DSA5101 Project Report

A Deep Dive into the HDB Resale Prices Dataset

Submitted by

Hon Cheng Hui (Allen)

A0296351Y

Department of Mathematics, MSc. DSML

Table of Contents

Introduction

As part of the project requirement for DSA5101, the HDB resale flat prices was obtained from https://data.gov.sg/datasets/d_8b84c4ee58e3cfc0ece0d773c8ca6abc/view for analysis. An ensemble of models was compared against each other to predict the 'resale_price'.

Exploratory Data Analysis

In a brief snippet of the dataset (Figure 1), this dataset has a few continuous and categorical variables. Upon checking the data types, there are 4 numerical/continuous data types and 7 categorical data types in the dataset. For simplicity 'lease_commence_date' has been used to determine the age of the resale flat price, instead of the 'remaining_lease'.

```
data= pd.read_csv(r"Resale flat prices based on registration date from Jan-2017 onwards.csv")
data
```

✓ 0.1s

	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
...	...	...	...	...	...	...	...	...	...	...	...
191662	2024-10	YISHUN	5 ROOM	504C	YISHUN ST 51	10 TO 12	113.0	Improved	2016	90 years 06 months	747000.0
191663	2024-10	YISHUN	5 ROOM	502D	YISHUN ST 51	04 TO 06	112.0	Improved	2018	92 years 04 months	725000.0
191664	2024-10	YISHUN	5 ROOM	644	YISHUN ST 61	07 TO 09	123.0	Improved	1988	62 years 09 months	728000.0
191665	2024-10	YISHUN	5 ROOM	870	YISHUN ST 81	04 TO 06	122.0	Improved	1987	62 years 04 months	690000.0
191666	2024-10	YISHUN	EXECUTIVE	654	YISHUN AVE 4	10 TO 12	164.0	Apartment	1992	66 years 09 months	1000000.0

191667 rows × 11 columns

Figure 1: Brief look at the dataset

Table 1: Data types from dataset

Variable	dtype
month	object
town	object
flat_type	object
block	object
street_name	object
storey_range	object
floor_area_sqm	float64
flat_model	object
lease_commence_date	int64
remaining_lease	object
resale_price	float64

The continuous data was displayed in a pair plot. A new column 'age' was created by subtracting the 'lease_commence_date' from the current year. From the data (Figure 2) we can see a clear correlation between 'resale_price' and 'floor_area_sqm'. Another interesting observation is the bimodal distribution of the 'age' data. This could strongly imply that resale flats generally have 2 periods of increased sales – once after their Minimum Occupancy Period (MOP) of 5 – 10 years, and another at around 40 years of age

(close to half of the 99-year lease period).

