Zillow House Price Predict Analysis*

Yaning Jin

April 17, 2024

This project is interested in the factors affecting housing prices on zillow. After exploratory analysis of the data, a multiple linear regression was constructed, and a log transformation was performed to build a model with better performance. The transformed housing factors can explain more than 80% of the changes in housing prices. However, the existence of residuals in the model does not fully meet the constraints of the normality assumption, so in the future, data processing, feature engineering and model improvement are needed to study the factors affecting Zillow's housing prices..

Table of contents

*Code and data are available at: https://github.com/Yuki010305/Zillow-House-Price-Predict-Analysis.git.

1 Introduction

One of a nation's key industries is housing, and the health of this sector influences the degree of economic growth of the nation. A person's home is frequently the most expensive item in their life. Zillow is an online real estate database provider that assesses property values and offers details on the houses you're interested in (Loukissas 2018). The Zestimate prediction system was developed by Zillow. It is based on a data collection of hundreds of millions of homes and provides a preliminary estimate of a property's worth by combining a particular algorithm with the features of each home and the state of the market(Rolli 2020). To improve the accuracy of the Zestimate system and provide people with a more trustworthy way to estimate home prices, Zillow launched the Zillow Awards competition in 2017. Our goal in this project is not to participate in a competition to minimize the logarithmic error between the estimated house price and the actual selling price. The main purpose of our exploration in this project is to explore the factors that affect Zillow home sales prices and how they affect home sales prices. Characteristics of the homes we select include number of bathrooms, number of bedrooms, square footage, number of rooms, year, tax value, land tax value, etc. The estimated model is

$$\log(\text{price}) = \beta_0 + \beta_1 \text{bathrooms} + \beta_2 \log(\text{squarefootage}) + ... + \beta_6 \text{taxValue}$$

For some variables we will also do some log transformations.

2 Data

2.1 Raw Data

In this research, we examine zillow data sourced using the zillbowR library (Rolli 2020). The dataset encompasses 90,275 records (89,499 records after cleaning), focusing on specific variables: bathroom number, bedroom number, year built, square feet, room number, tax value, etc. The apartments were built in a wide range of years, from as early as 1885 to as recently as 2015.

The Raw data collects 60 features of the house, and data quality is not high, there are many features with many null values, so 8 high-quality variables are selected from the variables for modeling and prediction.

2.2 Data Analysis tools

R (R Core Team 2020), a potent open-source statistical programming language, was used to analyze the data. To improve the effectiveness of our data operations, we used a collection of R packages from the tidyverse (Wickham et al. 2019), which is a collection of tools created for data science. The `dplyr` package (Wickham et al. 2022) offered a consistent collection of verbs that aid in filtering, summarizing, and organizing the dataset, while the `ggplot2` package (Wickham 2016) made it easier to create complex visualizations. Because of its quick and user-friendly data reading capabilities, the `readr` package (Wickham, Hester, and Bryan 2022) was used. `Knitr` (Goodrich et al. 2020) handled report production dynamically, allowing R code to be included into this document. `kableExtra`(Zhu 2021) was also used to create visually appealing and editable tables, which improved the way our results were presented.

2.3 Variable Description

Variable	Description
bathroomcnt	Number of bathrooms in home
bedroomcnt	Number of bedrooms in home
calculatedfinishedsquarefeet	Total finished living area of the home
roomcnt	Total number of rooms
yearbuilt	The Year the principal residence was built
taxvaluedollarcnt	The total tax assessed value of the parcel
landtaxvaluedollarcnt	The assessed value of the land area
price	price

2.4 Measurement

In this study, we leveraged data from the Zillow dataset, accessed through the Zillow API. The Zillow dataset provides comprehensive information on residential properties, enabling analyzes focused on housing market dynamics, property valuation, and spatial characteristics.

