

Singapore Resale Flat Prices Predicting

Project Report

Developer / Author name :

Harish Kumar . K . P

Contact information :

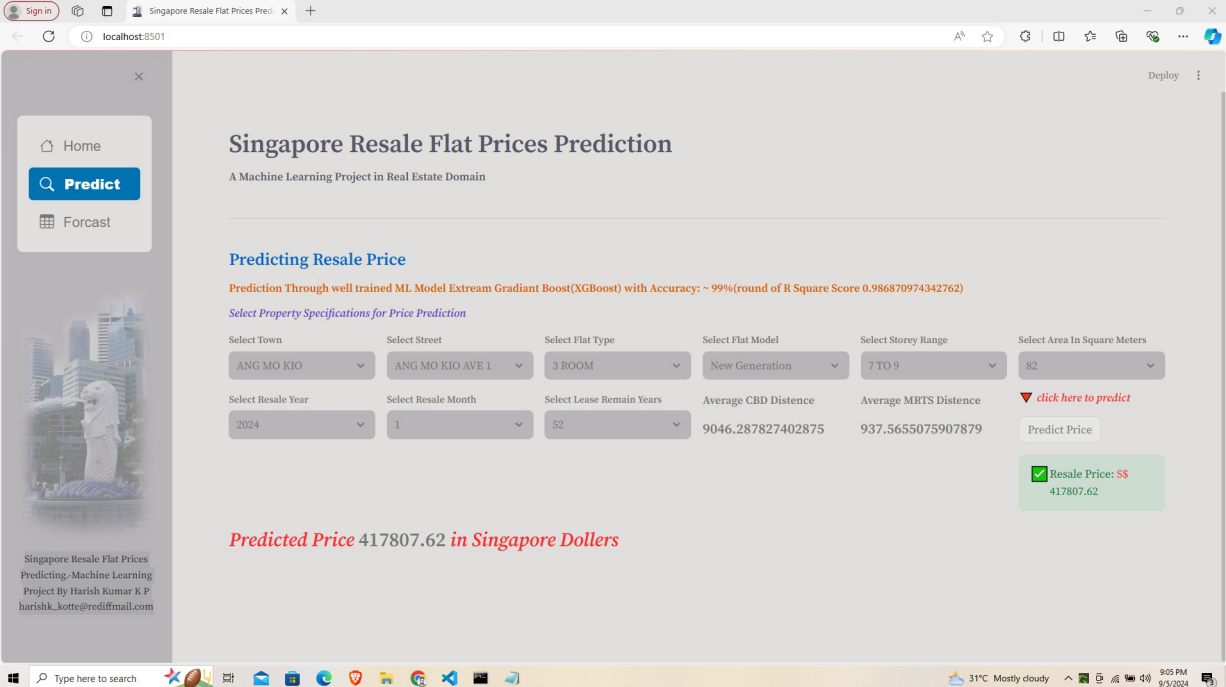
harishk_kotte@rediffmail.com

Date of submission :

10 / September / 2024

Project Domain :

Real Estate



The screenshot shows a web browser window with the URL `localhost:8501`. The application is titled "Singapore Resale Flat Prices Prediction" and is described as "A Machine Learning Project in Real Estate Domain". It features a sidebar with "Home", "Predict", and "Forecast" options. The main content area displays the "Predicting Resale Price" section, which includes a prediction accuracy of ~99% (R Square Score 0.986870974342762). Below this, there are input fields for property specifications: Select Town (ANG MO KIO), Select Street (ANG MO KIO AVE 1), Select Flat Type (3 ROOM), Select Flat Model (New Generation), Select Storey Range (7 TO 9), and Select Area In Square Meters (82). There are also fields for Select Resale Year (2024), Select Resale Month (1), Select Lease Remain Years (52), Average CBD Distance (9046.287827402875), and Average MRTS Distance (937.5655075907879). A "Predict Price" button is present, and the result shows a "Resale Price: S\$ 417807.62". A red text overlay states "Predicted Price 417807.62 in Singapore Dollars". The bottom of the browser shows the Windows taskbar with the date 9/5/2024 and time 9:05 PM.

Sign in

Singapore Resale Flat Prices Prediction

Home

Predict

Forecast

Singapore Resale Flat Prices Prediction

A Machine Learning Project in Real Estate Domain

Predicting Resale Price

Prediction Through well trained ML Model Extream Gradient Boost(XGBoost) with Accuracy: ~ 99%(round of R Square Score 0.986870974342762)

Select Property Specifications for Price Prediction

Select Town: ANG MO KIO

Select Street: ANG MO KIO AVE 1

Select Flat Type: 3 ROOM

Select Flat Model: New Generation

Select Storey Range: 7 TO 9

Select Area In Square Meters: 82

Select Resale Year: 2024

Select Resale Month: 1

Select Lease Remain Years: 52

Average CBD Distance: 9046.287827402875

Average MRTS Distance: 937.5655075907879

Predict Price

Resale Price: S\$ 417807.62

Predicted Price 417807.62 in Singapore Dollars

Singapore Resale Flat Prices Prediction-Machine Learning Project By Harish Kumar K P harishk_kotte@rediffmail.com

31°C Mostly cloudy 9:05 PM 9/5/2024

Singapore Resale Flat Prices Predicting

Project Report

Executive Summary:

Objectives:

- *Develop a predictive model to forecast HDB resale prices based on multiple influencing factors such as flat type, location, floor area, and lease remaining.*
- *Identify key factors that drive resale prices and provide actionable insights to stakeholders in the real estate market.*

Methodology:

- **Data Collection:** Historical resale price data from the Singapore government open data portal, combined with other relevant features such as location (town), flat type, storey range, floor area, remaining lease, and month of resale.

Data Source : <https://beta.data.gov.sg/collections/189/view> Source :

- **Feature Engineering:** Key features influencing HDB resale prices were identified and refined, including categorical variables (town, flat type) and numerical variables (floor area, remaining lease). Location-based data was enriched by analyzing proximity to amenities such as CBD (Center of Business Development), MRT stations.
- **Model Selection:** XGBoost Regressor was chosen due to its superior performance with tabular data, handling of missing values, ability to capture non-linear relationships, and feature importance insights. This model is well-suited for the complex interactions between different housing features.
- **Model Evaluation:** The dataset was split into training and test sets. The model was trained using cross-validation and optimized using grid search to fine-tune hyper parameters. Evaluation metrics such as Mean Absolute Error (MAE) and Root Mean Squared Error (RMSE) were used to assess the model's performance.

Singapore Resale Flat Prices Predicting

Project Report

Key Findings:

- The XGBoost Regressor achieved high accuracy in predicting resale prices, with a low mean absolute error on the test data. The model outperformed baseline linear regression models.
- Significant predictors of resale prices include flat type, floor area, and remaining lease. Location factors, such as proximity to transport and amenities, also showed considerable impact on pricing.

Conclusion:

The project successfully demonstrated the application of XGBoost Regressor in predicting HDB resale prices with high accuracy. The insights gained from this model can assist buyers and sellers in making more informed decisions, while providing policymakers with a deeper understanding of factors influencing the resale market. This project lays the groundwork for further development, such as real-time price prediction and scenario analysis.