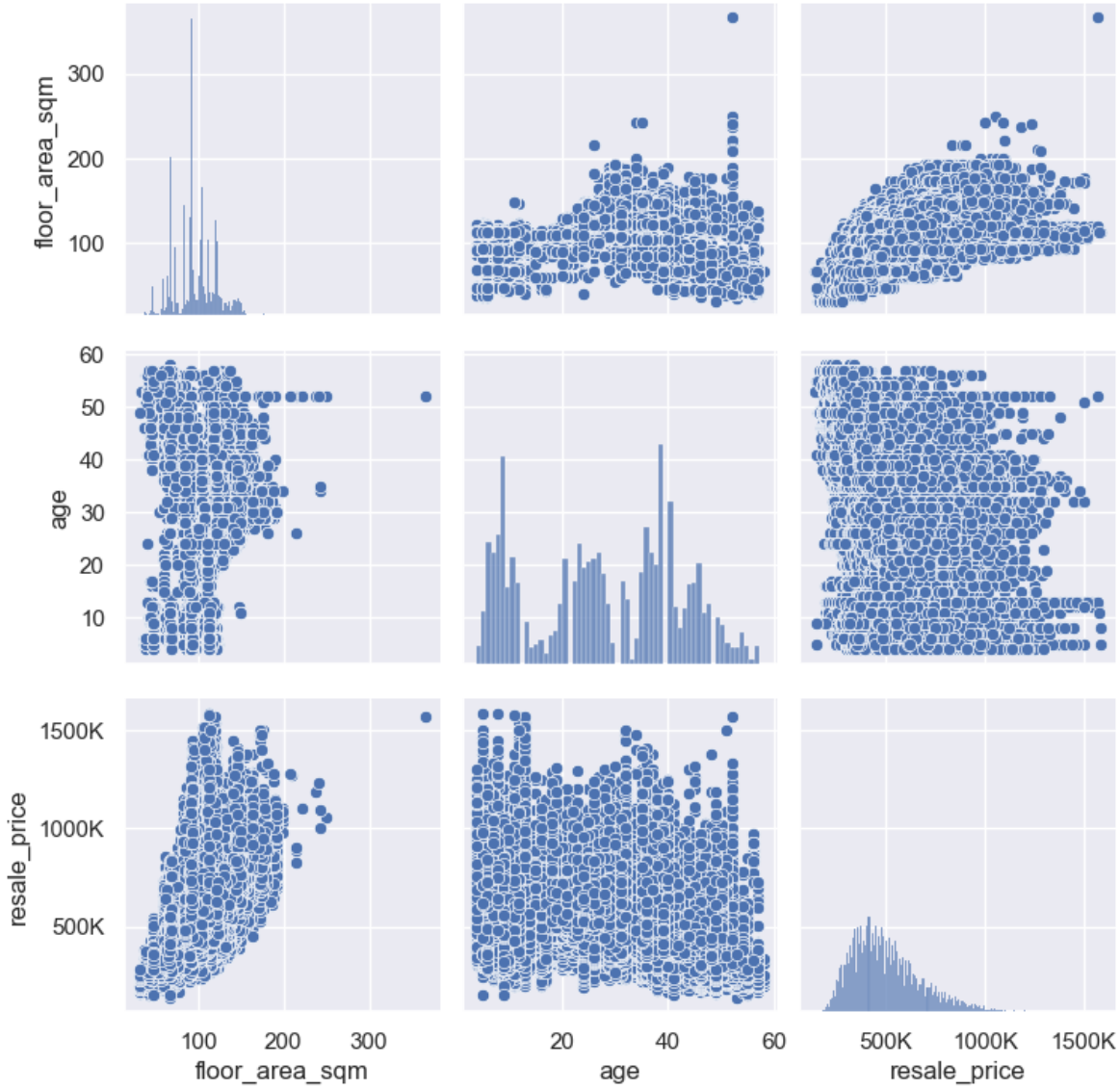

Figure 2:Pair plot of continuous variables of HDB dataset

For the categorical data, was sorted in ascending order by their median resale_price value. For data cleaning, variables were mostly one-hot encoded to be used in the models for resale_price prediction later.

For the 'flat_type' variable, as expected, the larger flat types are sold at a higher price, compared to the smaller flat types.