The Zillow dataset encompasses various attributes crucial for understanding residential properties' characteristics and market behavior. Specifically:

bathroomcnt: Indicates the number of bathrooms present in each residential property, influencing its utility and convenience for occupants. bedroomcnt: Represents the count of bedrooms within each residential unit, influencing its suitability for different household compositions. calculatedfinishedsquarefeet: Quantifies the total finished living area of the home, reflecting its size and spatial configuration. roomcnt: Denotes the total number of rooms in the residential unit, encompassing living spaces, bedrooms, and other functional areas, providing insights into the property's layout. yearbuilt: Specifies the year in which the principal

residence was constructed, offering insights into the age distribution of properties within the dataset. taxvaluedollarcnt: Indicates the total tax-assessed value of the parcel, serving as a key indicator of property valuation for taxation purposes. landtaxvaluedollarcnt: Reflects the assessed value of the land area associated with each residential parcel, contributing to the overall property valuation. price: Represents the price of the residential property, serving as a fundamental metric for market valuation and investment analysis. By analyzing these attributes, researchers can gain valuable insights into housing market trends, property valuation factors, and the impact of various features on real estate prices within the studied area.

To ensure the dataset's suitability for analysis, we performed several preprocessing steps. Select high-quality variables with few missing values and remove missing values. Additionally, we implemented data cleaning procedures to exclude observations with missing values, ensuring the completeness and reliability of our dataset for subsequent analyses.

For detailed descriptions of each variable, please refer to the table provided in Section Section 2.3.

3 Exploratory Data Analysis

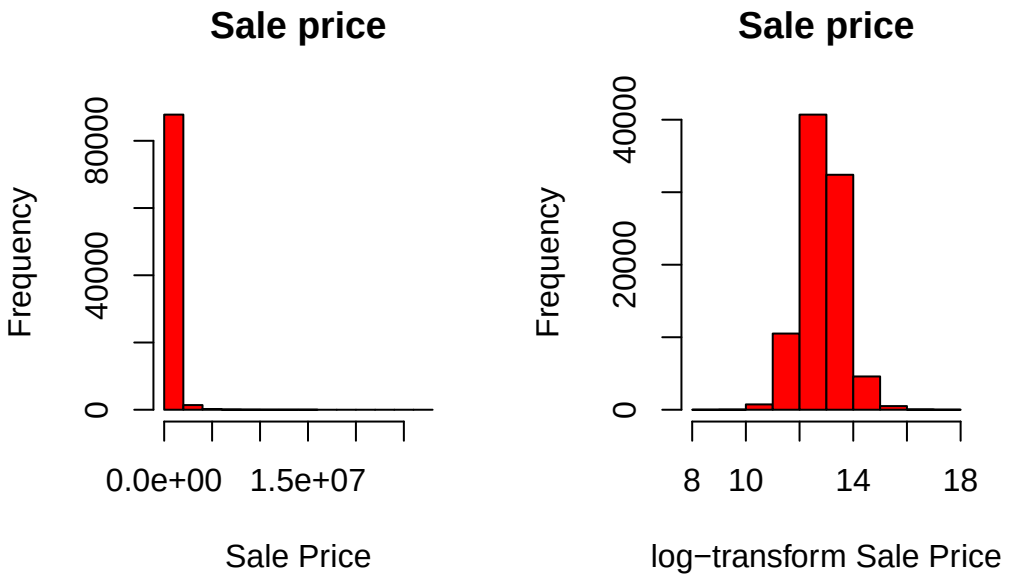

Figure 1: sale price and log sale price hist

Figure 1, house sales prices are right-skewed data. When building the model, we need to log-transform the house sales prices to make them conform to the normal distribution.

