Project 3 - HDB Price Prediction

PRABA
ZI HAN
DORA

BACKGROUND

- Established 1 February 1960, to solve Singapore's housing crisis
- Currently more than 80% of the population reside in HDB, out of which 90% own the flats.
- More than 1 million flats
- Government policies: 15 cooling measures since 2009
- External Effects : Effects of Covid-19 pandemic

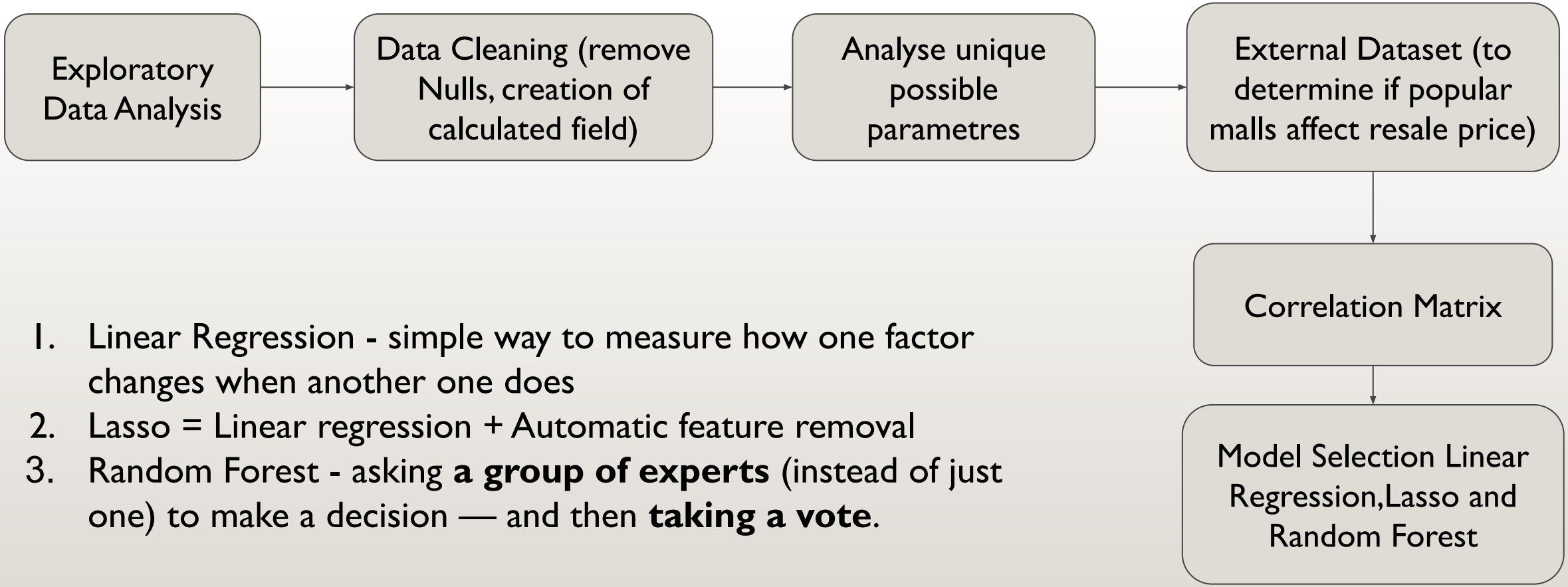
Limitations Vs. Relevance

Limitations	**Relevance**
External factors not considered (e.g. renovations nearby constructing MRT, government policies)	External factors maybe unpredictable - can be ignored. Have limited applicability.
Proximities to more than 1 MRT station	Accessibility to nearest public transport can be computed just by the nearest MRT distance. We focused on factors that are **measurable, consistent, and widely applicable**

WHY IS PREDICTING PRICES IMPORTANT

- Goal : Develop a machine learning model to predict HDB resale prices using key variables such as location, flat type, floor area, facilities and remaining lease.
- It is vital to have a data-driven methodology to be able to predict resale flat prices for the following :
  - Buyer/Sellers : Provides a guide to estimate fair prices based on market trends
  - Property Agents : Enhance accuracy in price setting and negotiations
  - Policy Makers : Supports housing policy evaluation and long-term planning through evidence based insights.

Data Timeline



1. Linear Regression - simple way to measure how one factor changes when another one does
2. Lasso = Linear regression + Automatic feature removal
3. Random Forest - asking **a group of experts** (instead of just one) to make a decision — and then **taking a vote**.