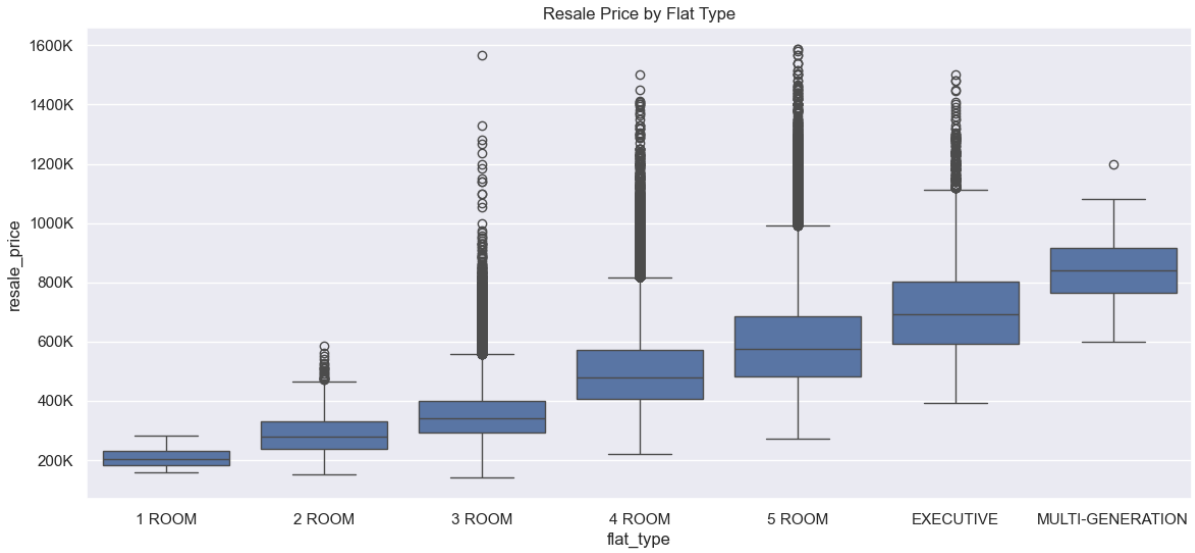

Figure 3: resale_price against flat_type

From the 'resale_price' against 'town' data, it is observed that the median resale price for resale HDBs is the highest in Bukit Timah.

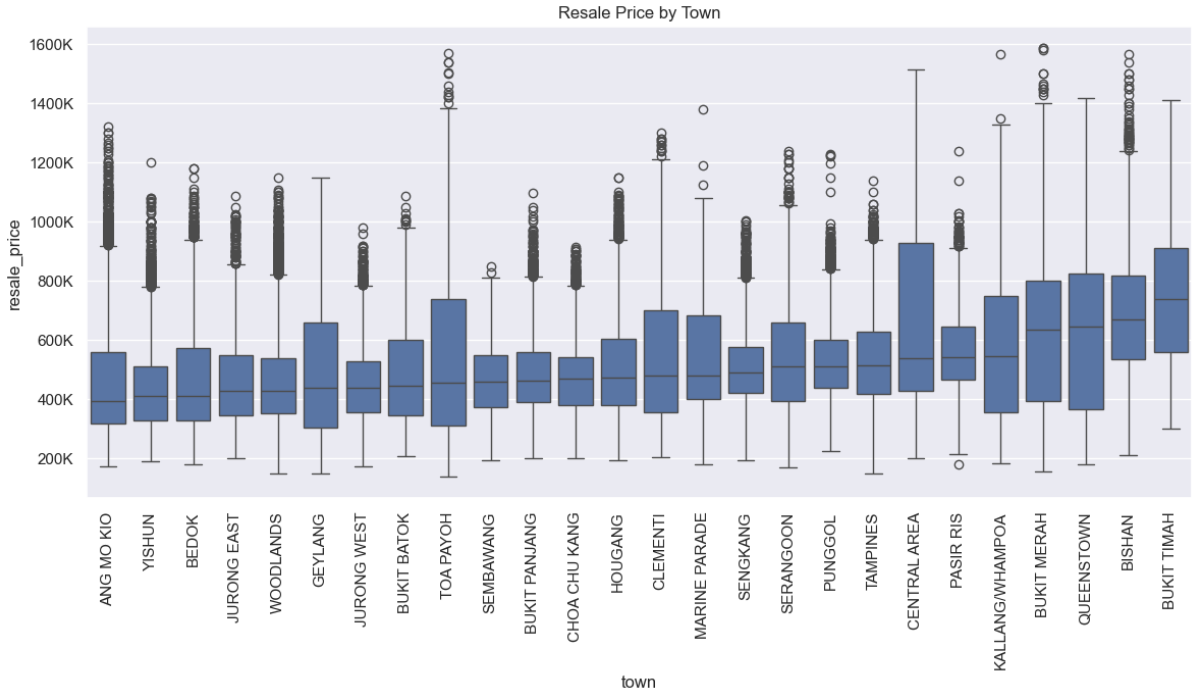

Figure 4: resale_price against town

As for the Flat Model with the highest median resale price, Type S2 is the clear winner here.

