House Price Prediction-A Machine Learning Approach

Supratik Sekhar Bhattacharya
A0228511M
School of Computing
National University of Singapore
e0674495@u.nus.edu

Abstract—**House price prediction is an important problem, especially in heavily developed housing markets such as Singapore. This work explores the nature of the data and uncovers hidden trends in it. The performance of over 40 classic baseline regression models on this task is tested. Tree-based models are found to perform well. A wide range of approaches are utilized in this work including dimensionality reduction, model explainability, etc. A couple of recommendation systems are developed and implemented. Insights are also determined from causal inference in machine learning & explainable artificial intelligence techniques.**

I. LITERATURE REVIEW

The workhorse random forest method is used by Hong et al (2020) to predict residential property prices in South Korea. Solatani et al (2022) introduce spatial-temporal considerations into the property price prediction problem. Spatiotemporal data often requires different kinds of analysis to the standard variety of problems. For instance, the standard linear models require modification to be utilized for time series problems. A case in point is methods such as ARIMA, ARIMAX, etc. which are designed for time series analysis. Spatiotemporal data have physical meanings as well as limited dimensions (Namely x,y,z, and t) which can make it difficult to use feature creation, dimensionality reduction, and other such methods to perform analysis restricting the degrees of freedom of the analysis that can be performed. The gold standard gradient boosting algorithm is used to perform house price prediction in Yang et al (2021) Maduri et al (2019) compare the performance of a number of different regression algorithms on a house price prediction dataset. Support vector machines can be hit or miss in terms of the out-of-sample generalization that they produce. Wu (2017) successfully used support vector regression for predicting house prices. Deep learning is leveraged to perform property price forecasting by Wang et al (2019) Testing for the consistency in relationships with relationships and importance across the top performing models is something important to look into, As evidenced by the case of L1 and L2 regularization (Cortes et al 2012) where it has been seen in the literature that small changes to the model such as adding in regularization can cause significant changes in working the model and the relationships between its parameters and the target even without a substantial change in model performance. From an interpretability point of view, it is important that we have insights from our analysis that enable us to make causal or at least confident predictions. (Molnar 2020) Apart from classical econometric (Huynh 2016) and linear models or tree-based approaches (Shimizu and Kaneko 202), Methods such as SHAP (Mokhtari 2019) and partial dependence plots (Johnson 2022) are important tools for explaining and interpreting machine learning models. Temur (2019) takes a time series approach to the house price forecasting problem and utilizes a blend of traditional time series models like ARIMA (Kalpakis 2001) with machine learning approaches such as LSTM (Yu 2019) to predict Turkish house prices. Like very many other assets, property markets are prone to exhibit bubbles, most famously, the 2008 housing bubble. (Baker 2008) Ayan and Eken (2021) leverage LSTM auto-encoders to detect price bubbles in the Istanbul housing market. In terms of predicting property prices in Singapore, it is important to take into consideration certain unique characteristics that one comes across. For instance, when it comes to predicting the price of Housing Development board properties. (Thong 2000) HDB properties have features such as of a maximum 99-year lease period which is somewhat unique.

2. INTRODUCTION

There is a known gulf between prediction and practice when it comes to house price prediction. Converting it into business values can be notoriously difficult. This phenomenon is perhaps most famously exemplified by the trajectory of Zillow's business which has run into many real-life challenges while giving good data for data science practitioners. The analysis begins by looking at the nature of the data. It is immediately clear that the house price is anything but normal in terms of its distribution. Given that this is the target variable itself data normalization cannot be done as is often done for the independent variables. The Weibull and Pareto distributions are found to be the best fits. The Weibull distribution has been extensively used in financial modeling. (Nadarajah and Kotz 2006 as well as Chen and Gerlach 2013) The Pareto distribution is the classic poster boy heavy tailed distribution. (Pareto 1991)

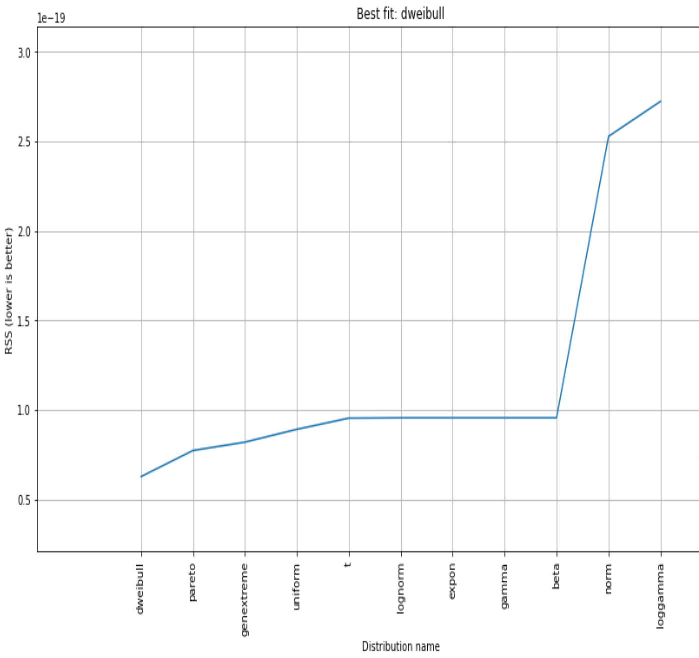


Fig 1-Distribution fits on price

Based on simple scatter plots, it is immediately apparent that without cleaning and filtering the trend lines are essentially invisible. It can be seen that outliers skew and snuff out the trends in the data. Without filtering out astronomically high house price homes and impossibly low ones there is no trend to be seen in the data. House prices that are off the charts are filtered out. The year built and the areas which sometimes seem fantastical are aligned with the test data as an anchor. The former has some future years and some of the areas are rather large and small too.

