GMR-47 A two-stage prediction model for house prices

Abstract

Predicting house prices is a challenging task that researchers from various fields (economics, statistics, politics, etc.) have attempted to answer. An accurate house prediction is useful not only to policymakers to improve their policies, but also to help sellers and buyers in the real estate market make well- informed decisions. Commonly, prediction models are trained on the whole dataset. However, as Azimlu et al [1] suggested, such models might not perform very well on dispersed data. They propose a new approach which first divides the whole dataset into smaller clusters, and then each cluster would be trained with an appropriate machine learning algorithm. It is approved to provide a more accurate prediction.

Materials and Methods

We will used a two-stage model for prediction.

In the first stage, we will divide the whole dataset into clusters based on their similarity. The clustering method we will use is K-means.
For each cluster, we will train with various machine learning algorithms (linear regression, Ridge, Lasso, ANN, etc.) to find the one with the highest performance using the Mean Normalized Absolute Error (MNAE) measure.

To avoid overfitting, we used cross-validation by splitting the data into training and validation sets (80-20).

Results

For whole dataset

	MAE	MNAE
Linear Regression	132,895.5146	0.2644
Ridge		
alpha = 1	132,746.0373	0.2641
alpha = 10	131,984.3302	0.2623
alpha = 100	130,877.6316	0.2560
Lasso		
alpha = 1	132,893.1768	0.26441
alpha = 10	132,872.3515	0.26438
alpha = 100	132,697.6504	0.2642
ANN	118,223.8212	0.2259

The ANN model produces the most accurate prediction with the MAE of $118, 223.82 and the MNAE of 0.2259

- For the Ridge and Lasso models, increasing alpha improves performance (reduces MAE/MNAE). But we must consider the trade-off between bias and accuracy. For that reason, in the clustering approach, we will use alpha = 10.
- As we expect, Ridge and Lasso models with alpha = 1 perform comparably with the Linear Regression model.

	Cluster0	Cluster1	Cluster2
# of samples	11,119	7,795	2,699
sqft_living	1,424.08	2,424.87	3,785.35
price	380,832.54	586,573.52	1,061,914.54

Observations:

- Individually, performance is improved across linear regression, Lasso, and Ridge. But it is worse for ANN.
- Cluster1 has the greatest accuracy, and Cluster2 has the worse accuracy.

Contact Information

Nguyen Thi Binh Nguyen <nnguy119@students.kennesaw.edu>
Brandon Bell <bbell31@students.kennesaw.edu>
Syanthan Reddy Ravula <sravula5@students.kennesaw.edu>
Hari Krishna Thota <hthota@students.kennesaw.edu>

References

[1] Fateme Azimlu, Shahryar Rahnamayan, and Masoud Makrehchi. 2021. House price prediction using clustering and genetic programming along with conducting a comparative study. In Proceedings of the Genetic and Evolutionary Computation Conference Companion (GECCO '21). Association for Computing Machinery, New York, NY, USA, 1809–1816. https://doi.org/10.1145/3449726.3463141

[2] Geerts, Margot, Seppe vanden Broucke, and Jochen De Weerdt. 2023. "A Survey of Methods and Input Data Types for House Price Prediction" ISPRS International Journal of Geo-Information 12, no. 5: 200. https://doi.org/10.3390/ijgi12050200

KENNESAW STATE UNIVERSITY
COLLEGE OF COMPUTING AND SOFTWARE ENGINEERING

Authors: Nguyen Thi Binh Nguyen; Brandon Bell; Syanthan Reddy Ravula; Hari Krishna Thota
Advisors: Dan Lo