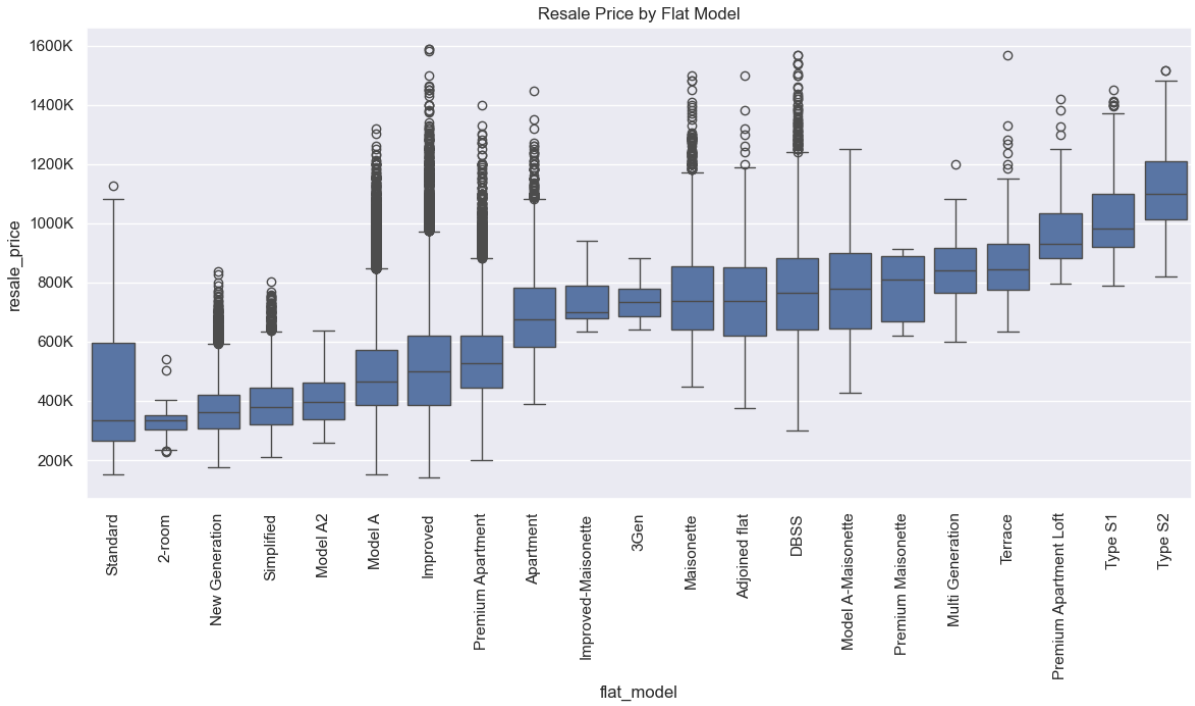

Figure 5: resale_price against flat_model

From the storey_range data, we can see that the median resale price of HDBs trends upwards the higher the flat is. Howevr an interesting point to note is the number of outliers in some of the lower floors.

