**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)

Table of Contents

[1. Project Overview 3](#_Toc178924499)

[2. Introduction 4](#_Toc178924500)

[3. Objectives 5](#_Toc178924501)

[4. Analysis Goals 6](#_Toc178924502)

[5. Detailed Aspects and Analysis of the Project 7](#_Toc178924503)

[5.1 Data Source and Preparation 7](#_Toc178924504)

[5.2 Exploratory Data Analysis (EDA) 14](#_Toc178924505)

[5.3 Model Development 22](#_Toc178924506)

[5.4 Model Evaluation 24](#_Toc178924507)

[5.5 Feature Importance and Interpretation 27](#_Toc178924508)

[5.6 Model Adjustments 27](#_Toc178924509)

[6. Stakeholders 34](#_Toc178924510)

[**7.** Benefits 35](#_Toc178924511)

[8. Use Cases 36](#_Toc178924512)

[9. Ethics and Governance 38](#_Toc178924513)

[9.1 Ethical Considerations 38](#_Toc178924514)

[9.2 Governance 39](#_Toc178924515)

[10. Conclusion 40](#_Toc178924516)

[10.1 Key Findings 40](#_Toc178924517)

[10.2 Implications 40](#_Toc178924518)

[10.3 Future Directions 41](#_Toc178924519)

# Project Overview

This project focuses on the application of machine learning to predict HDB resale prices and providing valuable insights for various stakeholders. It highlights the dual goals of developing a predictive model and offering decision support, making it clear and comprehensive.

# Introduction

The Housing and Development Board (HDB) flats are a cornerstone of Singapore’s public housing policy, providing affordable housing to over 80% of the population. With the dynamic nature of the real estate market, predicting HDB resale prices has become increasingly important 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.

# Objectives

* Develop a Predictive Model

Create a robust model to accurately predict HDB resale prices based on historical data.

* Identify Key Factors

Determine the most significant factors affecting resale prices, such as location, flat type, floor area, and remaining lease.

* Enhance Decision-Making

Provide insights to buyers, sellers, and policymakers to facilitate informed decision-making.

* Market Analysis

Analyse trends and patterns in the HDB resale market to understand its dynamics and future directions.

# Analysis Goals

* Data Cleaning and Preprocessing

Ensure the dataset is clean and ready for analysis by handling missing values, outliers, and transforming categorical variables into numerical formats.

* Exploratory Data Analysis (EDA)

Perform EDA to uncover initial insights and visualize relationships between variables.

* Model Development

Use various machine learning algorithms (e.g., linear regression, random forest, gradient boosting) to develop and validate the predictive model.

* Model Evaluation

Assess the model’s performance using metrics such as R-squared, Mean Absolute Error (MAE), and Root Mean Squared Error (RMSE).

* Feature Importance

Identify and rank the importance of different features in predicting resale prices.

# Detailed Aspects and Analysis of the Project

## Data Source and Preparation

* Dataset

This dataset includes resale price information for HDB flats in Singapore, covering the period from January 2017 to June 2024. It contains records from reliable sources such as the URA, and other relevant databases. The dataset is comprehensive, showing details such as transaction month, town, flat type, block, street name, storey range, floor area, flat model, lease commencement date, remaining lease, and resale price.

For this project, the dataset was downloaded from Kaggle using Python code, which originally sourced it from data.gov.sg.

[Singapore Resale Flat Prices (2017-2024) (kaggle.com)](https://www.kaggle.com/datasets/darrylljk/singapore-hdb-resale-flat-prices-2017-2024)

|  |  |
| --- | --- |
| A screenshot of a computer | Click below to see code shown on the left. This is to download the source datafile from Kaggle. |

A screenshot of a computer

Description automatically generated

To find out the content and manipulate the dataset, I imported the Pandas and Numpy libraries, and used the command code hdb\_df.info() (see above). This reveals that the dataset contains **181,262 entries** and **11 feature columns**. These columns are: 'month', 'town', 'flat\_type', 'block', 'street\_name', 'storey\_range', 'floor\_area\_sqm', 'flat\_model', 'lease\_commence\_date', 'remaining\_lease', and 'resale\_price'.