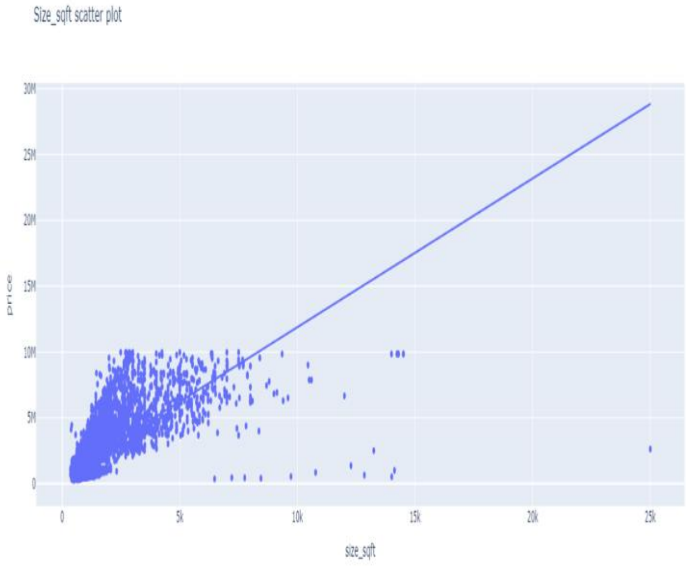


Fig-2 Size in square feet scatter plot

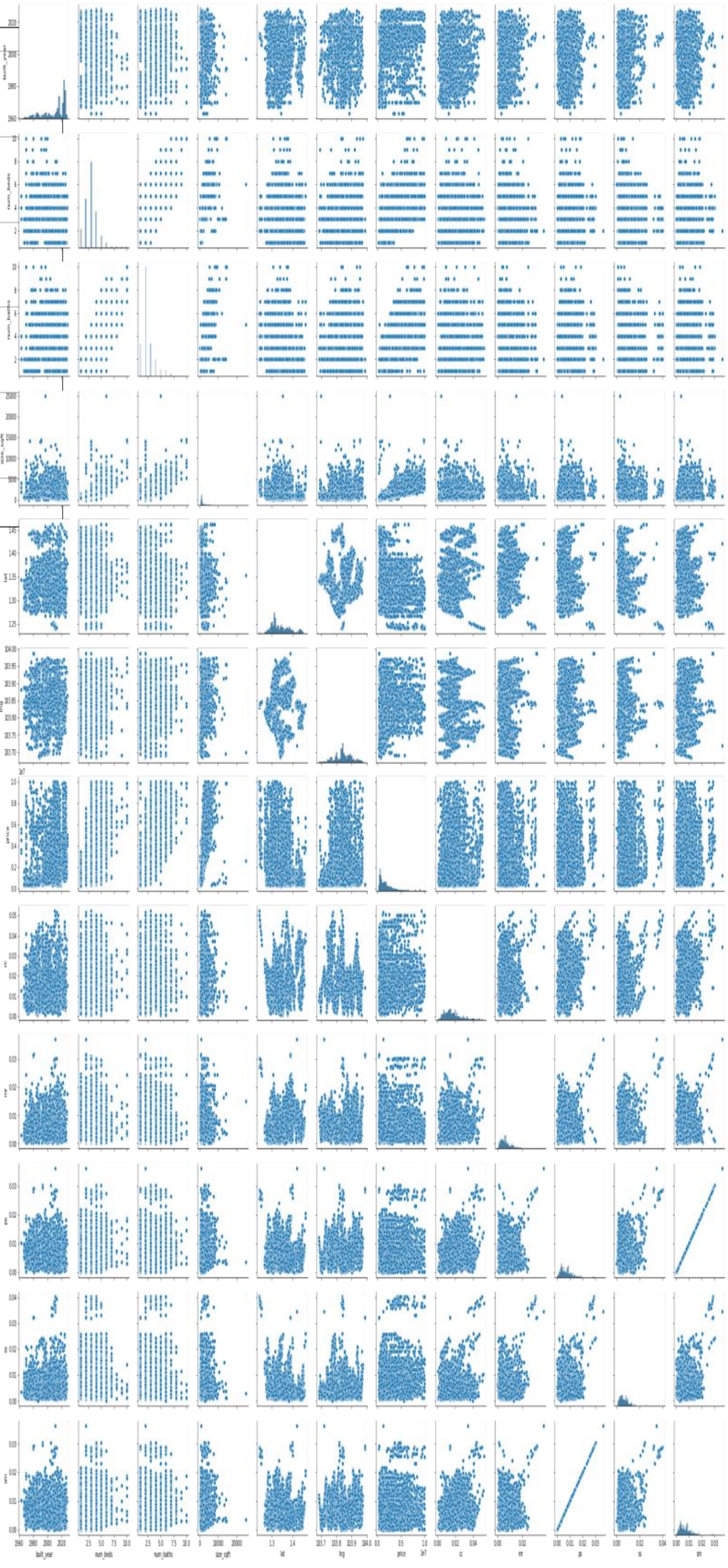


Fig 3-Pairwise plots between all numerical variables post outlier removal

For resolution and space reasons the plot cannot be made larger and the order of variables is as follows-built year, number of beds, number of baths, size in square feet, latitude, longitude, price, distance to a community center, distance to a MRT, distance to, distance to primary school, distance to secondary school and distance to a shopping mall. The maximum resolution plot is available in the results folder of the GitHub repository and in the notebook. After the filtering clear trend lines of price are visible in the scatter plots.

3. Materials and Methods

Library	Application
Lazypredict	Fitting regression models
Umap_learn	Creating Umap embedding's
Shap	SHAP values
Distfit	Fitting distributions
Sklearn	Making models, encodings etc.
Matplotlib and Plotly	Plotting
Numpy	Data wrangling
Pandas	Data wrangling & creating dummies
Category encoders	Target encoding

Table 1-Libraries used and their application

In terms of executing the methodology I briefly cover the libraries used in the implementation. Numpy and Pandas as always are used for working with the data in matrix and data frame form respectively. This includes working with the data frames, creating dummies etc. Plotly_express and matplotlib are used for plotting. Distfit allows us to estimate how good various probability distributions are as fits for the target variable. Umap_learn is used for generating Umap embeddings and is Umap typically considered the best for visualization. Sklearn as always is used for making machine learning models.