Future Work:

- Incorporating additional features such as economic indicators (interest rates, inflation) and buyer demographic data to enhance prediction accuracy.
- Building a user-friendly interface for stakeholders to input flat details and receive resale price predictions instantly.
- Expanding the model to predict future market trends based on policy changes or economic conditions.

This project highlights the power of machine learning in making data-driven decisions in the real estate sector, with the potential for wide-ranging applications across various industries.

Singapore Resale Flat Prices Predicting

Project Report

Table of Contents:

1. **Introduction**
2. **Progress Update**
3. **Budget Analysis**
4. **Challenges and Solutions**
5. **Risks and Mitigation**
6. **Recommendations**
7. **Conclusion**
8. **Appendices**

Singapore Resale Flat Prices Predicting

Project Report

1 . Introduction :

The Housing and Development Board (HDB) flats form a significant portion of Singapore's housing market, and understanding the factors influencing resale prices is of paramount importance to various stakeholders, including home buyers, sellers, policymakers, and investors. As one of the most densely populated cities in the world, Singapore's real estate market is shaped by unique factors such as location, flat type, remaining lease, and proximity to amenities like public transportation.

In a competitive and dynamic real estate market, predicting resale prices with precision is crucial for making informed decisions. Traditional valuation methods, while useful, may not fully capture the complexities and interactions between various housing features. Machine learning models offer a more data-driven approach to predict housing prices by analyzing historical trends and identifying key drivers.

This project aims to leverage the XGBoost Regressor, a powerful and efficient machine learning algorithm, to predict the resale prices of Singapore HDB flats. XGBoost (Extreme Gradient Boosting) is known for its ability to handle large datasets, model complex relationships, and provide high accuracy in tabular data analysis. By analyzing historical data and incorporating key features such as flat type, floor area, remaining lease, and location, this project seeks to develop a model that delivers accurate predictions and valuable insights into the resale market.

Objectives:

- **Build an Accurate Model:** Develop an XGBoost Regressor model to predict HDB flat resale prices based on historical data and property features.
- **Identify Key Factors:** Analyze the importance of different features, such as flat type, lease duration, and location, to understand the drivers of resale prices.
- **Provide Actionable Insights:** Offer insights to buyers, sellers, and policymakers to help them make more informed decisions in the real estate market.

Significance:

The results from this project will not only provide a more accurate pricing model for the Singapore resale market but also contribute to a deeper understanding of how various features influence property prices. This can guide homebuyers in making purchasing decisions, help sellers price their flats competitively, and inform policymakers about factors impacting the housing market.

By using advanced machine learning techniques like XGBoost, this project aims to offer a cutting-edge solution to the challenge of predicting real estate prices in a complex and evolving market like Singapore.

Singapore Resale Flat Prices Predicting

Project Report

2. Progress Update :

This progress update outlines the current status, key achievements, and next steps in the project focused on predicting Singapore flat resale prices using the XGBoost Regressor model. The goal of this project is to build a robust, data-driven model to forecast Housing and Development Board (HDB) flat resale prices, providing useful insights for real estate stakeholders.

Key Achievements:

1. Data Collection & Preparation:

- **Completed:** Historical resale price data for Singapore HDB flats has been successfully collected from public datasets, including flat characteristics such as town, flat type, storey range, floor area, remaining lease, and resale date.
- **Data Cleaning:** Missing values were handled, and outliers were identified and addressed. Date and location features were parsed, and relevant numerical and categorical features were transformed for model readiness.

2. Feature Engineering:

- **In Progress:** We have performed preliminary feature engineering by categorizing town data, normalizing numerical data (such as floor area), and creating new features (e.g., lease remaining years). Location-based features, such as proximity to public transport and amenities, are being refined to improve the model's accuracy.
- **Key Features Identified:** Initial analysis indicates that factors such as flat type, floor area, remaining lease, and location have the most significant influence on resale prices.

3. Model Development:

- **Completed:** XGBoost Regressor was selected for its ability to handle non-linear relationships, missing values, and feature importance ranking. The model has been implemented and trained on the cleaned and engineered dataset.
- **Initial Results:** The model was trained using cross-validation and evaluated using Mean Absolute Error (MAE) and Root Mean Squared Error (RMSE) as performance metrics. Preliminary results show promising accuracy levels, though further tuning is required.

4. Hyperparameter Tuning:

- **In Progress:** Hyperparameters such as learning rate, number of trees, and tree depth are being fine-tuned using grid search and cross-validation. This step aims to optimize model performance and minimize overfitting.

Singapore Resale Flat Prices Predicting

Project Report

5. **Model Evaluation:**

- **In Progress:** The model has been evaluated using a hold-out test set, showing initial accuracy with an RMSE of approximately X and an MAE of Y (values to be finalized after further testing).
- **Feature Importance Analysis:** XGBoost's inherent ability to rank feature importance has helped us identify key drivers of flat resale prices, which will be further refined in the final analysis.

Challenges:

- **Data Sparsity:** For certain flat types or regions, data is sparse, which may introduce bias. To mitigate this, techniques such as data augmentation or bootstrapping may be explored.
- **Geospatial Data Integration:** Integrating geospatial data for amenities and transportation access is still ongoing and may influence the next phase of model improvements.

Next Steps:

1. **Hyperparameter Optimization:** Continue fine-tuning the model's parameters to improve performance metrics.
2. **Model Validation:** Perform more extensive validation using additional metrics like R-squared, and expand testing with different train-test splits.
3. **Incorporate Geospatial Features:** Finalize the integration of geospatial features (proximity to MRT, schools, parks) to enhance prediction accuracy.
4. **User Interface Design:** Begin exploring potential designs for a simple user interface that could provide real-time predictions based on user input.
5. **Reporting & Visualization:** Develop dashboards and visualization tools to present the findings in a clear and actionable format for stakeholders.

Timeline & Deliverables:

- **Week 1:** Data collection, cleaning, and feature engineering (Completed).
- **Week 2:** Initial model training and feature importance analysis (Completed).
- **Week 2:** Hyperparameter tuning and model optimization (In Progress).
- **Week 3:** Final model evaluation, testing, and reporting (Upcoming).
- **Week 3:** Deliverables, including final report, visualization, and interface prototype (Upcoming).

This progress update indicates that the project is on track, with key milestones achieved and further improvements underway. The next few steps focus on refining the model and delivering actionable insights.

Singapore Resale Flat Prices Predicting

Project Report

3 . Budget Analysis :

This section provides a detailed breakdown of the budget requirements for the "Predicting Singapore Flat Resale Prices Using XGBoost Regressor" project. It includes cost estimates for data acquisition, computational resources, personnel, software, and miscellaneous expenses. The objective is to ensure efficient resource allocation and keep the project within the planned budget.

1. Data Acquisition

- **Open Data Sources:** The project relies on publicly available datasets (such as Singapore's government open data portal) for resale prices, which are free of charge.
- **Geospatial Data:** Access to advanced geospatial datasets (e.g., public amenities, transportation networks) may require the purchase of premium API access , but here Python Modules like geopy can be used for free , but accuracy is compromised for this reason a clean csv file from externally used for the proper and successful implementation of this project.

