NUS INVESTMENT SOCIETY

Quantitative Finance Department

Predicting Housing Prices in Singapore

Akshai Vengat, Hemashree A, Ivan Lee, Khairul Iman, Wang Kexin

June 2020

Table of Contents

Abstract

By collecting housing data from the government and OneMap API, we consolidated and augmented our housing resale dataset. This gave us access to a clearer dataset for housing in Singapore, from the period of 1990 to 2020. Afterwards, we applied regression techniques such as Lasso and Bayesian Ridge as well as deep learning methods such as Artificial Neural Networks to see if such a model would be better in predicting housing prices given a housing unit's location. Statistical plots such as QQplots, residual vs fitted values along with other parameters such as AIC, BIC and standard measurements of accuracy for machine learning models such as MSE, R-squared were used to evaluate the models.

Introduction

Being able to predict a property's price helps buyers and sellers alike, to make informed decisions. A wide range of factors affects a housing unit's price, such as the location of the property, area of the property, type of unit and even the economic conditions at the time of transaction. In the context of Singapore, there are two main types of housing - public and private, and the ability of these different types of housing to command premium prices differs. Hence, given that many factors are in play, accurate prediction of a property's price at a given time is a complex process. Public housing in Singapore was chosen as the focus of this paper as data on public housing was more complete and spanned over a larger time period from 1990 to 2020.

This study attempts to apply machine learning techniques to resale housing data provided by the Singapore Government and external organisations to predict the resale housing price given the details of a particular property. We approached this by applying pre-processing techniques to the time series data, and splitting the raw data in testing and training sets. After which, by applying various machine learning models and observing their accuracy and error results to predict which model would fit best for this task.

Literature Review

Lehner(2011) modelled housing prices in Singapore by applying spatial hedonic regression on 6 models: and in his study, Private sale combined, Private sale transaction, Private rental asking, HDB sale combined, HDB sale transaction, HDB rental asking. It was found that floor area plays a major factor in influencing prices of individual properties and a relatively high importance of the distance to the CBD in all markets. On the other hand, the year of construction of the property indicates a much lower impact in the private sale market than in private rental and HDB sale market and private housing prices are slightly less sensitive to the proximity to top secondary schools than HDB prices. His investigation also noted that a better replacement of distance from points of interest would be travel time or by generating accessibility indexes.

Lehner(2011) also determined 3 major factors that affect housing prices. This includes structural factors such as age of the property, floor area, floor level and presence of amenities such as swimming pool, garden, car parks and level of security, locational factors such as distance of the housing unit from points of interest such as the Central Business District(CBD), public transport, top schools, as well as shopping malls. Contractual factors such as housing tenure of a property also affects its price, whereby a freehold property is able to command higher premium prices compared to a leasehold property.

In another study conducted by Lu, Li, Qin, Yang, Goh (2017), a creative feature engineering was used to determine a suitable regression technique for house prices prediction. Ridge,

Lasso and Gradient boosting were found to be the most useful regression algorithms as Ridge and Lasso are commonly used to model cases with large numbers of features. In order to avoid overfitting, Ridge regression performs L2 regularization and Lasso regression performs L1 regularization. Hybrid regression was also conducted as it was found to be better than one specific regression algorithm. For both Ridge and Lasso, the minimum Root Mean Squared Error (RMSE) is 0.112276, for 160 features. Similar results were obtained for Gradient boosting regression. The best parameters for training data is when Subsample = 0.5, with 230 features. Finally, considering the Coupling effect among regression algorithms the best hybrid regression result for the test data is 0.11260 with 65% Lasso and 35% Gradient boosting.

Methodology

Overview

To form the final dataset, relevant data was obtained from various sources such as Data.gov, OneMap API and Singstat. The final dataset was then separated into training and testing sets, which were tested with Lasso and Bayesian Ridge regression models. Deep learning techniques such as Artificial Neural Networks are then used to compare the models' effectiveness in predicting a housing unit's price.