4. RESULTS AND DISCUSSION

4.1 TASK 1

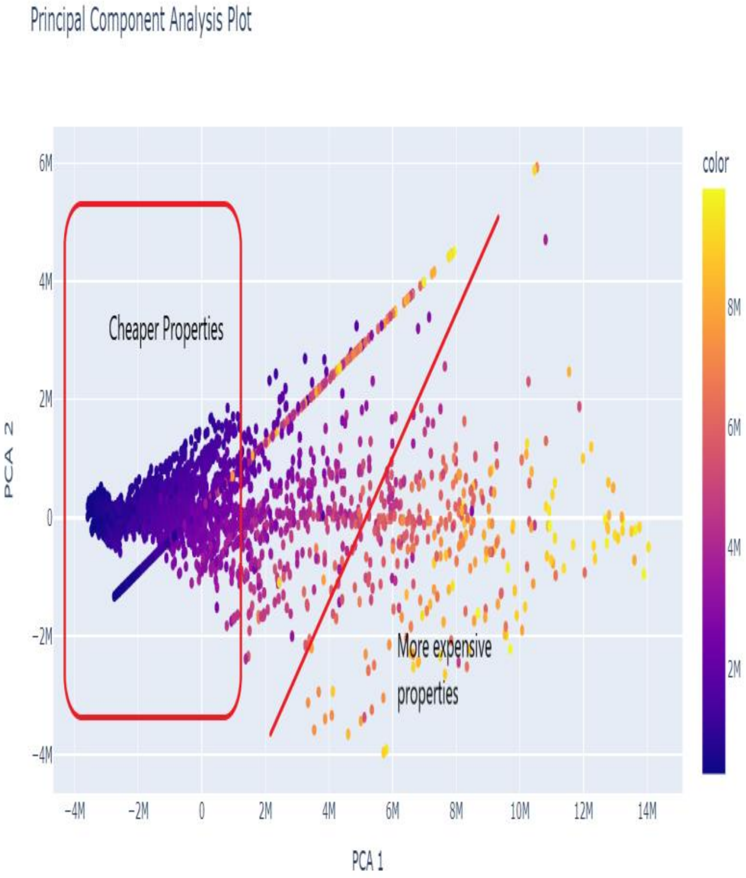


Fig 4-Principal component analysis (PCA) plot

Dimensionality reduction was performed for visualization purposes. Can be seen that higher-value properties cluster together in the PCA components which is an encouraging sign. It could be possible in future work to utilize clustering algorithms of various kinds to attempt to separate them out. PCA uses second-order moments while Independent component analysis does so for higher-order ones. An implementation of Fast ICA is also given in the notebook which results in a plot similar to the PCA plot.

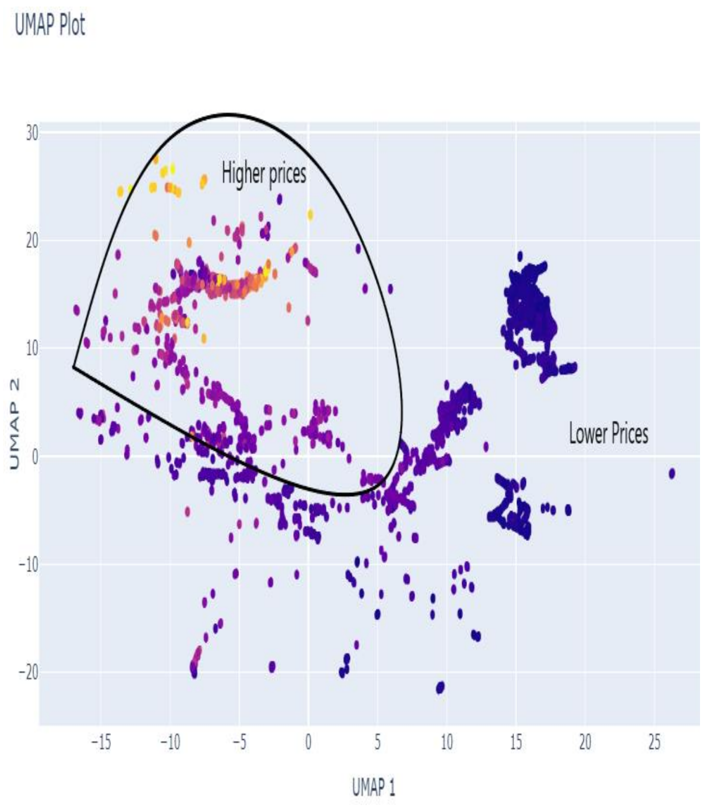


Fig 5-Uniform manifold approximation and projection (UMAP) plot

Neural network auto encoders were not found to give good separation in the specifications tested out. The best 'visual' clustering effects were found in the Umap embedding's.

