Project Report: Singapore Flat Resale Price Prediction

Introduction:

The Singapore Resale Flat Resale Price Prediction project focuses on analyzing a dataset from the Singapore Housing and Development Board (HDB). The primary objective is to predict the resale prices of properties based on various features. The project involves technologies such as Streamlit, Pandas, and popular machine learning libraries for model development.

Technologies Used:

1. Streamlit: Used for creating interactive and user-friendly web applications.

2. Pandas: Utilized for data manipulation, analysis, and exploration.

3. Data Wrangling: Applied techniques to clean and preprocess the dataset.

4. Machine Learning: Employed regression algorithms for predicting resale prices.

5. EDA (Exploratory Data Analysis): Conducted exploratory data analysis to gain insights into the dataset.

Domain:

The project falls under the domain of Real Estate.

Project Overview:

1. Data Collection:

Overview:

The dataset used in this project was sourced from the Singapore Housing and Development Board (HDB). It comprises information on resale flat prices and associated attributes such as flat type, block details, street names, floor area, flat model, lease commencement date, and more.

Steps:

- Imported necessary libraries, including Pandas for data manipulation and exploration.

- Read multiple CSV files corresponding to different periods and merged them to create a comprehensive dataset.

- The dataset spans various years, capturing changes and trends in the resale flat market over time.

2. Data Preprocessing and Wrangling:

Overview:

Data preprocessing is a crucial step to ensure that the dataset is clean, consistent, and suitable for machine learning model development. This phase involves handling missing values, converting categorical variables into numerical formats, performing feature engineering and feature selection.

Steps:

- Handled missing values by dropping columns with significant null values. Dropped ‘remaining_lease’ column because it contained 7,09,050 null values.

- Applied feature engineering to extract relevant information from columns such as 'month' and 'storey_range.'

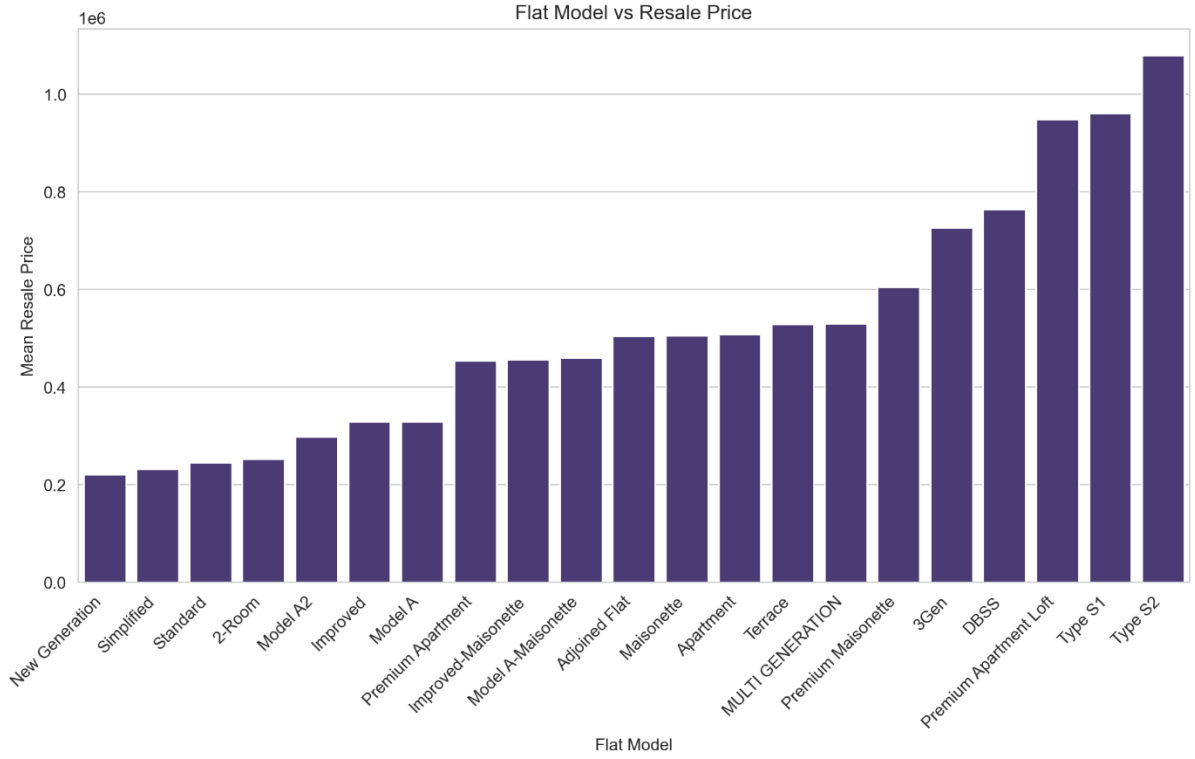
- Converted categorical variables into numerical representations using techniques like one-hot encoding and binary encoding.

