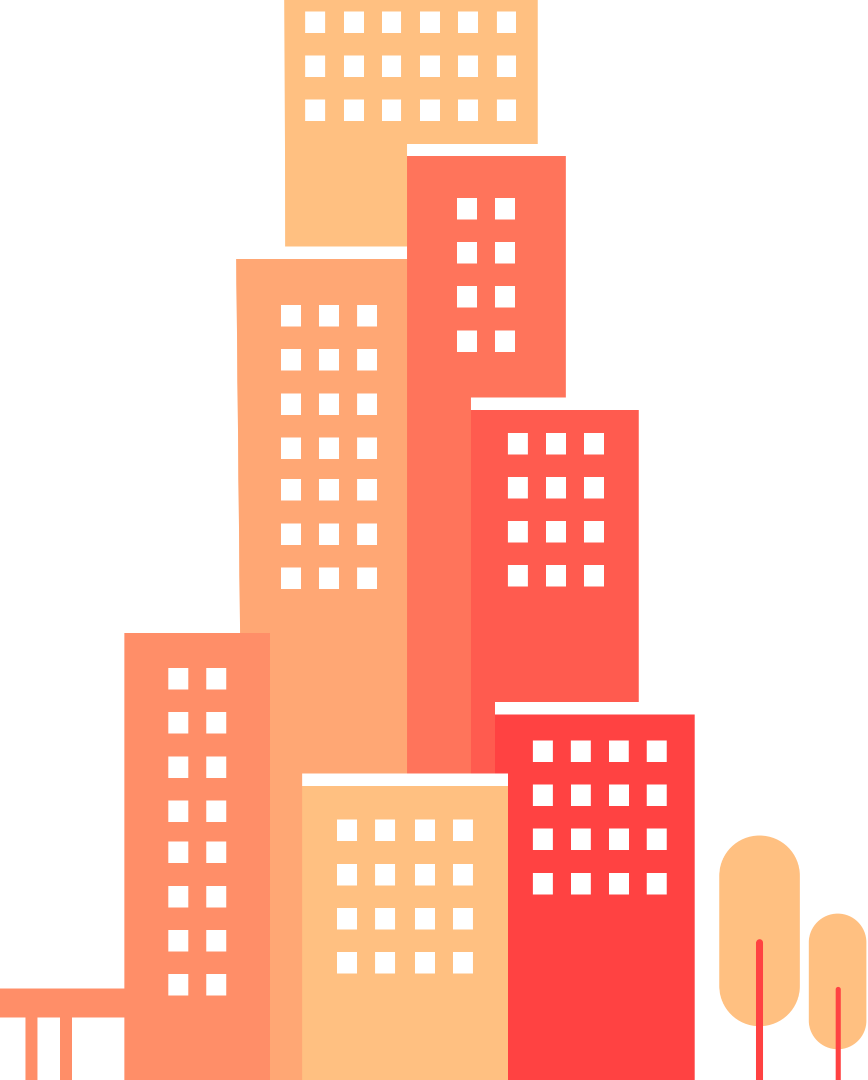


House Price Prediction

Predicting HDB Flat Resale Prices

-Ng Geok Teng-

Overview

Introduction

Problem Statement

Data

Introducing Data used

Exploratory Data Analysis

Understand the characteristics of the features

Modelling

Include exploring using Lasso for feature selection.

Discussion & Conclusion

Understand selected model performance and future exploration

Introduction:

Problem Statement

- An entrepreneur wanted to set up a new property agency in Singapore.
- She collected a list of flat-related data, but did not know how to use the data to predict HDB resale flat prices nor how to quantitatively understand how the data impact prices.

Objectives

- Develop a predictive model for the entrepreneur
- Show the relationship between key features and the price

