





Data Science Portfolio

Sivakorn Chongfeungprinya (Oak)





Data Science Portfolio

1. Personal Project

- HDB (Singapore Public Housing) Resale Price Predictor
 - Data pre-processing and model development: https://github.com/sivakornchong/hdb_project
 - Deployment: https://huggingface.co/spaces/sivakornchong/HDB_resale_predict

2. Company Contribution

- Construct several data pipeline transfer from SQL database to labelling platform, and back to model training.
 - Utilized SQL, pandas, python. Intermediate files used are .json and .csv.
- Computer Vision: classification and detection model training
 - Optimized model hyperparameters and input datasets , considering recall and precision trade-offs.

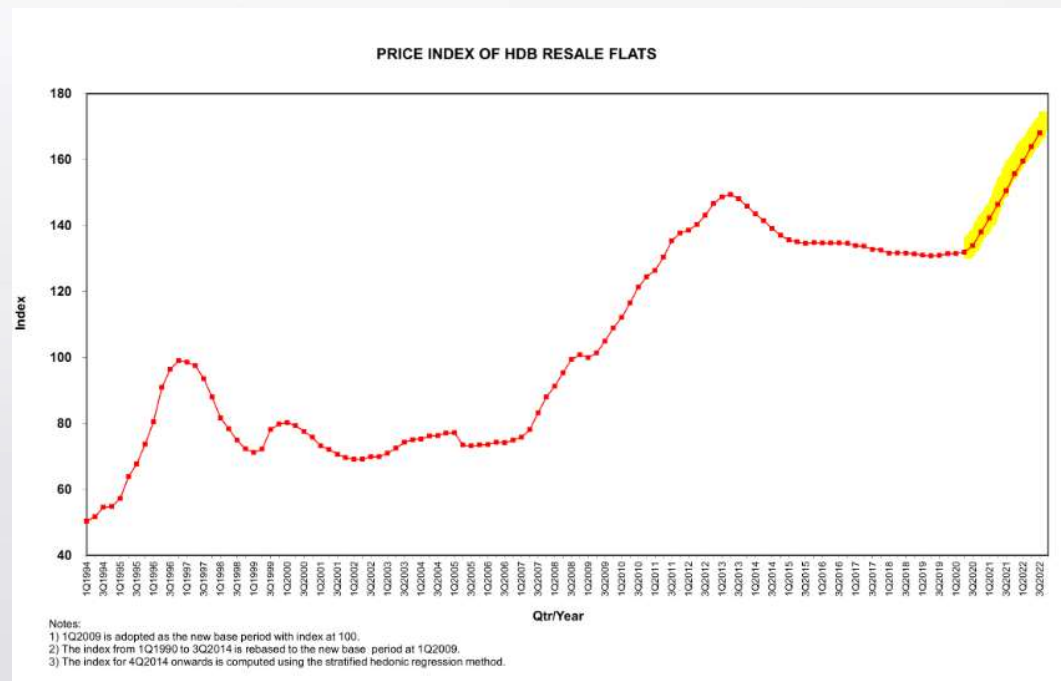
HDB Resale Price Predictor

- **Business Background:**

- Booming property market in Singapore has resulted in a spike in resale prices of HDB flats
- Many owners take opportunity to price their house at a large premium on the market.
- Difficult for buyers to know the intrinsic (sensible) price.

- **Project Objectives:**

- Develop a reliable housing price predictor for buyers to negotiate better in a hot seller market.



Source: [HDB](#)

HDB Resale Price Predictor - MVP

- [Link](#) to interface

Postal Code	<input type="text" value="651446"/>
Years since lease commencement (TOP)	<input type="text" value="5"/>
Town	<input type="text" value="BUKIT BATOK"/>
Floor	<input type="text" value="15"/>
Room	<input type="text" value="5 ROOM"/>

Input: A sample listing entered into platform.


Predicted House Price (\$)
<input type="text" value="856474"/>

Output: Model price

s\$ 838,000 Negotiable

3 2 1205 sqft s\$ 695.44 psf

Est. Repayment S\$ 2,323 /mo [Get Pre-Approved Now](#)

446A West Crest @ Bukit Batok
446A Bukit Batok West Avenue 8 651446 Bukit Batok Estate 

Nearby Stations

- 4 mins (330 m) to JE2 Tengah Park MRT
- 7 mins (550 m) to JE3 Bukit Batok West MRT