Raw data Collection

In Singapore, public housing is regulated and authorised by the Housing Development Board (HDB) where resale data is accurately tracked and published for public viewing. Data for both public and private housing within the years 1990-2020 was collected from various government sources such as Data.gov and the Urban Redevelopment Agency (URA) website.

However, key information such as specific unit numbers of flats and condominiums were omitted from the datasets in order to protect the homeowner's identity. They were instead replaced with a range of floors. Unit numbers of landed properties were also omitted. This reduces the model accuracy and feature engineering methods were thus used to generate more input data based on currently available dataset features to further improve model accuracy. The techniques used are outlined below.

Location data was extracted by combining block numbers and their corresponding street names, after which a search query was run on OneMap API for first location data, which provided a location's geolocation. Current data on public infrastructures such as public transport (bus stops, MRT and LRT stations) and shopping malls, the current infrastructure data was obtained from OneMap API as well. The absolute distance between the housing units and their nearby

public infrastructure was then compared using the haversine function, and were considered to be ‘near’ the particular housing unit if the absolute distance was within a particular cut-off distance. Yearly GDP values were collected from SingStat which were used as economic indicators of each year.

Data Analysis

From the final dataset, a number of parameters were identified as key features that affect housing prices. The specific creation details for each parameter are outlined in Appendix A.

Analysis of resale HDB prices

Figure 1 below shows a sample of the raw data collected from resale HDB prices from Data.gov.

	month	town	flat_type	block	street_name	storey_range	floor_area_sqm	flat_model	lease_commence_date	resale_price
0	1990-01	ANG MO KIO	1 ROOM	309	ANG MO KIO AVE 1	10 TO 12	31.0	IMPROVED	1977	9000
1	1990-01	ANG MO KIO	1 ROOM	309	ANG MO KIO AVE 1	04 TO 06	31.0	IMPROVED	1977	6000
2	1990-01	ANG MO KIO	1 ROOM	309	ANG MO KIO AVE 1	10 TO 12	31.0	IMPROVED	1977	8000
3	1990-01	ANG MO KIO	1 ROOM	309	ANG MO KIO AVE 1	07 TO 09	31.0	IMPROVED	1977	6000
4	1990-01	ANG MO KIO	3 ROOM	216	ANG MO KIO AVE 1	04 TO 06	73.0	NEW GENERATION	1976	47200

Figure 1. Part of HDB resale data

Consolidating this information from 1990 to 2020, we are able to come up with a dataset which tracks all resale housing transactions for public housing in Singapore.

Economic indicators

The state of the economy also affects the resale price of HDB houses. During recession, we expect housing prices to be lower relative to when the economy is booming. As a proxy of the state of the economy, we add in **yearly GDP** values into our dataset, using the GDP value of the year before the house is sold (prevent look-ahead GDP data).

Time variables

As seen in figure 1 above, resale prices vary with time and generally follow time series processes, hence variables that vary with time are also included to capture the time series process that resale prices follow. Some of these variables are:

Last known price of apartment

One of the problems with the housing price dataset is that observations are not consistent - each observation is based on particular units that were sold and there is no consistency in the frequency of such housing resales. It can be a year before someone buys a flat or a resale can happen in the following month in the same HDB building. As a result, we use the last known resale price of the apartment as an indicator of what housing prices will be.

Time since last known price

As stated above, frequency of resales do not have a pattern and hence the last known price can be a month ago, or the year before. We capture this information as well.

Year of sale of house

We expect a time trend in housing prices - as the country grows, we expect housing prices to rise in general, as seen in the rise of housing prices in general from 1990 to present.

