Predicting HDB Resale Prices in Singapore

Using regression models and neural networks

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 understand the **features influencing 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 score. We will seek to find the best performing model based on the lowest 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

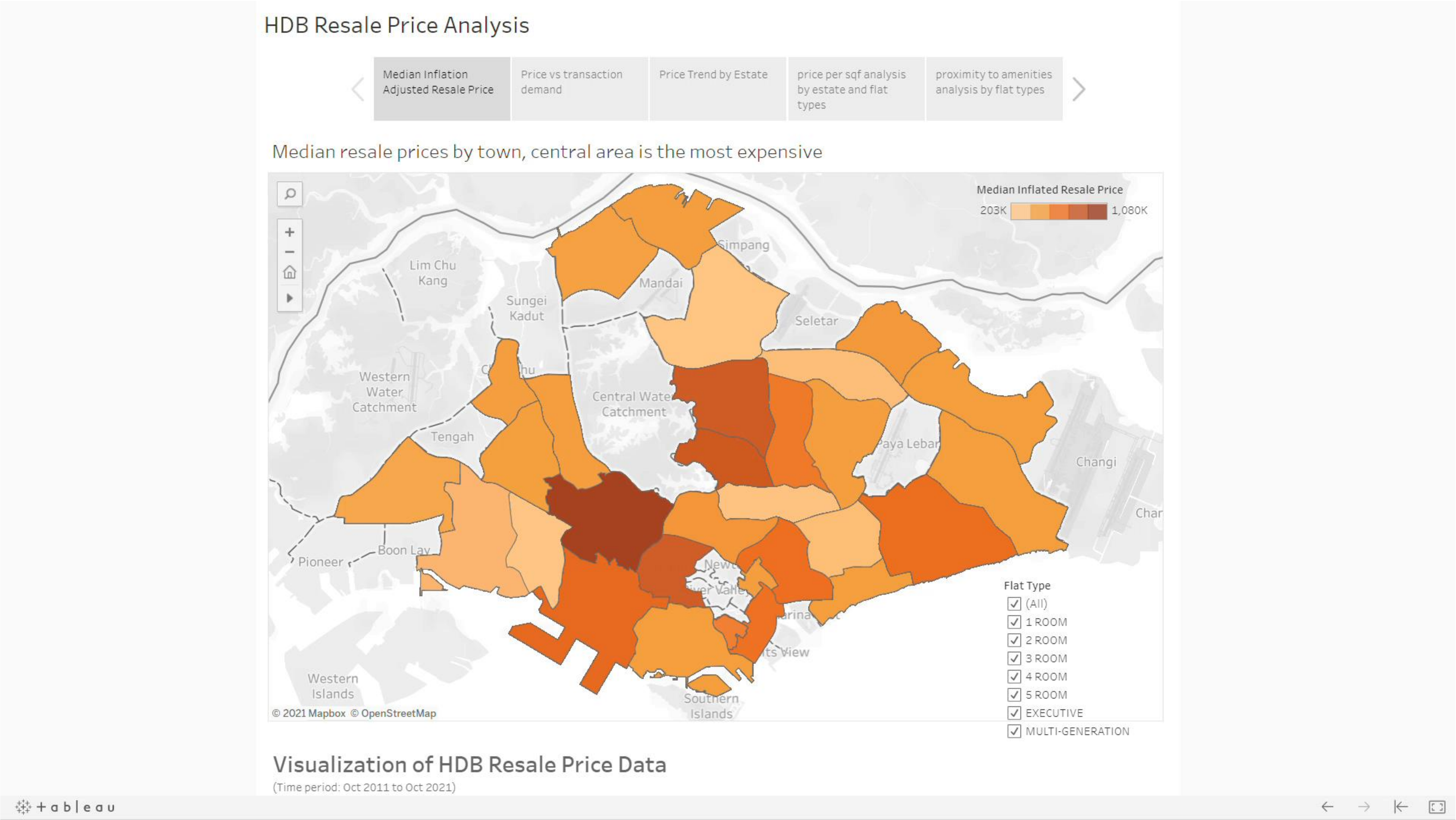
- Created new distance-based features such as:
  - HDB's proximity to nearest
    - MRT station
    - Primary school
    - Secondary school
    - Mall
    - CBD

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

Exploratory Data Analysis

Visualized with Tableau

Link to Tableau visualization [here]