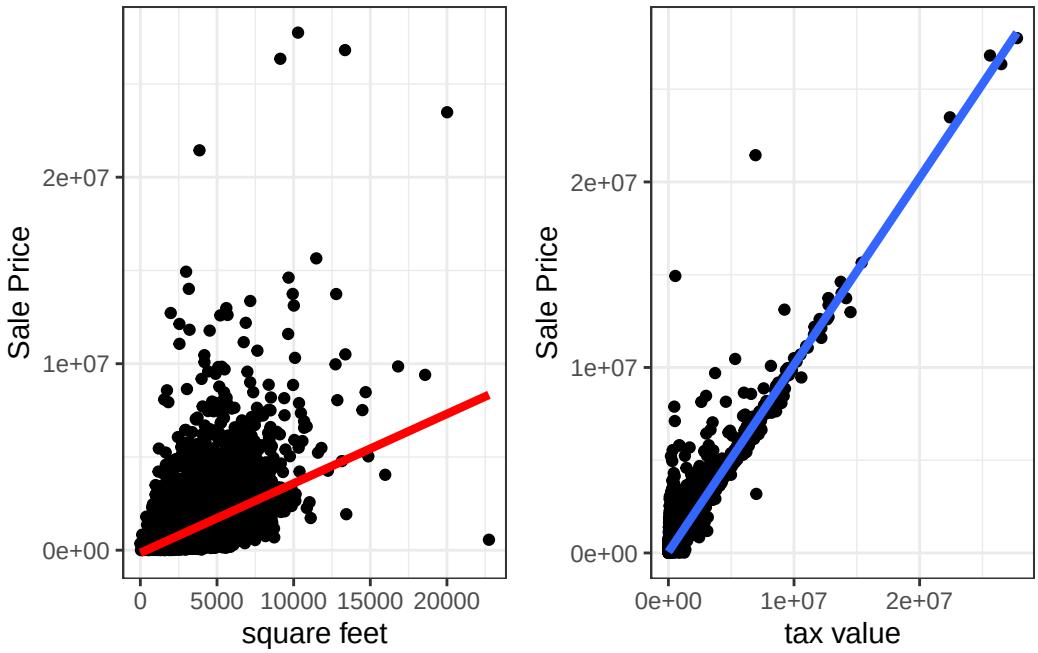

Figure 2: Bivariate visualization

Figure 2 shows that tax value has a certain positive impact on sale price, the positive correlation between tax value and sale price is very strong. Figure 2 also shows that square feet has a certain positive impact on sale price, the positive correlation between square feet and sale price is very strong.

4 Model

4.1 Model set-up

After data processing, the data set is a clean data set with 2929 observations and 10 house characteristic variables. In order to evaluate the performance of the model, we randomly split the analysis data set into a test set and a training set in a ratio of 75%:25%.

The first model we build is the full model. We then improve the model by removing insignificant variables.

The residual test found that both ends of the QQ graph deviated greatly from the straight line and were affected by special points such as outliers and leverage points. So, in order to further improve the performance of model fitting, we will delete special points from the training set.

After deleting special points such as level points, a new model was refitted.

