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
  - Primary and secondary 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

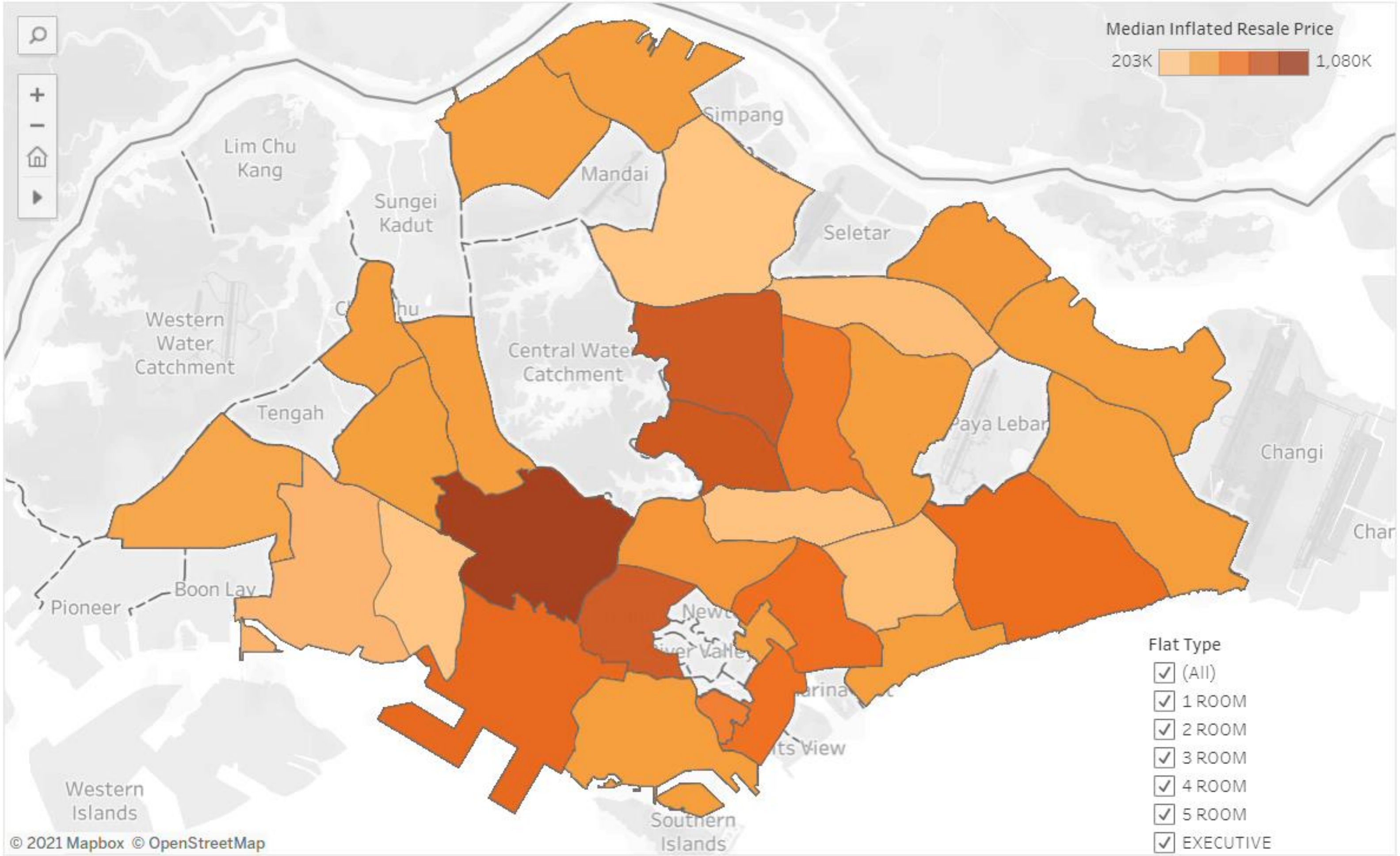
Exploratory Data Analysis

Visualized with Tableau

HDB Resale Price Analysis

Median Inflation Adjusted Resale Price	Price vs transaction demand	Price Trend by Estate	price per sqf analysis by estate and flat types	proximity to amenities analysis by flat types

Median resale prices by town, central area is the most expensive



Flat Type
- [x] (All)
- [x] 1 ROOM
- [x] 2 ROOM
- [x] 3 ROOM
- [x] 4 ROOM
- [x] 5 ROOM
- [x] EXECUTIVE
- [x] MULTI-GENERATION

Visualization of HDB Resale Price Data

(Time period: Oct 2011 to Oct 2021)