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
- Splitting data to training and test sets

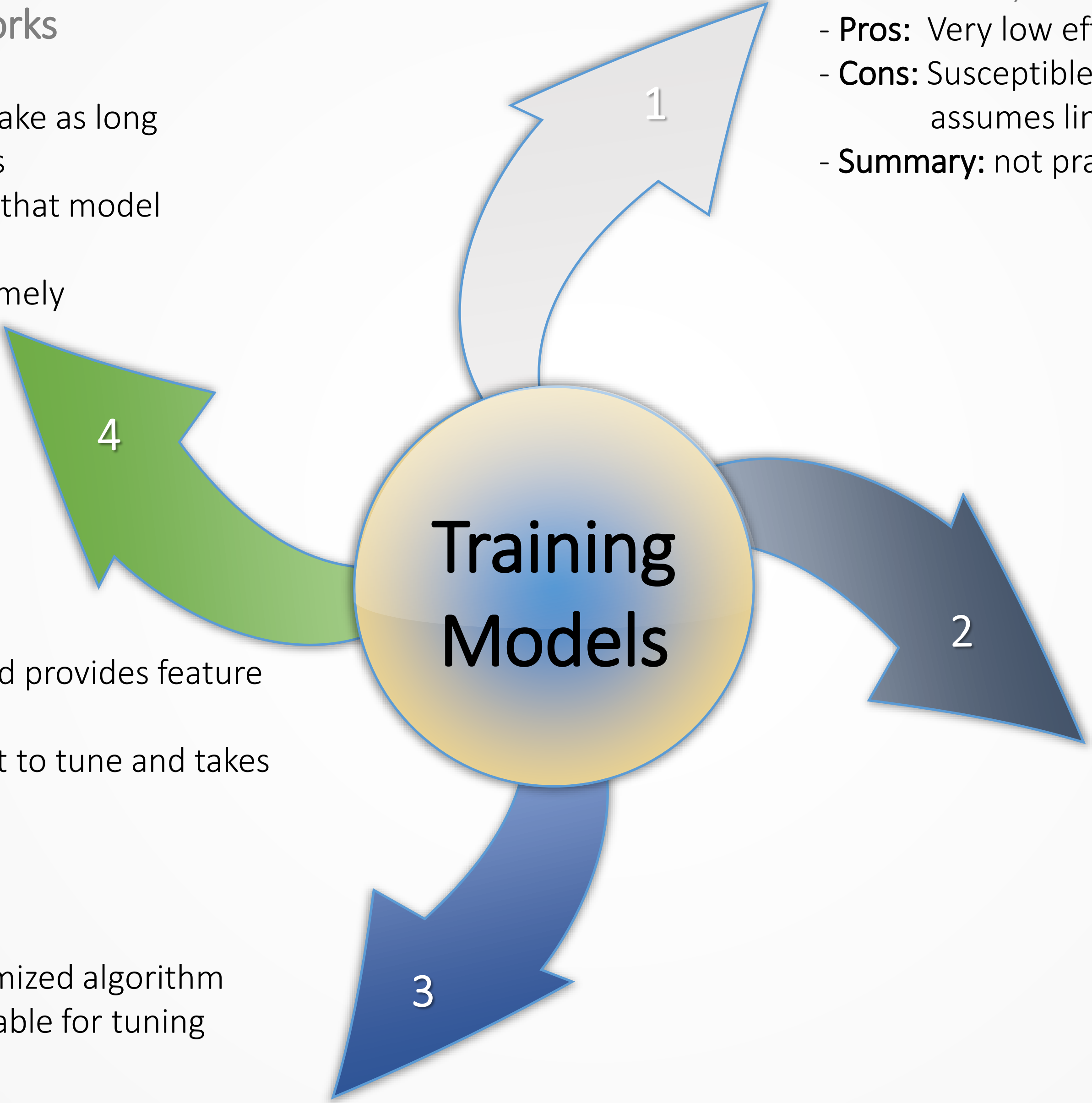
Training Models with GridSearchCV

- 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: 52,005

- **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: 52,005 and 52,792

- **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: 24,066 and 24,865

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

XG Boost

RMSE: 23,372

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

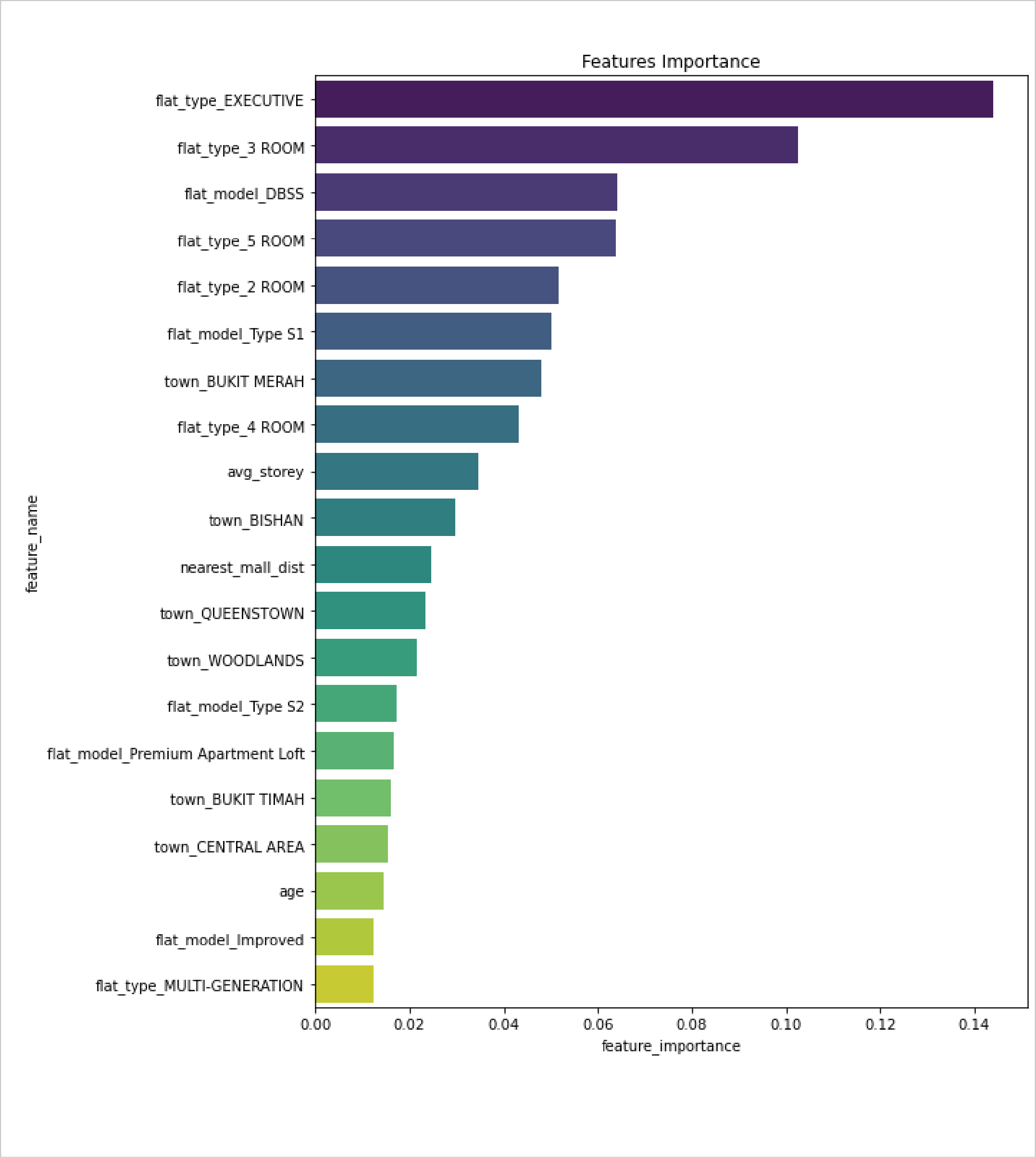
Feed forward neural networks

Test RMSE: 27,645 to 29,626

- **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 flat type and flat model, followed by location-based features like the town in which the HDB apartments were found in.
- Specifically, flat types like executive, 2-room, 3-room and 5-room were in the top 5 features.
- We noticed that towns located in the central part of Singapore made it into the list
- Interestingly, proximity to nearest mall was the only distance-based feature making into the list, which meant that proximity to malls had a greater influence in price than proximity to schools and mrt stations
- Age: HDBs are leasehold in nature, it would be natural that people would want to buy newer flats, thus an important feature



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

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

Update: model is now deployed on Heroku!