All feature columns, except for three, contains categorical datatypes. The features 'floor\_area\_sqm' and 'resale\_price' are of the float datatype, while the feature 'lease\_commence\_date' is of the integer data type.

Importantly, the entries to all the features are complete ie there are no null entries or missing values.

* Data Cleaning

Missing values

Since all entries are complete, there is no adjustment needed at this stage. This ensures a smooth transition to the next steps in the data preparation and analysis process.

Feature removal

In the initial stages of this data preparation, features deemed inconsequential to the resale price will be removed. The selection is currently based on a judgement call, leveraging my familiarity and understanding of the housing market. Later, a correlation analysis will be conducted to determine which of the remaining features significantly impact the resale price.

The features to remove are ‘street\_name’, ‘flat\_model’ and ‘block’.

Shown below is the code to remove the features stated (line 7). I have at this point also added the code ‘pd.set\_option('display.max\_columns', None)’ and imported matplotlib together with the seaborn libraries for data visualization. The code in line 1 is to allow full display of columns in the dataset. Also at this point, I have renamed the code file to ‘hdb\_dataset\_dropped\_features.ipynb’

|  |  |
| --- | --- |
|  | Click below to see code shown on the left. |

Determining outliers

The dataset is now left with 8 feature columns with 2, ‘floor\_area\_sqm’ and ‘resale\_price’, having numerical datatypes. Performing outlier analysis on these 2 features would reveal any outliers that exist within them. Outliers are caused by variability in the data and by a possible experimental measurement error. This would in turn result in various problems in the statistical analysis and may also cause a significant impact to the results ahead.

The boxplot analysis below for ‘floor\_area\_sqm’ indicates significant outliers, with many entries exceeding 160 square meters and falling below 45 square meters. Upon reviewing the smallest and largest floor areas for HDB flats in Singapore, which range from 45 to 186 square meters, it is reasonable to exclude entries below 45 square meters and above 186 square meters. This adjustment will help ensure the dataset accurately represents typical HDB flat

sizes.

Furthermore, there are only a total 976 out of 181262 entries (refer to second figure below), strengthening the rationale for their removal.

A screenshot of a computer

Description automatically generated

A screenshot of a computer

Description automatically generated

The figures below show the dataset after removal of these outliers, leaving 180286 rows.

|  |  |
| --- | --- |
| A screenshot of a computer  Description automatically generated  A screenshot of a computer  Description automatically generated | Click below to see code shown on the left. |

Below is the outlier analysis for ‘resale\_price’. The boxplot reveals numerous data points outside the upper bound of Q3 + 1.5 x IQR. Further investigation (see the second figure below) indicates a total of 3858 entries in this category. A check on the maximum transacted price (third figure below) reveals that the highest transaction is at $1588000, which although seemingly high, accurately reflects the current market trend, particularly on the higher end model. As a result, I have decided to retain all these outliers.

|  |  |
| --- | --- |
|  | Click below to see code shown on the left. |

|  |  |
| --- | --- |
|  |  |

* Feature Engineering

The dataset contains several categorical features that need to be converted to numerical data types or one-hot encoding.

For a start, I will convert the ‘remaining\_lease’ into months to ensure it is suitable for machine learning models. The ‘remaining\_lease’ feature is significant as it directly impacts buyer behaviour. Thereafter it will be checked for reasonableness, considering that Singapore’s HDB lease is 99 years or 1188 months.

These other categorical features such as ‘town’, ‘flat\_type’, and ‘storey\_range’, will be dealt with at a later stage.

The figure below shows the conversion of the ‘remaining\_lease’ feature from years-and-months format to a months-only format, resulting in a new feature named ‘lease\_remaining’, and the removal of ‘remaining\_lease’ column. The second figure (refer to figure below it) shows that none of the values in the ‘lease\_remaining’ exceeds 1188 months.

|  |  |
| --- | --- |
|  | Click below to see code shown on the left. |

With the redefinition of ‘remaining\_lease’, the features ‘month’ and ‘lease\_commence\_date’ have become redundant and are removed from the dataset (refer to figure below).

|  |  |
| --- | --- |
|  | Click below to see code shown on the left. |

## Exploratory Data Analysis (EDA)

* Descriptive Statistics

A screenshot of a computer

Description automatically generated

Key Insights

Floor Area:

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 Price:

The resale prices vary significantly, with a mean price of around $498,733.70. The prices are skewed towards higher values, as indicated by the maximum price of $1,588,000.