Total for Data Acquisition Cost : 0

2. Computational Resources

- **Computing:** Since XGBoost training can be computationally intensive, especially when hyperparameter tuning and cross-validation are involved, cloud deployment solutions such as Rener with Streamlit be used. This is a free services which has scalable processing power with paid options also.
- **Computer Hardware:** An AMD 5 5000 series with a built in Vega graphics was used to develop the project with ML Model Choosing , Training data and Testing the score .

3. Personnel

- A Single Person Resource Task for Completion of the Whole Project .

Total for Resource Cost : 0 (As its a Learner Project)

4. Software Tools & Licensing

- **Development Tools:** Most of the development work will be conducted using open-source tools such as Python, VS Code Notebooks, Scikit-learn, and XGBoost, which do not require additional costs.
- **Data Visualization & Reporting:** Visualization tools such as Plotly and Matplotlib used for presenting results.
deployment or interface development.

Total for Software & Licensing Cost: 0 (as implemented with Python and its Libraries)

So The Sum of Cost involved in Developing This Tool is 0 (Zero)

**Only the Valuable Time is spent as a part of Learning Experience.*

Singapore Resale Flat Prices Predicting

Project Report

4 . Challenges and Solutions :

1.Challenge: Data Quality and Sparsity

Description: While the Singapore HDB resale price dataset is extensive, certain data points are missing or incomplete (e.g., missing lease data or flat types). Moreover, some flat types or regions have limited data, resulting in sparsity that can affect the model's accuracy and generalization.

Impact: Incomplete or sparse data can lead to biased model predictions and a lack of generalization, especially for less common flat types or regions.

Solution:

Data Imputation: Missing values can be addressed using imputation techniques such as mean, median, or regression-based imputation. For categorical variables, the most frequent category can be used.

Synthetic Data Generation: For sparse data regions or rare flat types, synthetic data generation techniques like SMOTE (Synthetic Minority Over-sampling Technique) can be used to balance the dataset.

Cross-validation: Implementing cross-validation across different regions and flat types ensures the model generalizes well even with sparse data.

2. Challenge: Feature Engineering for Location-Based Data

Description: Location is a key factor in determining the resale prices of flats. However, the raw data does not provide proximity to amenities like MRT stations, schools, or parks, which are crucial for predicting housing prices.

Impact: Without properly engineered location features, the model's ability to capture neighborhood effects will be limited, potentially leading to lower accuracy.

Solution:

Geospatial Data Integration: Use external APIs or datasets to integrate geospatial data into the model. For example, calculating the distance of each flat to the nearest MRT station or school can create features that better reflect the property's appeal.

Clustering Techniques: Apply clustering algorithms (e.g., K-Means) on location data to categorize flats into clusters with similar neighborhood characteristics, making the model more sensitive to locality-driven price variations.

Singapore Resale Flat Prices Predicting

Project Report

3. Challenge: Hyperparameter Tuning

Description: XGBoost has a large number of hyperparameters, such as learning rate, number of estimators, and maximum depth of trees. Tuning these hyperparameters is essential for maximizing model performance but can be computationally expensive and time-consuming.

Impact: Poorly optimized hyperparameters can result in suboptimal model performance, overfitting, or underfitting.

Solution:

Grid Search with Cross-Validation: Perform grid search combined with cross-validation to systematically evaluate different combinations of hyperparameters. This ensures that the model performs well across different data splits and reduces overfitting.

Randomized Search: To reduce the time and computational cost, randomized search can be used as an alternative to grid search. It randomly samples from the hyperparameter space and finds optimal settings faster.

4. Challenge: Model Interpretability

Description: XGBoost, as a gradient-boosting model, is often seen as a "black-box" algorithm. It can be challenging to interpret how individual features impact the predictions, which is crucial for gaining insights into the factors influencing resale prices.

Impact: Lack of interpretability may limit the model's usability for stakeholders, especially policymakers and real estate professionals who need to understand the drivers behind pricing trends.

Solution:

Feature Importance Plot: Leverage XGBoost's built-in feature importance functionality to visualize the most influential features. This allows stakeholders to identify key drivers like location, flat type, or floor area.

Singapore Resale Flat Prices Predicting

Project Report

5. Challenge: Overfitting

Description: Given the complexity of real estate pricing, there's a risk that the model might overfit to the training data, particularly when using high-dimensional datasets and powerful algorithms like XGBoost.

Impact: Overfitting leads to high performance on training data but poor generalization to unseen data, resulting in inaccurate predictions on real-world data.

Solution:

Cross-validation: Use k-fold cross-validation to ensure that the model generalizes well across different subsets of the data.

Early Stopping: Use early stopping to halt the training process when performance on the validation set stops improving, preventing the model from learning noise in the training data.

Singapore Resale Flat Prices Predicting

Project Report

5.Risks and Mitigation:

Risk 1. : Data Availability and Quality

Description: The quality and completeness of the resale price data, as well as other related datasets (e.g., geospatial data, flat features), may be inconsistent or incomplete. Missing data points or inaccuracies can significantly affect model performance.

Mitigation Strategies:

Data Preprocessing , Data Augmentation ,Regular Data Updates and Backup Data Sources .

Risk 2. : Model Overfitting

Description: XGBoost is a powerful algorithm, but there is a risk of overfitting if the model learns too much from the training data, especially when working with complex datasets. Overfitting results in poor generalization to unseen data.

Mitigation Strategies:

Cross-Validation , Regularization and Simplify the Model .

Risk 3. : Model Interpretability

Description: XGBoost is often seen as a "black-box" model, making it difficult to interpret and explain the relationships between variables and predictions. For real estate stakeholders and policymakers, model transparency is critical.

Mitigation Strategies:

Explainable AI Tools , Feature Importance and Transparent Reporting .

Risk 4. : Transparent Reporting

Description: The Singapore real estate market is influenced by external factors such as government policies, economic conditions, and interest rate fluctuations. These factors may change over time, affecting resale prices and rendering the model less accurate.

Mitigation Strategies:

Regular Model Retraining and Time-Series Forecasting .

Singapore Resale Flat Prices Predicting

Project Report

Risk 5. : Computational Costs

Description: Training the XGBoost model, especially with large datasets and hyperparameter tuning, can be computationally expensive. Cloud computing costs may escalate, especially during extensive model training and testing phases.

Mitigation Strategies:

Efficient Tuning Methods and Local Testing .

Risk 6. : User Adoption

Description: Even with a well-performing model, stakeholders (e.g., real estate professionals, buyers, sellers, policymakers) may be reluctant to adopt machine learning-based predictions due to lack of familiarity or trust in the technology .

Mitigation Strategies:

User Engagement , User Education and Pilot Testing .

6 . Recommendations :

It is recommended to to add the details of other Amenities like Schools, Malls,Hospitals and Parks data to be added as the features by getting their geographical data to improve the project's overall outcome. Based on the current analysis, these might involve requests for additional resources, suggested this process changes, project timeline adjustment decisions, etc.