Singapore

HDB resale flat prices up 10.3% in 2022, slower than 12.7% increase in 2021

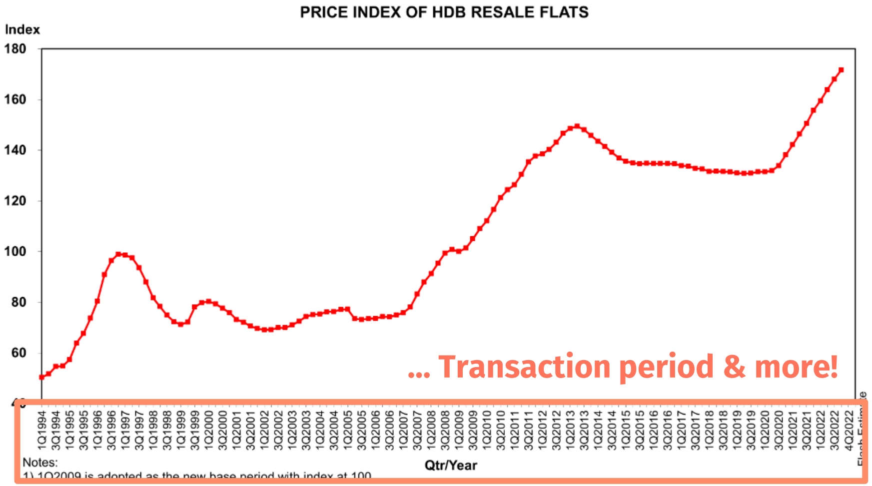
THE STRAITS TIMES

Price growth of HDB resale flats slows in December, analysts expect prices to stabilise in 2023

Locations...

Flat types...

TOWNS	1-ROOM	2-ROOM	3-ROOM	4-ROOM	5-ROOM	EXECUTIVE
ANG MO KIO	-	*	$365,500	$516,500	$800,000	*
BEDOK	-	*	$355,000	$475,000	$680,000	$820,000
BISHAN	-	-	*	$640,000	$855,000	$1,045,000
BUKIT BATOK	-	*	$353,000	$500,000	$720,000	$790,900
BUKIT MERAH	*	*	$368,000	$765,000	$875,000	-
BUKIT PANJANG	-	*	$386,500	$471,900	$610,000	$750,000
BUKIT TIMAH	-	-	*	*	*	*
CENTRAL	-	*	$460,000	$680,000	*	-



Notes:
1) 1Q2009 is adopted as the new base period with index at 100.
2) The index from 1Q1990 to 3Q2014 is rebased to the new base period at 1Q2009.
3) The index from 4Q2014 onwards is computed using the stratified hedonic regression method.
4) The 4Q2022 index is a flash estimate.

Sources: 1. CNA, 2. ST, 3. HDB stats

Data

77 Data Features

Location

Address, postal, town name, street name, planning area, longitude & latitude

Facilities

Presence of malls, hawkers, primary & secondary schools, transportation

Block-related

Block number, block age, building age, max level, number of units sold

Unit-related

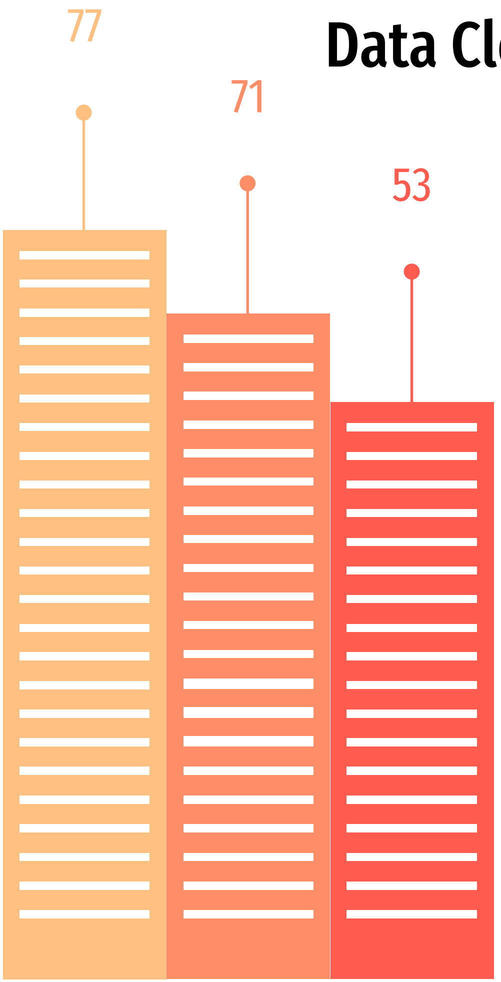
Floor area, flat model, flat type, storey

Transaction

Transaction year, month, resale price

Full list of data in [Kaggle Challenge page]

Data Cleaning and Feature selection



- There is 77 features originally
- Missing values were addressed, Duplicate data were confirmed absent
- Similar features and redundant features were removed **(Number of features left: 71 features)**
- Data values and Data types were checked and corrected appropriately (e.g. Converting Boolean feature to ‘0’ and ‘1’)
- Further selection of features after careful analysis **(Number of features left: 52 features)**