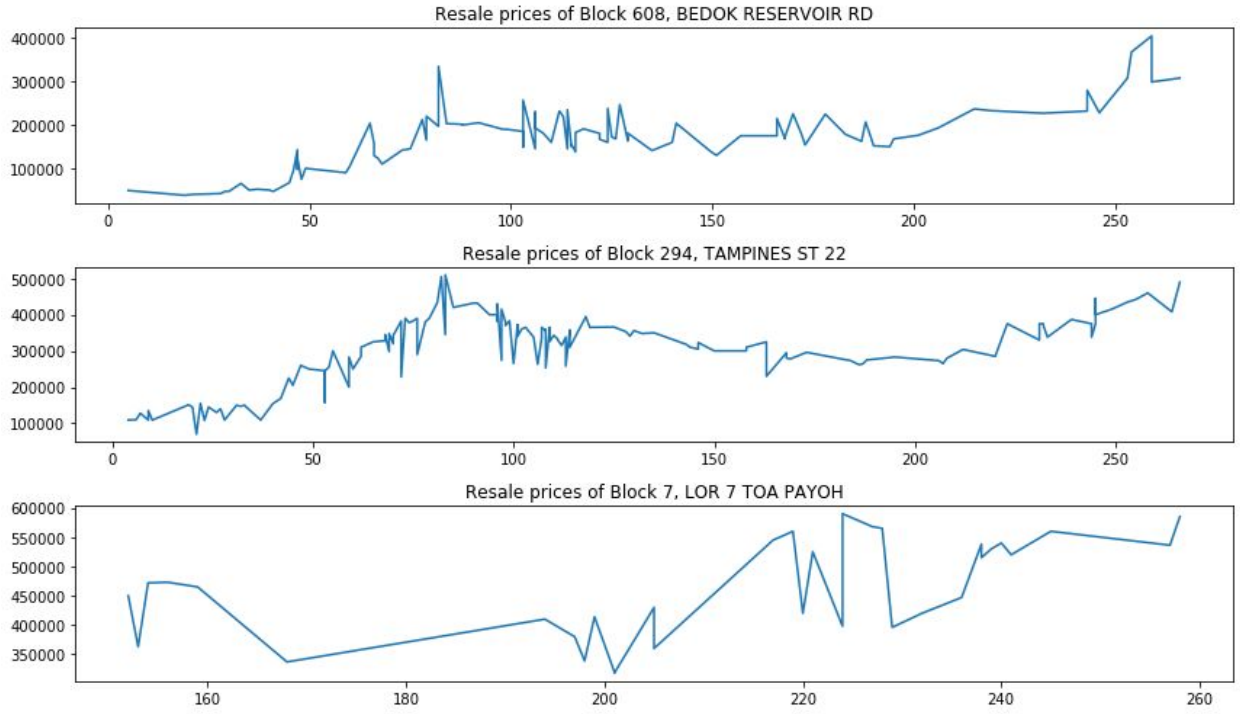
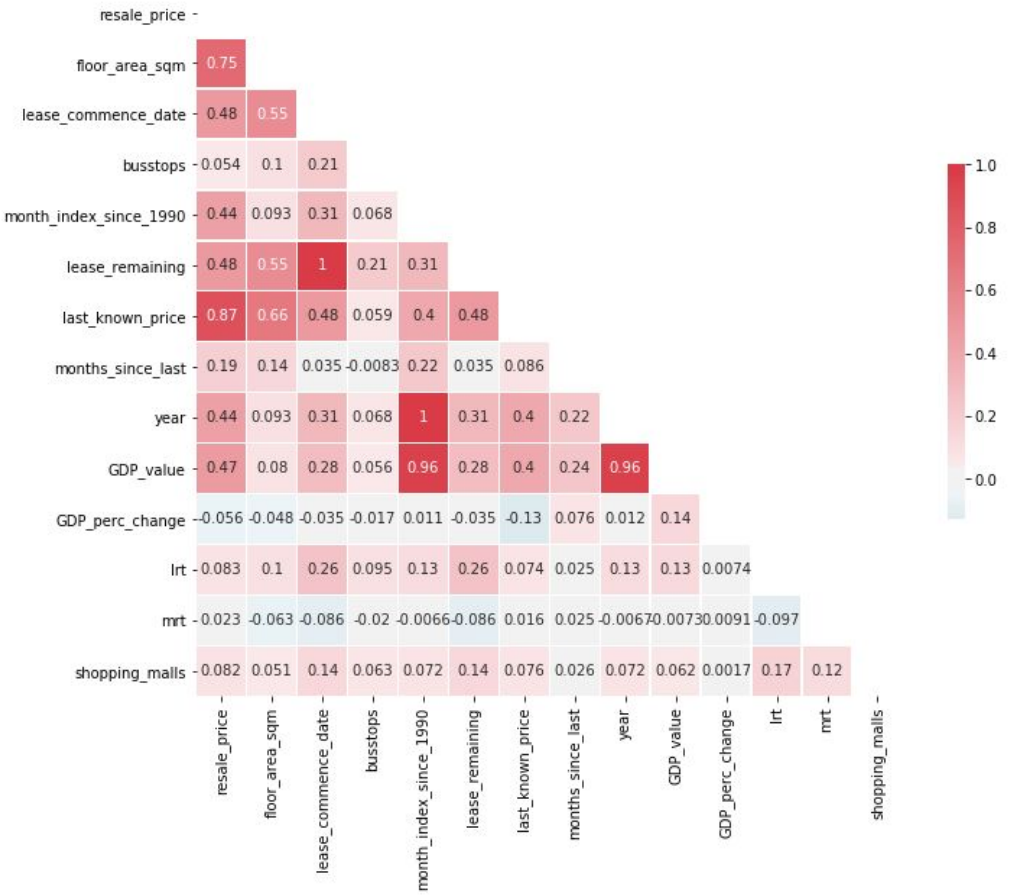