The classic prediction metrics for a regression task are defined as follows

$$\text{Root mean square error (RMSE)}=\sqrt{\sum_{i=1}^{N}\frac{(\widehat{y}_i-y_i)^2}{N}}$$

$$R^2 = \frac{\sum_{i=1}^{N}(\widehat{y}_i-\overline{y})^2}{\sum_{i=1}^{N}(y_i-\overline{y})^2}$$

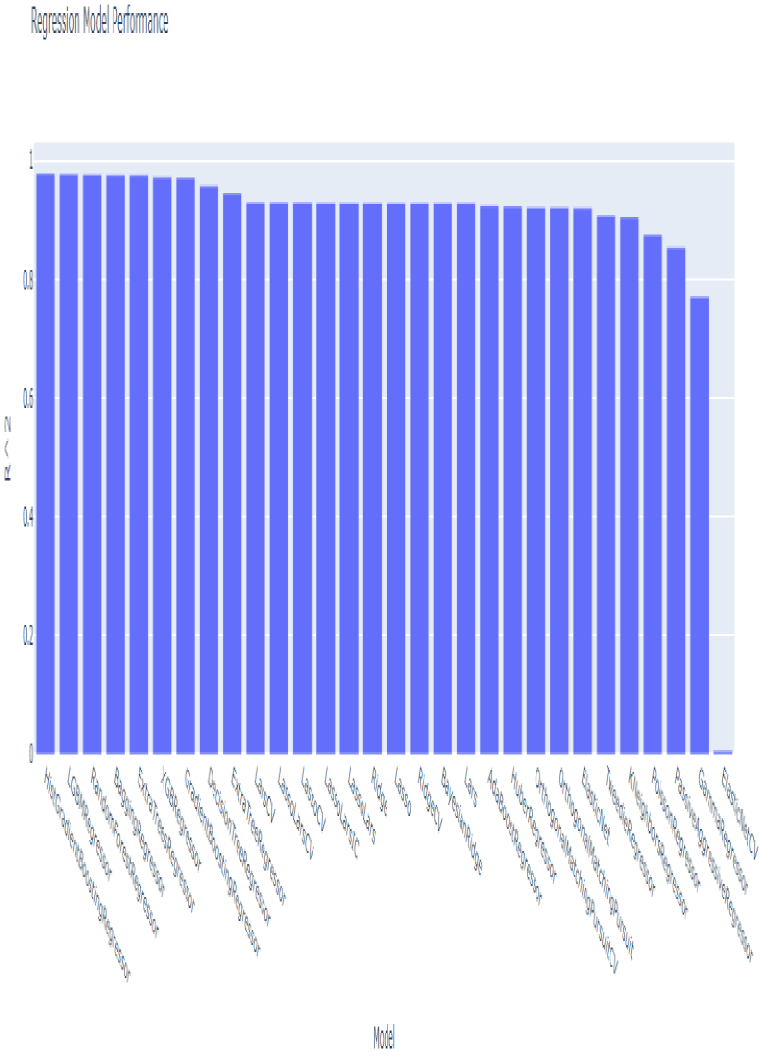


Fig 6-Model performances

Once the rationalizing is done the whole suite of regression models are utilized for the house price prediction task. To determine the best baseline models 42 Machine learning algorithms in all are fitted on the data and their performance determined. The top 30 model results are displayed in figure 5. The top performing models are found to be light gradient boosting and histgradientboosting. This is followed by random forest, bagged trees, extratrees regressor, extreme gradient boosting, gradient boosting regressor etc. Tree based models perform the best as a class of models.

Number of estimators	Number of leaves
Maximum depth	Learning rate
Objective	Minimum split gain
Minimum child weight	Minimum child samples
Alpha (L1 regularization coefficient)	Lambda (L2 regularization coefficient)
Boosting type	Subsample for bin

Table 2-Light gradient boosting parameters

Grid search cross validation is used for hyper-parameter optimization of the light gradient boosting model. Grid search divides the search space into a grid with equal sized parts and tests out each combination of hyper-parameters. Note that the difference between hyper-parameters and parameters from a model point of view is that the latter are updated during training and the former are not. Hence, testing out a new hyper-parameter combination requires the model to be run again.

4.2 TASK 2

Coming to task 2, two implementations are provided. For the first, K nearest neighbors are used to retrieve the nearest neighbors to that of any row. The function returns the top k nearest neighbors.

Euclidean	Braycurtis	Canberra
Chebyshev	Cityblock	Correlation
Cosine	Minkowski	Seuclidean

Table 3-The 9 vector distance metrics usable for the recommendation engine

The recommendations can also be made using 9 different vector distance metrics in the literature. Following this hierarchical clustering is performed and the plot is displayed to demonstrate by visualization the differences that arise from the choice of the distance metric. It can be further tuned by techniques such as weighting the columns by their relative importance. It might be deemed for example that the price is relatively more important than all other variables and that might be given a larger weight in such a setting or anything in that vein. In terms of design choices, multiple types of linkages exist including ward, single, average, etc. which could be tried out. The distance matrices obtained from different metrics can also be used for various other types of clustering algorithms In terms of practical real-time implementations there are other aspects to it too. Cold start is something that needs to always be taken into consideration. A typical approach to alleviating the cold start problem is to simply display the most popular or trending (such as high-traffic) properties. Another approach that could be used if user profile data exists is to show properties that are popular among similar users. Typically this works to the effect of something like finding k most similar people and finding what they like most and displaying those. Systems often help to have some randomization in the system. In terms of maximizing payoffs over time, an epsilon greedy type strategy could be applied. Wherein most users are shown the best guess at that instant in time with some space for exploration and trying out other options the rest of the time. Other well-known methods like Thompson sampling exist too. (Agrawal and Goyal 2012)

4.3 TASK 3

As pertains to task 3, I look at the insights gleaned from the explainable AI and causal inference techniques as applied to the data.