Data: How is Missing Data Addressed

```
Columns with missing values:
                     col  num_nulls  perc_null
45  Mall_Nearest_Distance        829       0.01
46        Mall_Within_500m      92789       0.62
47         Mall_Within_1km      25426       0.17
48         Mall_Within_2km       1940       0.01
50      Hawker_Within_500m      97390       0.65
51       Hawker_Within_1km      60868       0.40
52       Hawker_Within_2km      29202       0.19
```

829 Flats have no record of any nearby mall:

- `fillna(4000)` `(replace NaN with 4km)` for Mall_Nearest_Distance because generally most MRT stations has a mall, and the maximum distance a flat is away from nearest MRT station is 3.54km.

Analysis discovered that the data contained NaN is because there is **zero** mall/ hawker within specified distance.

`fillna(0)` for features Mall_Within_500km, 1km and 2km

Data: How are features filtered out

Features that are similar

Example: `mid_storey` and `mid` are the same thing

Example: `hdb_age`, `year_completed` and `lease_comence_date` share strong correlation (r≈1) as they are referring to similar thing

Feature that is redundant

Example: `residential` is a feature with boolean value if resale flat has residential units in the same block. The column only contains one value same for all flats

Feature that show no significant effect in resale price through stats-model OLS

Further 5 features were excluded out as they were found to have P|t|>0.05 in OLS analysis

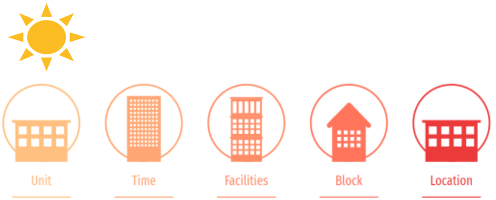
	coef	std err	t	P>|t|
Have_market_hawker	7682.9828	1.57e+04	0.489	0.625
multigen_sold	-178.9652	509.901	-0.351	0.726
3room_rental	-322.8696	548.610	-0.589	0.556
Hawker_Within_500m	225.3956	326.432	0.690	0.490
bus_stop_nearest_distance	2.2208	2.850	0.779	0.436

Result in **52** features for model training

Exploratory Data Analysis (EDA)

Exploratory Data Analysis (EDA)