Lease Remaining:

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.

* Visualization

Graphical visualizations are powerful tools in data analysis, enabling the identification of trends and patterns that support informed decision-making. In this project, however, the primary purpose of the graphs is to validate the data. For instance, we expect that the price will increase as the floor size increases across all towns, as shown below.

|  |  |
| --- | --- |
|  | Click below to see code shown on the left. |

In the next analysis, the histogram (refer to figure below) provides a clear visual representation of the distribution of HDB resale prices across different flat types.

|  |  |
| --- | --- |
|  | Click below to see code shown on the left. |
|  | |

Here are the observations:

Price Distribution

The resale prices for 3 ROOM and 4 ROOM flats are more concentrated in the lower price ranges, while EXECUTIVE and MULTI-GENERATIONAL flats have a wider spread extending into higher price range.

Market Trends

The higher count of 4 ROOM flats in the mid-price range suggests that they are quite popular, possibly to a balance of size and affordability.

Investment Insights

For potential investors or buyers, understanding these distributions can help make informed decisions based on budget and flat type preferences.

Policy Implications

This data could be useful for policymakers to understand housing affordability and demand trends, potentially guiding future housing policies.

From another perspective, when comparing the mean resale prices across various towns by flat types, it becomes evident that certain areas, such as Queenstown, the Central Areas, Bishan, and Ang Mo Kio, consistently have higher prices compared to other towns. This may explain the right-tail distribution of the mean prices across all flat types as shown in the previous graphs. This trend for the mean resale prices across various towns is illustrated in the charts below.

|  |  |
| --- | --- |
|  | Click below to see code shown on the left. |

|  |  |
| --- | --- |
|  |  |

Several factors could contribute to the higher resale prices in towns like Queenstown, the Central Areas, Bishan, and Ang Mo Kio, and within generally accepted expectations.

Location and Accessibility

Proximity to the City Centre: Areas closer to the Central Business District (CBD) and city centre, like Queenstown and the Central Areas, tend to have higher property values due to their prime location.

Transportation Links

These towns often have excellent public transport links, including MRT stations and bus services, making them highly accessible.

Amenities and Facilities

Educational Institutions: Presence of reputable schools and educational institutions can drive up property prices as families seek to live near good schools.

Healthcare Facilities

Proximity to hospitals and clinics adds to the desirability of these areas.

Shopping and Entertainment: Availability of shopping malls, restaurants, and entertainment options enhances the attractiveness of these towns.

Market Demand

High Demand: These towns may have a higher demand due to their desirable attributes, leading to increased competition and higher prices.

Limited Supply

Limited availability of flats in these prime locations tends to drive prices up.

* Correlation Analysis

To prepare for the correlation analysis, the remaining categorical features (‘town’, ‘flat\_type’, and ‘storey\_range’) are converted into numerical data types (refer to the figure below). The ‘town’ feature, which has a larger number of categories (see the second figure below), will additionally be one-hot encoded later in preparation for machine learning. The third and fourth figures below display the converted categorical features.

|  |  |
| --- | --- |
| A screenshot of a computer  Description automatically generated | Click below to see code shown on the left. |

To identify features that are correlated with the resale price, a heat map is generated. The features ‘flat\_type’ and ‘floor\_area\_sqm’ show strong positive correlations with ‘resale\_price’, while ‘lease\_remaining’ and ‘storey\_range’ exhibit moderate positive correlations. Interestingly, the ‘town’ feature has a weak negative correlation index of –0.045.

This is unexpected as graphs of mean resale prices across towns indicate that some towns command a premium. One possible explanation for this could be outliers or anomalies in the data that affect the correlation coefficient. A few towns with extremely high or low prices might distort the overall correlation.

