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..

Extract of the questions from Gebru et al. (2021).

Motivation

1. *For what purpose was the dataset created? Was there a specific task in mind? Was there a specific gap that needed to be filled? Please provide a description.*
    - This dataset was created to study the factors that affect Zillow home prices, with a focus on determining how factors such as bathroom, square footage, tax value, etc. affect Zillow home prices.
2. *Who created the dataset (for example, which team, research group) and on behalf of which entity (for example, company, institution, organization)?*
    - The dataset was created by Zillow as data of kaggle competition.
3. *Who funded the creation of the dataset? If there is an associated grant, please provide the name of the grantor and the grant name and number.*
    - The dataset creation was funded by Zillow.
4. *Any other comments?*
    - N/A

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

Composition

1. *What do the instances that comprise the dataset represent (for example, documents, photos, people, countries)? Are there multiple types of instances (for example, movies, users, and ratings; people and interactions between them; nodes and edges)? Please provide a description.*

   - The instances represent evaluations of zillow houses and their selling prices, including their bathroom number, bedroom number.

2. *How many instances are there in total (of each type, if appropriate)?*

   - There are 90,275 instances in the dataset.

3. *Does the dataset contain all possible instances or is it a sample (not necessarily random) of instances from a larger set? If the dataset is a sample, then what is the larger set? Is the sample representative of the larger set (for example, geographic coverage)? If so, please describe how this representativeness was validated/verified. If it is not representative of the larger set, please describe why not (for example, to cover a more diverse range of instances, because instances were withheld or unavailable).*

   - The dataset is a sample of year 2016 zillow of all houses sold. The representativeness of the sample.

4. *What data does each instance consist of? "Raw" data (for example, unprocessed text or images) or features? In either case, please provide a description.*

   - Each instance consists of features such as the bedroom number, tax value, year built and price et al.

5. *Is there a label or target associated with each instance? If so, please provide a description.*

   - The target variable is price, which indicates zillow sales price.

6. *Is any information missing from individual instances? If so, please provide a description, explaining why this information is missing (for example, because it was unavailable). This does not include intentionally removed information, but might include, for example, redacted text.*

   - Some instances may have missing information due to incomplete records or unavailability of data at the time of evaluation.

7. *Are relationships between individual instances made explicit (for example, users' movie ratings, social network links)? If so, please describe how these relationships are made explicit.*

   - The dataset does not explicitly make relationships between individual instances.

8. *Are there recommended data splits (for example, training, development/validation, testing)? If so, please provide a description of these splits, explaining the rationale behind them.*

    - specific data splits are recommended for this dataset. And use the test set to evaluate the model.

9. *Are there any errors, sources of noise, or redundancies in the dataset? If so, please provide a description.*

    - The dataset may contain errors or noise due to inaccuracies in the evaluation process or inconsistencies in record-keeping.

10. *Is the dataset self-contained, or does it link to or otherwise rely on external resources (for example, websites, tweets, other datasets)? If it links to or relies on external resources, a) are there guarantees that they will exist, and remain constant, over time; b) are there official archival versions of the complete dataset (that is, including the external resources as they existed at the time the dataset was created); c) are there any restrictions (for example, licenses, fees) associated with any of the external resources that might apply to a dataset consumer? Please provide descriptions of all external resources and any restrictions associated with them, as well as links or other access points, as appropriate.*

    - No.

11. *Does the dataset contain data that might be considered confidential (for example, data that is protected by legal privilege or by doctor-patient confidentiality, data that includes the content of individuals' non-public communications)? If so, please provide a description.*

    - No.

12. *Does the dataset contain data that, if viewed directly, might be offensive, insulting, threatening, or might otherwise cause anxiety? If so, please describe why.*

    - No.

13. *Does the dataset identify any sub-populations (for example, by age, gender)? If so, please describe how these subpopulations are identified and provide a description of their respective distributions within the dataset.*

    - No.

14. *Is it possible to identify individuals (that is, one or more natural persons), either directly or indirectly (that is, in combination with other data) from the dataset? If so, please describe how.*

    - No.

15. *Does the dataset contain data that might be considered sensitive in any way (for example, data that reveals race or ethnic origins, sexual orientations, religious beliefs, political*

opinions or union memberships, or locations; financial or health data; biometric or genetic data; forms of government identification, such as social security numbers; criminal history)? If so, please provide a description.