Unit - Flat area, Flat Model, Flat types, Flat Storey

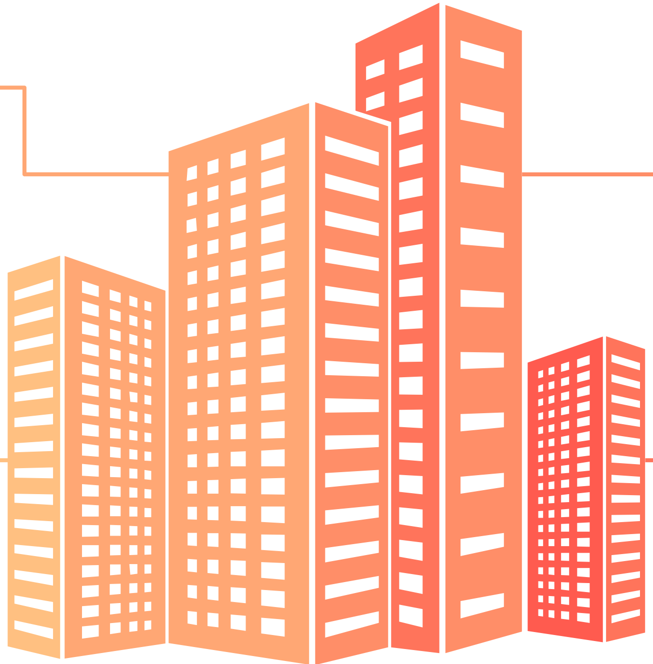


Floor Area

Larger floor area,
higher resale price

Flat Storey

Higher the flat storey,
higher the resale price



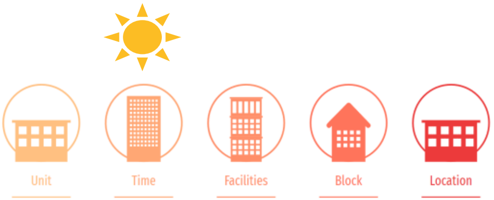
Flat types

Flat types that have greater floor area has higher resale price

Flat Model

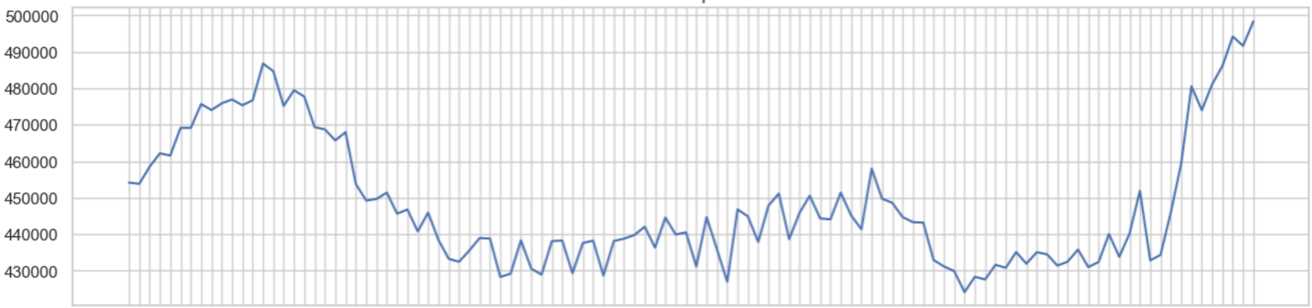
Flat Model prices not necessarily depends on its floor area

Time – Transaction Year Month



Mean Price Across the months

(Jan2012-Dec2021)

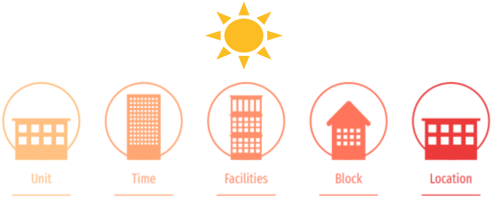


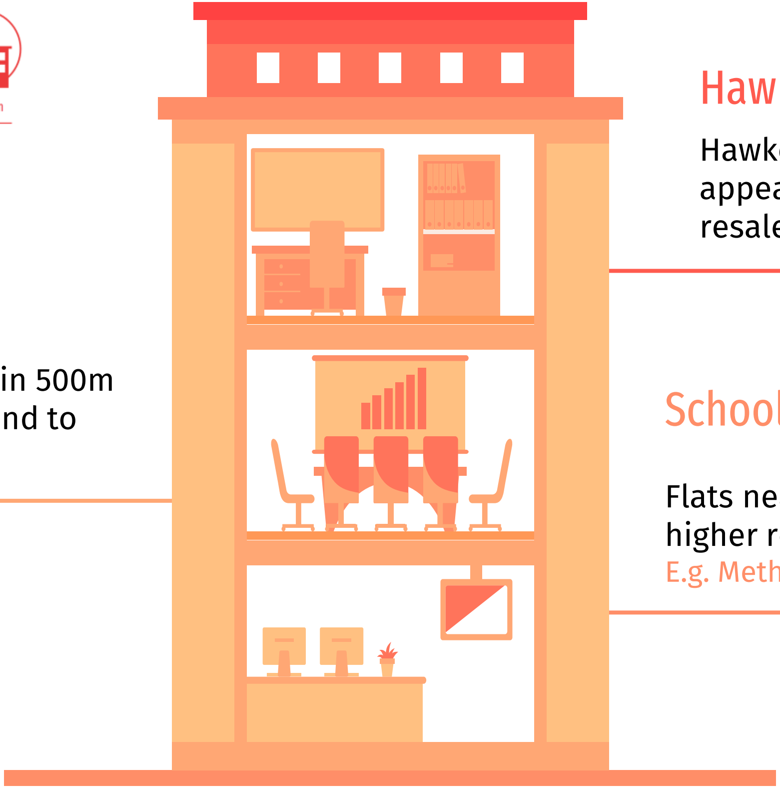
Observation & Remarks

- The fluctuation in HDB resale prices over the years demonstrated poor consistency in seasonality and trend.
- In fact, the prices reflects impact of key events across the years, such as cooling measures[1] implemented by the government in 2013 and 2018.

