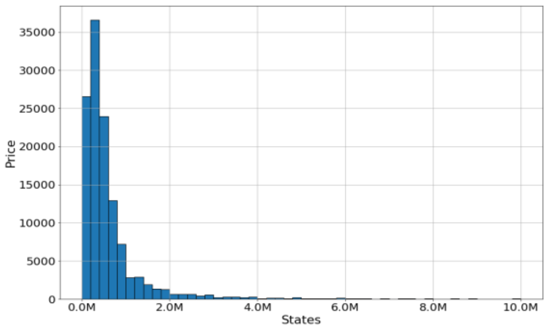

Fig. 3. Original price distribution [Owner-draw]

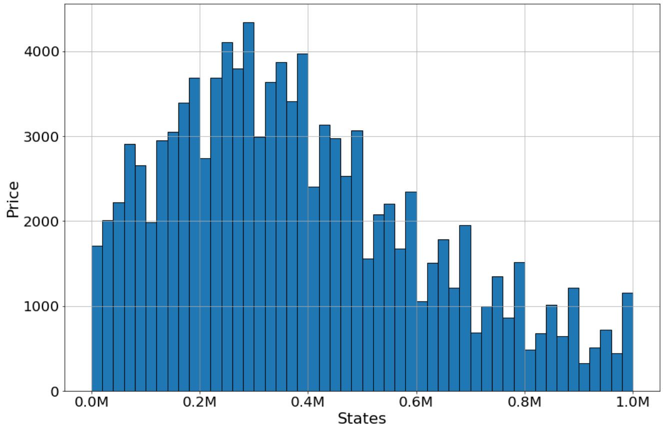

Fig. 4. Price distribution after removing outliers [Owner-draw]

4 Use Supervised Model for Price Prediction

One of the key features on Zillow.com is the Zestimate, an estimated price of listed property that Zillow's proprietary machine learning model generates.[6] Here we tried to implement a similar model to generate price estimation based on the public dataset. By definition, supervised learning is used when we have training data of input/output pairs and intend to predict the outputs of new inputs. In this case, our model aims to predict house prices with the available features in the dataset, so our target variable y is the unformatted price, and independent variables x are all the house features.

We chose to use XGBoost as our implementation for a variety of reasons. XGBoost is known for its speed and performance, which sets it apart from its peers and makes the de-facto baseline in machine learning application across industries. In addition, XGBoost has an array of parameters to config, which enables in-depth parameter-tuning as the next step to improve the performance further. It also has a very mature community supporting its future enhancement.

Our regressor uses XGBRegressor with parameters 'hist' as the tree method and RMSE as the evaluation metric. The pre-processed dataset was split into train and test datasets, whose ratio was 3:1. After 100 epochs, RMSE settled around 129,000, as showed in Table 1. Since RMSE are in the same unit as unformatted price, it is a reasonable number considering housing prices are typically in the range of hundreds of thousands of dollars.

Table 1. RMSE Results for Different Epoch Settings [Owner-draw]

Epoch	RMSE
0	336k
10	135k
30	130k
50	129k
100	129k

When we plot all the predicted prices against the actual unformatted price, we see a linearly upward trend, which proves our predictions are mostly on point. However, we do notice that some of the predicted prices are negative, which is due to the extrapolatory nature of the tree based model, but these negative values only make up less than 10% of all predictions.

If we consider this when designing a commercial solution (such as Zestimate), it is reasonable to include additional rules that limit such results from being displayed. In fact, not all listings on Zillow have a Zestimate, it could be due to the lack of comparable listings, but it is also possible that undesired predictions are curbed with business logic, such as no negative prices should be displayed.

XGboost also analyzes feature importance. Unsurprisingly, we discover that address zip code is the most critical factor when it comes to houses in the US and Canada. It is common sense that the location of a property is significant since it provides safety and accessibility to local amenities. Still, out of all the location features, address zip code appears to be the most representative.

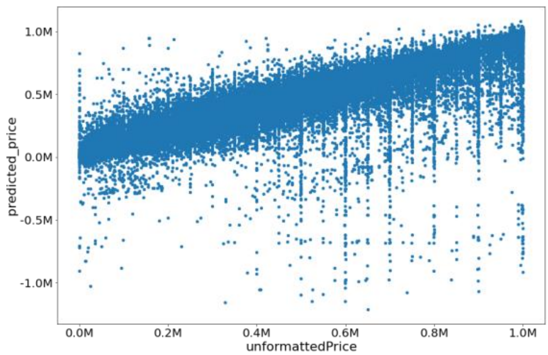

Fig. 5. Predicted price vs real Price [Owner-draw]

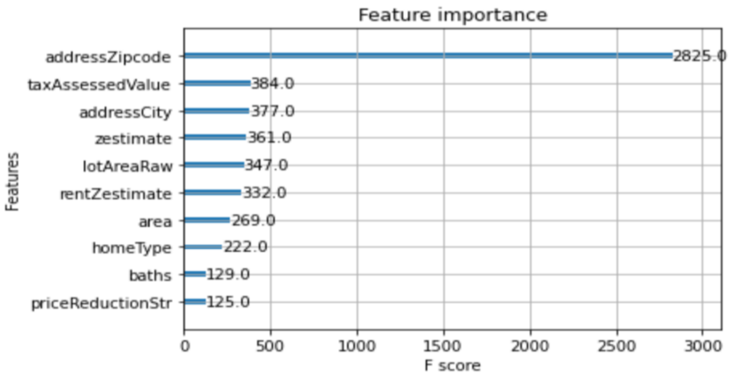

Fig. 6. Feature importance [Owner-draw]

5 Housing segmentation using unsupervised Learning

The second part of our research involved unsupervised learning. Unlike supervised learning, unsupervised learning does not use input/output pairs. Its goal is to identify patterns and structural properties within the data. In our research, we cluster the houses into six groups based on similarities in their features.