4.1.1 Model justification

4.1.1.1 A1:Linearity of the Relationship

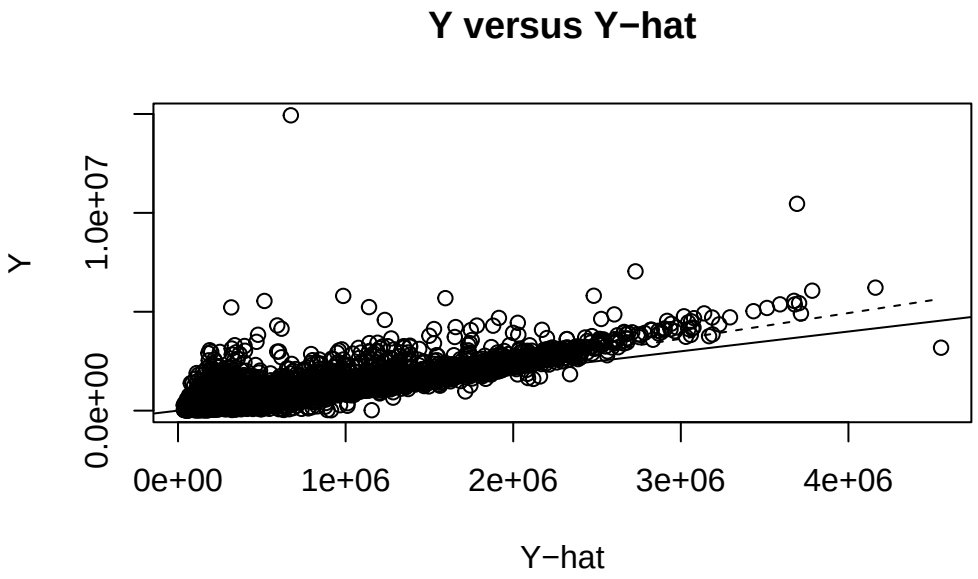

Figure 3: Y versus Y-hat

Figure 3 fits the relationship between the predicted value and the actual value. You can see that these points are almost on or close to the line, so we can say that a linear relationship is satisfied.

4.1.1.2 A2.Covariance of Errors

The residual plot Figure 4 does not show any pattern, we can conclude that the residual terms are independently distributed across different predicted value ranges, consistent with the independence assumption of linear regression analysis.

4.1.1.3 A3.Common Error variance

The residual plot Figure 4 is not a uniform distribution. As the predicted value increases, the variance decreases, so the assumption of constant variance is not completely established.