1: https://stackedhomes.com/editorial/singapore-cooling-measures-history/#gs.sld3js

Facilities - School, Transport, Mall, Hawker





Hawker & Mall

Hawker & Mall related features appeared to have weak correlation with resale price (|r|<0.2)

Transport

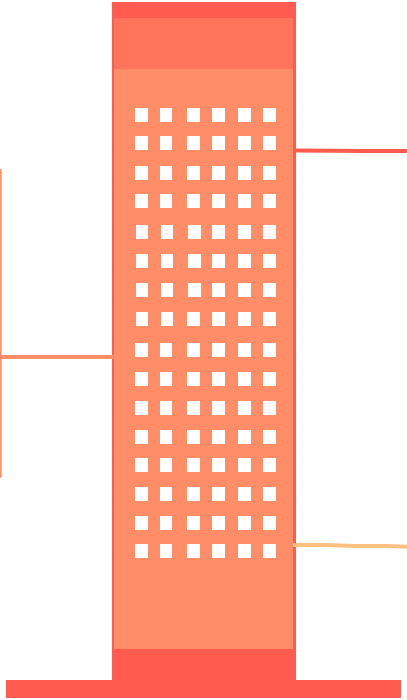
- All flats have a bus stop within 500m
- Flats near specific schools tend to have higher resale prices:

School

Flats near specific schools tend to have higher resale prices:
E.g. Methodist Girls' School

Exploratory Data Analysis (EDA)

Block- Rent, Dwelling Units,HDB Age, Hawker, Max Floor levels



Flats-sold

Flats with higher number of executive or 5-room sold in the building appear to have higher resale-price

Max Floor Level

Units in blocks with higher Max Floor Level appear to have higher resale price (r=0.496)

HDB Age

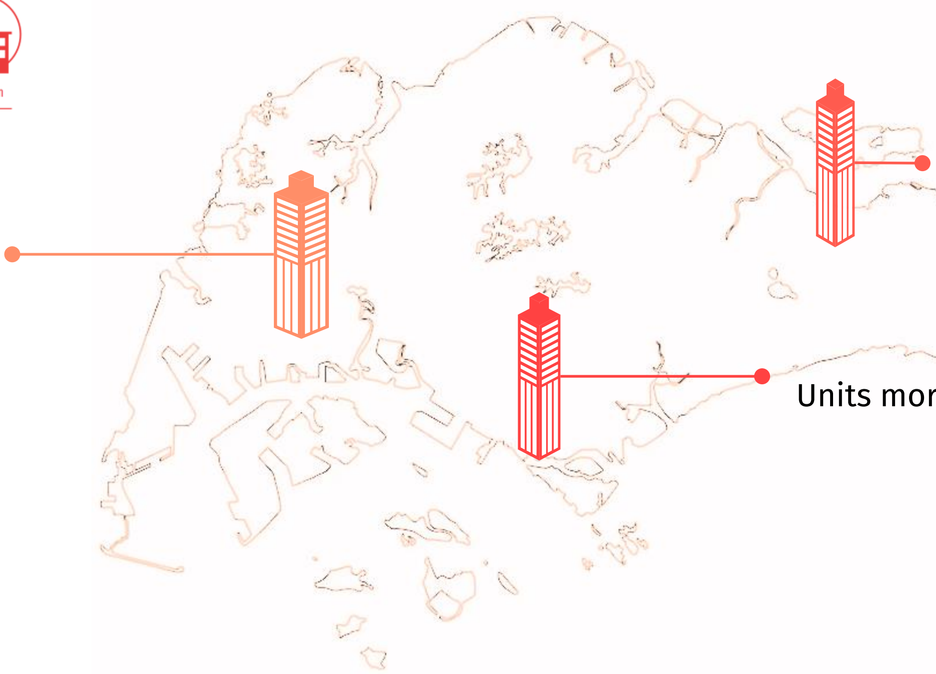
Units in older HDB Blocks appear to have lower resale price (r = -0.35)

Exploratory Data Analysis (EDA)

Location-Longitude, Latitude, Town, Planning Area

Planning area and Town

- Resale prices range vary greatly among the areas.
- Meanwhile, some particular town/ planning areas observed distinctly low or high resale price



Longitude

Units more towards East, higher resale prices

Latitude

Units more towards the South, higher resale prices

Modelling

Modelling Approach

Boolean features

Numerical represented as ‘1’ and 0’

Categorical features

Apply OneHotEncoding

Numerical features

Apply StandardScaler

Regression Models

Linear Regression,
Ridge Regression, Lasso Regression

Modelling Approach

8 Boolean features

32 Numerical features

11 Categorical features

Among these categorical features,
2 of them have notably
high number of unique elements:

- `address`: 9157
- `bus_stop_name`: 1657

This may post computational memory issue…

Models	Algorithm	Features used
A	Linear Regression	Exclude `address` Resulting **2855** number of features post-processing for modelling
B	Ridge Regression	
C	Lasso	Include `address` Resulting **11964** number of features post-processing for modelling