Figure 1. Housing prices of 3 random units with time

Analysis

The following is a heatmap of the correlation matrix of the final dataset.



From the correlation matrix, the proximity of flats to the various facilities appears to have a poor correlation with resale prices, with only a correlation coefficient of 0.083, 0.023, 0.054 and 0.082 for "mrt", "lrt", "busstops" and "shopping malls" respectively. However, this could be due to a limitation during the extraction of location information. Complete data on infrastructure was only available for the current year of 2020 as such all past housing data could only be compared to infrastructure existing in 2020, instead of infrastructure available during the year of transaction.

"Year" and "month_index_since_1990" are perfectly correlated and that is because the latter is formed by looking at the number of months since 1990. "Year" is removed from the dataset used for model fitting.

The remaining variables are kept to be fit into regression models. While some of these variables do not seem to be strongly correlated to resale prices, we believe there is a theoretical relationship between these variables and "resale_prices", as explained in Appendix A.

Our final processed dataset would look as such:

Model Fitting

To build datasets that would fit into machine learning methods, we consider the following:

For categorical data such as type of apartments, a boolean type would be added for each distinct apartment type, converting the categorical string data type to a numerical one. Categorical data with features which are already numerical in value was converted to a value between 0 and 1, before being validated against the predictive models.

One-hot encoding is applied for categorical data as well.
For ordinal variables, scikit-learn's LabelEncoding was employed to encode variables numerically in a specific order such that values that should be of higher value would be encoded as such.

Time-series Modelling
Knowing past values of resale prices would allow us to treat the data as time series, allowing us to apply time-series methods to the dataset.

However, as the dataset includes data from various entities, or HDB flats, over the years, it wouldn't be computationally feasible to fit a time-series model for each unique hdb flat, which amounts to 8142 different time-series models.

Coupled with the fact that data isn't consistent -- data is not in set intervals (eg, monthly), a time-series model (Autoregression, GARCH, etc.) is not feasible.