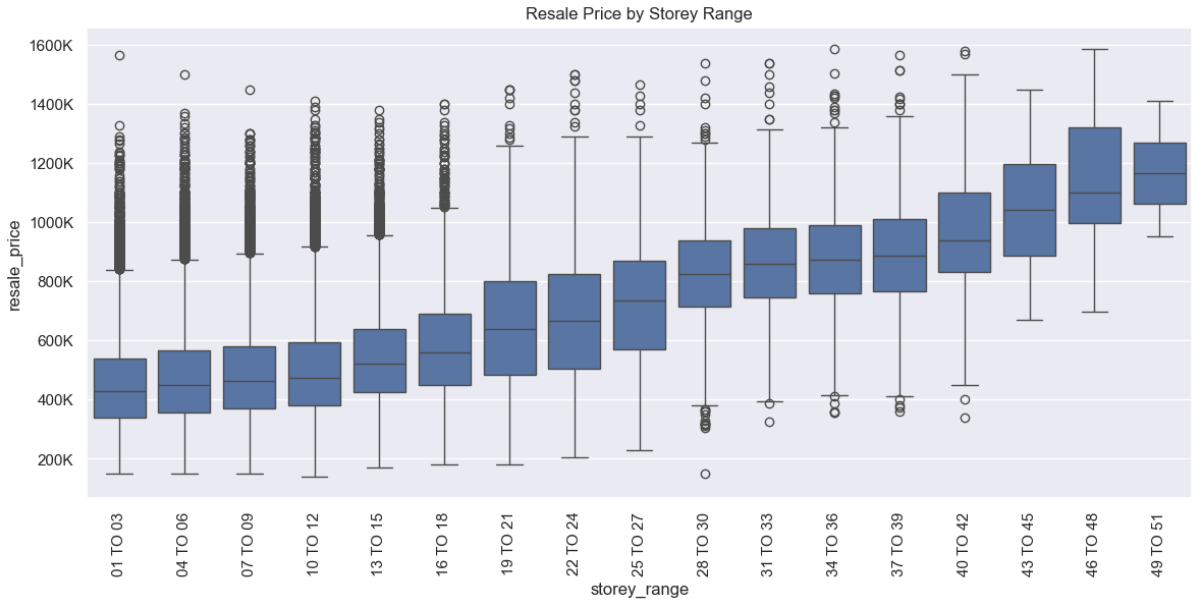

Figure 6: resale_price against storey_range

Methodology

Data Cleaning

Categorical variables like 'town', 'storey_range', and 'flat_model' were one-hot encoded. The categorical variable 'flat_type' was ordinally encoded as there was meaningful sequence to the order. We assigned the value of 1to 5 for the corresponding number of rooms, and the value of 6 and 7 for the 'EXECUTIVE and 'MULTI-GENERATION' types of flats. Here we are making an assumption that the 'MULTI-GENERATION' class weighs more than the 'EXECUTIVE' class.

As the dataset is too large to fit into main memory, a random sampling technique with a fraction of 0.1 was used to sample the dataset.

Polynomial Regression

Polynomial regression is an extension of linear regression that allows a model to capture non-linear relationships. A polynomial model with degree 2 can be expressed as:

$$\hat{f}(\mathrm{x}) = \beta_0 + \beta_1 x + \beta_2 x^2 + \epsilon$$

Where β_0 refers to the bias term polynomial regression model, β_1 is the coefficient of the linear term, β_2 is the coefficient of the quadratic term, and ϵ is the error term.

Elastic Net

Elastic Net is a regularization method that combines both Lasso and Ridge regression. The loss function of Elastic Net can be written as:

$$L_{Elastic\ Net} = \frac{1}{2n}||y - Xw||_2^2 + \alpha \cdot \lambda_1 \cdot \left|||w|\right|_1 + \frac{1}{2} \cdot \alpha \cdot (1 - \lambda_1) \cdot ||w||_2^2$$

Where $\frac{1}{2n}||y - Xw||_2^2$ represents the mean squared error between the true output values y and the predicted values Xw, $\alpha \cdot \lambda_1 \cdot \left|||w|\right|_1$ represents the L1 regularization term on the weights w, and $\frac{1}{2} \cdot \alpha \cdot (1 - \lambda_1) \cdot ||w||_2^2$ represents the L2 regularization term of the weights.

As the dataset contains non-linear data, this report combines both a polynomial model with Elastic Net to fit non-linear data, while utilizing regularization to prevent overfitting.

Kernel Ridge Regression

In this report, various kernel ridge regression (KRR) models were used to predict the resale_price of resale HDBs. From the $\ell^2 - regularized$ least squares problem, the best approximator function $\hat{f}$ and the least squares solution $\widehat{w}$ can be written as:

$$\hat{f}(\mathrm{x}) = \sum_{j=0}^{M-1} \widehat{w}_j{}^T \phi_j(x) \;=\; \phi(x)^T \widehat{w}$$

$$\widehat{w} = (\phi^T\phi + \lambda I_M)^{-1}\phi^T y$$

Where M is the total number of basis functions and $\phi_j(x)$ is the j-th basis function evaluated at x for the first equation. For the second equation, $\phi^T\phi$ refers to the matrix product of the design matrix with its transpose, resulting in an M x M matrix, λ refers to the regularization parameter, I_M is an M x M identify matrix, and y which is the vector of outputs for each training example.

The closed form solution for the least squares solution $\widehat{w}$ can be rewritten as:

$$\widehat{w} = \phi^T(\phi\phi^T + \lambda I_N)^{-1}y$$

Subsequently, $\hat{f}(\mathrm{x})$ can be rewritten as:

$$\hat{f}(\mathrm{x}) = \phi(x)^T\phi^T(\phi\phi^T + \lambda I_N)^{-1}y$$

$$\hat{f}(\mathrm{x}) = \sum_{i=1}^{N} \phi(x)^T\phi^T\,\alpha = \sum_{i=1}^{N} \alpha\, k(x_i, x)$$

Where $\alpha = (\phi\phi^T + \lambda I_N)^{-1}y$ and $\phi(x)^T\phi^T$ has been reformulated to $k(x_i, x)$ using the kernel trick. Using this new formulation, we can easily choose from multiple different kernels, such as the linear kernel $\mathrm{k}(x, x') = x^T x'$, the polynomial kernel $\mathrm{k}(x, x') = (1 + x^T x')^m, m > 0$ and the radial basis function (rbf) kernel $\mathrm{k}(x, x') = \exp\left(-\frac{||x-x'||^2}{2s^2}\right), s > 0$ for more complex non-linear relationships. Additionally, KRR should have no issues handling high-dimensional data as it uses the kernel trick to avoid transforming the data into higher-dimensions.