Due to the size of the data post-processing, unable to run Linear Regression or RidgeCV or LassoCV with `address` included

Model Performance Evaluation

Models	Algorithm	Features used	R2 score	RMSE score
A	Linear Regression	Exclude `address` Resulting **2855** number of features post-processing for modelling	Train:0.944 Test:0.941	Train: 3966.9 Test: 34668.3
B	Ridge Regression *(Utilise GridSearchCV to explore different alpha values)*		Train:0.944 Test:0.941	Train: 33954.6 Test: 34644.8
C	Lasso Regression	Include `address` Resulting **11964** number of features post-processing for modelling	Train:0.956 Test:0.951	Train: 30249.5 Test: 31768.7

Best Performing Model

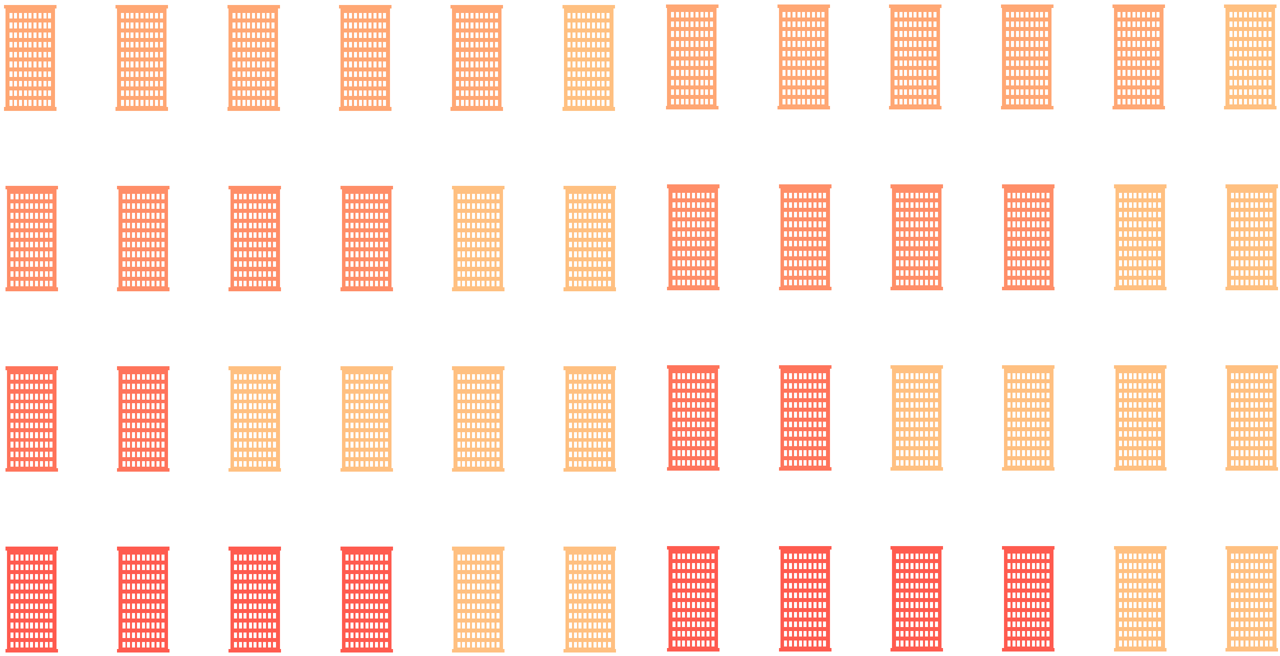
Discussion & Conclusion

Model C Performance

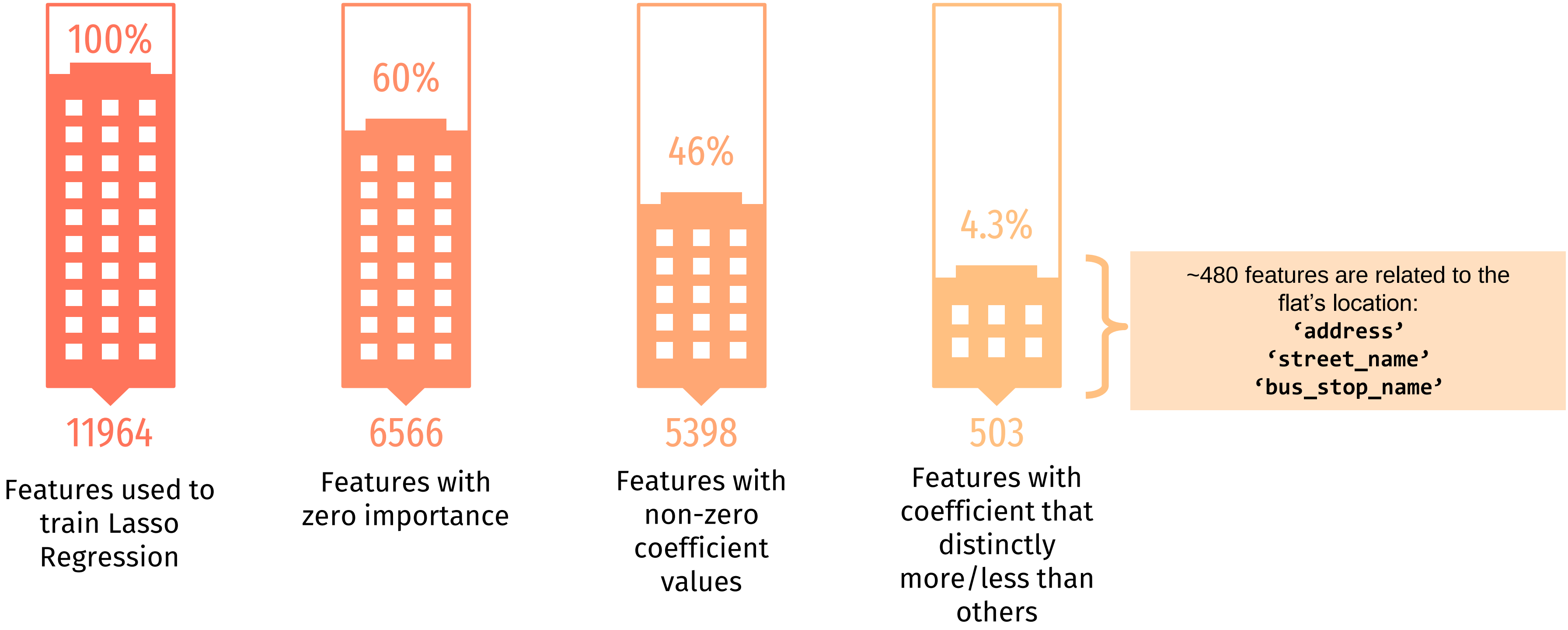
Metric:

R2: 0.95
RMSE: 31769

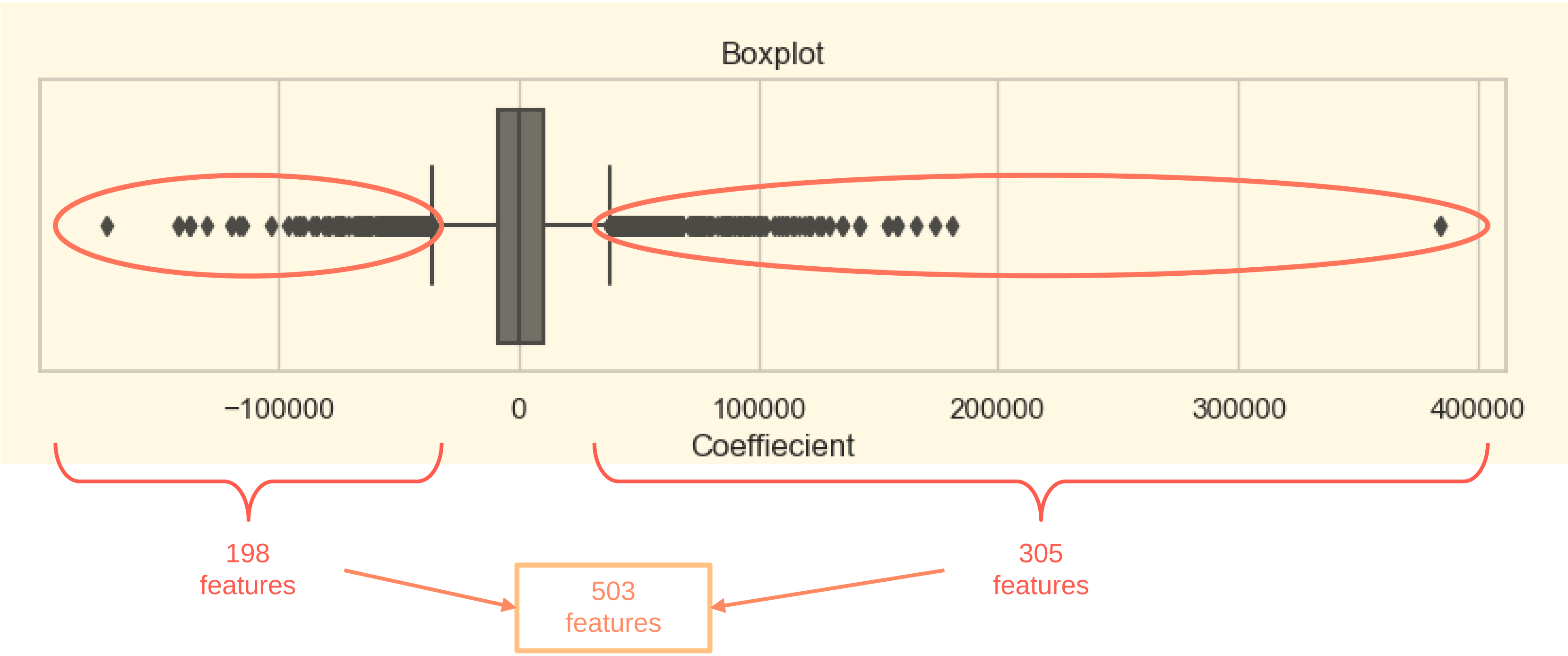
Given a predicted resale price $*X*,
the true resale price within the range ~$*X* +/- 32,000