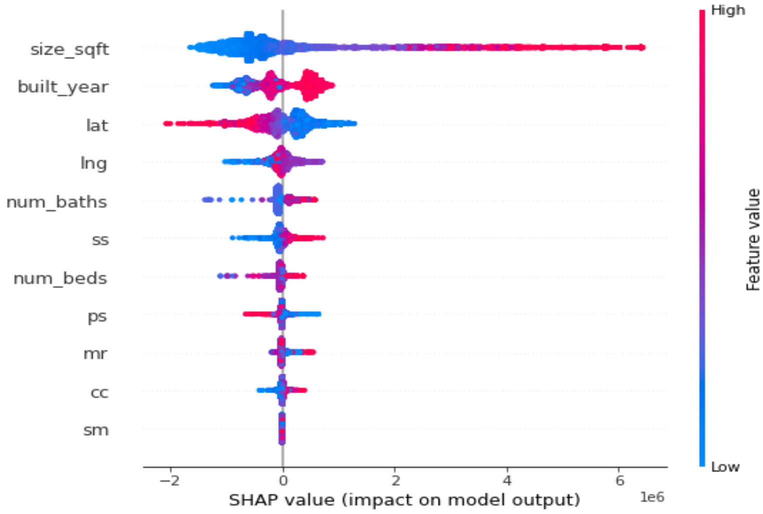
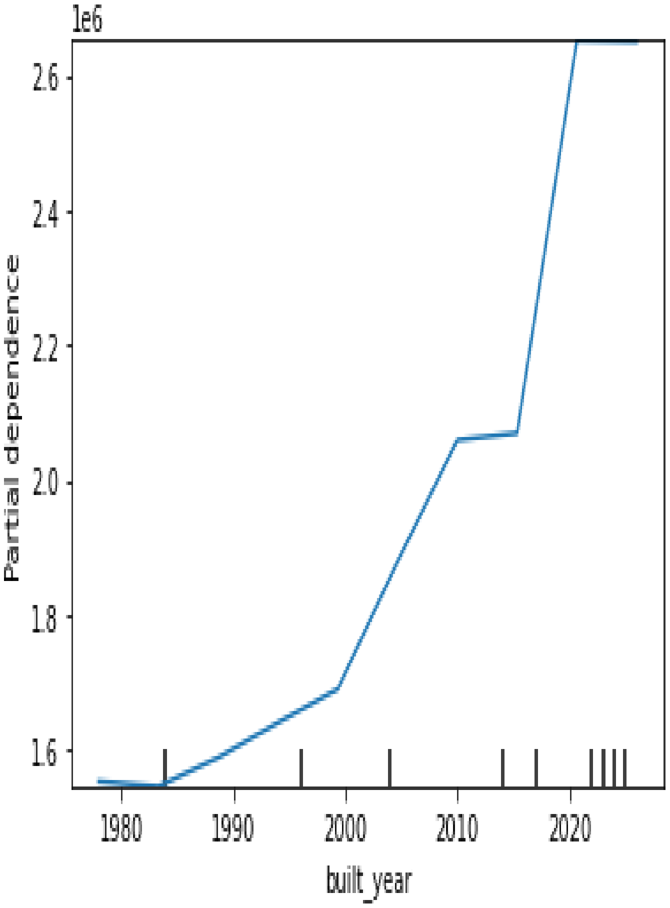
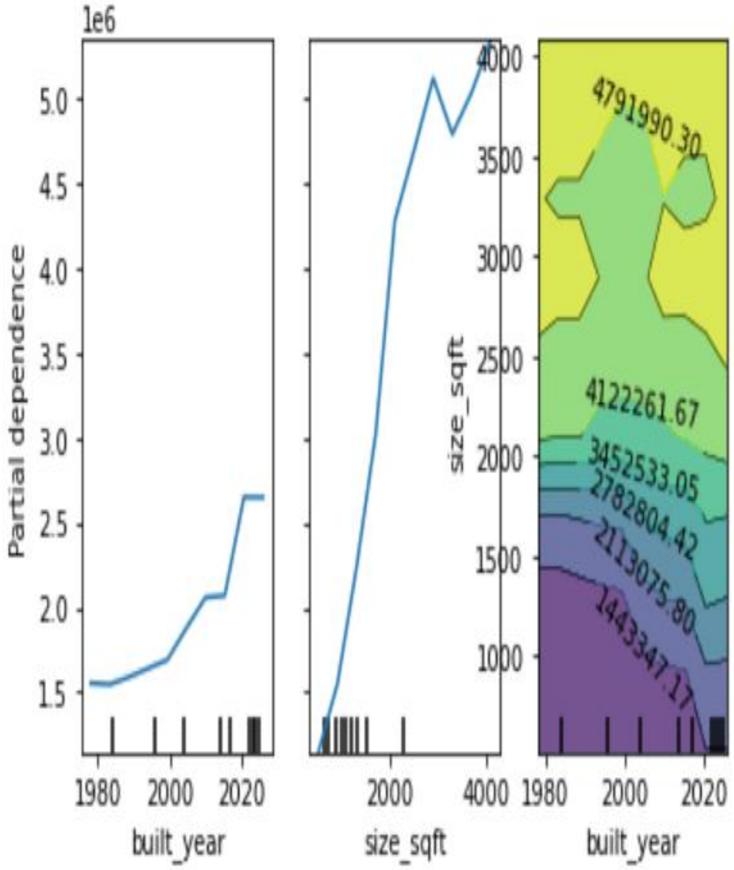


Fig 7-SHAP values





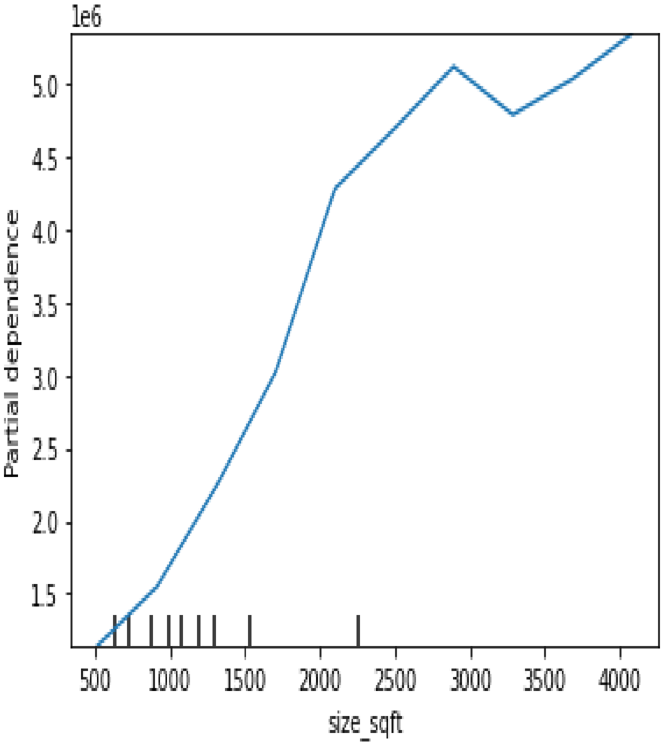


Fig 8, 9, 10-Partial dependence plots

Partial dependence plots and SHAP values enable us to peek inside the working of the. Black box models. The area of the property and the built_year are found to be the most consistently important variable. Positive relationships are also found with the number of baths. The distance to community centers and MRT stations are also found to have a bearing on the price.

