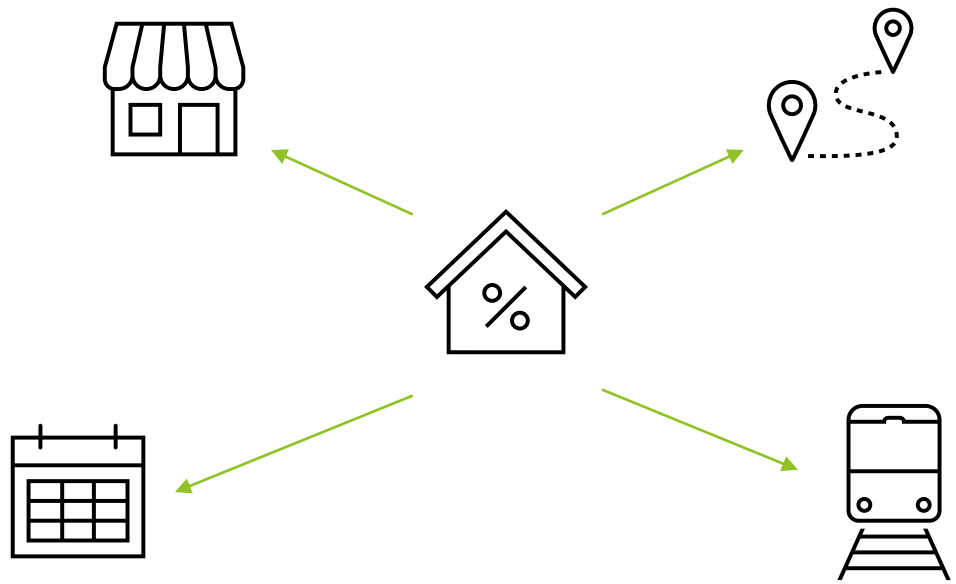
INFO – 6105 Data Science
Engineering Methods and Tools

Statistical Analysis and Prediction of Real Estate Prices

- By Prachi Navale

Why Predict Real Estate Prices?

- Objective: Use machine learning models to accurately predict housing prices and uncover key influencing factors



Dataset and Features

	No	X1 transaction date	X2 house age	X3 distance to the nearest MRT station	X4 number of convenience stores	X5 latitude	X6 longitude	Y house price of unit area
0	1	2012.917	32.0	84.878820	10	24.98298	121.540240	37.9
1	2	2012.917	19.5	306.594700	9	24.98034	121.539510	42.2
2	3	2013.583	13.3	561.984500	5	24.98746	121.543910	47.3
3	4	2013.500	13.3	561.984500	5	24.98746	121.543910	54.8
4	5	2012.833	5.0	390.568400	5	24.97937	121.542450	43.1
...	...	...	...	...	...	...	...	...
409	410	2013.000	13.7	1083.885689	0	24.94155	121.533361	15.4
410	411	2012.667	5.6	90.456060	9	24.97433	121.543100	50.0
411	412	2013.250	18.8	390.969600	7	24.97923	121.539860	40.6
412	413	2013.000	8.1	104.810100	5	24.96674	121.540670	52.5
413	414	2013.500	6.5	90.456060	9	24.97433	121.543100	63.9

414 rows × 8 columns

- Sourced from the UC Irvine Machine Learning Repository
- Key Attributes:
  - Age of the Property
  - Distance to MRT Station
  - Number of nearby convenience stores
  - Geographical location (latitude/longitude)