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

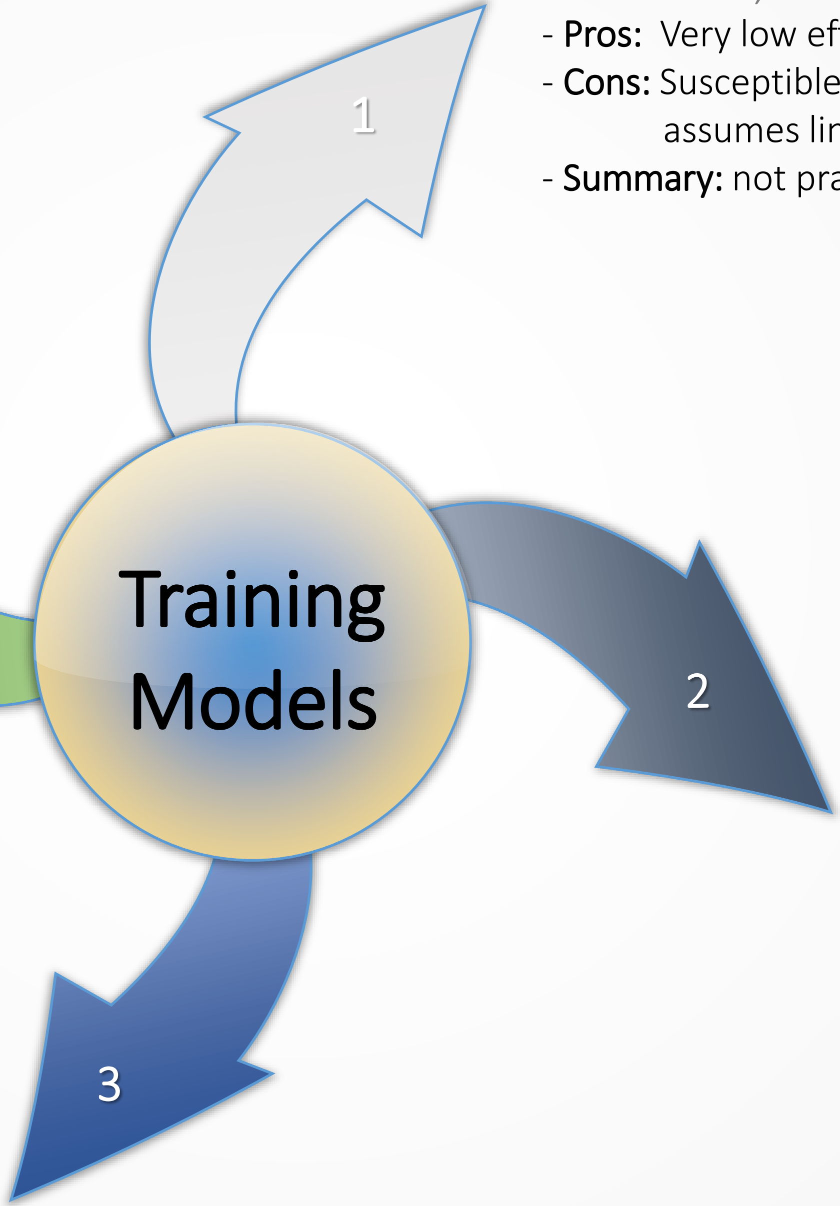
Training Models

- Linear Regression
- Lasso Regression
- Ridge Regression
- Random Forest
- ExtraTrees
- XG Boost
- Feed Forward Neural Network
  - Vanilla model
  - Weight Decay
  - Dropout

Evaluation

- RMSE is our guiding metric

Model Evaluation



Linear Regression

Test RMSE: 48,357

- **Pros:** Very low effort, extremely fast to train
- **Cons:** Susceptible to overfitting and only assumes linear relationship
- **Summary:** not practical in most situations

Regularized Linear Regression

Lasso and Ridge

Test RMSE: 49,822 and 48,361

- **Pros:** Low effort, easy to train
  Alpha hyperparameter helps to reduce likelihood of overfitting
- **Cons:** may increase the bias

Tree-based models

Random Forest & ExtraTrees

Test RMSE: 23,908 and 37,114

- **Pros:** Reasonable RMSE score and provides feature importance for intepretability
- **Cons:** tends to overfit, high effort to tune and takes long time to train

XG Boost

RMSE: 23,342

- **Pros:** medium effort, highly optimized algorithm with many hyperparameters available for tuning
- Best RMSE score

