Report Summary

Data Loading

We concatenated all the csv files together as they have the same number of columns.

	month	town	flat_type	block	street_name	storey_range	floor_area_sqm	flat_model	lease_commence_date	resale_price	remaining_lease
0	1990-01-01	ANG MO KIO	1 ROOM	309	ANG MO KIO AVE 1	10 TO 12	31.0	IMPROVED	1977	9000.0	NaN
1010	1990-01-01	KALLANG/WHAMPOA	3 ROOM	44	BENDEMEER RD	04 TO 06	63.0	STANDARD	1981	31400.0	NaN
1009	1990-01-01	KALLANG/WHAMPOA	3 ROOM	20	ST. GEORGE'S RD	04 TO 06	67.0	NEW GENERATION	1984	66500.0	NaN
1008	1990-01-01	KALLANG/WHAMPOA	3 ROOM	14	KG ARANG RD	04 TO 06	103.0	NEW GENERATION	1984	77000.0	NaN
1007	1990-01-01	KALLANG/WHAMPOA	3 ROOM	46	OWEN RD	01 TO 03	68.0	NEW GENERATION	1982	58000.0	NaN

Several columns are text based, such as town, flat_type, street_name, storey_range, flat_model, remaining_lease, while the rest are numeric. Some columns are not useful due to how specific and non-informative, such as the block number and the street_name as there are too many and too specific to specific units and buildings.

Data Preprocessing

We tried various preprocessing steps for each feature. Some are used for the final features while others may be removed after testing.

Features	**Preprocessing**	**Comments**
resale_price (Target)	Divided by 10000 to narrow down the range of values	
flat_model	Convert all names to lower alphabet for standardization	Tried to group different flat models as one hot and label encoding but it only seems to worsen results
storey_range	We get the average of the storey. Additionally, we grouped the storey into: - Low floor (<5 - Mid floor (5-9) - High floor (>=10)	Tried to get the lower/higher end of the storey, but it does not seem to help much
remaining_lease	From the years and months left on the lease, we standardize the lease into just months. Also, we get the flat age by getting the difference between the date and lease_commence_date column.	I tried to normalize the flat age by a multiple of 5, but it did not seem to improve results
flat_type	Remove a random ‘-‘ in some of the text to standardize output	
town		Tried to get the region from the town by grouping them into East,

		North, North-East, Central, and West region, but this makes the results poorer
year	Get the year from the date	Does not seem important as we end up using the most recent data only
month	Get the month from the date	Tried getting cyclic encoding for month but it does nothing

Data splitting

As this is a time-series problem, we must split the data by date, where the testing data is the latest data. Also, house prices have tended to change a lot in recent years. Consequently, the latest data is the best representation of the latest trends. Hence, while we have data from 1990 to 2020, using old data from before 2000 seems like a bad idea.

For the testing set, we kept the latest data, which is between 2019-2020. Meanwhile, for the testing data, we take the latest n years of data, and test our results to see how recency of training data affects performance

Pipeline

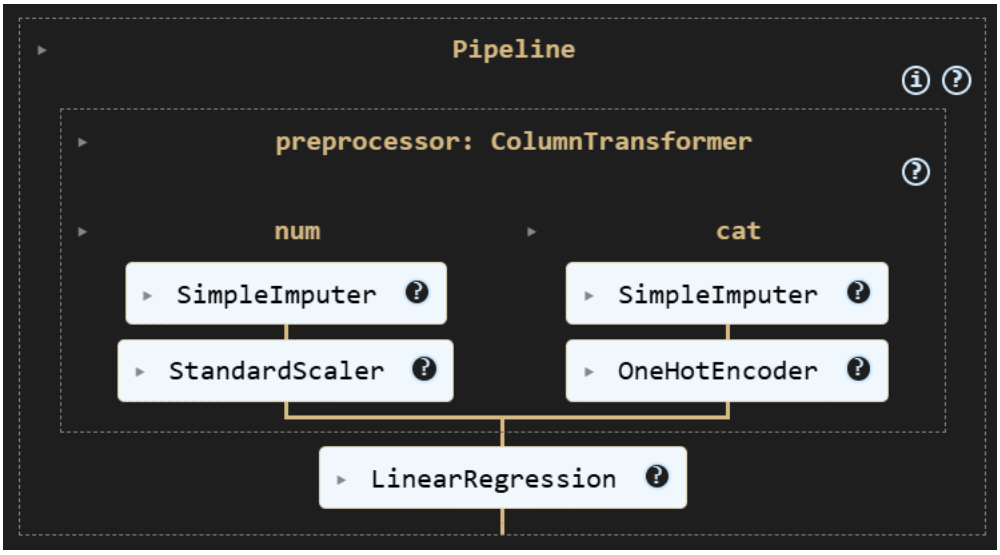
As some of the features are text based, one hot encoding is required:

- flat_type - 1 ROOM, 2 ROOM, 3 ROOM, etc
- flat_model - improved, model a, etc
- storey_range_avg - mid floor, high floor, etc
- region - SENGKANG, GEYLANG, etc)

