Predictive Modelling of HDB Resale Prices

Leveraging Machine Learning for Market Insights and Decision Support

An AI ML Capstone Project

By: Roslan M Amin (Sep 2024)

Outline

- Introduction
- Machine Learning Approach
- Conclusion
- Applications
- Stakeholder
- Benefits

Introduction

- HDB flats are central to Singapore’s public housing policy, offering affordable homes to over 80% of the population.

- With a rapidly evolving real estate market, accurately predicting HDB resale prices has become increasingly essential for buyers, sellers, and policymakers.

- This project aims to develop a predictive model for HDB resale prices using historical data and advanced machine learning techniques.

- By understanding the factors influencing resale prices, stakeholders can make informed decisions, ensuring a fair and transparent market.

Machine Learning Approach

- Data Acquisition
- Cleaning and Preprocessing
- Exploratory Data Analysis (EDA)
- Model Development
- Model Evaluation

The Dataset

- The dataset is obtained from Kaggle, which originally sourced it from data.gov.sg

  Singapore Resale Flat Prices (2017-2024) (kaggle.com)

- It contained 181262 entries with 11 feature columns

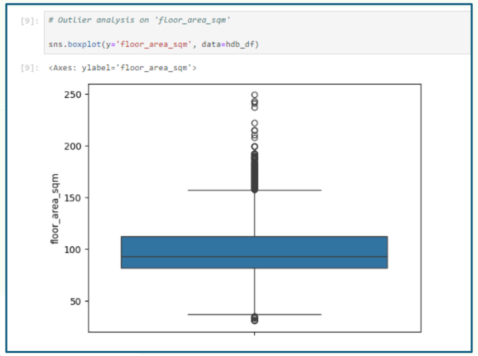
- All but 3 features are categorical, presenting an opportunity to convert them into numerical data types for ML and analysis

Data Cleaning & Preparation

- Missing values
    - All entries are complete, there is no missing value handling were needed at this stage.

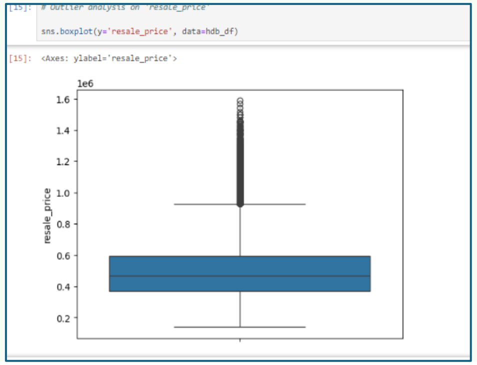
- Feature removal
    - Features such as ‘street_name’, ‘flat_model’ and ‘block’ which were deemed inconsequential in predicting resale price were removed

- Determining outliers
    - Floor sizes above 186 square metres and below 45 square metres were removed
    - Resale price above $927500 were observed but retained to reflect market trend reality



Boxplot to the Left

Outliers above 186 and below 45 sq m



Boxplot to the Left

Outliers above $927500

Data Cleaning & Preparation… con’t

- Feature engineering
  - Categorical features conversion to numerical datatypes
  - ‘remaining_lease’ converted to months (fig. 1)
  - ‘flat_type’ and ‘storey_range’ converted to 1- 6 and 1-17 respectively (fig. 2)
  - ‘towns’ were one-hot encoded (fig. 3)

[21]:
```python
# redefining 'remaining_lease' feature from years and months into months.
# creating a new feature column 'lease_remaining'

def convert_to_months(remaining_lease):
    parts = remaining_lease.split()
    years = 0
    months = 0

    for i in range(len(parts)):
        if 'year' in parts[i]:
            years = int(parts[i-1])
        elif 'month' in parts[i]:
            months = int(parts[i-1])

    total_months = years * 12 + months
    return total_months

hdb_df['lease_remaining'] = hdb_df['remaining_lease'].apply(convert_to_months)

hdb_df = hdb_df.drop('remaining_lease', axis=1)

hdb_df
```