This leaves us with a huge problem - resale prices generally increase with time and we have no way of modelling that. We minimize this by introducing predictor variables that have a time series process as well, to establish a relationship between resale price and these variables in regression models.

Hence, we consider the following methods. Note that only important details about the methods are explained and specific theoretical details are beyond the scope of the paper.

Possible Models

Lasso

Lasso is a variant of the standard linear regression to prevent overfitting of the model. Lasso is similar to Linear Regression but uses L1 Regularisation. Instead of minimizing MSE, the loss function also has an added penalty, which is the weighted sum of the coefficients. L1 loss function has a constraint as such:

$$\sum_{i=1}^{M}(y_i - \hat{y}_i)^2 = \sum_{i=1}^{M}\left(y_i - \sum_{j=0}^{p} w_j \times x_{ij}\right)^2 + \lambda \sum_{j=0}^{p} |w_j|$$

Cost function for Lasso regression

Lasso models indirectly perform model selection as well. Depending on the magnitude of lambda, some of the coefficient values can be 0, effectively removing the effect of the variable in the prediction of resale prices.

Bayesian Ridge

The Bayesian Ridge extends on ridge regression, which uses L2 penalty instead of L1 penalty as shown above. The ridge minimizes the following equation:

$$\min_{w} ||Xw - y||_2^2 + \alpha||w||_2^2$$

A Bayesian view of ridge regression is obtained by noting that the minimizer of Ridge Regression can be considered as the posterior mean of a model where the coefficients follow a gaussian distribution (prior). Compared to the OLS (ordinary least squares) estimator, the coefficient weights are slightly shifted toward zeros, which stabilises them. Hyperparameters are trained to maximise log marginal likelihood.

Neural Networks

Neural Networks, specifically Artificial Neural Network, is used to fit a non-linear model for resale prices using the available variables. Neural Networks apply gradient descent to iteratively update its weights so as to minimise its defined loss function, in this case MSE. Neural Networks are generally known as black-box models that perform exceptionally well in their tasks - forecasting - as the dataset grows. As our dataset is large enough, we use Neural Networks and compare against Lasso and Bayesian Ridge models as well.

A subset of the dataset is then fitted into the 3 different statistical algorithms for comparison for model validity and performance.

Fitted Model

These models are then fit to a subset of the dataset (about 75% of observations) while the remaining is treated as unseen data for model evaluation performance. The following models are fit using k-fold cross valuation (k=4).

Model Assumptions

Neural Network models do not have any underlying assumptions in its algorithm. However, there are some assumptions for linear models such as Lasso. In this section, we look to challenge the assumptions of Linear Models - Linearity, Heteroskedasticity and Normality. We can look at both in residual plots.