Despite it, this feature will be retained for further analysis. I plan to make adjustments at a later stage to improve the low correlation index observed in the data (refer to ‘Model Adjustments’ under the ‘Feature Importance and Interpretation’ section).

|  |  |
| --- | --- |
|  | Click below to see code shown on the left. |

* One-hot encoding

The steps below (refer to the figure and code) demonstrate the one-hot encoding of the ‘town’ feature and the repositioning of the ‘resale\_price’ column to the end. The dataset now contains thirty-one columns, incorporating all twenty-six towns. The dataset is now prepared for the machine learning phase.

|  |  |
| --- | --- |
|  | Click below to see code shown on the left. |

## Model Development

* Algorithm Selection

In this section, I will explore various machine learning algorithms, in particular the Linear Regression, Decision Trees, and Random Forest, to predict resale prices. The reasons for selecting these algorithms are as follows:

Linear Regression

* Simplicity: It’s a straightforward algorithm that is easy to implement and interpret.
* Efficiency: Works well with smaller datasets and provides a good baseline for comparison with more complex models.
* Interpretability: The coefficients can give insights into the relationship between features and the target variable.

Decision Trees

* Non-linearity: Can capture non-linear relationships between features and the target variable.
* Interpretability: The tree structure is easy to visualize and understand, making it clear how decisions are made.
* Feature Importance: Can provide insights into the importance of different features in predicting the target variable.

Random Forest

* Robustness: Combines multiple decision trees to reduce overfitting and improve generalization.
* Accuracy: Often provides better predictive performance compared to individual decision trees.
* Feature Importance: Aggregates feature importance from multiple trees, giving a more reliable measure of feature significance.

These algorithms offer a good mix of simplicity, interpretability, and predictive power, making them suitable for initial exploration and comparison in this machine learning project.

* Preparation for Model Training

In this process of feature selection for multiple linear regression, the target variable `resale\_price` is isolated from the dataset to serve as the dependent variable, while all other columns are designated as independent variables for the model.

The independent variables, which represent various factors influencing resale prices, are stored in a new DataFrame (`x`) by dropping the `resale\_price` column. Meanwhile, `resale\_price` is set as the output variable (`y`) for the regression analysis.

This setup allows the model to use the input features to predict resale prices in a housing dataset, facilitating the next steps of splitting the data for training and testing, fitting the model, and making predictions (refer to figure below).

|  |  |
| --- | --- |
|  | Click below to see code shown on the left. |

* Model Training

The code below demonstrates the steps involved in training the model. It starts by importing the required libraries and evaluation metrics, then proceeds to instantiate the model using Linear Regression. The model is then fitted to the training data, and finally, the results are generated.

|  |  |
| --- | --- |
|  | Click below to see code shown on the left.    Note: The value may vary with each iteration |

## Model Evaluation

* Performance Metrics

In this section, the performance of this regression model is evaluated to determine its efficacy in predicting the target variable accurately and reliably. To assess the quality of this model, I used three key metrics: R-squared (R²), Mean Absolute Error (MAE), and Root Mean Squared Error (RMSE). By analysing these metrics, we can gauge the accuracy of the predictions, identify potential overfitting, and understand the model's strengths and weaknesses in capturing the underlying data patterns.

Key results

Mean Absolute Error (MAE): $73,883.89

This metric indicates the average magnitude of errors in predictions. In this case, the average difference between the predicted and actual values is around $73,484.42.

R-squared Score (R²): 72%

The R-squared score indicates how well the model captures the variability of the target variable. A score of 72% means that 72% of the total variance in resale price can be explained by the feature variables, suggesting that the model provides a reasonably good fit to the data.

Root Mean Squared Error (RMSE): $91,456.30

RMSE represents the square root of the average of the squared differences between the predicted and actual values. An RMSE of $91,002.31, which is larger than the MAE, indicates that some predictions are significantly farther from the actual values. This suggests that the dataset may have considerable variability, or improvements might be needed, such as exploring different model types to reduce prediction errors.

* Model Comparison

Additionally, I trained the model using Decision Tree Regression and Random Forest algorithms. The goal is to compare and evaluate which model performs best for predicting HDB resale prices.