Label encoding is tested for flat_type and storey_range_avg, but they appear to worsen the results. Meanwhile, other features such as month, floor_area_sqm, storey_range_avg, remaining_lease_month, and flat_age, are simple numerical values that can be used directly. Initially, there are plans to normalize them by rounding them to the nearest factor to make them more standardized (for instance, making all the area a fact of 5). However, after testing, it appears that its impact is nominal.

	flat_model	storey_range_height	region	flat_type	month	floor_area_sqm	storey_range_avg	remaining_lease_month	flat_age
0	improved	mid floor	SENGKANG	5 ROOM	1	110.0	5	NaN	13.0
1	model a2	low floor	SEMBAWANG	4 ROOM	1	86.0	2	NaN	14.0
2	model a	mid floor	SEMBAWANG	4 ROOM	1	90.0	8	NaN	15.0
3	model a	mid floor	SEMBAWANG	4 ROOM	1	90.0	5	NaN	11.0
4	improved	high floor	QUEENSTOWN	5 ROOM	1	117.0	20	NaN	3.0
...	...	...	...	...	...	...	...	...	...
79218	model a	high floor	GEYLANG	4 ROOM	12	102.0	11	939.0	20.0
79219	new generation	mid floor	GEYLANG	4 ROOM	12	92.0	5	697.0	40.0
79220	simplified	mid floor	GEYLANG	3 ROOM	12	64.0	8	790.0	33.0
79221	improved	high floor	GEYLANG	3 ROOM	12	65.0	11	686.0	41.0
79222	model a	high floor	GEYLANG	3 ROOM	12	90.0	14	792.0	33.0

Thus, we have a parallel preprocessing pipeline (see below), that consists of a regular standard scaler for numeric features, while a one hot encoder for category data.



For this project, it is a regression problem as we are trying to predict the housing price, which is a continuous target. Meanwhile the model selected is the Linear Regressor as it is the fastest and yields decent results. While Random Forest Regressor or XGBoost Regressor may potentially yield more results, due to the high dimensionality of the problem due to the one hot encoding, as well as limited computation time and sources, the Linear Regressor is used. Grid search CV is used to determine the best parameters in the Linear Regressor.

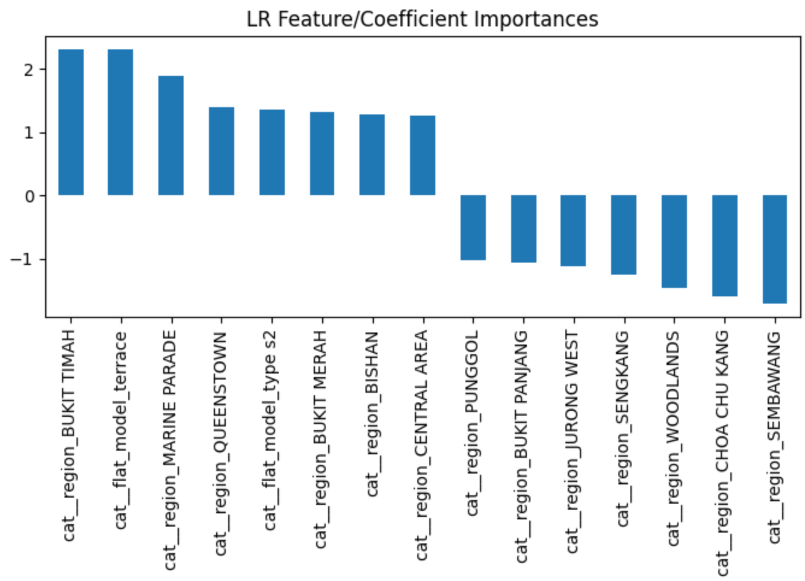
Performance

Training data	Testing data	R2 (training)	R2	MAE	MSE
1990-2018	2019-2020	0.666	0.49	0.77	1.196
2000-2018	2019-2020	0.618	0.562	0.724	1.028
2010-2018	2019-2020	0.784	0.821	0.488	0.42

2015-2018	2019-2020	0.864	0.854	0.446	0.343
2017-2018	2019-2020	0.868	0.857	0.444	0.336
2018	**2019-2020**	**0.869**	**0.859**	**0.439**	**0.33**

From the results, it appears our assumption that the latest results should be represented by the latest trend is correct. While we have a lot of historic data, most old data are no longer useful, and should be discarded. This is because old data is not reliable as the housing price is not stationery and tends to follow an upward trend in general.

Discussion



By determining the coefficient of the Linear Regressor, we can determine the most important features. From what we see, the area where the house is located is the most important thing in determining the price. This makes sense as popular locations near the Central Area like Bukit Timah and Marine Parade are favored and thus more expensive (high positive coefficient) over areas like Sembawang and Choa Chu Kang (high negative coefficient). This makes sense as property is always about location and convenience. Additionally, flat models such as Terrace and Type S2 are more expensive.

Conclusion

For this project, we used a simple model to model the housing price in Singapore. From our experiments, using the latest data for prediction is the most accurate. Additionally, by looking at feature importance, it appears that location is the most important factor when determining housing prices, over housing size and storey level. Nonetheless, those features are also important as they affect the prices, just less significantly.