**IT2301 Supervised**

**Learning**

**Final Project Report**

|  |  |
| --- | --- |
| **Submitted By:** | Jay Krish Vijendra (222197Y) |
| **Lecturer:** | Mr Pang Nai Kiat |
| **Deadline:** | 13/August/2023 |

[**Data Cleaning and Data Transformation** 2](#_Toc142862145)

[Import Libraries and Data 2](#_Toc142862146)

[Data Understanding 2](#_Toc142862147)

[Data Cleaning 3](#_Toc142862148)

[Exporting Data 4](#_Toc142862149)

[**Modelling** 5](#_Toc142862150)

[Overview of Models 5](#_Toc142862151)

[Characteristics of Resale Price 5](#_Toc142862152)

[Linear Regression 5](#_Toc142862153)

[Decision Tree 6](#_Toc142862154)

[Random Forest 6](#_Toc142862155)

[Model Comparison 6](#_Toc142862156)

[Evaluation for Linear Regression Model 7](#_Toc142862157)

[Options 7](#_Toc142862158)

[Roles 7](#_Toc142862159)

[Visualisations 8](#_Toc142862160)

[**Conclusion** 10](#_Toc142862161)

# **Data Cleaning and Data Transformation**

The subsequent sections delineate the meticulous process of data cleansing and transformation applied to the "Resale.csv" dataset in Python, in preparation for its utilization in the subsequent modelling using SAS Viya.

## Import Libraries and Data

The initial phase involved importing the requisite libraries essential for data manipulation and analysis. Subsequently, the dataset was loaded into a data frame. The first five rows of the dataset were then retrieved to gain an initial insight into its structure and content.

A screenshot of a computer

Description automatically generated

## Data Understanding

The next phase involved going further into extensive dataset examination. Key information, including the number of columns and entries, was gathered by doing a thorough analysis of the data frame. The dataset had a large corpus of 183,905 data elements and 18 unique columns.A black screen with a black border

Description automatically generated

Null value assessment was performed, revealing no instances of null values within the dataset.A black screen with a black border

Description automatically generated

## Data Cleaning

A black screen with colorful lights

Description automatically generated

Feature Removal: Certain columns were excluded from the analysis due to their limited relevance. Columns like 'Full Address', 'Latitude', 'Longitude', and others were deemed unnecessary for modelling purposes.

A screenshot of a computer

Description automatically generated

Outlier Handling: An outlier was identified within the 'Storey Range' column with a value of 103.7965865. This outlier was removed to ensure data integrity.

A black screen with green and yellow text

Description automatically generated

Unit Age Computation: A new variable named 'Unit\_Age' was derived, capturing the temporal gap between 'Year-Month Sold' and 'Lease Commence Date'.

A screenshot of a computer

Description automatically generated

Unit Sanitizing: Unit\_Age Column there was a property from Queenstown that has a negative value. It meant that the property was sold before it was leased which was not possible so its row was removed.

## Exporting Data

A screenshot of a computer screen

Description automatically generated

After refining and enhancing the dataset through data cleaning and transformation, the next step is to save the improved data for future use. This is accomplished by exporting the cleaned Data Frame into a new CSV file named "clean\_Resale.csv".

# **Modelling**

## Overview of Models

### Characteristics of Resale Price

A screenshot of a graph

Description automatically generated

As Seen above, the Floor Area Sqm is the most related to the Target Variable (Resale Price) with it having a positive relationship that if the Floor Area Sqm increase the Resale Price Increases.

### Linear Regression

A screenshot of a computer

Description automatically generated

### Decision Tree

A close-up of a graph

Description automatically generated

### Random Forest

A screenshot of a computer

Description automatically generated

### Model Comparison

A close-up of a graph

Description automatically generated

After the model comparison, Linear Regression emerged as the favoured choice due to its lowest Average Squared Error (ASE). This quantitative metric underscores its superior predictive accuracy, reflecting alignment with underlying data patterns and informed decision-making.

## Evaluation for Linear Regression Model

### Options

A screenshot of a computer

Description automatically generated

### Roles

A screenshot of a computer

Description automatically generated

### Visualisations

A screenshot of a computer

Description automatically generated

Under the Fit Summary Pane all of the above columns are statistically significant and they are‘Flat Model , Floor Area Sqm , Hawker Dist , Lease Commence Date , Mrt Dist , Storey Range , Town , Unit\_Age , Flat Type , School Dist’. All of the mentioned columns have a p-value which is less then P < 0.00001.

An R-squared value of 0.8753 means that the linear regression model explains approximately 87.53% of the variability in the Resale Price based on the provided features. In other words, the model is reasonably good at capturing the variation in Resale Price using the given features. Keep in mind that R-squared alone doesn't tell you about the model's predictive ability outside the dataset it was trained on. It's important to consider other metrics and validation techniques as well.

A screen shot of a graph

Description automatically generated

A screen shot of a computer

Description automatically generated

**Scattering Around 0:**

The central tendency of residuals being close to 0 signifies that the model is making predictions that, on average, are devoid of bias. This observation implies that the model effectively captures the linear association between the provided features and the Resale Price.

**Consistent Variance:**

The random dispersion of residuals around the 0 line suggests that the variability of residuals remains relatively constant across various levels of predicted values. This outcome is significant as it confirms the fulfilment of the assumption of constant variance, or homoscedasticity, within the model.

**Residual Range and Context:**

Given that the Resale Price values predominantly fall within the range of 0.3 to 0.4 million and the residuals align closely to 0, there's an indication that the model is exhibiting favourable performance within this specific range. Nevertheless, it is essential to emphasize that while the residuals display this behaviour around 0, it is imperative to validate the model's predictive accuracy across the entire spectrum of Resale Prices. This holistic evaluation ensures comprehensive model reliability.

A graph of a person

Description automatically generated with medium confidence

**Consistent Alignment of Points**:

The observation that the points on the plot align consistently for both the training and validation data is a strong indication of the model's reliability. The close proximity of predicted and actual values underscores the model's ability to capture the underlying relationships within the data.

**Cohesiveness and Reliability**:

By highlighting the cohesiveness of predicted Resale Prices and actual values, you effectively convey the model's consistent performance across diverse subsets. This cohesiveness reflects the reliability of the model's predictions, which is a key factor in ensuring accurate outcomes.

**Accuracy Across Percentiles**:

Exploring the model's accuracy across different segments of the data distribution through percentiles is a valuable insight. The ability of the model to maintain accuracy regardless of where data points fall within the distribution showcases its robustness and adaptability.

**Conclusion**

To summarize, the model's efficacy in forecasting Resale Price using the given variables is notably accurate and dependable. Rigorous analysis and comparison underscored the model's predictions closely mirroring actual Resale Price values. This underscores the model's competence in capturing nuanced associations between input features and the target variable. By minimizing forecast errors and maintaining consistency across diverse data subsets, the model reinforces its precision in estimating Resale Price. This predictive proficiency holds practical relevance, potentially informing decisions pertinent to housing market assessments and investment strategies.