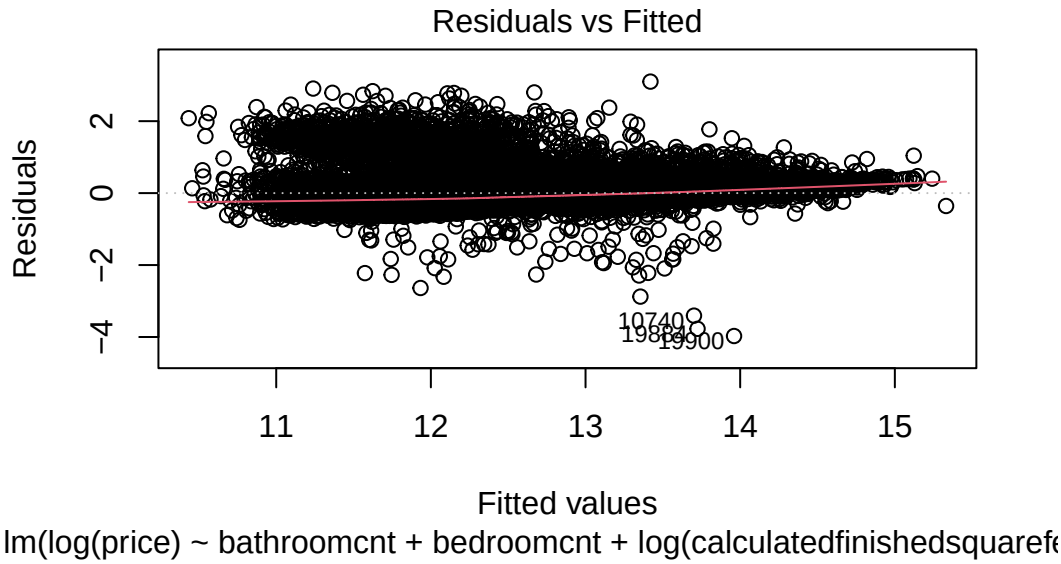

Figure 4: Residual Plot

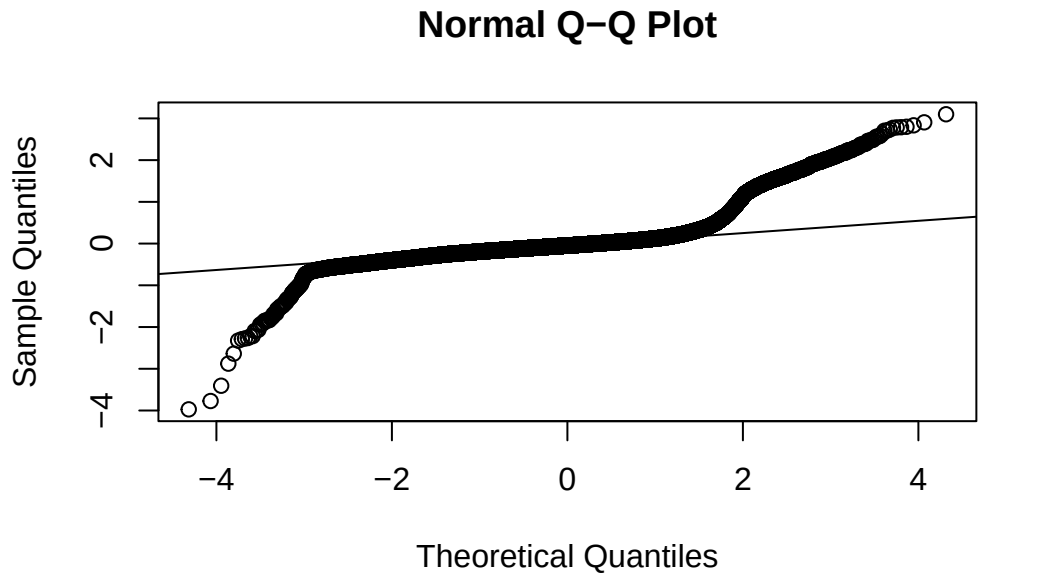

Figure 5: Normal Q-Q plot

4.1.1.4 A4. Normality of Error

There are obvious deviations at both ends of the QQ graph Figure 5, which may indicate that the data does not conform to the normal distribution in this area.

5 Results

Estimating the value of a home is a common difficulty. Therefore, a lot of work has already been completed. Lee (Lee 2016) make an effort to develop multivariate regression models based on house datasets and assess the models using maximum information coefficient statistics based on anticipated values and home prices. When a moderate to big data sample size is employed, Nghiep et al. (Nguyen and Cripps 2001) examined two methods: multiple regression analysis (MRA) and artificial neural networks (ANN). Based on the prediction performance, ANN performs better than MRA. An artificial neural network (ANN) model was created after Limsombunchai (Limsombunchao 2004) that ANNs are more accurate in predicting property values than hedonic regression models.

The final model is:

$$\begin{aligned} log(price) = 2.7 &- 0.007 bathroomcnt - 0.006 bedroomcnt - 0.2 log(squarefeet) \\ &- 0.015 roomcnt - 0.00025 yearbuilt + 0.67 log(taxvaluedollarcnt) - 0.055 log(landtaxvaluedollarcnt) \end{aligned}$$

Adjusted R-squared is 0.8082, indicating that the model can explain approximately 80.83% of the variance of the target variable (logarithm of house prices).

The final model shows that the number of bathrooms, bedrooms, rooms, year taxes and fees of the building, square feet, etc. all have a significant impact on Zillow's housing prices. For every additional bathroom, the house price increases by 1.017% per unit. For every additional bedroom, the house price actually decreases by 0.01%. In addition, the newer the building is, the housing prices actually decrease. For every 1% unit increase in taxes and fees, housing prices increase by 0.67% unit.These findings explain the relationship between house-related attributes and house prices.

6 Discussion

6.1 Discussion

Model Interpretability and Generalization While the developed model provides valuable insights into the factors influencing housing prices, its interpretability is hindered by the transformation of data and the exclusion of certain categorical variables due to memory and time

constraints. To address this, future research could explore methods to maintain model interpretability while optimizing predictive performance. Techniques such as feature selection and dimensionality reduction may help streamline the model without sacrificing interpretability.

Addressing Residual Normality Assumption The observed violations of the normality assumption in the model's residuals raise concerns regarding its robustness and generalizability. To mitigate this issue, researchers could explore alternative modeling approaches or employ robust regression techniques that are less sensitive to deviations from normality. Additionally, conducting thorough diagnostic checks and sensitivity analyses can help identify potential sources of bias and improve the model's reliability.

Enhancing Model Performance Through Cross-Validation The suggestion to partition training datasets into smaller subsets and perform cross-validation before generating test files holds promise for improving model performance and generalization. By systematically evaluating the model's performance across different subsets of data, researchers can gain insights into its stability and identify areas for refinement. Implementing rigorous cross-validation protocols can also enhance the model's credibility and ensure its applicability across diverse datasets.