- Conducted exploratory data analysis (EDA) to visualize data distributions and relationships.

The below figure shows the Year-wise Trend of Resale Prices:

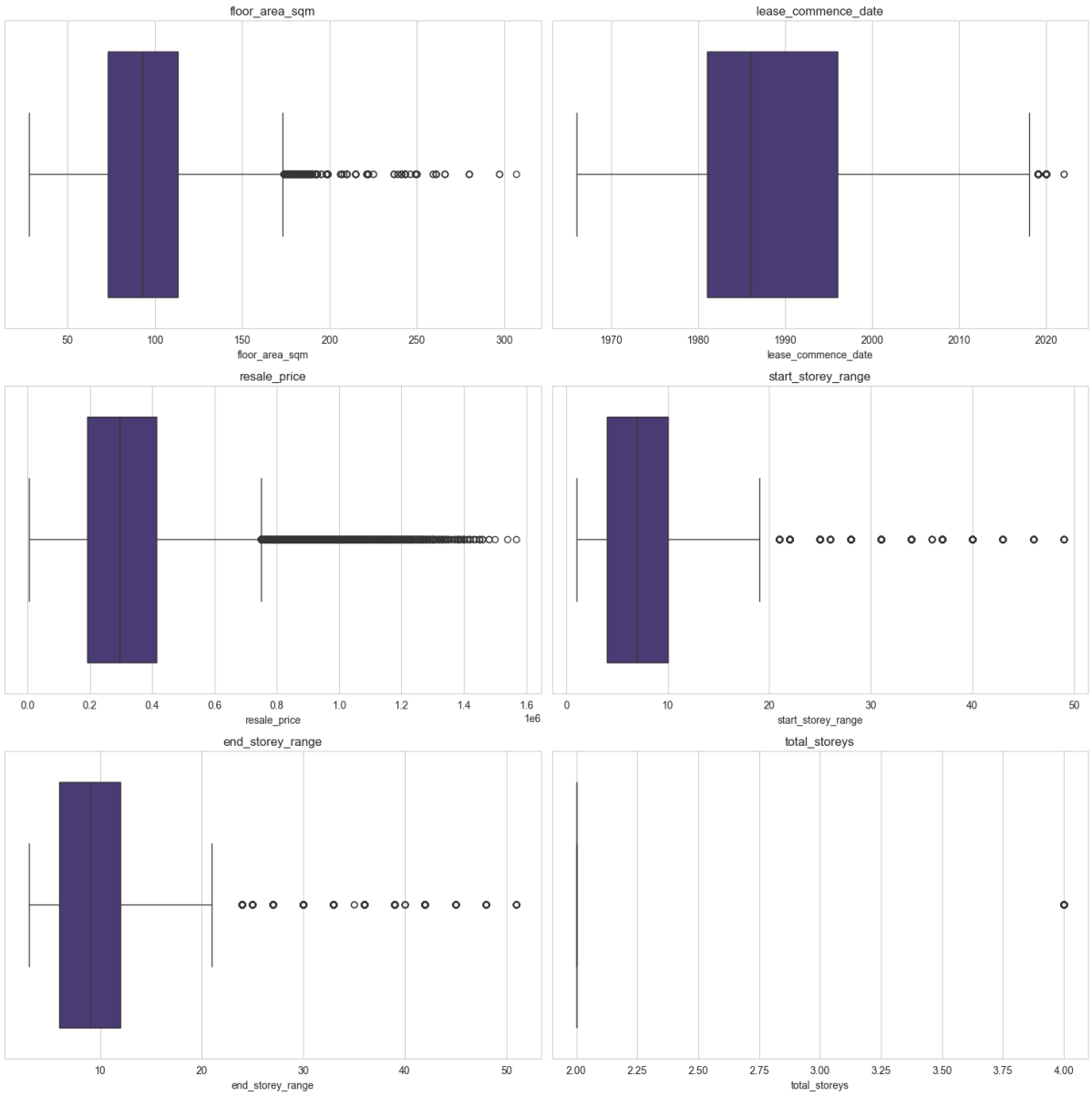
There is a steady increase in the Resale price from 1990 to 2023.