- No.

16. *Any other comments?*

    - No.

Collection process

1. *How was the data associated with each instance acquired? Was the data directly observable (for example, raw text, movie ratings), reported by subjects (for example, survey responses), or indirectly inferred/derived from other data (for example, part-of-speech tags, model-based guesses for age or language)? If the data was reported by subjects or indirectly inferred/derived from other data, was the data validated/verified? If so, please describe how.*

    - The data for each instance was obtained through zillow, an airbnb property sales organization. This information is observable and is reported by zillow. This data was then validated to ensure accuracy and reliability.

2. *What mechanisms or procedures were used to collect the data (for example, hardware apparatuses or sensors, manual human curation, software programs, software APIs)? How were these mechanisms or procedures validated?*

    - The data collection involved administrative procedures by zillow to ensure consistent and valid data collection.

3. *If the dataset is a sample from a larger set, what was the sampling strategy (for example, deterministic, probabilistic with specific sampling probabilities)?*

    - The dataset is not a sample but a comprehensive collection of evaluations from the zillow in year 2016.

4. *Who was involved in the data collection process (for example, students, crowdworkers, contractors) and how were they compensated (for example, how much were crowdworkers paid)?*

    - The data collection was conducted by zillow. No additional compensation was provided for data collection as it falls under their regular job responsibilities.

5. *Over what timeframe was the data collected? Does this timeframe match the creation timeframe of the data associated with the instances (for example, recent crawl of old news articles)? If not, please describe the timeframe in which the data associated with the instances was created.*

    - The data was collected over 2016.

6. *Were any ethical review processes conducted (for example, by an institutional review board)? If so, please provide a description of these review processes, including the outcomes, as well as a link or other access point to any supporting documentation.*

   - While specific ethical review processes for this dataset are not mentioned, it is assumed that zillow to follow standard for data collection, ensuring compliance with legal and ethical standards, information will not be matched to specific houses.

7. *Did you collect the data from the individuals in question directly, or obtain it via third parties or other sources (for example, websites)?*

   - The data was collected directly by Zillow, so I collected it directly from the person involved.

8. *Were the individuals in question notified about the data collection? If so, please describe (or show with screenshots or other information) how notice was provided, and provide a link or other access point to, or otherwise reproduce, the exact language of the notification itself.*

   - This is a public data source and the persons concerned will not be notified about the data collection.

9. *Did the individuals in question consent to the collection and use of their data? If so, please describe (or show with screenshots or other information) how consent was requested and provided, and provide a link or other access point to, or otherwise reproduce, the exact language to which the individuals consented.*

   - The dataset does not involve personal data, so there is no mechanism for individuals to revoke consent for the use of this data as it is part of regulatory compliance.

10. *If consent was obtained, were the consenting individuals provided with a mechanism to revoke their consent in the future or for certain uses? If so, please provide a description, as well as a link or other access point to the mechanism (if appropriate).*

    - An impact analysis specifically for this dataset was not described.

11. *Has an analysis of the potential impact of the dataset and its use on data subjects (for example, a data protection impact analysis) been conducted? If so, please provide a description of this analysis, including the outcomes, as well as a link or other access point to any supporting documentation.*

    - Additional comments or concerns regarding the data collection process can be directed to zillow Open Data through the contact information provided on the portal, and mobile phone verification is required before collection.

12. *Any other comments?*

    - No.

Preprocessing/cleaning/labeling

1. *Was any preprocessing/cleaning/labeling of the data done (for example, discretization or bucketing, tokenization, part-of-speech tagging, SIFT feature extraction, removal of instances, processing of missing values)? If so, please provide a description. If not, you may skip the remaining questions in this section.*
   - Yes, the dataset was preprocessed and cleaned. Since many of our 60 features contain a large number of missing values, in order to ensure the completeness and reliability of the analysis, I only selected 8 key factors with relatively few missing values as predictor variables.
2. *Was the "raw" data saved in addition to the preprocessed/cleaned/labeled data (for example, to support unanticipated future uses)? If so, please provide a link or other access point to the "raw" data.*
   - The original raw dataset remains stored in its initial directory labeled 'data/raw_data' within our project repository. This allows for reference or re-analysis should future research require access to the unprocessed information.
3. *Is the software that was used to preprocess/clean/label the data available? If so, please provide a link or other access point.*
   - The software used for preprocessing, namely the R programming language along with packages 'readr', 'dplyr', is publicly available.
4. *Any other comments?*
   - Following the cleaning process, we only retained the essential columns necessary for our analysis to streamline the dataset and focus on the variables of interest. The filtered dataset with these columns is available for review and further analysis, ensuring transparency and reproducibility of our study.