[21]:

	month	town	flat_type	storey_range	floor_area_sqm	lease_commence_date	resale_price	lease_remaining
1	2017-01	ANG MO KIO	3 ROOM	01 TO 03	67.0	1978	250000.0	727
2	2017-01	ANG MO KIO	3 ROOM	01 TO 03	67.0	1980	262000.0	749
3	2017-01	ANG MO KIO	3 ROOM	04 TO 06	68.0	1980	265000.0	745
4	2017-01	ANG MO KIO	3 ROOM	01 TO 03	67.0	1980	265000.0	749
5	2017-01	ANG MO KIO	3 ROOM	01 TO 03	68.0	1981	275000.0	756
...	...	...	...	...	...	...	...	...
181257	2024-06	YISHUN	5 ROOM	10 TO 12	112.0	2018	738000.0	1112
181258	2024-06	YISHUN	5 ROOM	07 TO 09	122.0	1988	680000.0	754
181259	2024-06	YISHUN	EXECUTIVE	10 TO 12	146.0	1986	790000.0	732
181260	2024-06	YISHUN	EXECUTIVE	10 TO 12	146.0	1988	1000000.0	752
181261	2024-06	YISHUN	EXECUTIVE	04 TO 06	146.0	1988	1000000.0	752

180286 rows × 8 columns

fig. 1

[30]:
```python
from sklearn.preprocessing import LabelEncoder

# Initialize the LabelEncoder
le = LabelEncoder()

# Fit and transform the categorical variable
hdb_df['flat_type'] = le.fit_transform(hdb_df['flat_type'])
hdb_df
```

[30]:

	town	flat_type	storey_range	floor_area_sqm	resale_price	lease_remaining
1	ANG MO KIO	1	01 TO 03	67.0	250000.0	727
2	ANG MO KIO	1	01 TO 03	67.0	262000.0	749
3	ANG MO KIO	1	04 TO 06	68.0	265000.0	745
4	ANG MO KIO	1	01 TO 03	67.0	265000.0	749
5	ANG MO KIO	1	01 TO 03	68.0	275000.0	756
...	...	...	...	...	...	...
181257	YISHUN	3	10 TO 12	112.0	738000.0	1112
181258	YISHUN	3	07 TO 09	122.0	680000.0	754
181259	YISHUN	4	10 TO 12	146.0	790000.0	732
181260	YISHUN	4	10 TO 12	146.0	1000000.0	752
181261	YISHUN	4	04 TO 06	146.0	1000000.0	752

180286 rows × 6 columns

[31]:
```python
hdb_df['storey_range'] = le.fit_transform(hdb_df['storey_range'])
hdb_df
```

[31]:

	town	flat_type	storey_range	floor_area_sqm	resale_price	lease_remaining
1	ANG MO KIO	1	0	67.0	250000.0	727
2	ANG MO KIO	1	0	67.0	262000.0	749
3	ANG MO KIO	1	1	68.0	265000.0	745
4	ANG MO KIO	1	0	67.0	265000.0	749
5	ANG MO KIO	1	0	68.0	275000.0	756
...	...	...	...	...	...	...
181257	YISHUN	3	3	112.0	738000.0	1112
181258	YISHUN	3	2	122.0	680000.0	754
181259	YISHUN	4	3	146.0	790000.0	732
181260	YISHUN	4	3	146.0	1000000.0	752
181261	YISHUN	4	1	146.0	1000000.0	752

180286 rows × 6 columns

fig. 2

[34]:
```python
# one-hot encoding for 'town' feature and move the 'resale_price' feature position in the column to last

hdb_df['town'] = le.inverse_transform(hdb_df['town'])
# hdb_df

# hdb_df = pd.get_dummies(hdb_df, columns=['town'])
# hdb_df
```

[35]:
```python
hdb_df = pd.get_dummies(hdb_df, columns=['town'])
hdb_df = hdb_df.astype(int)
col = hdb_df.pop('resale_price')
hdb_df.insert(len(hdb_df.columns), 'resale_price', col)
hdb_df
```