Model C Performance



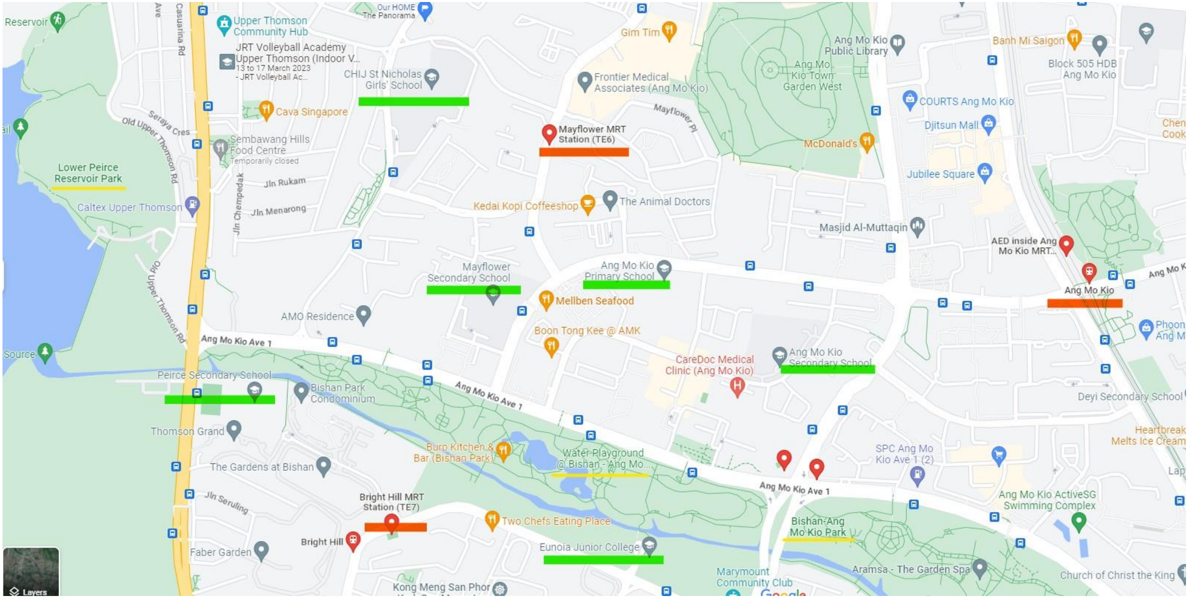
Model C Performance



Conclusion

Example: AMK Ave 2

- Multiple MRT train stations, schools and parks in the area
- Coefficient: 173358 (among top 20 coefficient values)
- In other words, if all else constant, a flat from ANG MO KIO AVE 2 would have $173,358 higher in resale price



- We can confidently recognise that location has important influence on flat resale price
- It account for various facilities available in the vicinity
- However, as seen in earlier EDA, other potential significant factors include key events like implementation of cooling measures.
- As such, in order to maintain accurate prediction of the HDB prices, it will require periodic training of the models with more recent data and explore expert's recognized pricing factors as well.

End