The models are evaluated using Mean Absolute Error (MAE), R-squared (R²), and Root Mean Squared Error (RMSE) metrics to determine which performs best on the same test set (refer to figures below).

|  |  |
| --- | --- |
|  | Click below to see code shown on the left.    Note: The value may vary with each iteration |

The results from the three iterations show a clear progression in model performance:

Linear Regression Model

The Mean Absolute Error (MAE) is $73,663.38, with an R-squared score of 72% and a Root Mean Squared Error (RMSE) of $91,239.11. This indicates that the model's predictions are relatively far from the actual values, and the R-squared score suggests that about 72% of the variance in the resale prices is explained by the model

.

Decision Tree Regressor Model

There's a noticeable improvement, with the MAE decreasing to $47,623.40 and the R-squared score rising to 83%. The RMSE also improves to $69,734.81. This suggests that the model is better at capturing the underlying patterns in the data, resulting in more accurate predictions.

Random Forest Regressor Model

The results continue to improve, with the MAE further reduced to $40,324.28 and the R-squared score increasing to 89%. The RMSE decreases to $57,064.70, indicating that the model has become significantly more accurate and is able to explain a larger portion of the variance in resale prices.

Overall, the consistent decline in MAE and RMSE, along with the increase in R-squared scores, demonstrates effective model enhancement across the tests, suggesting that the techniques or parameters being adjusted are yielding positive results.

## Feature Importance and Interpretation

* Feature Importance

Based on the correlation analysis, ‘flat\_type’ and ‘floor\_area\_sqm’ appear to be the most important feature for predicting resale prices due to its strong positive correlation.

The features ‘storey\_range’ and ‘lease\_remaining’ also contribute to predicting resale prices but to a lesser extent.

* Implications for Model Building

When building predictive models, priorities should be given to the ‘flat\_type’ and ‘floor\_area\_sqm’ because to the strong correlation with the target variable, ‘resale\_price’. Features with weak correlations that might exhibit less predictive power could be considered for exclusion.

## Model Adjustments

Given the unexpectedly low correlation between ‘town’ and ‘resale\_price’ shown in the heatmap, which contrasts with findings that towns closer to amenities tend to have higher resale prices, I will now attempt to remove all the outliers that were retained in the earlier analysis.

Analysing Outliers by Flat Types

This analysis aims to identify the onset of outliers within the dataset categorized by ‘flat\_type’. Once identified, the records containing these outliers will be removed. This adjustment is intended to evaluate whether the model’s performance can be enhanced by excluding these extreme values.

|  |  |
| --- | --- |
|  | Click below to see code shown on the left. |

The boxplot figures above illustrate the outliers for each flat type, marking the initial data points identified for removal. The subsequent figure displays the filtered dataset, with all records containing outliers excluded. As a result, the dataset now comprises 173,851 rows, down from the original 180,286 rows. The figures below show the boxplot and the reconstructed histogram with the removed outliers.

|  |  |
| --- | --- |
|  | Click below to see code shown on the left. |

A quick comparison of the filtered versus unfiltered data (see figures below) reveals that a significant amount of information has been removed in the filtered dataset. Notably, floor levels from the 43rd storey onwards are missing. This omission does not accurately represent the real situation.

|  |  |
| --- | --- |
|  |  |

The figure below illustrates the correlation after adjusting for outliers. The correlation indices between flat\_type, floor\_area\_sqm, storey\_range, and lease\_remaining in relation to resale\_price have all increased slightly. However, the correlation between town and resale\_price has weakened. While this aligns with expectations, removing the town feature is not realistic, as this feature should impact resale prices.

|  |  |
| --- | --- |
|  | Click below to see code shown on the left. |

The categorical data in the adjusted dataset was subsequently converted to numerical data types, and the model retrained.