Model Selection Parameters

Purpose : To model the linear relationship between HDB resale prices and independent variables that affect HDB resale prices

Evaluation : RMSE-evaluates the accuracy of predictive models based on the average magnitude of the errors

Feature selection : Lasso (Least Absolute Shrinkage and Selection Operator) / RFE (recursive feature elimination)

Addresses overfitting and feature redundancy

Algorithm : Linear Regression: aids in deployment

Preliminary Analysis

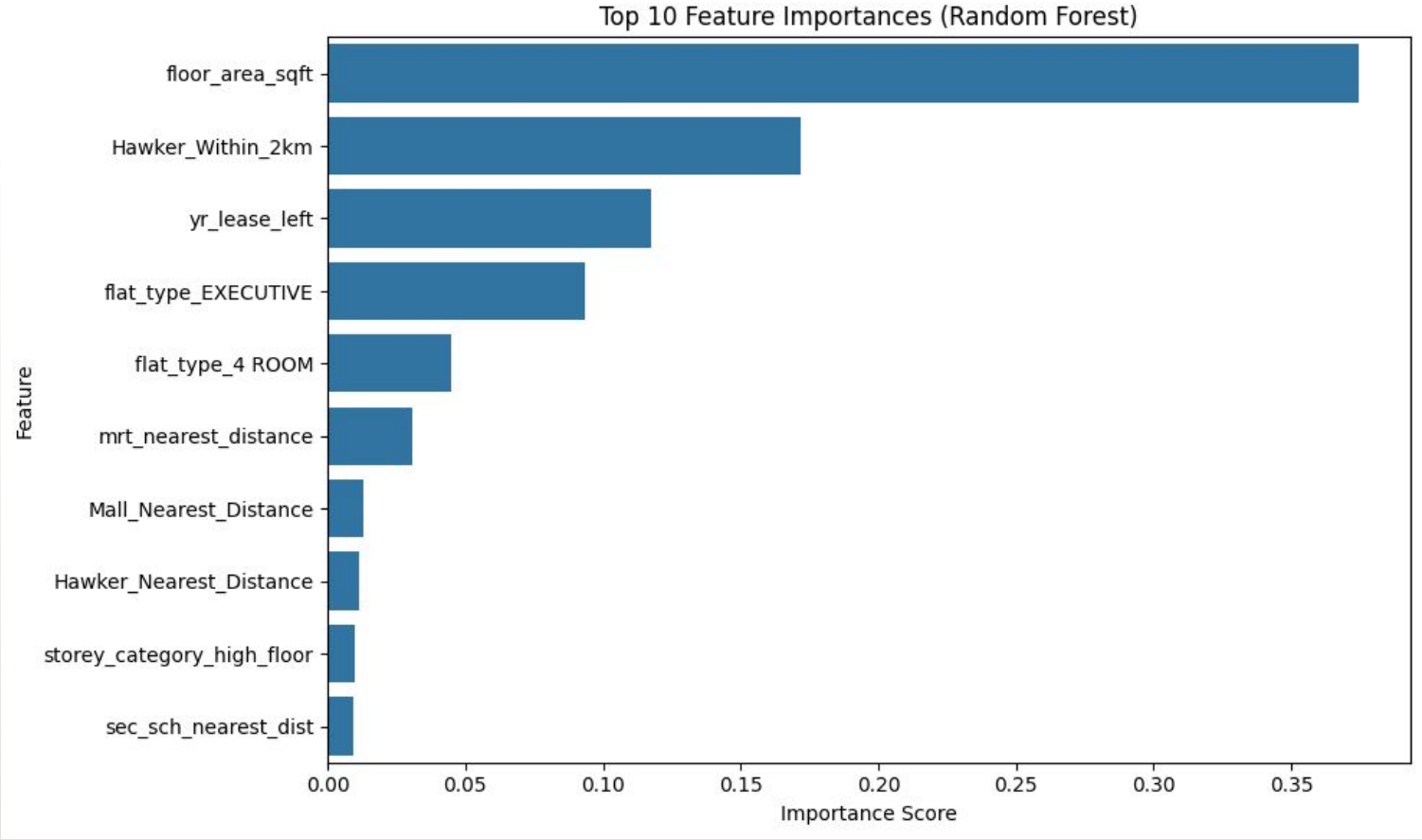
- Initial Linear Regression (No Standardization)
  RMSE of 55,382, with 67 features : moderate predictive accuracy
- Lasso-based Feature Selection Model
  To facilitate auto-selection of the most impactful features
  RMSE of 76,357 but with only 9 features