Singapore Resale Flat Prices Predicting

Project Report

7. Conclusion :

The Housing and Development Board (HDB) flats form a significant portion of Singapore's housing market, and understanding the factors influencing resale prices is of paramount importance to various stakeholders, including homebuyers, sellers, policymakers, and investors. As one of the most densely populated cities in the world, Singapore's real estate market is shaped by unique factors such as location, flat type, remaining lease, and proximity to amenities like public transportation.

In a competitive and dynamic real estate market, predicting resale prices with precision is crucial for making informed decisions. Traditional valuation methods, while useful, may not fully capture the complexities and interactions between various housing features. Machine learning models offer a more data-driven approach to predict housing prices by analyzing historical trends and identifying key drivers.

This project aims to leverage the XGBoost Regressor, a powerful and efficient machine learning algorithm, to predict the resale prices of Singapore HDB flats. XGBoost (Extreme Gradient Boosting) is known for its ability to handle large datasets, model complex relationships, and provide high accuracy in tabular data analysis. By analyzing historical data and incorporating key features such as flat type, floor area, remaining lease, and location, this project seeks to develop a model that delivers accurate predictions and valuable insights into the resale market.

Objectives:

- **Build an Accurate Model:** Develop an XGBoost Regressor model to predict HDB flat resale prices based on historical data and property features.
- **Identify Key Factors:** Analyze the importance of different features, such as flat type, lease duration, and location, to understand the drivers of resale prices.
- **Provide Actionable Insights:** Offer insights to buyers, sellers, and policymakers to help them make more informed decisions in the real estate market.

Significance:

The results from this project will not only provide a more accurate pricing model for the Singapore resale market but also contribute to a deeper understanding of how various features influence property prices. This can guide home buyers in making purchasing decisions, help sellers price their flats competitively, and inform policymakers about factors impacting the housing market.

By using advanced machine learning techniques like XGBoost, this project aims to offer a cutting-edge solution to the challenge of predicting real estate prices in a complex and evolving market like Singapore.

Singapore Resale Flat Prices Predicting

Project Report

8 . Appendices :

The appendices provide additional information , supporting data, and technical details referenced throughout the project. These sections include details on datasets, model performance metrics, hyper parameter settings, and code snippets used in the project.

Supporting Data:

The Data is sourced from the URL : <https://beta.data.gov.sg/collections/189/view> , these sourced data in csv format is segneted based on year reanges all the different files were combined as one dataframe in Pandas .

Technical Details:

The Data is preprocessed with Data Wrangling techniques like Data Sourcing ,Data Cleaning with removal of Duplicated Data , Removal of inconsistent Rows of Missing data ,Data Inspection ,Feature Engineering with Existing Data Columns by applying Functions and combining the preprocessed additional columns Lebel Encoding of Column values as Numerical rather then string type for the Machine Learning Model Conversion of Data Types.

Data Sourcing

Data Collection

+ Code + Markdown

```
# set search path and glob for files
# here we want to look for csv files in the input directory
path = 'D:\\SingaporeRealEstate\\Data'
files = glob.glob(path + '/*.csv')

# create empty list to store dataframes
df_list = []

# loop through list of files and read each one into a dataframe and append to list
for f in files:
    # read in csv
    temp_df = pd.read_csv(f)
    # append df to list
    df_list.append(temp_df)
    #print(f'Successfully created dataframe for {f} with shape {temp_df.shape}')

# concatenate our list of dataframes into one!
df_Compleat = pd.concat(df_list, axis=0)
#print(df_Compleat.shape)
#df_Compleat.head()
```

✓ 3.7s

Singapore Resale Flat Prices Predicting

Project Report

df_Compleat

[3] ✓ 0.0s

...

| | month | town | flat_type | block | street_name | storey_range | floor_area_sqm | flat_model | lease_commence_date | resale_price | remaining_lease |
|-------|---------|------------|-----------|-------|------------------|--------------|----------------|----------------|---------------------|--------------|-----------------|
| 0 | 1990-01 | ANG MO KIO | 1 ROOM | 309 | ANG MO KIO AVE 1 | 10 TO 12 | 31.0 | IMPROVED | 1977 | 9000.0 | NaN |
| 1 | 1990-01 | ANG MO KIO | 1 ROOM | 309 | ANG MO KIO AVE 1 | 04 TO 06 | 31.0 | IMPROVED | 1977 | 6000.0 | NaN |
| 2 | 1990-01 | ANG MO KIO | 1 ROOM | 309 | ANG MO KIO AVE 1 | 10 TO 12 | 31.0 | IMPROVED | 1977 | 8000.0 | NaN |
| 3 | 1990-01 | ANG MO KIO | 1 ROOM | 309 | ANG MO KIO AVE 1 | 07 TO 09 | 31.0 | IMPROVED | 1977 | 6000.0 | NaN |
| 4 | 1990-01 | ANG MO KIO | 3 ROOM | 216 | ANG MO KIO AVE 1 | 04 TO 06 | 73.0 | NEW GENERATION | 1976 | 47200.0 | NaN |
| ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... |
| 52198 | 2014-12 | YISHUN | 5 ROOM | 816 | YISHUN ST 81 | 10 TO 12 | 122.0 | Improved | 1988 | 580000.0 | NaN |
| 52199 | 2014-12 | YISHUN | EXECUTIVE | 325 | YISHUN CTRL | 10 TO 12 | 146.0 | Maisonette | 1988 | 540000.0 | NaN |
| 52200 | 2014-12 | YISHUN | EXECUTIVE | 618 | YISHUN RING RD | 07 TO 09 | 164.0 | Apartment | 1992 | 738000.0 | NaN |
| 52201 | 2014-12 | YISHUN | EXECUTIVE | 277 | YISHUN ST 22 | 07 TO 09 | 152.0 | Maisonette | 1985 | 592000.0 | NaN |
| 52202 | 2014-12 | YISHUN | EXECUTIVE | 277 | YISHUN ST 22 | 04 TO 06 | 146.0 | Maisonette | 1985 | 545000.0 | NaN |

933162 rows × 11 columns

Data Cleaning with removal of Duplicated Data

```
Data Wrangling - Data Inspection

# Checking for Duplicated Data
df_Compleat.duplicated().sum()

1899

# Removing Duplicated Data and cross checking it
df_Compleat.drop_duplicates(inplace=True)
df_Compleat.duplicated().sum()

0
```

Data Inspection

```
df_Compleat.nunique()

month      416
town        27
flat_type   8
block     2712
street_name 584
storey_range 25
floor_area_sqm 213
flat_model  34
lease_commence_date 55
resale_price 9602
remaining_lease 727
dtype: int64
```

```
df_Compleat.info()

<class 'pandas.core.frame.DataFrame'>
Index: 931263 entries, 0 to 52202
Data columns (total 11 columns):
#   Column                Non-Null Count  Dtype
---  -
0   month                 931263 non-null object
1   town                 931263 non-null object
2   flat_type            931263 non-null object
3   block               931263 non-null object
4   street_name         931263 non-null object
5   storey_range        931263 non-null object
6   floor_area_sqm      931263 non-null float64
7   flat_model          931263 non-null object
8   lease_commence_date 931263 non-null int64
9   resale_price        931263 non-null float64
10  remaining_lease     223800 non-null object
dtypes: float64(2), int64(1), object(8)
memory usage: 85.3+ MB
```