Distribution

1. *Will the dataset be distributed to third parties outside of the entity (for example, company, institution, organization) on behalf of which the dataset was created? If so, please provide a description.*
   - The dataset is distributed to the public, which allows for both commercial and non-commercial use without restrictions.
2. *How will the dataset be distributed (for example, tarball on website, API, GitHub)? Does the dataset have a digital object identifier (DOI)?*
   - The dataset can be downloaded directly from kaggle. While there is no mention of a DOI, the portal provides a direct download link for the dataset.
3. *When will the dataset be distributed?*

   - The dataset is already available for distribution as it is posted on the open data portal.

4. *Will the dataset be distributed under a copyright or other intellectual property (IP) license, and/or under applicable terms of use (ToU)? If so, please describe this license and/ or ToU, and provide a link or other access point to, or otherwise reproduce, any relevant licensing terms or ToU, as well as any fees associated with these restrictions.*

   - As mentioned, the dataset allows for worldwide, royalty-free, and perpetual use.

5. *Have any third parties imposed IP-based or other restrictions on the data associated with the instances? If so, please describe these restrictions, and provide a link or other access point to, or otherwise reproduce, any relevant licensing terms, as well as any fees associated with these restrictions.*

   - No additional IP-based restrictions from third parties are imposed on the data, but you need to verify your mobile phone number to participate in the competition before you can download the data.

6. *Do any export controls or other regulatory restrictions apply to the dataset or to individual instances? If so, please describe these restrictions, and provide a link or other access point to, or otherwise reproduce, any supporting documentation.*

   - For any additional questions or comments about the dataset, the Open Data team can be reached via their contact details on the portal.

7. *Any other comments?*

   - No

Maintenance

1. *Who will be supporting/hosting/maintaining the dataset?*

   - zillow's data team will be responsible for supporting, hosting and maintaining the dataset.

2. *How can the owner/curator/manager of the dataset be contacted (for example, email address)?*

   - The Open Data team can be contacted through the official Open Data portal.

3. *Is there an erratum? If so, please provide a link or other access point.*

   - Any errata or updates to the dataset are likely to be communicated through zillow, where users can find the latest information and corrections.

4. *Will the dataset be updated (for example, to correct labeling errors, add new instances, delete instances)? If so, please describe how often, by whom, and how updates will be communicated to dataset consumers (for example, mailing list, GitHub)?*

- The dataset will not be updated and is historical sales data.

5. *If the dataset relates to people, are there applicable limits on the retention of the data associated with the instances (for example, were the individuals in question told that their data would be retained for a fixed period of time and then deleted)? If so, please describe these limits and explain how they will be enforced.*

    - This dataset primarily relates to house, so it does not directly pertain to individuals. Therefore, retention limits specific to personal data may not be applicable.

6. *Will older versions of the dataset continue to be supported/hosted/maintained? If so, please describe how. If not, please describe how its obsolescence will be communicated to dataset consumers.*

    - Older versions of the dataset may be archived and accessible through the Open Data Portal.

7. *If others want to extend/augment/build on/contribute to the dataset, is there a mechanism for them to do so? If so, please provide a description. Will these contributions be validated/verified? If so, please describe how. If not, why not? Is there a process for communicating/distributing these contributions to dataset consumers? If so, please provide a description.*

    - While zillow encourages the use of open data to help improve its models, contributions to official datasets are generally the responsibility of its data team. Any extensions or contributions need to be coordinated with the Open Data team and can only contribute to the missing filling of their data.

8. *Any other comments?*

    - No.

References

Gebru, Timnit, Jamie Morgenstern, Briana Vecchione, Jennifer Wortman Vaughan, Hanna Wallach, Hal Daumé Iii, and Kate Crawford. 2021. "Datasheets for Datasets." *Communications of the ACM* 64 (12): 86–92.