Predicting HDB Resale Prices in Singapore

Using regression models and neural networks to predict resale prices

Agenda

Problem Statement

Data Collection

Feature Engineering

Exploratory Data Analysis

Modeling Process

Model Evaluation

Feature Importance

Conclusion and Reflections

Next steps

Problem Statement

Problem Description

In recent months, HDB resale prices has been steadily increasing and has become a common discussion topics especially amongst first time young buyers for affordability

I am curious to find out **what features influence HDB resale prices** and help potential buyers find out if the current asking prices of HDBs are reasonable by **using regression models to predict HDB prices.**

Why is this a problem?

The increasing cost of living comes to mind for young Singaporeans looking to purchase a home and start a family. This model would serve a guide for them as part of their home purchase decision making process.

How will we tackle the problem

Regression models:

1. Linear
2. Lasso
3. Ridge
4. Random Forest
5. ExtraTrees
6. XGBoost
7. Neural Networks

How will we evaluate the results

Success Metrics: Model performance will be guided by RMSE. We will seek to find the best performing model based on the lowest RMSE score.

Structured data

Over <u>200,000 rows and 10 features</u>

	month	town	flat_type	block	street_name	storey_range	floor_area_sqm	flat_model	lease_commence_date	remaining_lease	resale_price
0	2017-01	ANG MO KIO	2 ROOM	406	ANG MO KIO AVE 10	10 TO 12	44.0	Improved	1979	61 years 04 months	232000.0
1	2017-01	ANG MO KIO	3 ROOM	108	ANG MO KIO AVE 4	01 TO 03	67.0	New Generation	1978	60 years 07 months	250000.0
2	2017-01	ANG MO KIO	3 ROOM	602	ANG MO KIO AVE 5	01 TO 03	67.0	New Generation	1980	62 years 05 months	262000.0
3	2017-01	ANG MO KIO	3 ROOM	465	ANG MO KIO AVE 10	04 TO 06	68.0	New Generation	1980	62 years 01 month	265000.0
4	2017-01	ANG MO KIO	3 ROOM	601	ANG MO KIO AVE 5	01 TO 03	67.0	New Generation	1980	62 years 05 months	265000.0

Sourced from data.gov.sg

Data Collection

Calling OneMap API

Scraping longitudes and latitudes of:

- the transacted HDBs
- Nearby amenities such as:
  - MRT stations
  - Schools
  - malls

Feature Engineering

- Created new distance-based features such as:
  - HDB's proximity to nearest
    - MRT station
    - Primary school
    - Secondary school
    - Mall
    - CBD

- Other created features include:
  - Year transacted
  - Resale price per sqf
  - Average storey (floor)
  - Age of HDB
  - Inflated adjusted resale price

Exploratory Data Analysis

Link to Tableau Public

Modeling Process

Baseline Model

- Using mean resale price from train dataset
- Train RMSE: 144,232
- Test RMSE: 142,975

Preprocessing

- Standard Scaler
- Dummify categorical features:
  - Town
  - Flat type
  - Flat model
  - Year transacted

Training Models

- Linear Regression with 1 feature
- Linear Regression with all features
- Lasso Regression
- Ridge Regression
- Random Forest Regressor
- ExtraTrees Regressor
- XG Boost Regressor
- Feed Forward Neural Network
  - Vanilla
  - Weight Decay
  - Dropout