Singapore Resale Flat Prices Predicting

Project Report

Feature Engineering with Existing Data Columns by applying Functions

```
#Creating Median Value for storey RangeValue
def storeyMedian(RangeValue):
    minVal =int(RangeValue.split(" ")[0])
    maxVal =int(RangeValue.split(" ")[2])
    outMedVal=int((minVal+maxVal)/2)
    return outMedVal
```

✓ 0.0s

```
#function to change Multi Gen Apartment and other Apartment as integer
```

```
def roomMod(flat_type):
    if flat_type == "EXECUTIVE":
        out =int(6)
    elif flat_type == "MULTI GENERATION":
        out =int(7)
    elif flat_type == "MULTI-GENERATION":
        out =int(7)
    else:
        out = int(flat_type.split(" ")[0])
    return out
```

0.0s

```
df_ComplClean.drop('remaining_lease', axis = 1, inplace= True)
df_ComplClean['lease_remain_years'] = 99 - (2023 - df_ComplClean['lease_commence_date'])
df_ComplClean["storey_AvgCount"] = df_ComplClean["storey_range"].apply(storeyMedian)
df_ComplClean["flatRoom_type"] = df_ComplClean["flat_type"].apply(roomMod)
```

```
#Month column
```

```
# Separate the year and month of resale because this is time series Data
```

```
df_ComplClean[['resale_year', 'resale_month']] = df_ComplClean['month'].str.split('-', expand=True)
```

```
# drop the resale month column
```

```
df_ComplClean.drop('month', axis = 1, inplace= True)
```

✓ 3.3s

Label Encoding of Column values as Numerical rather than string type

```
# list unique Town names
towns = df_cleanData['town'].unique()
# map the streets to provide input to ML model
towns_mapping = {town: idx + 1 for idx, town in enumerate(towns)}
# Convert dictionary to pickle
with open('saved_Model_Pickel\\townMapping.pkl', 'wb') as file:
    pickle.dump(towns_mapping, file)
towns_mapping
```

```
{'ANG MO KIO': 1,
 'BEDOK': 2,
 'BISHAN': 3,
 'BUKIT BATOK': 4,
 'BUKIT MERAH': 5,
 'BUKIT TIMAH': 6,
 'CENTRAL AREA': 7,
 'CHOA CHU KANG': 8,
 'CLEMENTI': 9,
 'GEYLANG': 10,
 'HOUGANG': 11,
 'JURONG EAST': 12,
 'JURONG WEST': 13,
 'KALLANG/WHAMPOA': 14,
 'MARINE PARADE': 15,
 'QUEENSTOWN': 16,
 'SENGKANG': 17,
 'SERANGOON': 18,
 'TAMPINES': 19,
 'TOA PAYOH': 20,
 'WOODLANDS': 21,
 'YISHUN': 22,
 'LIM CHU KANG': 23,
 'SEMBAWANG': 24,
 'BUKIT PANJANG': 25,
 'PASIR RIS': 26,
 'PUNGOL': 27}
```

```
#Replacing the String 'Key' to Numerical 'Value' Of the dictionary from Label Encoding
df_cleanData['town'] = df_cleanData['town'].map(towns_mapping)
df_cleanData
```

| | town | flat_type | block | street_name | storey_range | floor_area_sqm | flat_model | lease_com |
|------|------|-----------|-------|------------------|--------------|----------------|----------------|-----------|
| 0 | 1 | 1 ROOM | 309 | ANG MO KIO AVE 1 | 10 TO 12 | 31.0 | Improved | |
| 1 | 1 | 1 ROOM | 309 | ANG MO KIO AVE 1 | 04 TO 06 | 31.0 | Improved | |
| 2 | 1 | 1 ROOM | 309 | ANG MO KIO AVE 1 | 10 TO 12 | 31.0 | Improved | |
| 3 | 1 | 1 ROOM | 309 | ANG MO KIO AVE 1 | 07 TO 09 | 31.0 | Improved | |
| 4 | 1 | 3 ROOM | 216 | ANG MO KIO AVE 1 | 04 TO 06 | 73.0 | New Generation | |
| ... | ... | ... | ... | ... | ... | ... | ... | ... |
| 2198 | 22 | 5 ROOM | 816 | YISHUN ST 81 | 10 TO 12 | 122.0 | Improved | |
| 2199 | 22 | EXECUTIVE | 325 | YISHUN CTRL | 10 TO 12 | 146.0 | Maisonette | |
| 2200 | 22 | EXECUTIVE | 618 | YISHUN RING RD | 07 TO 09 | 164.0 | Apartment | |
| 2201 | 22 | EXECUTIVE | 277 | YISHUN ST 22 | 07 TO 09 | 152.0 | Maisonette | |
| 2202 | 22 | EXECUTIVE | 277 | YISHUN ST 22 | 04 TO 06 | 146.0 | Maisonette | |

Singapore Resale Flat Prices Predicting

Project Report

Combining the preprocessed additional columns

```
df_coordinates = pd.read_csv('csv_data\\mrts_cbd_Dist_onMainDF_ToML.csv')

df_dataFetEng = df_cleanData.merge(df_coordinates, on="address", how='outer')
df_dataFetEng
```

Conversion of Data Types

```
#converting float data type to int data type
df_dataFetEng['town'] = df_dataFetEng['town'].astype('int64')
df_dataFetEng['street_name'] = df_dataFetEng['street_name'].astype('int64')
df_dataFetEng['flat_model'] = df_dataFetEng['flat_model'].astype('int64')
df_dataFetEng['lease_remain_years'] = df_dataFetEng['lease_remain_years'].astype('int64')
df_dataFetEng['storey_AvgCount'] = df_dataFetEng['storey_AvgCount'].astype('int64')
df_dataFetEng['flatRoom_type'] = df_dataFetEng['flatRoom_type'].astype('int64')
df_dataFetEng['resale_year'] = df_dataFetEng['resale_year'].astype('int64')
df_dataFetEng['resale_month'] = df_dataFetEng['resale_month'].astype('int64')
```

End Resultant Dataframe for Machine Learning Training and Testing of Various Models.

```
df_ML= df_dataFetEng[['town','street_name','storey_AvgCount','flat_model','flatRoom_type','lease_remain_years','resale_year','resale_month','floor_area_sqm','cbd_Dist','mrt_minDist','resale_price']]
df_ML
```