Chosen Model

- With standardisation: Lasso and RFE gave models with 67 features

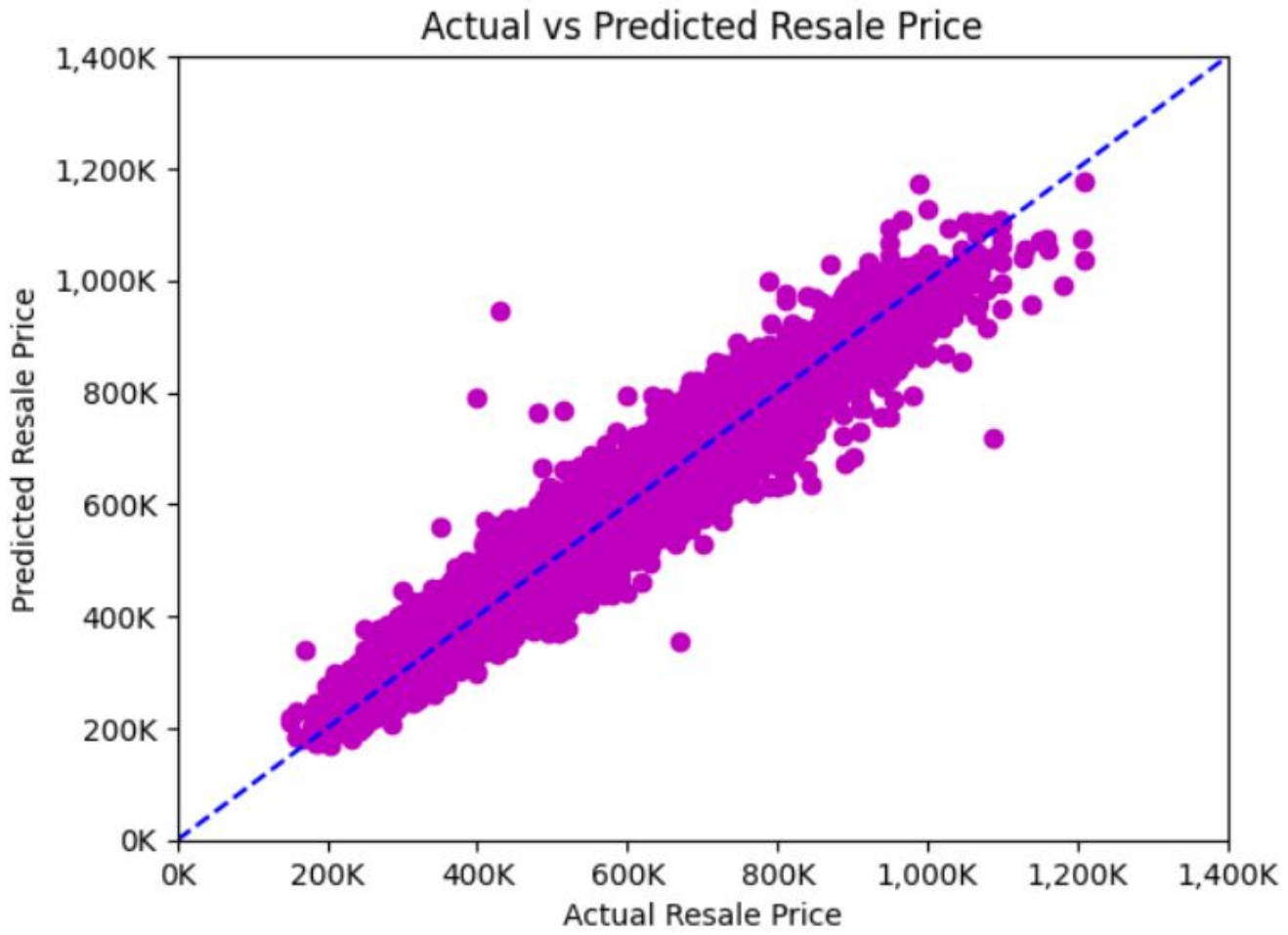
- Random Forest
  RMSE of 28,975 (lowest: significantly outperforming Linear Regression RMSE of 55,382), 67 features