Future Directions in Feature Engineering Future iterations of the model could benefit from more sophisticated feature engineering techniques aimed at capturing the nuanced relationships between housing attributes and prices. Exploring advanced feature transformation methods, such as polynomial features or interaction terms, may help uncover hidden patterns and improve the model's predictive accuracy. Additionally, incorporating domain knowledge and expert insights into the feature selection process can enrich the model's explanatory power and enhance its utility for real-world applications.

Handling Missing Data Addressing missing data is critical for ensuring the reliability and validity of the model's predictions. Researchers should carefully consider imputation strategies and explore techniques for incorporating missing variables as predictors in the model. By systematically addressing missing data issues, researchers can enhance the model's robustness and improve its performance on new datasets.

In summary, while the developed model provides valuable insights into the factors influencing housing prices, there are several avenues for future research to enhance its interpretability, robustness, and predictive accuracy. By addressing these challenges and leveraging advanced modeling techniques, researchers can develop more reliable and actionable models for estimating housing values.

6.2 Weaknesses and next steps

However, the residuals of our model do not fully satisfy the normality assumption. Such violations may reduce the model's predictive accuracy on new data sets. Another limitation of this model is that transforming the data makes the model less interpretable.

To start making these models better in the future, divide each training dataset into smaller training tests and perform some cross-validation before generating the test files. This might enhance performance, or at the very least increase the predictability of the model's output. Although most categorical variables cannot be included in the model because of memory and time restrictions when running the model, the model may be enhanced with improved feature engineering. These variables can be included in the model by being divided into smaller groups. Making some interactive words is an additional thought. It is crucial to classify missing variables as predictors once all missing values have been imputed, as was previously noted.

References

Goodrich, Ben, Jonah Gabry, Imad Ali, and Sam Brilleman. 2020. *Rstanarm: Bayesian Applied Regression Modeling via Stan*. https://mc-stan.org/rstanarm.

Lee, Roger. 2016. *Software Engineering, Artificial Intelligence, Networking and Parallel/Distributed Computing*. Vol. 653. Springer.

Limsombunchao, Visit. 2004. "House Price Prediction: Hedonic Price Model Vs. Artificial Neural Network."

Loukissas, Yanni. 2018. "All the Homes: Zillow and the Operational Context of Data." In *Transforming Digital Worlds: 13th International Conference, iConference 2018, Sheffield, UK, March 25-28, 2018, Proceedings 13*, 272–81. Springer.

Nguyen, Nghiep, and Al Cripps. 2001. "Predicting Housing Value: A Comparison of Multiple Regression Analysis and Artificial Neural Networks." *Journal of Real Estate Research* 22 (3): 313–36.

R Core Team. 2020. *R: A Language and Environment for Statistical Computing*. Vienna, Austria: R Foundation for Statistical Computing. https://www.R-project.org/.

Rolli, Channamallikarjun Siddaramappa. 2020. "Zillow Home Value Prediction (Zestimate) by Using XGBoost."

Wickham, Hadley. 2016. *Ggplot2: Elegant Graphics for Data Analysis*. Springer-Verlag New York. https://ggplot2.tidyverse.org.

Wickham, Hadley, Mara Averick, Jennifer Bryan, Winston Chang, Lucy D'Agostino McGowan, Romain François, Garrett Grolemund, et al. 2019. "Welcome to the tidyverse." *Journal of Open Source Software* 4 (43): 1686. https://doi.org/10.21105/joss.01686.

Wickham, Hadley, Romain François, Lionel Henry, and Kirill Müller. 2022. *Dplyr: A Grammar of Data Manipulation*. https://CRAN.R-project.org/package=dplyr.

Wickham, Hadley, Jim Hester, and Jennifer Bryan. 2022. *Readr: Read Rectangular Text Data*. https://CRAN.R-project.org/package=readr.

Zhu, Hao. 2021. *kableExtra: Construct Complex Table with 'Kable' and Pipe Syntax*. https://CRAN.R-project.org/package=kableExtra.