| | town | street_name | storey_AvgCount | flat_model | flatRoom_type | lease_remain_years | resale_year | resale_month | floor_area_sqm | cbd_Dist | mrt_minDist | resale_price |
|--------|------|-------------|-----------------|------------|---------------|--------------------|-------------|--------------|----------------|-------------|-------------|--------------|
| 0 | 14 | 143 | 5 | 1 | 3 | 55 | 1990 | 1 | 68.0 | 1351.713661 | 143.728945 | 56000.0 |
| 1 | 14 | 143 | 11 | 1 | 3 | 55 | 1990 | 3 | 68.0 | 1351.713661 | 143.728945 | 46000.0 |
| 2 | 14 | 143 | 8 | 1 | 3 | 55 | 1990 | 5 | 74.0 | 1351.713661 | 143.728945 | 53800.0 |
| 3 | 14 | 143 | 8 | 1 | 3 | 55 | 1990 | 9 | 68.0 | 1351.713661 | 143.728945 | 45000.0 |
| 4 | 14 | 143 | 17 | 1 | 3 | 55 | 1990 | 10 | 68.0 | 1351.713661 | 143.728945 | 55000.0 |
| ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... |
| 931259 | 5 | 488 | 23 | 1 | 5 | 92 | 2023 | 8 | 112.0 | 2525.251132 | 182.864524 | 1325000.0 |
| 931260 | 5 | 488 | 20 | 1 | 5 | 92 | 2023 | 12 | 112.0 | 2525.251132 | 182.864524 | 1400000.0 |
| 931261 | 5 | 488 | 26 | 3 | 4 | 92 | 2024 | 3 | 92.0 | 2525.251132 | 182.864524 | 1188000.0 |
| 931262 | 5 | 488 | 35 | 1 | 5 | 92 | 2024 | 6 | 112.0 | 2525.251132 | 182.864524 | 1588000.0 |
| 931263 | 5 | 488 | 11 | 3 | 4 | 92 | 2024 | 7 | 92.0 | 2525.251132 | 182.864524 | 1100000.0 |

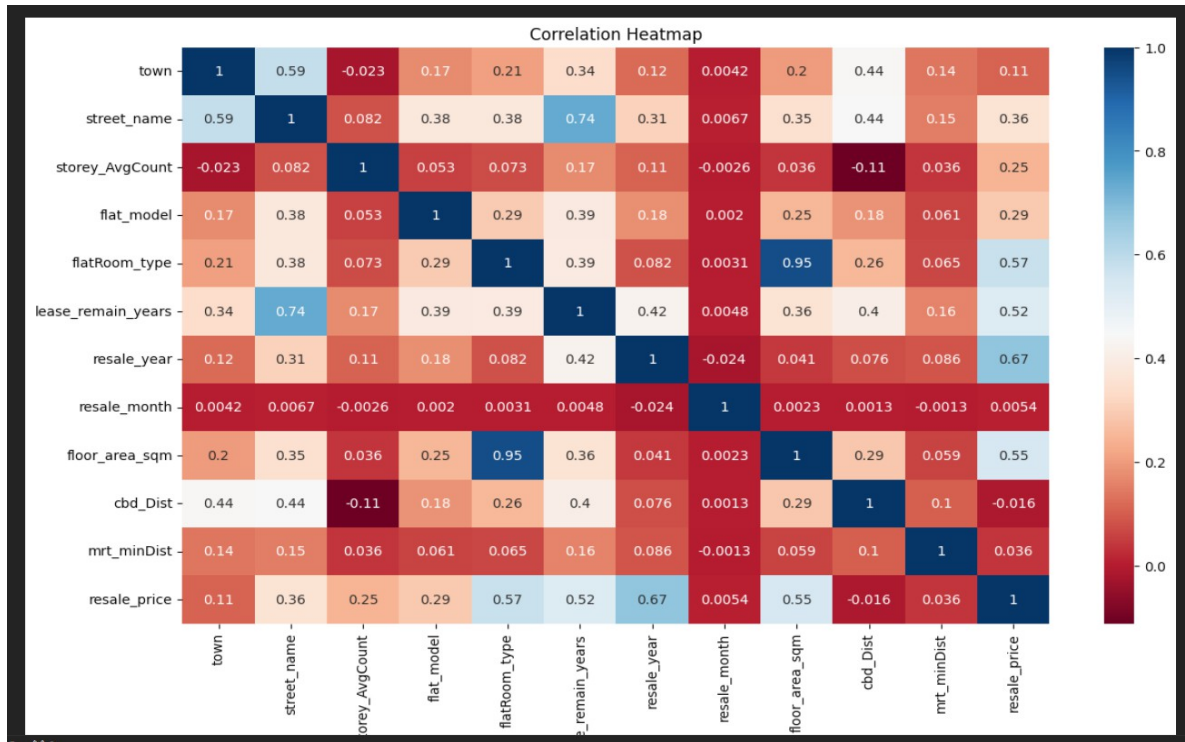
911285 rows x 12 columns

```
df_ML.describe()
```

| | town | street_name | storey_AvgCount | flat_model | flatRoom_type | lease_remain_years | resale_year | resale_month | floor_area_sqm | cbd_Dist | mrt_minDist | resale_price |
|-------|---------------|---------------|-----------------|---------------|---------------|--------------------|---------------|---------------|----------------|---------------|---------------|--------------|
| count | 911285.000000 | 911285.000000 | 911285.000000 | 911285.000000 | 911285.000000 | 911285.000000 | 911285.000000 | 911285.000000 | 911285.000000 | 911285.000000 | 911285.000000 | 9.112850e+05 |
| mean | 13.714430 | 239.842771 | 7.717321 | 3.534285 | 4.03883 | 64.594542 | 2006.433680 | 6.555463 | 96.428858 | 12203.383542 | 765.813417 | 3.264257e+05 |
| std | 7.835607 | 158.593026 | 4.844280 | 3.343698 | 0.93953 | 10.596130 | 9.383019 | 3.402373 | 25.514571 | 4366.739885 | 411.154110 | 1.709534e+05 |
| min | 1.000000 | 1.000000 | 2.000000 | 1.000000 | 1.000000 | 42.000000 | 1990.000000 | 1.000000 | 31.000000 | 592.121638 | 36.079525 | 5.000000e+03 |
| 25% | 8.000000 | 114.000000 | 5.000000 | 1.000000 | 3.000000 | 57.000000 | 1999.000000 | 4.000000 | 73.000000 | 9319.382184 | 459.992482 | 2.000000e+05 |
| 50% | 13.000000 | 219.000000 | 8.000000 | 3.000000 | 4.000000 | 62.000000 | 2005.000000 | 7.000000 | 94.000000 | 12919.010173 | 698.538883 | 3.020000e+05 |
| 75% | 21.000000 | 367.000000 | 11.000000 | 3.000000 | 5.000000 | 72.000000 | 2014.000000 | 9.000000 | 114.000000 | 15502.328503 | 982.061433 | 4.202000e+05 |
| max | 27.000000 | 582.000000 | 50.000000 | 21.000000 | 7.000000 | 96.000000 | 2024.000000 | 12.000000 | 366.700000 | 20225.103698 | 3646.118943 | 1.588000e+06 |

Singapore Resale Flat Prices Predicting

Project Report



Hyperparameter Tuning

```
model = LinearRegression()

# hyperparameters
param_grid = {
    'fit_intercept': [True, False],
    'copy_X': [True, False],
    'n_jobs': [-1],
    'positive': [True, False]
}

# gridsearchcv
grid_search = GridSearchCV(estimator=model, param_grid=param_grid, cv=5)
grid_search.fit(X_train_scaled, y_train)
print("Best hyperparameters:", grid_search.best_params_)
best_model = grid_search.best_estimator_
y_pred = best_model.predict(X_test_scaled)

# evaluation metrics
mse = mean_squared_error(y_test, y_pred)
mae = mean_absolute_error(y_test, y_pred)
rmse = np.sqrt(mse)
r2 = r2_score(y_test, y_pred)
print(" ")
print('Mean squared error:', mse)
print('Mean Absolute Error', mae)
print('Root Mean squared error:', rmse)
print(" ")
print('R-squared:', r2)
```