Evaluation

- RMSE is our guiding metric

Model Evaluation

	model	train_r2_score	train_rmse	test_r2_score	test_rmse	fit_time (min)
0	Linear Regression (1 feature)	0.424661	109401.40	0.431538	107794.18	0.00
1	Linear Regression (all features)	0.886708	48546.76	0.885599	48357.07	0.02
2	Lasso Regression (all features)	0.879362	50096.00	0.878561	49822.21	0.46
3	Ridge Regression (all features)	0.886702	48547.97	0.885582	48360.57	3.68
4	Random Forest Regression (all features)	0.986806	16567.43	0.972037	23907.57	31.30
5	ExtraTrees Regression (all features)	0.935992	36490.27	0.932610	37114.43	32.00
6	XGBoost Regression (all features)	0.981277	19735.78	0.973344	$\boxed{23342.20}$	$\boxed{7.74}$
7	Vanilla Neural Network	0.966433	26425.05	0.963494	27316.57	1.37
8	Neural Network (w/ Batch Normalization)	0.963086	27711.24	0.960833	28294.81	1.65
9	Neural Network with Weight Decay	0.968299	25680.34	0.964727	26851.39	1.72
10	Neural Network with Dropout	0.963221	27660.65	0.961004	28232.98	2.22

Best model: XG Boost

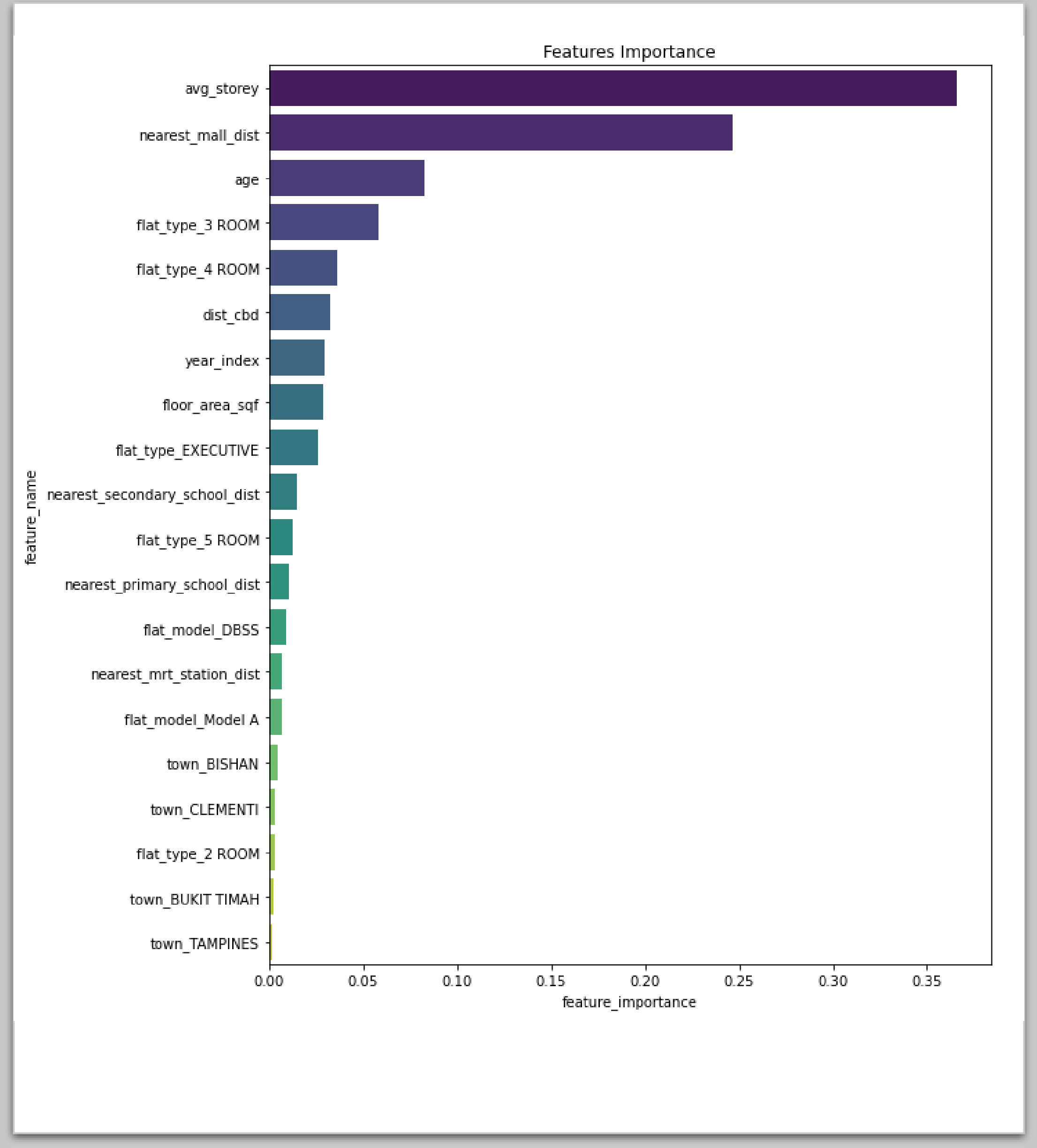
Hyperparameters:

- Number of trees: 250
- Learning rate: 0.1
- Max depth: 9
- Column sample ratio by tree: 0.5
- Subsample: 0.8

- Using RMSE as our guiding metric, XG Boost is the best model with the lowest test RMSE score
- Linear regression and Ridge regression performed closely with similar test RMSE. On the other hand, Lasso regression performed worse compared to the above 2
- We also note that the runtime for training XG Boost is much faster compared to Random Forest and ExtraTrees
- Neural networks (NN) not preferred in this case. Apart from not performing in terms of RMSE, its most disadvantage is its black box nature

Feature Importance

- Overall, we had more physical housing features such as floor level, age, flat type, floor area which were ranked higher in importunate compared to distance-based features
- Avg_storey: Floor level of the HDB unit had the most influence in price
- Proximity to mall and CBD were most important distanced-based features
- Age: HDBs are leasehold in nature and natural that people would want to buy newer flats
- year_index: Demand for resale HBDs were unusually high in 2020 and 2021 due to the longer delays in BTO projects and cheaper housing loans from low interest rates



Conclusion and reflections

Tying back to the problem statement

- Using the best model, we discussed the important features which influenced a HDB's resale price
- We successfully ran different regression models in HDB price prediction and concluded that XG Boost performed the best

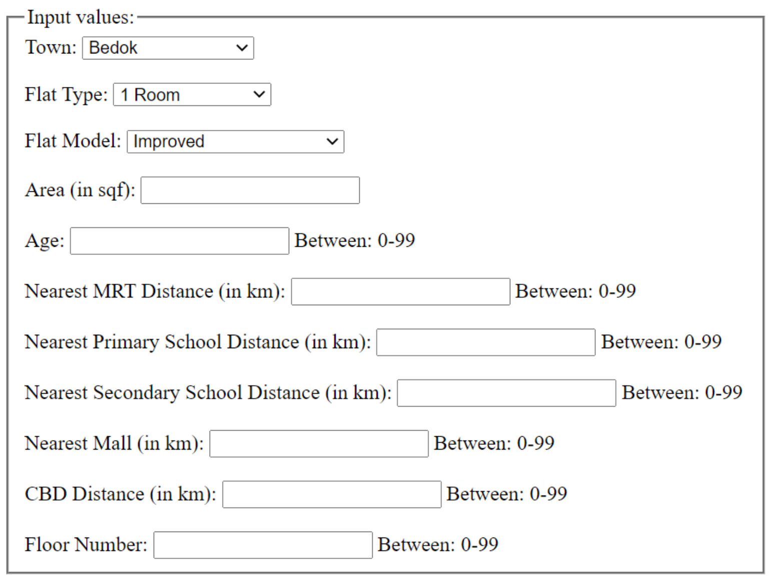
Model limitations

- HDB prices are driven by macro economic factors like status of the economy and inflation as well as housing policies

Reflections

- Good planning is half the battle
- Understanding the computing resources that you have and plan for contingencies

Next steps/how we can improve



Retrain

Retrain model with more data!
Try other models like support vector regression

Retune

Further tuning on hyperparameters!
- GridSearchCV and RandomizedSearchCV

Collect

Scrape private residential data (condos)
- Explore the data and test our model's predictive capability

Deploy

Using Flask to develop a web application