LI (Lasso) and L2 (Ridge) regularization:

$$L1=\sum_{i=0}^{N}(y_i - X_i\beta)^2+\alpha * \sum_{i=0}^{N}\left|\beta_i\right|$$

$$L2=\sum_{i=0}^{N}(y_i - X_i\beta)^2+\lambda * \sum_{i=0}^{N}\beta_i^2$$

Finally L1 and L2 regularization when performed show stability of the results giving us some confidence that the models make sense. The plots can be seen in the notebook and GitHub repository. Stable coefficient values are seen over a wide range of the regularization coefficients.

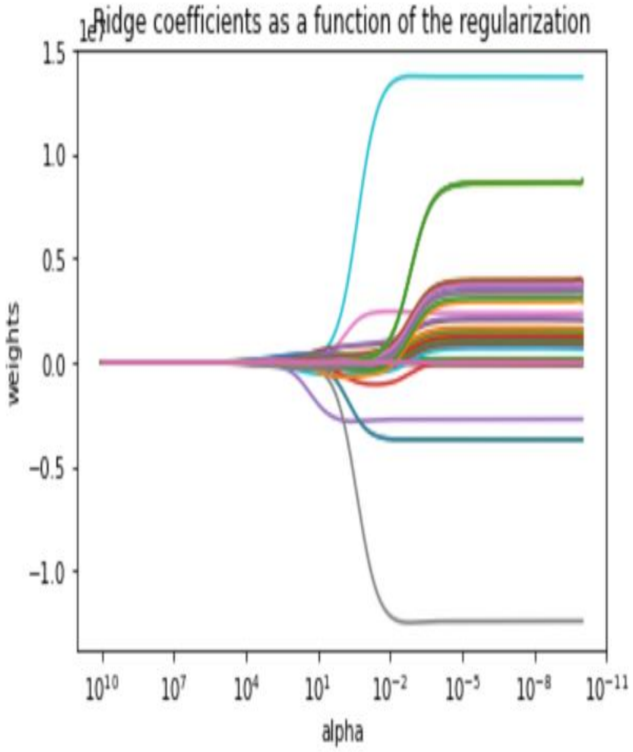


Fig 11-Ridge plot

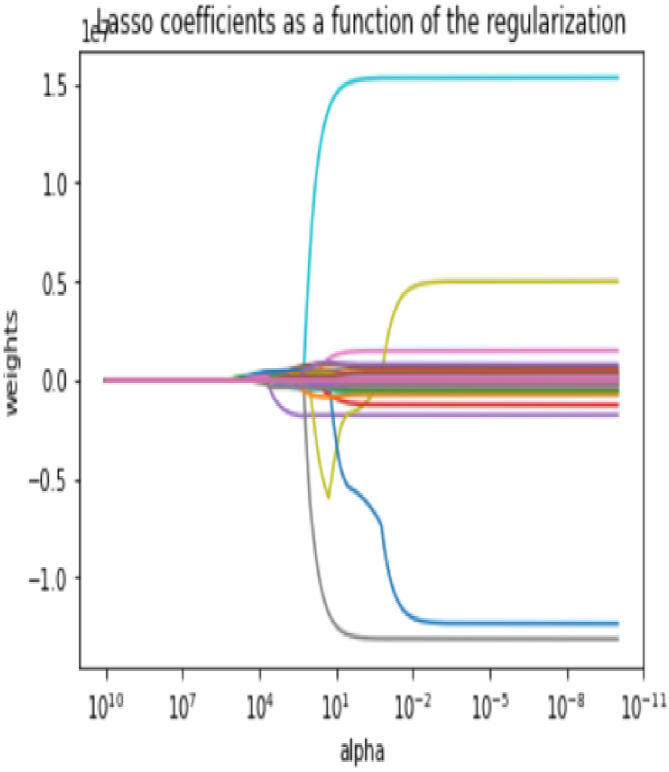


Fig 12-Lasso plot

References

[1] 1. Yu, Y., Si, X., Hu, C., & Zhang, J. (2019). A review of recurrent neural networks: LSTM cells and network architectures. Neural computation, 31(7), 1235-1270.

[2] 2. Soltani, A., Heydari, M., Aghaei, F., & Pettit, C. J. (2022). Housing price prediction incorporating spatio-temporal dependency into machine learning algorithms. Cities, 131, 103941.

[3] 3. Hong, J., Choi, H., & Kim, W. S. (2020). A house price valuation based on the random forest approach: the mass appraisal of residential property in South Korea. International Journal of Strategic Property Management, 24(3), 140-152.

[4] 4. Madhuri, C. R., Anuradha, G., & Pujitha, M. V. (2019, March). House price prediction using regression techniques: a comparative study. In 2019 International conference on smart structures and systems (ICSSS) (pp. 1-5). IEEE.

[5] 5. Molnar, C. (2020). Interpretable machine learning. Lulu. com.

[6] 6. Thong, J. Y., Yap, C. S., & Seah, K. L. (2000). Business process reengineering in the public sector: the case of the Housing Development Board in Singapore. Journal of Management Information Systems, 17(1), 245-270.