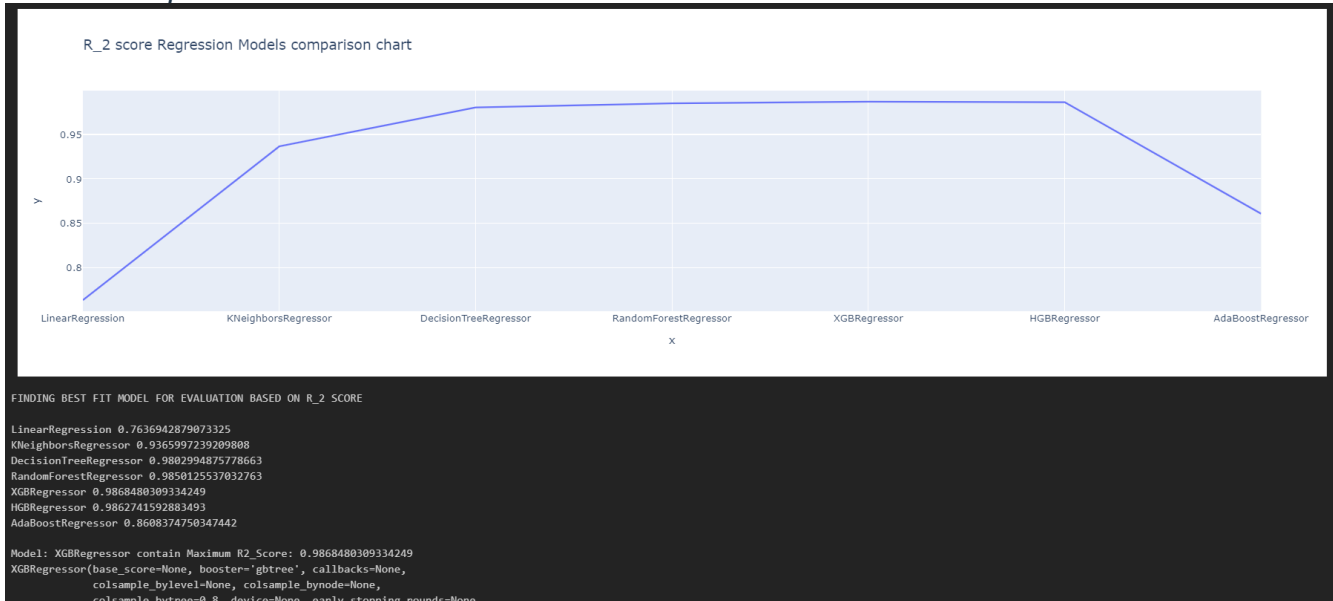
Best hyperparameters: {'copy_X': True, 'fit_intercept': True, 'n_jobs': -1, 'positive': False}

Mean squared error: 0.07304240906065877
Mean Absolute Error 0.19912869872882558
Root Mean squared error: 0.2702635918148406

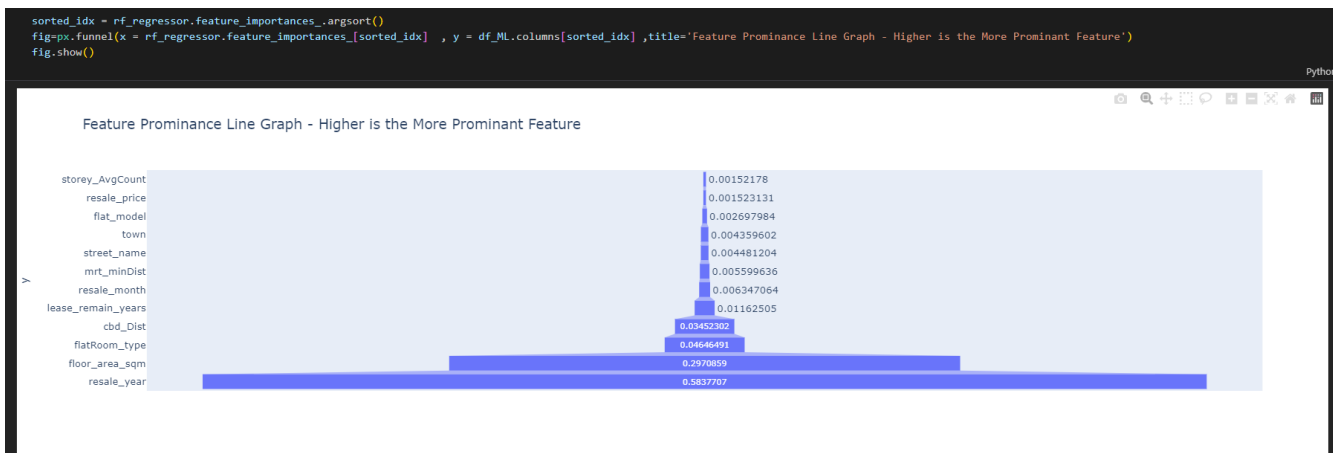
Singapore Resale Flat Prices Predicting

Project Report


Model Comparison Chart



Features Comparison Chart



Application Prediction Results



Singapore Resale Flat Prices
Predicting-Machine Learning
Project By Harish Kumar K P
harishk.kotte@rediffmail.com

Select Resale Year
2024

Select Resale Month
1

Select Lease Remain Years
52

Average CBD Distance
9046.287827402875

Average MRTS Distance
937.5655075907879

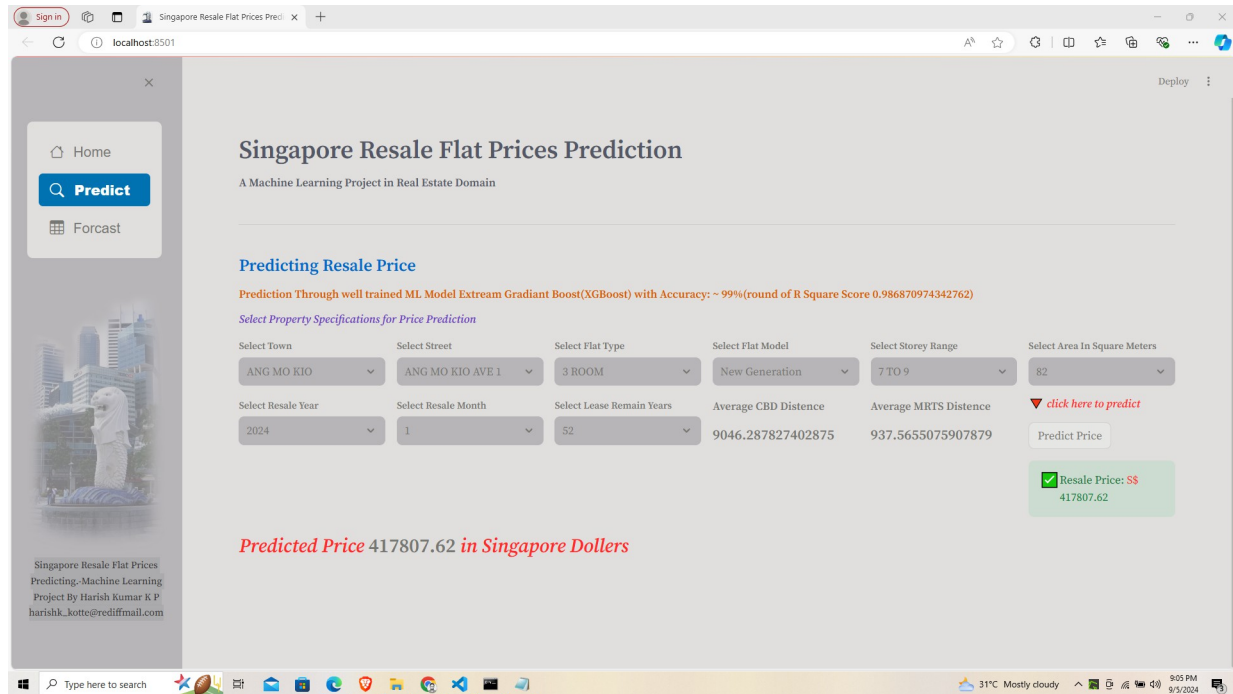
[click here to predict](#)
Predict Price