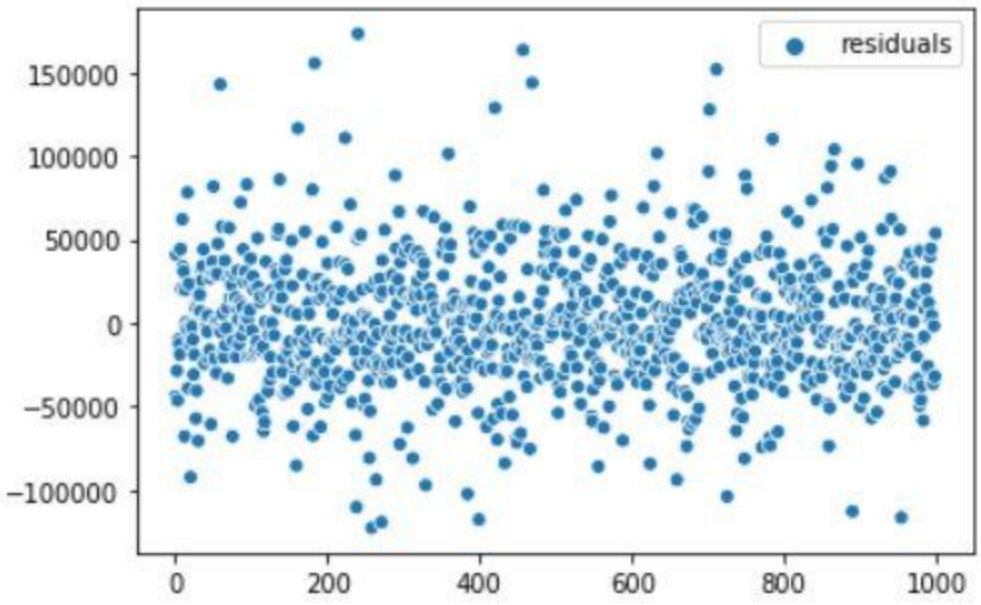


Figure 2. Scatterplot of residuals, 1000 datapoints

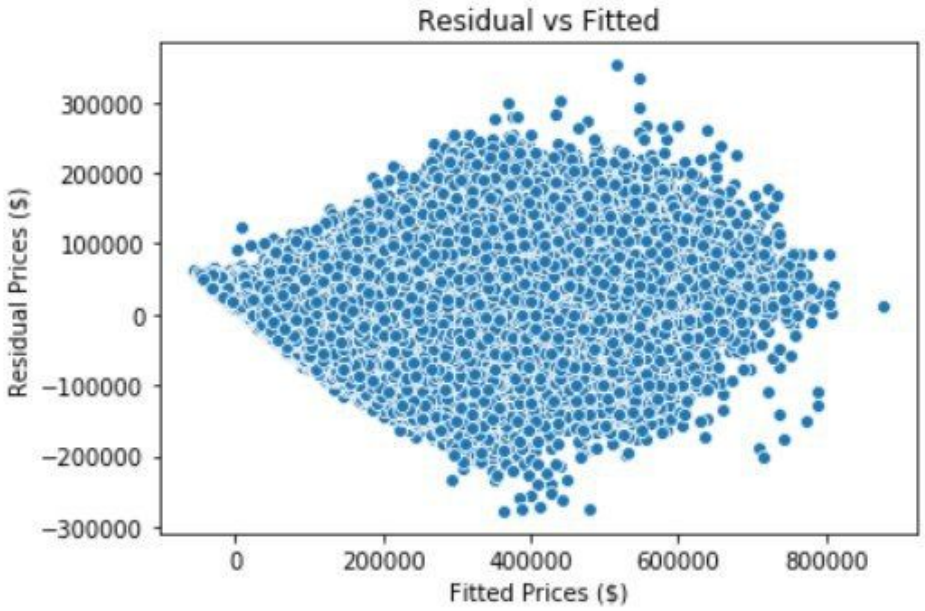


Figure 3. Residual vs Fitted

Looking at Figure 2, the scatterplot show a random distribution of residuals. It does not suggest any non-linear relationship, and the full residual vs fitted plot (similar to figure 3.) does not show much indication that a nonlinear relationship would fit much better. While the mean of the residuals do not seem to be 0, this is to be expected as regularization adds bias to the models. 1000 data points are shown to aid visualization in the randomness of the residuals. This proves that our model satisfies the Linearity, Normality and Hetreoskedascity assumptions of linear models.

Model Performance

Models are then used to predict on observations not used for estimation of parameters (test data) for all 3 models. Being a regression problem, we look for the lowest MSE among these models. The model that achieves the lowest MSE would be the most effective model in predicting housing prices. The following table shows the performance metrics for test data.

Note that while MSE is not included, RMSE (Root MSE) can be used to determine the best model as MSE is a monotonic function - the model with lowest MSE also has the lowest RMSE. One advantage of using RMSE is that the values are of the same scale as the predicted values. Mean Absolute Error (MAE) is also recorded as it gives an intuitive understanding of the effectiveness of the model (difference in magnitude between predicted and actual on average) but since MAE is not minimized when performing model fitting, it is not used to determine the model of best fit. MAPE calculates the average percentage deviation of the predicted value from the actual value.