[7] 7. Cortes, C., Mohri, M., & Rostamizadeh, A. (2012). L2 regularization for learning kernels. arXiv preprint arXiv:1205.2653.

[8] 8. Yang, L., Liang, Y., Zhu, Q., & Chu, X. (2021). Machine learning for inference: Using gradient boosting decision tree to assess non-linear effects of bus rapid transit on house prices. Annals of GIS, 27(3), 273-284.

[9] 9. Wu, J. Y. (2017). Housing price prediction using support vector regression.

[10] 10. Wang, F., Zou, Y., Zhang, H., & Shi, H. (2019, October). House price prediction approach based on deep learning and ARIMA model. In 2019 IEEE 7th International Conference on Computer Science and Network Technology (ICCSNT) (pp. 303-307). IEEE.

[11] 11. Shin, E. K., Kim, E. M., & Hong, T. H. (2021). Real Estate Price Forecasting by Exploiting the Regional Analysis Based on SOM and LSTM. The Journal of Information Systems, 30(2), 147-163.

[12] 12. Ayan, E., & Eken, S. (2021). Detection of price bubbles in Istanbul housing market using LSTM autoencoders: A district-based approach. Soft Computing, 25(12), 7957-7973.

[13] 13. Temur, A. S., Akgün, M., & Temur, G. (2019). Predicting housing sales in Turkey using ARIMA, LSTM and hybrid models.

[14] 14. Kalpakis, K., Gada, D., & Puttagunta, V. (2001, November). Distance measures for effective clustering of ARIMA time-series. In Proceedings 2001 IEEE international conference on data mining (pp. 273-280). IEEE.

[15] 15. Baker, D. (2008). The housing bubble and the financial crisis. Real-world economics review, 46(20), 73-81.

[16] 16. Huynh, V. N., Kreinovich, V., & Sriboonchitta, S. (Eds.). (2016). Causal inference in econometrics. Springer.

[17] 17. SHIMIZU, N., & KANEKO, H. (2021). Constructing regression models with high prediction accuracy and interpretability based on decision tree and random forests. Journal of Computer Chemistry, Japan, 20(2), 71-87.

[18] 18. Mokhtari, K. E., Higdon, B. P., & Başar, A. (2019, November). Interpreting financial time series with SHAP values. In Proceedings of the 29th Annual International Conference on Computer Science and Software Engineering (pp. 166-172).

[19] 19. Johnson, P. M., Barbour, W., Camp, J. V., & Baroud, H. (2022). Using machine learning to examine freight network spatial vulnerabilities to disasters: A new take on partial dependence plots. Transportation Research Interdisciplinary Perspectives, 14, 100617.

[20] 20. Nadarajah, S., & Kotz, S. (2006). The modified Weibull distribution for asset returns. Quantitative Finance, 6(6), 449-449.

[21] 21. Chen, Q., & Gerlach, R. H. (2013). The two-sided Weibull distribution and forecasting financial tail risk. International Journal of Forecasting, 29(4), 527-540.

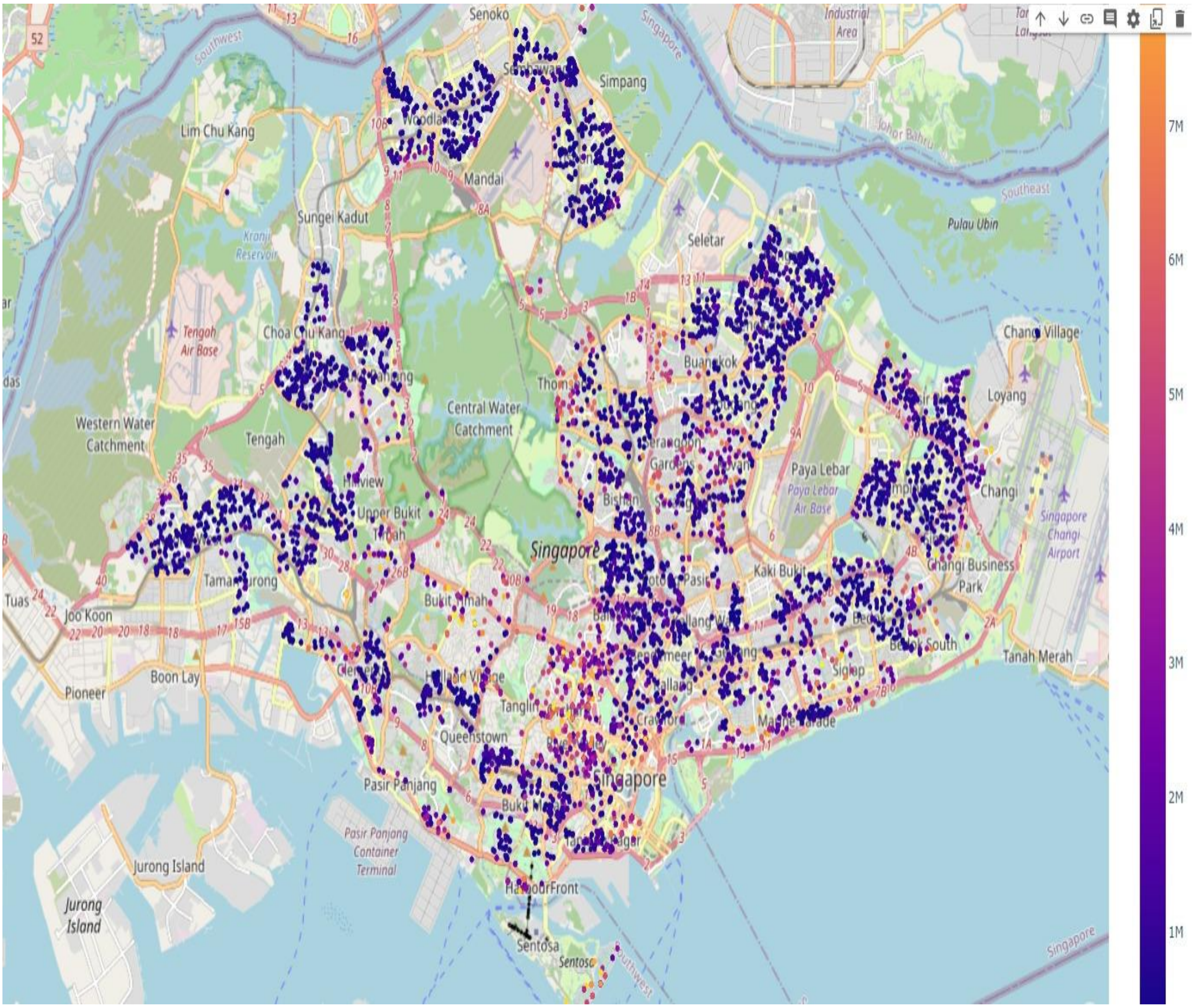
[22] 22. Pareto, V. (1991). The rise and fall of the elites: an application of theoretical sociology. transaction publishers.