[35]:

	flat_type	storey_range	floor_area_sqm	lease_remaining	town_ANG MO KIO	town_BEDOK	town_BISHAN	town_BUKIT BATOK	town_BUKIT MERAH	town_BUKIT PANJANG	town_BUKIT TIMAH
1	1	0	67	727	1	0	0	0	0	0	0
2	1	0	67	749	1	0	0	0	0	0	0
3	1	1	68	745	1	0	0	0	0	0	0
4	1	0	67	749	1	0	0	0	0	0	0
5	1	0	68	756	1	0	0	0	0	0	0
...	...	...	...	...	...	...	...	...	...	...	...
181257	3	3	112	1112	0	0	0	0	0	0	0
181258	3	2	122	754	0	0	0	0	0	0	0
181259	4	3	146	732	0	0	0	0	0	0	0
181260	4	3	146	752	0	0	0	0	0	0	0
181261	4	1	146	752	0	0	0	0	0	0	0

180286 rows × 31 columns

fig. 3

Exploratory Data Analysis

Key Insights

- The majority of the flats have a floor area between 82 sqm and 112 sqm, with a mean of 97.34 sqm. This suggests that most flats are medium-sized.

- Resale prices vary significantly, with a mean price of around $498,733.70. The prices are skewed towards higher values. The maximum price recorded is $1,588,000.

- The lease remaining for most flats is quite high, with a median of 74.5 years. This indicates that the majority of the flats have a substantial amount of lease remaining, which is a positive factor for potential buyers.

[33]:
```
hdb_df.describe()
```

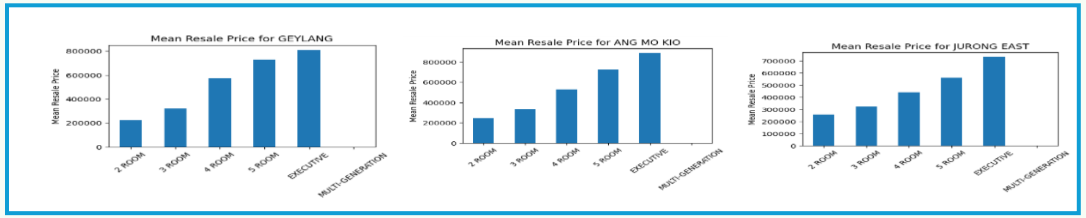
[33]:

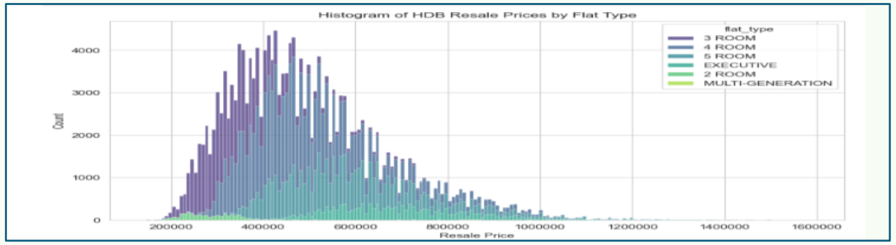
	floor_area_sqm	resale_price	lease_remaining
count	180286.000000	1.802860e+05	180286.000000
mean	97.344420	4.987337e+05	895.699084
std	23.628913	1.718718e+05	167.015426
min	45.000000	1.400000e+05	498.000000
25%	82.000000	3.700000e+05	759.000000
50%	93.000000	4.688880e+05	894.000000
75%	112.000000	5.930000e+05	1060.000000
max	186.000000	1.588000e+06	1173.000000