	Lasso	Bayesian Ridge	ANN
RMSE	38781.06	38939.23	35840.67
MAPE	2.30%	2.322%	7.691%
MAE	28825.79	28882.63	26310.55
Explained Variance	89.75%	89.73%	92.90%

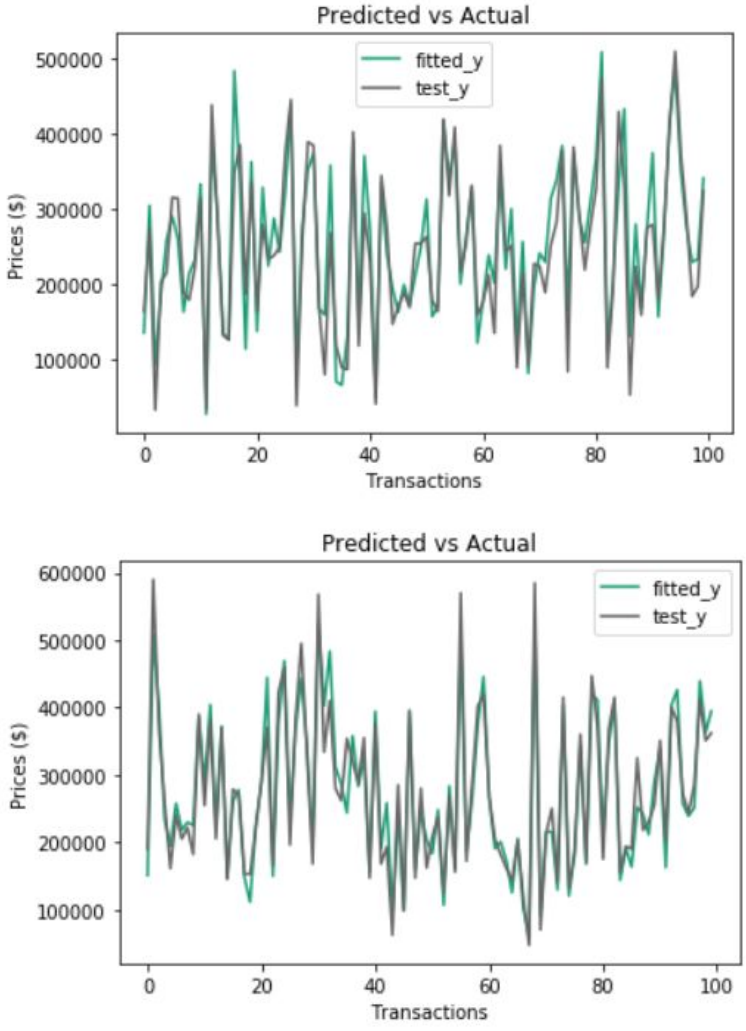


Figure 4. Random sets of 100 data points chosen for fitted vs actual for the bayesian ridge model.

Figure 4 shows two plots of 100 data points of predicted values against actual values. A subset of 100 data points is chosen to ensure visibility of the plot. As observed, the model gives close

predictions to the actual values, regardless of the location of the flat. Refer to Appendix B for more plots from Lasso Model and Neural Network model.

According to the results, the Neural Network model has the lowest RMSE and hence would be the model that would best be able to predict housing resale prices. This isn't a surprise as the former models have a linear constraint and housing resale prices are unlikely to follow a linear trend.

Limitations

Data Collection

Our data collection methodology is incomplete. The public OneMap API only provides facilities information at the time which data was collected and does not apply throughout the timeframe in the dataset. This explains the extremely low values of correlation between the facilities (bus stops, mrt, lrt and shopping malls) and resale price individually. Despite the limitations, we chose to keep it as variables in predicting resale prices as we believe the availability of these resale prices would be a good enough approximation. For future explorations, more for each dataset period, a specified infrastructure dataset could be obtained for better accuracy.