Principal Component Analysis (PCA)

PCA is a dimensionality reduction technique used to transform aa dataset into a lower-dimensional space while maximizing the variance. Apart from reducing the computation time needed to evaluate the models, PCA may improve model performance by reducing multicollinearity and reducing noise by removing less significant components.

K-means Clustering

K-means is a clustering algorithm used for partitioning a dataset into clusters based on their Euclidean distances between each data point x_i and the nearest centroid c_k (Figure 7). To determine the number of clusters, the elbow method was used, and the

dataset was clustered into 8. As part of feature engineering, the clusters were used as predictors and were one-hot encoded as categorical data.

Algorithm 8: K-means Clustering Algorithm

Data: $\mathcal{D} = \{x_i\}_{i=1}^{N}$, $x_i \in \mathbb{R}^d$ for all i
Hyperparameters: K (number of clusters); stopping criterion
Initialize: $Z \in \mathbb{R}^{K \times d}$
while *stopping criterion not reached* **do**
  Update R: $r_{ik} = \begin{cases} 1 & k = \arg\min_j \|x_i - z_j\|^2 \\ 0 & \text{otherwise.} \end{cases}, i = 1, \ldots, N;$
  Update Z: $z_k = \frac{\sum_{i=1}^{N} r_{ik} x_i}{\sum_{i=1}^{N} r_{ik}}, k = 1, \ldots, K;$
end
return *cluster centers* Z, *cluster assignments* R

Figure 7: K-means Clustering Algorithm from the Lecture Slides

Evaluation Metric

K-fold Cross Validation

K-fold cross validation (cv) is a robust way for evaluating the performance of a model by dividing the dataset into k equal-sized folds or partitions. For the sake of reducing computational time, k was chosen to be 3 for this project, and the dataset was split into k-equal folds. Next, the model is trained 3 times, with each time using a different fold as the validation set, and the remaining k-1 folds as the training set. Finally, after k-iterations where each fold has been used once as the validation set, the R^2 score was used and averaged across all folds.

For the models that have parameters that can be tuned, GridSearchCV was used to find the optimal parameters. The search space is listed as:

```
# Define parameter grids for ElasticNet and RBF KernelRidge
param_grids = {
'elastic_net': {
'alpha': [0.1, 1.0, 10.0],
'l1_ratio': [0.1, 0.5, 0.9],
},
'rbf_kernel': {
'alpha': [0.01, 0.1, 1.0],
'gamma': [0.01, 0.1, 1.0],
}
}
```

Results and Discussion

The Mean R^2 scores of the various models were tabulated in Table 2. The baseline models from Elastic Net, Polynomial Kernel, and RBF Kernel perform relatively well with a Mean R^2 score of at least 0.75. Across the Predicted against True 'resale_price' scatterplots (Figure 8, Figure 9, Figure 10, Figure 11), there is a strong linear relationship between the variables, indicating that

the models are generally capturing the relationship between the features and the output variable.

Introducing PCA reduces the R^2 score across all the models, suggesting that dimensionality reduction could be discarding too many important features.

Implementing K-means alone improves the R^2 score across all the models. This suggests that generally, the K-means clusters are capturing some patterns in the data that the prior models using the prior features could not. Utilizing K-means clusters with the Polynomial Kernel resulted in the highest Mean R^2 score of 0.8294 in this study.

Finally, implementing PCA + K-means actually reduced the Mean R^2 score for Elastic Net and the Polynomial Kernel, but showed an increase in the R^2 score for the RBF Kernel when compared to the baseline. This may suggest that adding PCA with the K-means clusters could have removed some components with lower variance that are still important for prediction in Elastic Net and Polynomial Kernel. Additionally, RBF kernel relies less on the directionality of the features but rather the Euclidean relationships in the transformed space. Hence the RBF kernel is not affected by the PCA step.