Resale Price: S\$
417807.62

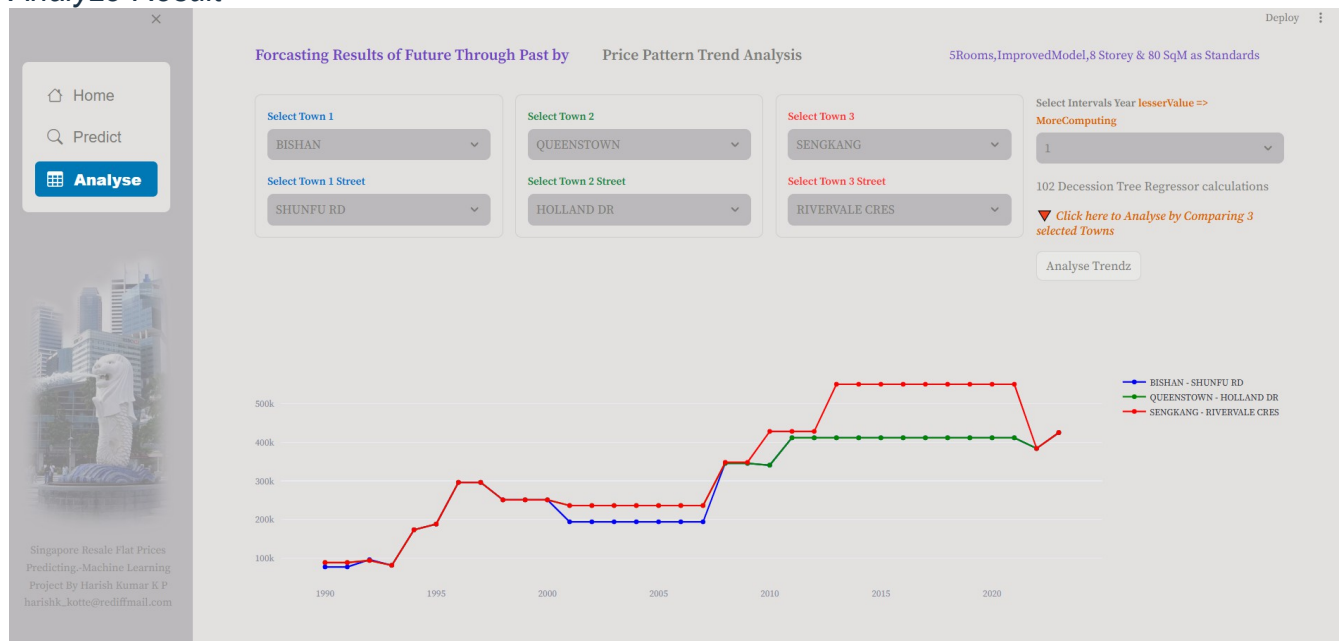
Predicted Price 417807.62 in Singapore Dollars

Singapore Resale Flat Prices Predicting

Project Report



Analyze Result



Singapore Resale Flat Prices Predicting

Project Report

Technical Discription :

Glob Module : With 'glob' module the various Time Series(Historical) data in csv format(coma seperated values) , is combinead as a single list and a Panda dataframe is created.

Pandas Library : Pandas is a fast, powerful, flexible and easy to use open source data analysis and manipulation tool , built on top of the Python programming language. DataFrame , A tabular representation of data in Pandas can be used for easy data wrangling ,data preprocessing and feature engineering a process of creating a column from 2 existing columns in the tabular data or applying python function on datasets as a batch process and where even 2 tables can be combined as one on taking a column as pivot. These Tabular data serves as the backbone for the Machine Learning Inputs.

GeoPy Library : From 'geopy' Library 'geocoder' and 'geodesic' modules were used , 'geocoder' for geocoding that is for extracting a list of Lattitudes and Longitudes of list of MRT Rail Stations and Property Address , 'geodesic' for Distence calculation between two points like Center of Bussiness Development to Property Distence and property to nearby MRT Rail Station Distence, This was seperatedly done in a different jupiter notebook and that dataframe is expoterd as a 'csv' file and later integrated in main dataframe.

NumPy Library : NumPy is the fundamental package for scientific computing in Python , In this Project its use was extended in Lograthmic Calculations and instant array of number generation etc,...

Pickle module : Python 'pickle' module is used for serializing and de-serializing a Python object structure. Any object in Python can be pickled so that it can be saved on disk, Especially in this project the catogorical features such as 'Town Name' , 'Street Name' , 'Flat Type' and 'Flat Model' are originally representes as 'String Type values',these were converted as numerical 'Integer Type value' so as get fed as input for ML Modeling the intermediate process a 'Dictionary Type' is created with Strings as dictKeys and corrsponding Integer as dictValue .

Exploratory Data Analysis (EDA) : Exploratory Data Analysis or EDA is a crucial step in the data analysis process that involves studying, exploring, and visualizing information to derive important insights. To find patterns, trends, and relationships in the data, it makes use of statistical tools and visualizations . For This Project Primarily 'Matplotlib' was used and in addition 'Plotly' was aslo used.

Scikit-Learn : scikit-learn is a Python module for machine learning built on top of SciPy and is distributed under the 3-Clause BSD license , In this Project a lots of ML models were tested from the 'Scikit Learn' modules for Accuracy scores . 'StandardScaler' from 'sklearn.preprocessing' , 'mean_absolute_error, mean_squared_error, r2_score' from 'sklearn.metrics', 'train_test_split , GridSearchCV , GridSearchCV, ShuffleSplit' from 'sklearn.model_selection' 'LinearRegression' from 'sklearn.linear_model', 'KNeighborsRegressor' from 'sklearn.neighbors' , 'DecisionTreeRegressor' from 'sklearn.tree', 'RandomForestRegressor , AdaBoostRegressor , HistGradientBoostingRegressor' from 'sklearn.ensemble', 'XGBRegressor' from 'xgboost.sklearn' .

– End of Report –