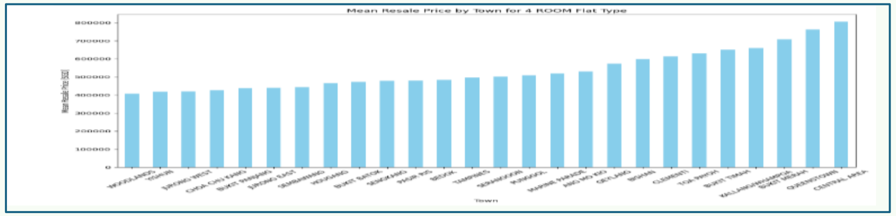
[ ]:

Exploratory Data Analysis… con’t

- Graphs (right) show variation in resale prices across flat types across towns
- The larger the flat type, the higher are the prices





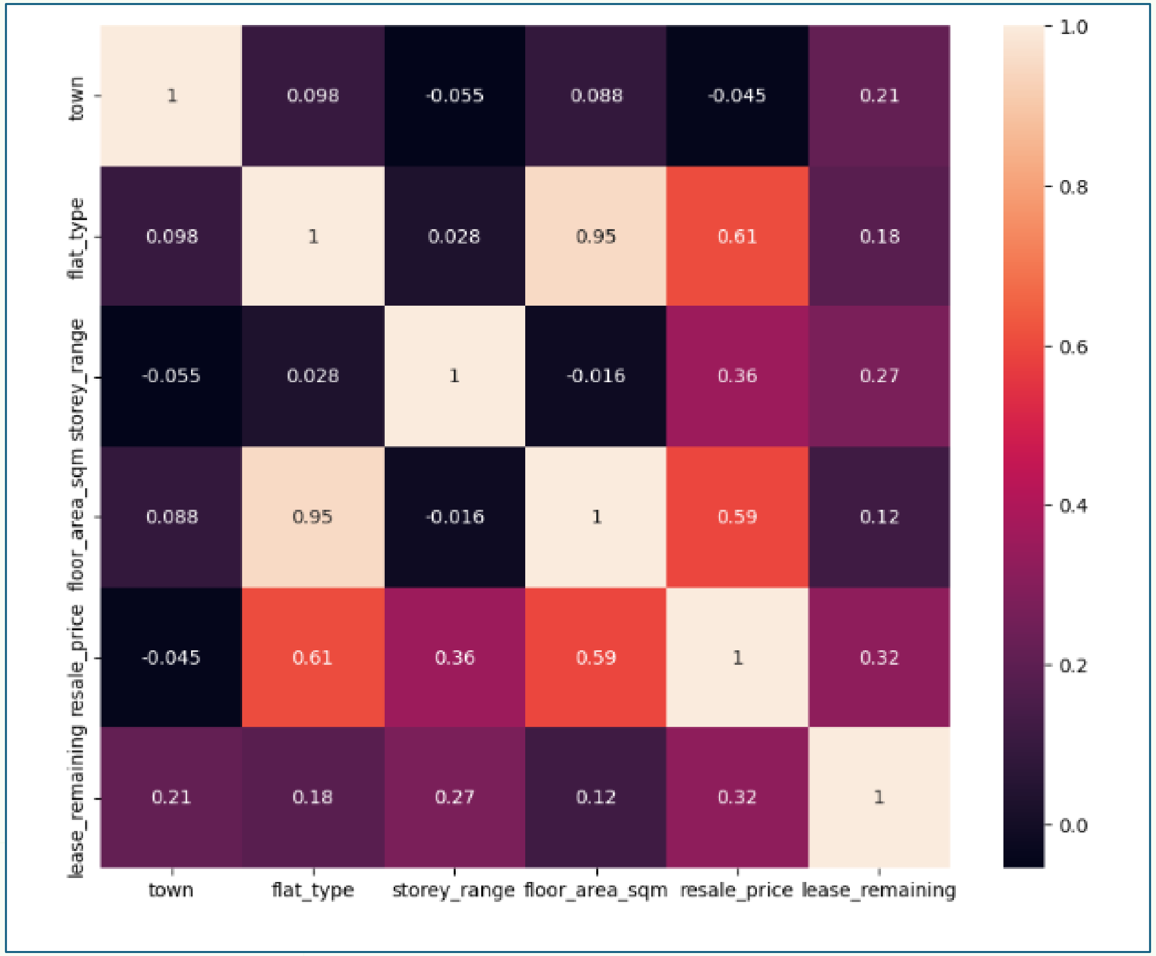


- The histogram (left) depicts the distribution of resale prices across various flat types. The data is skewed to the right, indicating the presence of outliers for higher-end prices
- Chart (left) comparing mean resale prices across towns by flat type