Table 2: Results of models with various feature engineering

	Mean R^2 score (cross-val)		
	ElasticNet	Polynomial Kernel	RBF Kernel
Baseline	0.7982	0.8015	0.7671
PCA	0.6838	0.6717	0.7767
K-means	0.8263	0.8294	0.7927
PCA + K-means	0.7929	0.7919	0.8126

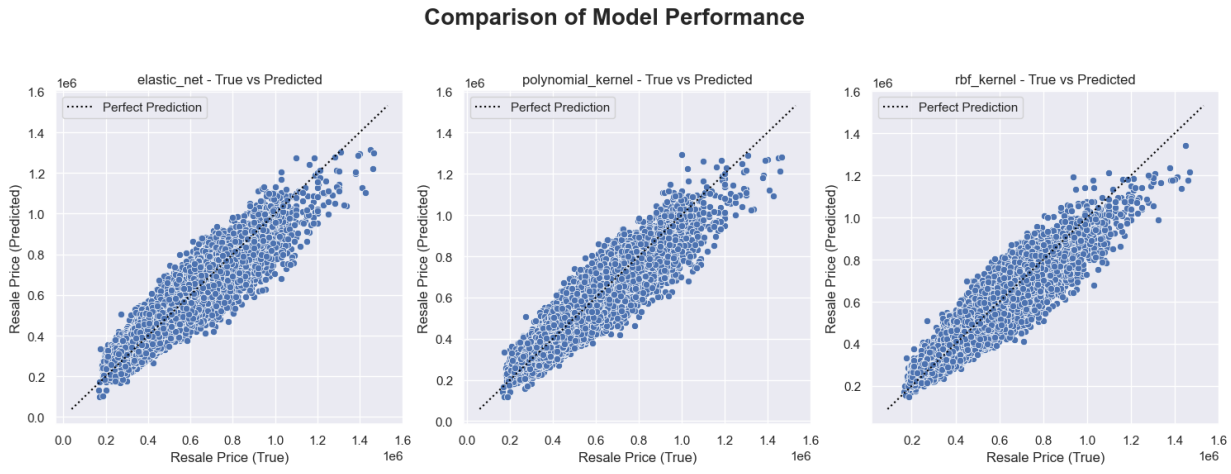

Figure 8: Comparison of Baseline model performance

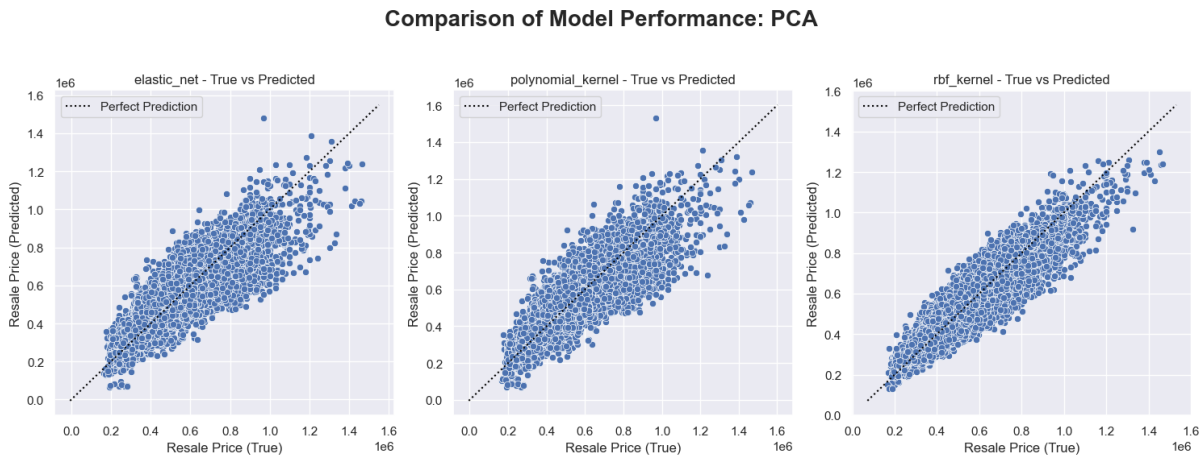

Figure 9:Comparison of Model Performance after PCA

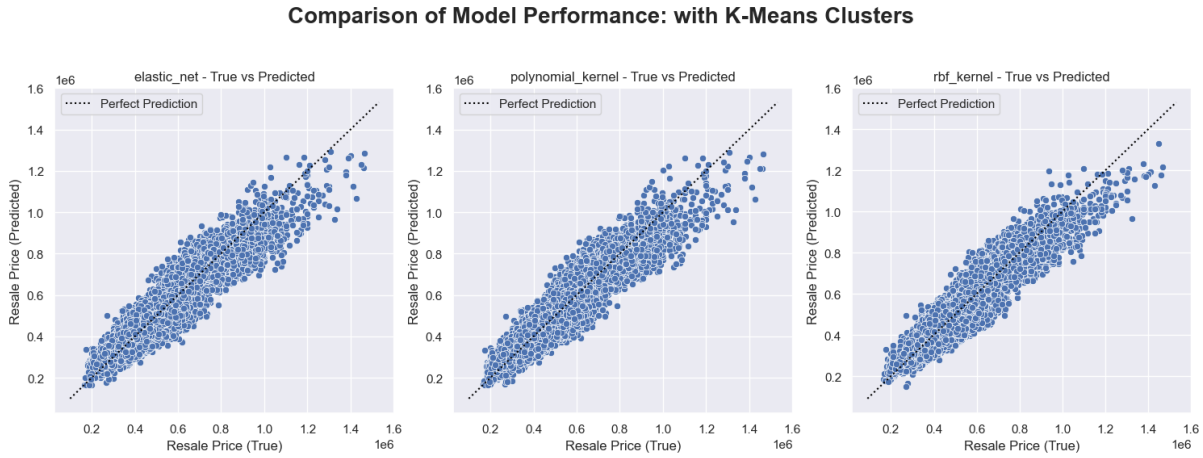

Figure 10: Comparison of Model Performance using K-Means Clusters as features

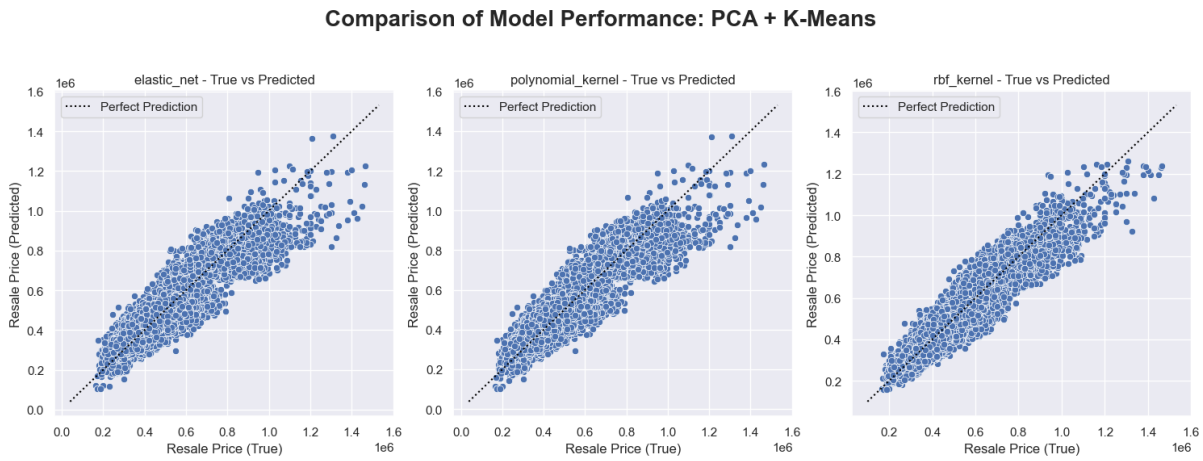

Figure 11: Combining all methods used in this study - PCA and K-Means Clustering

Conclusion & Future Work

In conclusion, this report presents a deep dive into the HDB Resale Prices dataset and implements a predictive model to accurately predict the resale price of HDBs.

Comparing across models, the Polynomial Kernel with the added feature from K-means clustering had the highest R^2 score of 0.8294, followed closely by the Elastic Net model with K-means with an R^2 score of 0.8263, and lastly the RBF Kernel with PCA + K-means with an R^2 score of 0.8126. This highlights the importance of feature engineering and model comparison while evaluating various models.

In the future, other clustering algorithms like Gaussian Mixture Models can be considered and added to the features to see if the models improve.