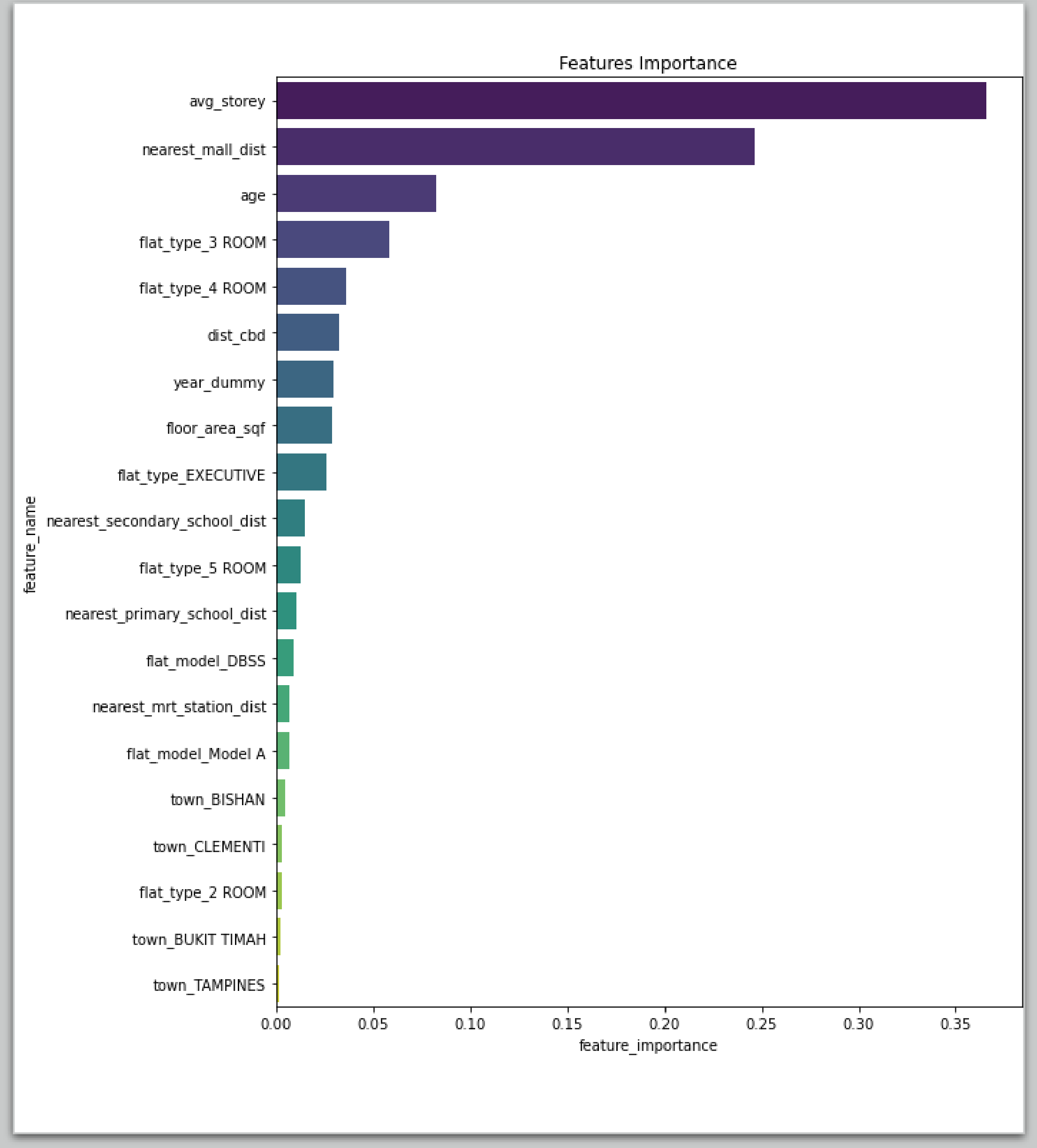
Feed forward neural networks

Test RMSE: 26,851 to 28,233

- **Pros:** Medium effort, doesn't take as long as the tree-based models
- **Cons:** Black box nature means that model isn't easily intepretable
- **Summary:** best used for extremely complex datasets

Feature Importance

- Overall, we had more physical housing features such as floor level, age, flat type, floor area which were ranked higher in importance 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_dummy: 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
- RandomizedSearchCV (didn't have time to try this out)

Collect

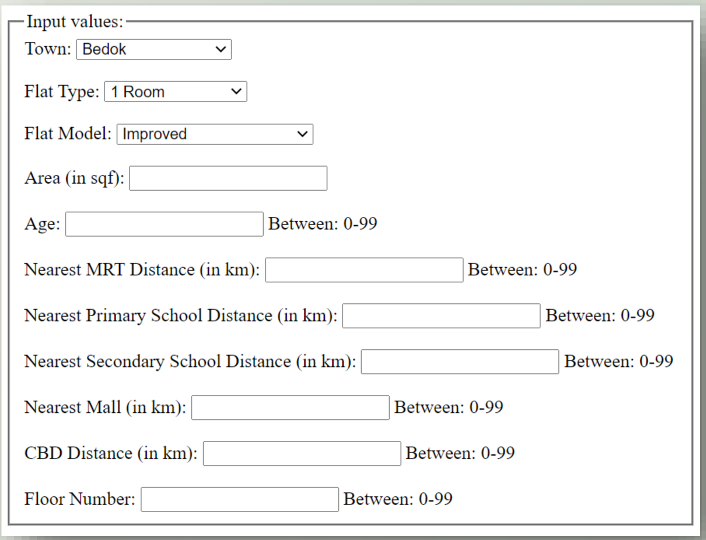
Scrape private residential data (condos)

- Explore the data and test our model's predictive capability

Deploy

Using Flask to develop a web application and deploy model online

- Allow users to input some HDB features and see a price estimate



Input values:

Town: Bedok

Flat Type: 1 Room

Flat Model: Improved

Area (in sqf):

Age: Between: 0-99

Nearest MRT Distance (in km): Between: 0-99

Nearest Primary School Distance (in km): Between: 0-99

Nearest Secondary School Distance (in km): Between: 0-99

Nearest Mall (in km): Between: 0-99

CBD Distance (in km): Between: 0-99

Floor Number: Between: 0-99