The below figure shows the Resale price in different Flat model types:



The 3Gen, DBSS, Premium Apartment Loft, Type S1, and Type S2 have the highest Average resale price among other flat model types.

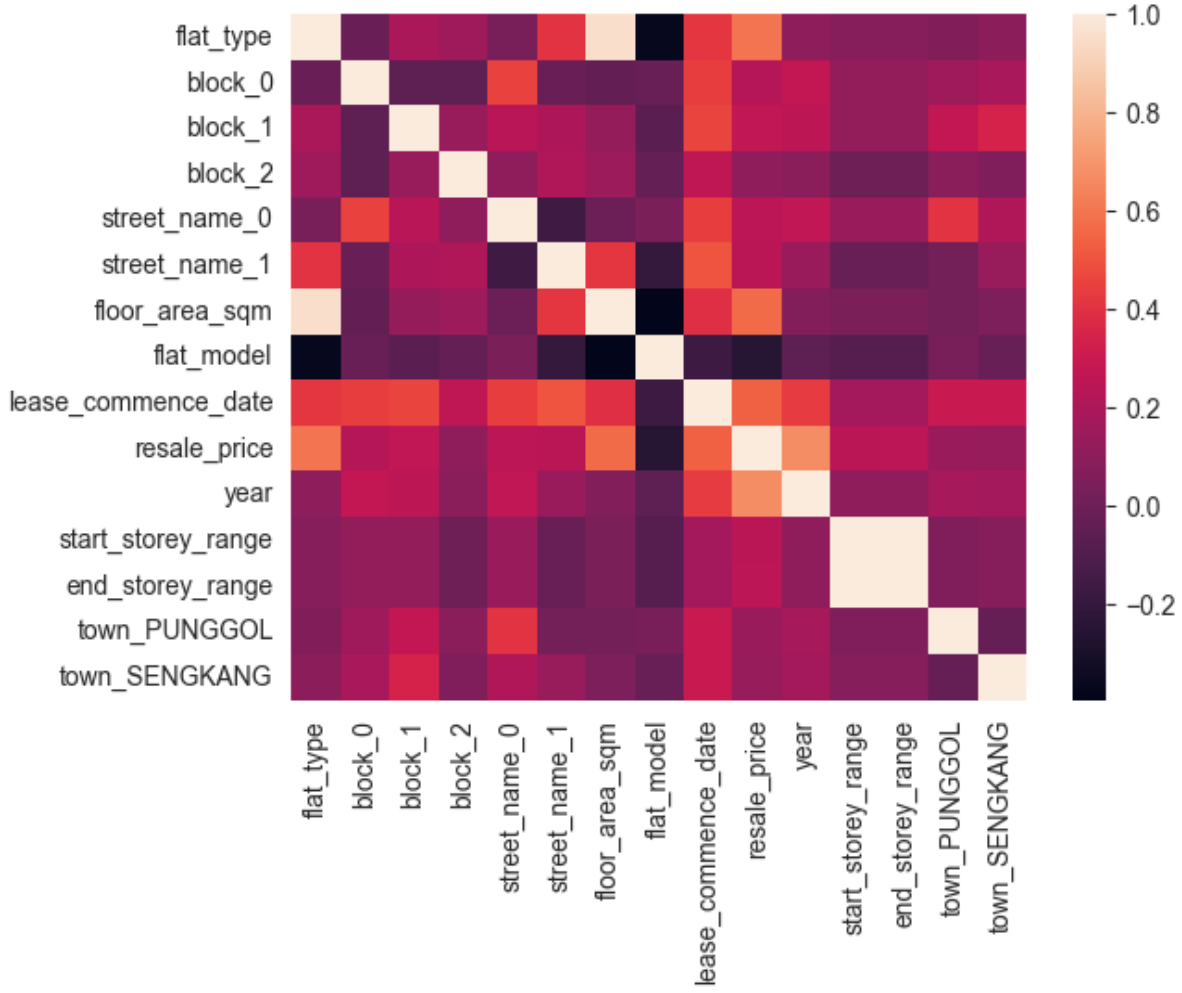
- Used Box plots to detect outliers in the dataset. Below figure shows the presence of outliers in different columns:



As there are large values present in the dataset, the box plot identified them as outliers, However, in the Real Estate industry, there are few properties with high resale price because it would be a huge house or apartment. So, I considered them as valid outliers and proceeded with them for model development.

- Performed Feature selection to select the most relevant features using correlation technique.

Used heatmap to determine the strength of correlation between each feature and removed the features with no correlation to the target variable.



After performing all the transformation steps, the cleaned dataset is stored in a CSV file.

3. Model Development:

Overview:

Model development involves choosing suitable machine learning algorithms, training the models, and fine-tuning them to achieve optimal performance. For this project, regression models were employed to predict resale prices based on property features.

Steps:

- Selected regression algorithms, including Elastic Net, Random Forest, Decision Tree, Gradient Boosting, LightGBM, and XGBoost.

- Split the dataset into training and testing sets for model evaluation.

- Utilized GridSearchCV to fine-tune hyperparameters and enhance model performance.

- Evaluated models using metrics such as Mean Absolute Error, Mean Squared Error, Root Mean Squared Error, and R-squared.

Out of all the models, XGBoost Regressor performed better in terms of overall metrics such as MAE, MSE, RMSE, and R2 score.

Metrics:

```
Mean Absolute Error (MAE):  31010.584877603516
Mean Squared Error (MSE):  2091790605.4283772
Root Mean Squared Error (RMSE):  45736.09740050387
R-squared (R2):  0.925205906901622
```

4. Model Deployment:

Overview:

Model deployment involves making the machine learning model accessible to end-users. In this project, a Streamlit web application was developed to create an interactive interface for users to input property details and obtain resale price predictions.

Steps:

- Imported the necessary libraries, including Streamlit, for web application development.

- Created input fields in the Streamlit app for users to input relevant property details.

- Integrated the trained regression model into the Streamlit app.

- Users can interact with the app, input property details, and receive predicted resale prices.

5. Conclusion:

Overview:

The conclusion summarizes the key findings and outcomes of the project, highlighting the success of combining data analysis, machine learning, and web application development.

Key Points:

- Successfully implemented a regression model for predicting resale flat prices.

- Streamlit web application hosted on Render provides a user-friendly interface for model interaction.

- Project contributes to the real estate domain by offering a practical tool for price prediction.

6. Findings and Future Work:

Overview:

This section discusses the insights gained from the project and outlines potential areas for future enhancements.

Key Points:

- Identified that XGBoost performed the best among the regression models.

- Future work may include incorporating additional features, further optimizing the model, and expanding the dataset for more robust predictions.

The project seamlessly integrates data analysis, machine learning, and web development, providing a comprehensive solution for predicting resale flat prices in Singapore.