LiveTour

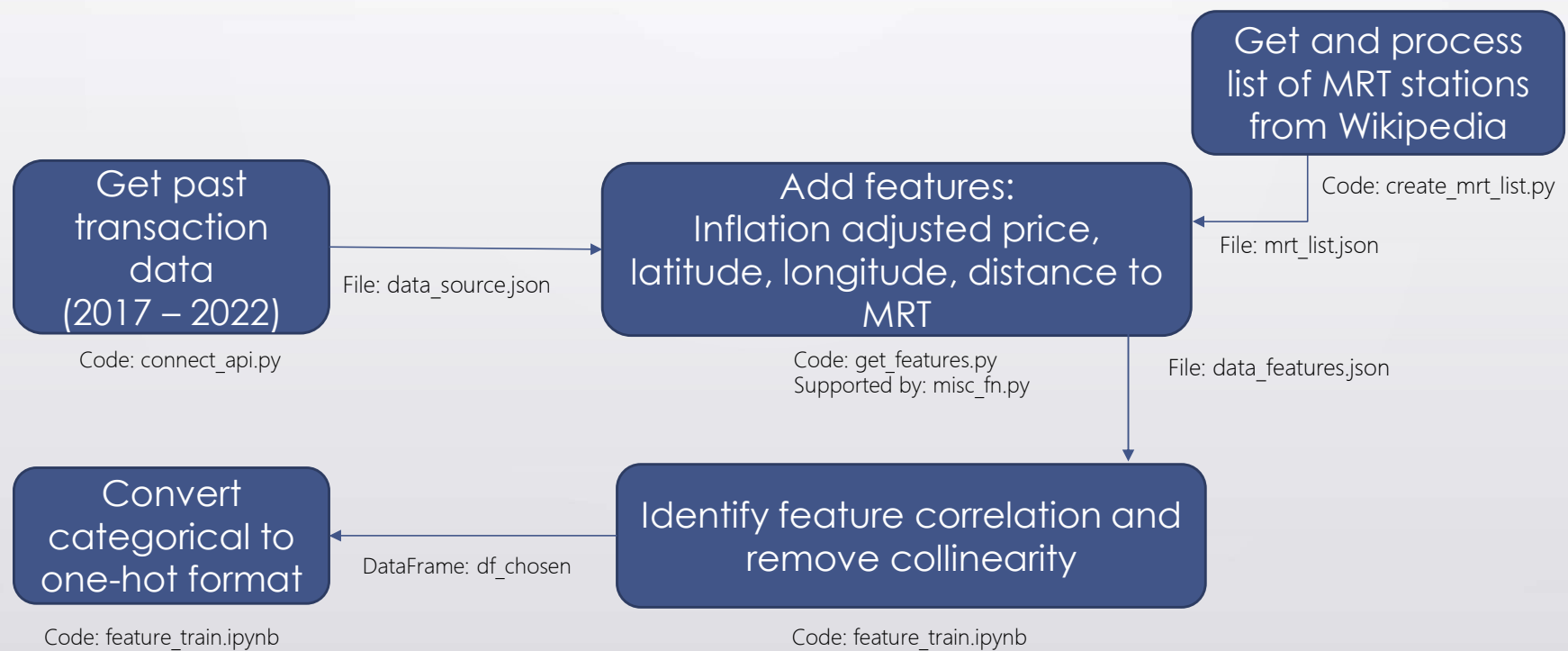
Check out this listing via video call. Get in touch with the agent to arrange.

Details

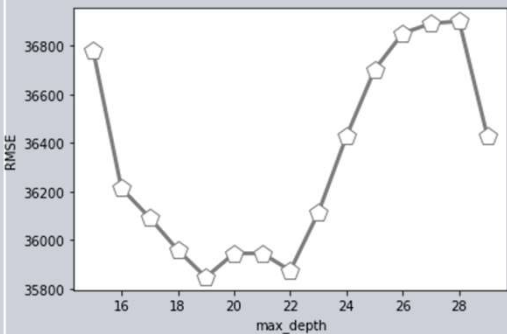
Property Type	Floor Size
5A HDB For Sale	1205 sqft
Developer	PSF
Housing & Development Board (HDB)	S\$ 695.44 psf
Furnishing	Floor Level
Unfurnished	High Floor
Tenure	TOP
99-year Leasehold	2017

[Actual listing](#)

Data Pre-processing



Model Selection

Model name	Root Mean Square Error (RMSE)	Explanation
Linear Regression	61774	Fast and basic model
Ridge Regression	61767	Variation of linear regression that uses shrinkage, suitable for model with heavy collinearity.
Gradient Boosting	51027	Out of the three complex models: - Random Forest is the best performing for this dataset. - Its RMSE is lower than decision tree with optimized maximum depth parameter. - Gradient boosting might have been affected by overfitting, resulting in higher RMSE than expected.
Random Forest	28464	
Decision Tree		

Findings

- Deployed model is the Random Forest Regressor.
- The best decision tree model and the random regressor model has **similar** rankings in model parameters
- Based on top and bottom features
 - The individual town names are not as important as room types and other continuous parameters.
 - **Postal code** is the most important.

	importance
Postal	0.251385
flat_num_3 ROOM	0.245478
age_transation	0.141285
flat_num_EXECUTIVE	0.108566
flat_num_5 ROOM	0.078541
distance_mrt	0.045479
storey_height	0.029064
town_BISHAN	0.023681
flat_num_2 ROOM	0.016861
flat_num_4 ROOM	0.014835

Top Features

town_MARINE PARADE	0.000621
town_CHOJA CHU KANG	0.000510
town_CLEMENTI	0.000440
town_WOODLANDS	0.000395
town_TOA PAYOH	0.000390
town_GEYLANG	0.000322
town_PUNGGOL	0.000234
town_PASIR RIS	0.000232
town_YISHUN	0.000154
flat_num_1 ROOM	0.000151

Bottom Features



Conclusion and Next Steps

- Model file size is **large** (813MB), resulting in long load and run time (3-4 secs)
 - Parameters could be minimized if not important.
 - The town names contributed to 26 out of 38 model input parameters due to one-hot processing of categorical inputs.
 - Based on findings, the town names could possibly be **replaced** by a new parameter calculating distance to city center and existing parameter for postal code sufficiently.
 - To test the performance in the next model iteration.
- In model training phase, Geopy package was used to calculate distance for each house (130,000) to each MRT (80). This resulted in **long processing time** (5 hours).
 - Consider using **multi-processing** modules.
- Expansion of features:
 - Model could output list of recent transactions to the webpage. **Further visualization and analytics** could be done based on the list and shown on the page.