Here are the results (refer to figures below) of the MAE, R², and RMSE scores after iterating with Linear Regression, Decision Tree Regression, and Random Forest models:

|  |
| --- |
| * **Linear Regression**:   MAE: $ 70484.87  R²: 70%  RMSE: $ 85795.73   * **Decision Tree Regression**:   MAE: $ 46558.21  R²: 81%  RMSE: $ 67678.2   * **Random Forest**:   MAE: $ 39457.82  R²: 87%  RMSE: $ 55631.28 |

|  |  |
| --- | --- |
|  | Click below to see code shown on the left.    Note: The value may vary with each iteration |

When comparing the results before and after removing outliers (refer to figures below), improvements in the MAE and RMSE scores are evident. However, the R² scores have decreased slightly by 1 to 2%. The decision to retain or remove outliers is debatable, but I recommend keeping the outliers to ensure the results remain realistic and to capture the real situation in the market.

|  |  |
| --- | --- |
| **Before Adjustment** | **After Adjustment** |
| * **Linear Regression**:   MAE: $ 74448.84  R²: 71%  RMSE: $ 91973.83   * **Decision Tree Regression**:   MAE: $ 47785.42  R²: 83%  RMSE: $ 69962.65   * **Random Forest**:   MAE: $ 40828.95  R²: 89%  RMSE: $ 57943.24 | * **Linear Regression**:   MAE: $ 70484.87  R²: 70%  RMSE: $ 85795.73   * **Decision Tree Regression**:   MAE: $ 46558.21  R²: 81%  RMSE: $ 67678.2   * **Random Forest**:   MAE: $ 39457.82  R²: 87%  RMSE: $ 55631.28 |

# Stakeholders

* Homebuyers and Sellers

Individuals looking to buy or sell HDB flats can use the model to estimate fair prices.

* Real Estate Agents

Professionals can leverage the insights to provide better advice to their clients.

* Policymakers

Government bodies can use the analysis to understand market trends and implement effective housing policies.

* Financial Institutions

Banks and lenders can assess the value of properties more accurately for mortgage purposes.

* Researchers and Academics

Scholars can use the findings for further research in real estate economics and urban planning.

# Benefits

* Transparency

Enhances transparency in the HDB resale market by providing data-driven price estimates.

* Informed Decisions

Empowers stakeholders with accurate information to make better decisions.

* Market Stability

Helps in stabilizing the market by reducing price speculation and volatility.

* Policy Formulation

Assists policymakers in crafting data-driven housing policies to address market needs.

# Use Cases

The predictive model for HDB resale prices can be adapted for various applications in the real estate and urban planning sectors. Here are some potential applications:

* Rental Price Prediction

Predict rental prices for HDB flats based on similar features used for resale price prediction. This may help tenants and landlords set fair rental prices and understand market trends.

* Property Valuation for Insurance

Provide accurate property valuations for insurance purposes. It ensures that properties are adequately insured, reflecting their true market value.

* Urban Planning and Development

Assist urban planners in identifying areas with high growth potential and planning infrastructure development accordingly. This will support strategic development and efficient allocation of resources.

* Mortgage Risk Assessment

Help financial institutions assess the risk associated with mortgage lending by predicting future property values. It reduces the risk of loan defaults and improves the accuracy of loan-to-value ratios.

* Investment Analysis

Aid real estate investors in identifying lucrative investment opportunities by predicting future price trends. It enhances investment decision-making and portfolio management.

* Policy Formulation

Provide data-driven insights to policymakers for crafting housing policies and regulations. It ensures policies are aligned with market realities and address housing affordability and availability.

* Community and Social Services Planning

Assist in planning community services and amenities based on predicted population growth and housing demand. It may help improve the quality of life by ensuring adequate provision of services and facilities.

* Environmental Impact Assessment

Evaluate the environmental impact of new housing developments by predicting changes in property values and population density. It helps support sustainable development and environmental conservation efforts.

* Real Estate Market Analysis

Provide comprehensive market analysis reports for real estate agencies and developers. It helps enhance market understanding and strategic planning for new projects.

* Customized Real Estate Solutions

Develop personalized real estate solutions for clients based on their preferences and predicted market trends. It improves customer satisfaction and service quality.

# Ethics and Governance

Ensuring ethical practices and robust governance is essential when creating and implementing predictive models, particularly in sensitive domains such as housing.

## Ethical Considerations

* Bias and Fairness

Historical data may contain biases that can be perpetuated by the model, leading to unfair predictions. Implementing fairness-aware algorithms and regularly audit the model for biases would ensure diverse and representative data to mitigate bias.

* Transparency and Explainability

Complex models can be difficult to interpret, making it hard for stakeholders to understand how predictions are made. Using interpretable models or techniques to explain model predictions would provide clear documentation and visualizations to enhance transparency.