Observation

- Resale price movements are in line with expectation
- Information used for dataset validation more than for gathering insights

Correlation Analysis



Heatmap

- ‘flat_type’ and ‘floor_area_sqm’ show strong positive correlations
- ‘lease_remaining’ and ‘storey_range’ exhibit moderate positive correlations
- Interestingly, the ‘town’ feature has a weak negative correlation index of –0.045

Model Development

Algorithm Selection

- Linear Regression
  Simplicity, Efficiency and Interpretability
- Decision Tree
  Ability to capture non-linear relationship, interpretability and ability to provide insights
- Random Forest
  Robust, Accurate and ability to give more reliable measure from aggregated key features

Model Evaluation

- Mean Absolute Error
- R-Squared Score
- Root Mean Squared Error

Model Adjustment

- Adjustment to remove price outliers

Results

Before Adjustment	After Adjustment
▪ **Linear Regression**: MAE: $ 74448.84 R^2: 71% RMSE: $ 91973.83	▪ **Linear Regression**: MAE: $ 70484.87 R^2: 70% RMSE: $ 85795.73
▪ **Decision Tree Regression**: MAE: $ 47785.42 R^2: 83% RMSE: $ 69962.65	▪ **Decision Tree Regression**: MAE: $ 46558.21 R^2: 81% RMSE: $ 67678.2
▪ **Random Forest**: MAE: $ 40828.95 R^2: 89% RMSE: $ 57943.24	▪ **Random Forest**: MAE: $ 39457.82 R^2: 87% RMSE: $ 55631.28

Observation

- Improvements in the MAE and RMSE scores
- R^2 scores have decreased slightly by 1 to 2%
- Retention or removal of outliers debatable
- Keep the outliers to ensure results remain realistic and to capture the real situation in the market

Conclusion

- The HDB Resale Prediction Project has successfully demonstrated the potential of ML) and AI in accurately forecasting resale prices.

- By leveraging a diverse set of features, including location, flat type, floor area, and transaction history, the model achieved a relatively high degree of predictive accuracy.

- The model’s performance metrics indicate a robust ability to predict resale prices, with a mean absolute error (MAE) of $ 40828.95, R-squared (R^2) of 89% and a root mean square error (RMSE) of $ 57943.24.

The Applications

- Rental Price Prediction
  - Helps tenants and landlords set fair rental prices and understand trends like resale pricing approach
- Valuation for Insurance
  - Ensures properties are adequately insured based on their true market value
- Mortgage Risk Analysis
  - Helps financial institutions assess risks associated by predicting property values
- Investment Analysis
  - Helps investors find opportunities by predicting future market trends

The Stakeholder

- Homebuyers and Sellers
  - Model can be used to estimate fair prices
- Real Estate Agents
  - Leverage insights from model to provide better advice to clients
- Policymakers
  - The analysis can be used to understand market trends and implement policies
- Financial Institutions
  - Assess value of properties more accurately
- Researchers and Academics
  - Further research in real estate economics and urban planning

The Benefits

- Transparency
  - Data-driven price estimates enhance transparency in the resale market
- Informed Decision
  - Empowers stakeholders with accurate information for decision making
- Market Stability
  - Reduced price speculation helps stabilize market
- Policy Formulation
  - Assist policymakers in crafting data-driven housing policies to address market needs


brain
genius
creativity
search
WORK
TEAM
PLAN
IDEA
Idea
WEB
PLAN
BUSINESS

Thank You!