[23] 23. Kolenikov, S., & Angeles, G. (2004). The use of discrete data in PCA: theory, simulations, and applications to socioeconomic indices. *Chapel Hill: Carolina Population Center, University of North Carolina*, *20*, 1-59.

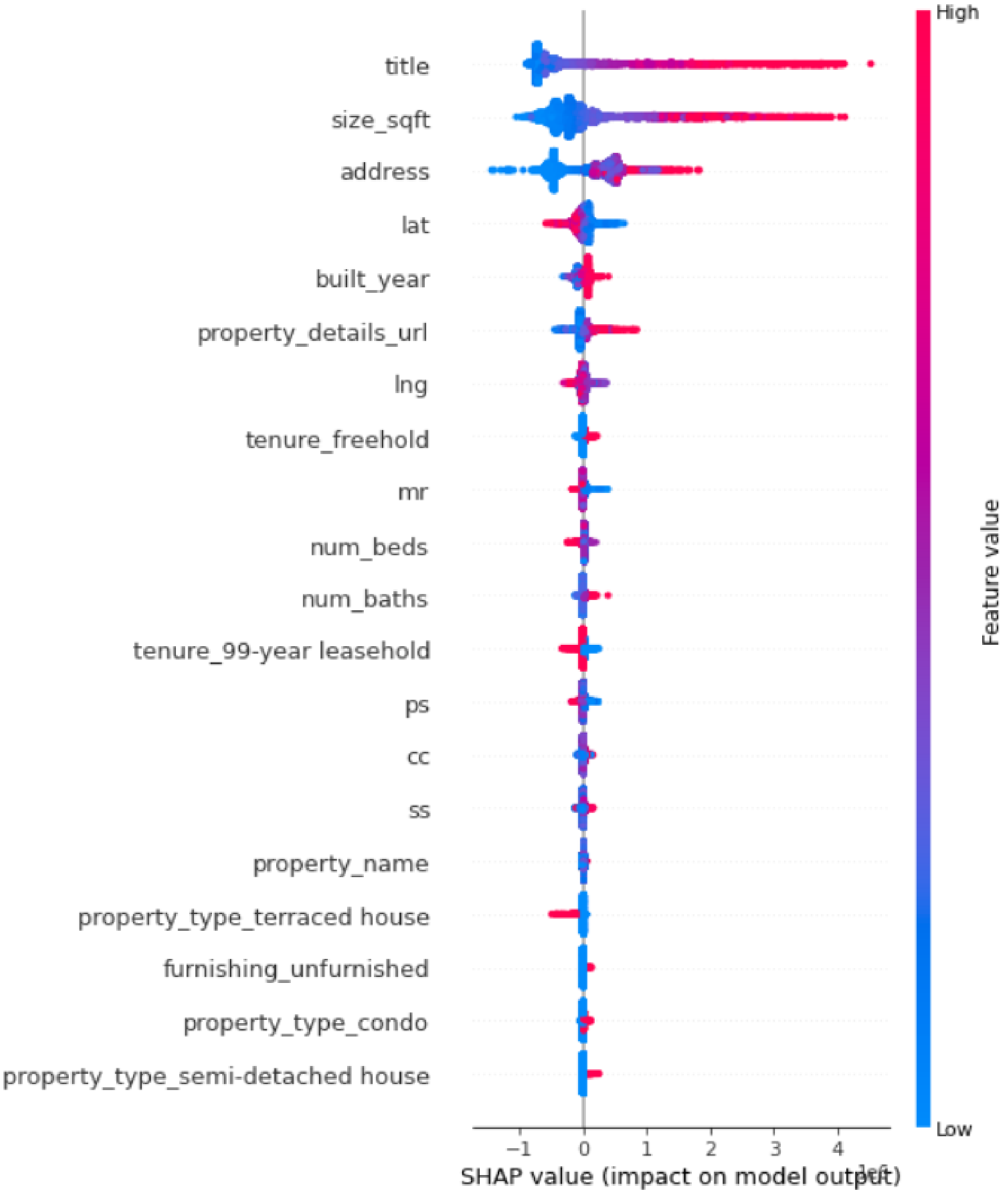
[24] 24, Kamishima, T., & Akaho, S. (2011, October). Personalized pricing recommender system: Multi-stage epsilon-greedy approach. In *Proceedings of the 2nd International Workshop on Information Heterogeneity and Fusion in Recommender Systems* (pp. 57-64).

[25] 25, Agrawal, S., & Goyal, N. (2012, June). Analysis of thompson sampling for the multi-armed bandit problem. In *Conference on learning theory* (pp. 39-1). JMLR Workshop and Conference Proceedings.

APPENDIX



Appendix Fig 1-Singapore properties by price range



Appendix Fig 2-SHAP values including categorical and target encoding variables