Approach to Analysis

DATA PREPROCESSING

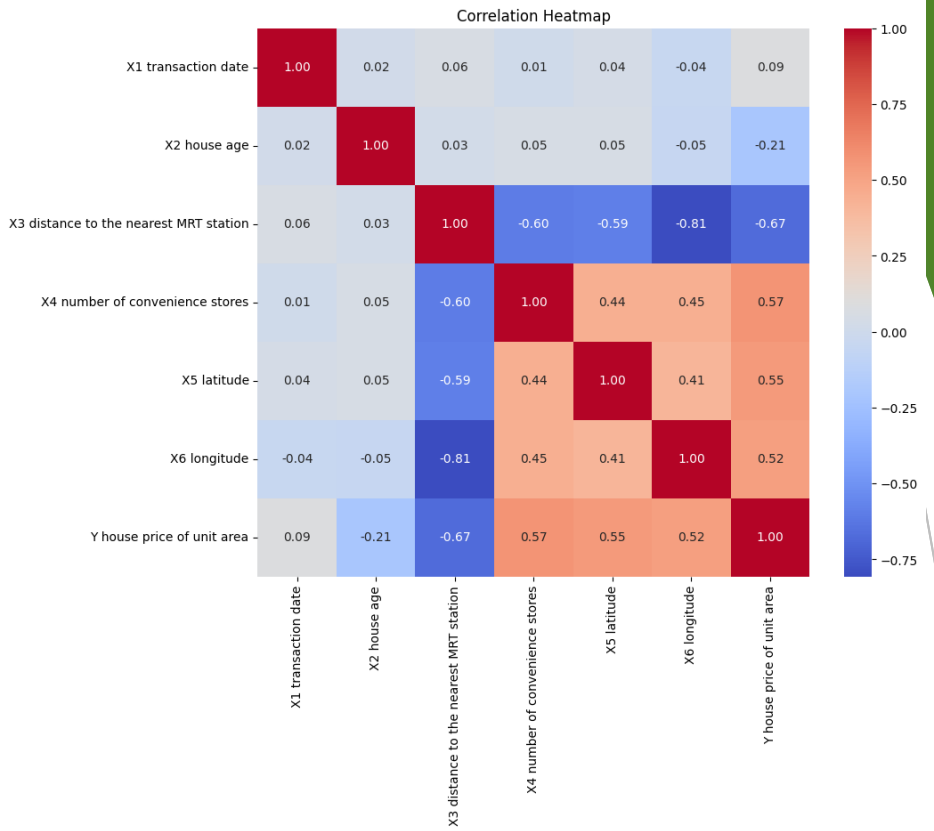
EXPLORATORY DATA ANALYSIS

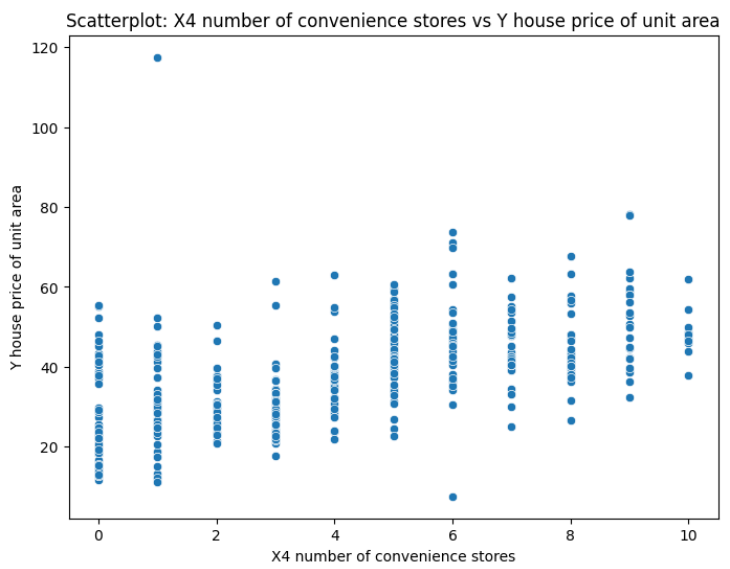
MODEL TRAINING

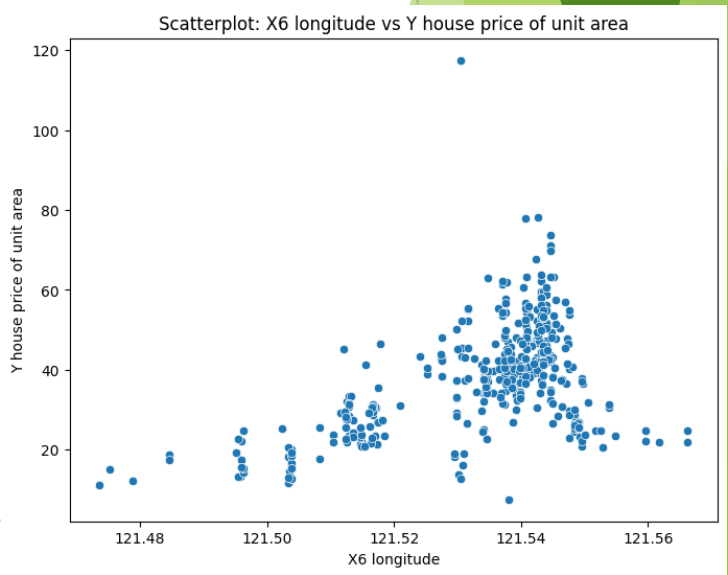
EVALUATION

Key Insights from EDA

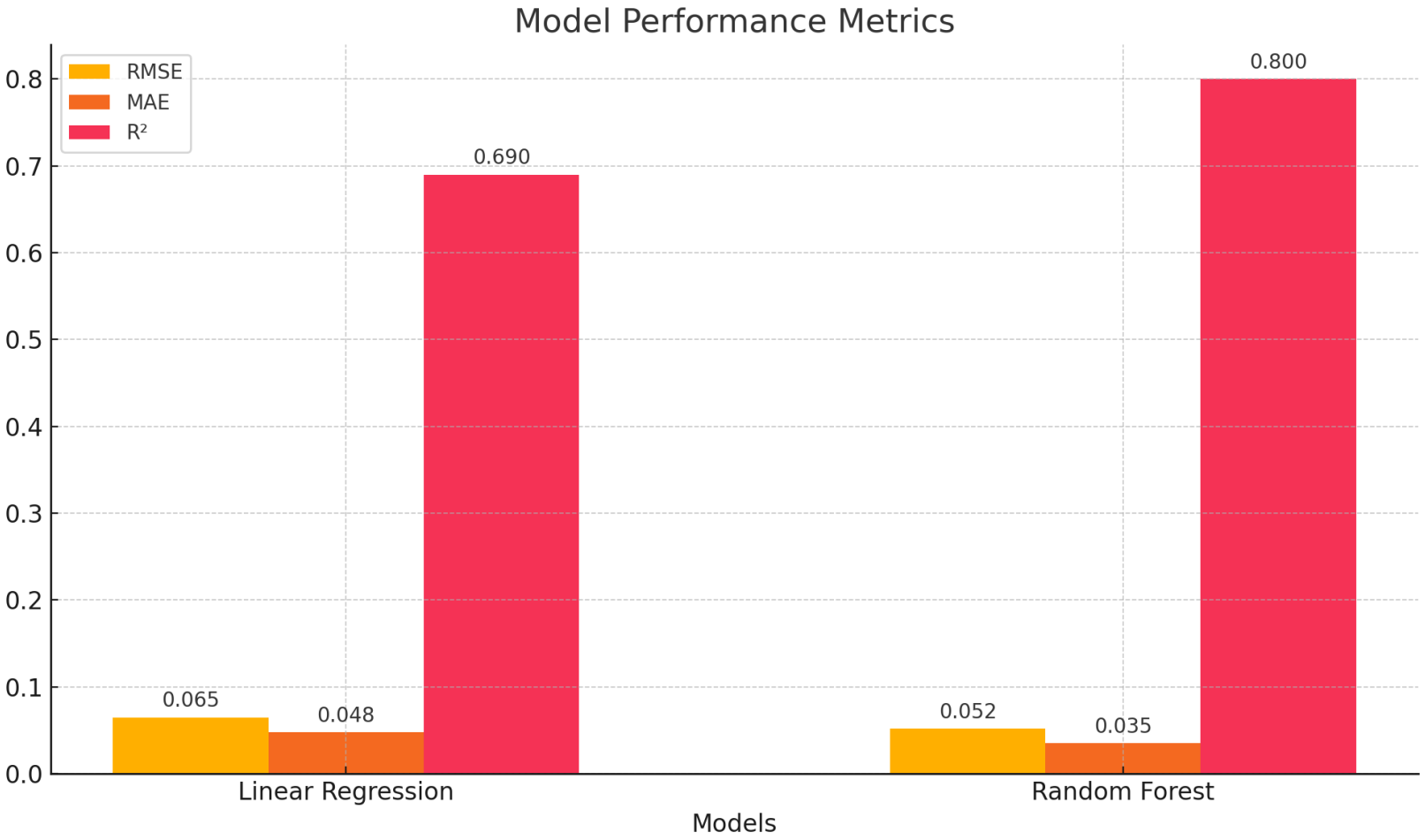
- Strong negative correlation between house prices and distance to MRT stations (r=−0.67)
- More nearby convenience stores = higher property value
- Geospatial clustering of high-priced properties in desirable neighborhoods







Comparison of Models



Practical Insights and Limitations

- Proximity to public transport significantly impacts property value
- Neighborhood amenities play a crucial role in valuation
- Limited features in the dataset (e.g., no crime rates or school quality)
- Risk of overfitting in Random Forest

Conclusion and Next Steps

Key Takeaway

Random Forest achieved the highest R^2 value of 0.80, demonstrating superior predictive power

Insights can guide urban planners, investors and policymakers

Future Directions

Add features like crime rates, school quality and economic indicators and collect temporal data for market trend analysis

Explore advanced models like Gradient Boosting and Neural Networks with larger datasets and richer diversity

Thank You !