* Privacy and Data Protection

Handling personal and sensitive data can lead to privacy concerns. Anonymizing data and comply with data protection regulations like the Personal Data Protection Act (PDPA) in Singapore would ensure secure data storage and access controls.

* Informed Consent

Using data without the knowledge or consent of individuals can be unethical. Obtaining informed consent from data subjects and be transparent about how their data could be used.

* Impact on Stakeholders

Predictions can influence market behaviour and individual decisions, potentially leading to unintended consequences. Conducting impact assessments to understand the potential effects on different stakeholders and implement measures could mitigate negative impacts.

## Governance

* Regulatory Compliance

Ensure the project complies with relevant laws and regulations, such as housing policies and data protection laws. This can be done with regular reviews of legal requirements and consultation with legal experts to ensure compliance.

* Ethical Guidelines

Develop and adhere to a set of ethical guidelines for the project. This can be done by establishing a code of ethics that outlines principles such as fairness, transparency, and accountability. Also to ensure all team members are aware of and follow these guidelines.

* Stakeholder Engagement

Engage with stakeholders throughout the project to understand their needs and concerns. This can be done by conducting regular consultations and feedback sessions with stakeholders, including homebuyers, sellers, policymakers, and community representatives.

* Accountability and Oversight

Establish mechanisms for accountability and oversight to ensure ethical conduct. This can be done by forming an ethics review board or committee to oversee the project and address any ethical issues that arise. Implement regular audits and reviews of the model and its outcomes.

* Continuous Monitoring and Evaluation

Continuously monitor and evaluate the model to ensure it remains fair and effective. Implement ongoing monitoring processes to track the model’s performance and impact. Adjust as needed to address any issues.

Addressing these ethical considerations and governance aspects would ensure that the project is conducted responsibly and ethically, benefiting all stakeholders involved.

# Conclusion

The HDB Resale Prediction Project has successfully demonstrated the potential of machine learning (ML) and artificial intelligence (AI) in accurately forecasting HDB 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.

## Key Findings

* Predictive Performance and 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 (R²) of 89% and a root mean square error (RMSE) of $ 57943.24.

* Feature Importance

Flat type and floor area emerged as the most significant predictors, highlighting the importance of flat type and size in determining resale values.

* Biases

Although not examined in this project, the dataset may contain historical biases, such as location preferences or economic conditions that have evolved over time.

Also, some machine learning algorithms might inherently favour certain features over others, leading to biased predictions. This is shown in the model iteration exercise earlier.

## Implications

* Policy Making

The insights gained from this project can inform policymakers in designing housing policies that are data-driven and equitable.

* Market Transparency

Enhanced price prediction models can increase market transparency, helping buyers and sellers make more informed decisions.

* Ethical Considerations

It’s essential to address ethical concerns such as data privacy, bias, and transparency in applications.

Data privacy is a critical ethical issue because it involves safeguarding individuals’ personal information from unauthorized access and misuse. Protecting data privacy helps maintain trust, respects individuals’ rights, and prevents potential harm from data breaches or misuse.

Implementing strategies like anonymizing data by removing identifiable personal information, using aggregated data instead of detailed individual data, and conducting regular audits can significantly enhance the privacy, fairness, and transparency of the dataset while still allowing for valuable insights and analysis.

## Future Directions

* Model Improvement

Future work could focus on incorporating additional features, such as economic indicators and environmental factors, to further enhance predictive accuracy.

* Ethical Frameworks

Developing and implementing robust ethical frameworks to govern the use of AI in real estate is essential to ensure fairness and accountability.

* Stakeholder Engagement

Engaging with a broader range of stakeholders, including residents, policymakers, and ethicists, will be crucial in refining the model and its applications.

In conclusion, the HDB Resale Prediction Project has successfully demonstrated the potential of machine learning (ML) and artificial intelligence (AI) in accurately forecasting HDB 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.

To effectively address ethical and governance challenges, it is crucial to maintain ongoing efforts. By continuously improving both the model and its ethical framework, we can ensure that AI-driven insights have a positive impact on the housing market and the user community.