We have to prepare the dataset for clustering. We only keep the numerical values and fill in all the missing values with -999 since the PCA and K-means algorithms we will perform only deal with numbers and do not work with missing values.

We first perform PCA (Principle Component Analysis). PCA is a method to reduce dimensionality by transforming the data into a new coordinate system where the variation in data can be explained with fewer variables. PCA helps to minimize the error by reducing the number of features. However, its disadvantage is the low interpretability of the new principal components since they are linear combinations of the original variables. We use Python's sklearn.pca to implement the algorithm. We choose parameter n_component = 0.95, which means we want a new feature space that can explain 95% of the variance of the original feature space. As a result, we obtain a dataset with two columns representing the first and the second principal component, respectively.

Then we perform K-means on the new dataset. K-means is one of the simplest and most commonly used clustering algorithms. It works by alternating between assigning each data point to the closest cluster center and re-calculating the cluster center as the mean of its group of data points. It will repeat until the assignments no longer change with iteration. To perform the algorithm, we used the sklearn.kmeans with six as parameter, and we obtained clusters of sizes 119307, 88, 548, 1, 4493, and 8.

We then use TSNE to visualize the high-dimensional data. As shown in Figure 7, we can see a noticeable cluster among the majority (the largest cluster of 119307 records).

Lastly, we use the silhouette score to evaluate our clustering result. Silhouette score has a range of -1 to 1, and a high value indicates that the data point is well-matched with its cluster and poorly matched with its neighboring clusters. We received a silhouette score of 0.85.

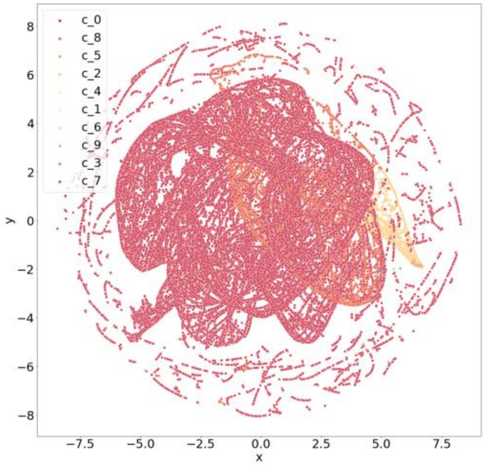

Fig. 7. TSNE high dimensional visualization [Owner-draw]

6 Conclusion and Application

We achieved both of our objectives. We built a model where we can predict housing prices based on its features, and we segmented the houses into 6 clusters based on their likeness.

Price prediction models have many applications in the real world. As mentioned before, Zillow's Zestimate is similar to our model. Our model can provide an estimate of the value of a property upon listing that serves as a reference for both the buyer and seller. The model can be valuable in the house-flipping business as well. House flipper are often unsure whether the changes they're making is profitable. The model can help predict whether adding extra features will be worth it.

Housing segmentation also has many applications. A clustering algorithm will not only suggest similar listings to the ones that the buyer is interested in but also optimize sales efforts and prioritize resources for the sellers since real estate agents use different strategies to reach other target customers. We believe these applications have great potential in the real estate industry and will significantly help both buyers and sellers have a smoother experience.

References

1. Dowell, Earlene K.P. "Remote Working, Commuting Time, Life Events All Affect Home Buyers' Decisions." *Census.gov*, 13 Apr. 2022,

https://www.census.gov/library/stories/2021/10/zillow-and-census-bureau-data-show-pandemics-impact-on-housing-market.html#:~:text=working%20from%20home.-,Zillow%20found%20that%20nearly%20two%20million%20renters%20unable%20to%20afford,pandemic%20hit%20the%20United%20States

2. Olick, Diana. "1 In 5 Home Sellers Are Now Dropping Their Asking Price as the Housing Market Cools." *CNBC*, CNBC, 3 Sept. 2022, https://www.cnbc.com/2022/09/02/more-home-sellers-drop-their-asking-price-as-the-housing-market-cools.html.
3. Pros and Cons of Using Opendoor to Sell Your Home." *Raleigh, NC Real Estate & Homes for Sale - Raleigh Realty*, https://raleighrealtyhomes.com/blog/pros-and-cons-of-using-opendoor-to-sell-your-home/.
4. 100,000+ Zillow Properties: US+ Canada." *Kaggle*, 18 July 2022, https://www.kaggle.com/datasets/dataranch/100000-zillow-properties-us-canada.
5. Reinstein, Ilan. "XGBoost, a Top Machine Learning Method on Kaggle, Explained." *KDnuggets*, https://www.kdnuggets.com/2017/10/xgboost-top-machine-learning-method-kaggle-explained.html.
6. "What Is a Zestimate? Zillow's Zestimate Accuracy." *Zillow*, 25 July 2022, https://www.zillow.com/z/zestimate/.