It is also due to the lack of completeness in the data that time series models cannot be employed in modelling data that follow a time series process. If the dataset had monthly values, time series modelling as well as more complex Neural Networks models such as CNN, RNN and LSTMs might be able to better predict housing prices.

Limits In Neural Network models

Non-linear models such as neural networks that are optimised through the process of gradient descent do not have underlying assumptions in the models and only aim to minimise MSE. The nodes for each hidden layer are variable and do not directly correspond to the variables used in the datasets. As a result, Neural Network models are limited when it comes to economic studies, such as finding out the effectiveness of any variable in predicting housing prices, or the strength of relationship between the variables and the resale prices.

Conclusion

We looked at the problem of collecting housing information over time in Singapore and tried to model them using regression methods given their constraints. These models achieve satisfactory results, being about 2% away from their predicted values and being able to explain about 90% of the variance in housing prices. While data isn't perfect and hence limiting on the

arsenal of models available for the problem, we expect a good relationship between the variables as well as housing resale prices. The model should improve further with future data diligently collected at regular time intervals.

References

Kim, S.-J., Koh, K., Lustig, M., Boyd, S., & Gorinevsky, D. (2007). An Interior-Point Method for Large-Scale -Regularized Least Squares. *IEEE Journal of Selected Topics in Signal Processing*, *1*(4), 606–617. doi: 10.1109/jstsp.2007.910971

Mackay, D. J. C. (1992). Bayesian Interpolation. *Maximum Entropy and Bayesian Methods*, 39–66. doi: 10.1007/978-94-017-2219-3_3

A Beginner's Guide to Neural Networks and Deep Learning. (n.d.). Retrieved from https://pathmind.com/wiki/neural-network

Micheal A. Nielsen, "Neural Networks and Deep Learning", Determination Press, 2015.

Appendix A

Potential Feature	Details for Creation
Flat Model	The floor design is based on when the flat that was built. For example, older generations of flats would tend to have larger square areas compared to newer generations of flats for the same flat type.
Flat Type	Flat type determines the number of rooms and amenities inside the house. The more amenities there are, the higher the resale price of the house. For example, a 5 room has a study room while a 4 room flat does not. The presence of extra space/room will thus allow the unit to have a higher value.
Floor Area (Square meters)	Indicator of size of the living space. The larger the HDB flat, the more expensive the resale price of the house.

Town	Geographic location of the town will affect the price of the flats as some locations are more popular than others and the land price is also higher. For example, areas like Toa Payoh generally have higher resale prices as it is closer to CBD, where most commute to work.
Storey Range	Given the same building, people generally tend to prefer staying in higher floors as there is less noise and less insects will appear. To protect the privacy of the sellers and owners of the apartment, a numerical range is used as a proxy. We take the average value of the range and treat this variable as a numeric variable.
Lease Remaining	While data only has lease commence date, the number of years remaining can be easily computed. With a lower remaining lease period, the price of the property would be valued lesser compared to another property with a longer lease period.

The following variables are the presence of nearby facilities. With more nearby facilities, we expect that house resale prices would be higher as they would be valued more highly. We used the latitude and longitude of these landmarks as well as the individual HDBs to calculate the distance and count the number of these nearby facilities. This information is freely accessed from the OneMap API which is publicly available.

Nearby bus stops	Bus Stops within walking distance of flats should be counted. We used a 200m radius as a cut-off for counting busstops within walking distance.
Nearby Shopping Centers	Nearby shopping centers should not be too far and take too long to reach. We expect residents to be able to reach their nearest shopping centers in 10 minutes, which would translate to about 1km if we were

	to travel by transport.
Nearby Train Stations	We only count MRT and LRT stations which are within 500m based on absolute distance from latitude and longitude from the housing unit.

Appendix B

Plots of predicted vs actual