Top 10 Features

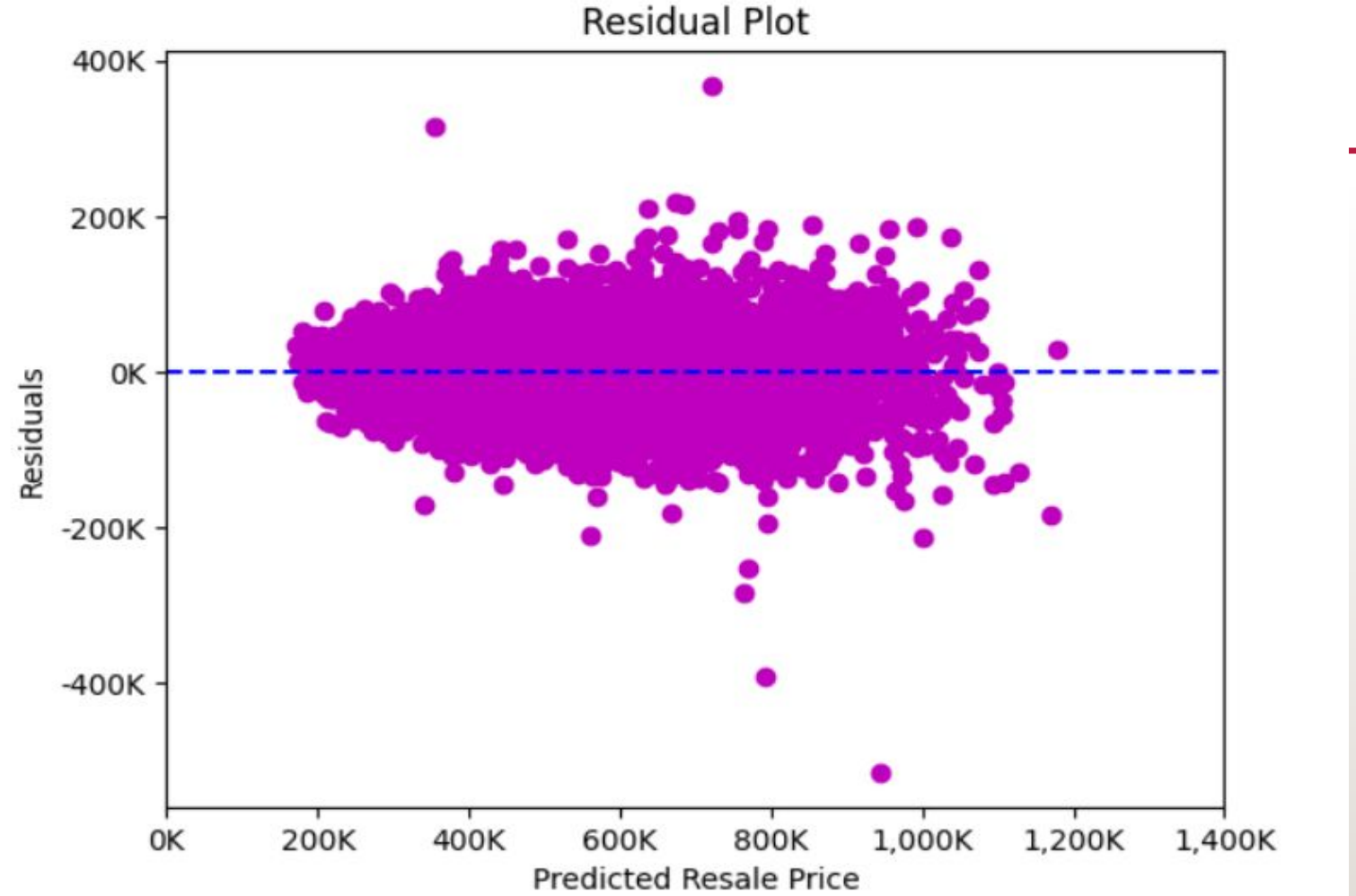


- Floor Area (sqft) emerged as the strongest predictor based on Random Forest, followed by the proximity to hawker centres and balance lease years.
- Key structural and accessibility-related features driving price variation .

Actual vs predicted sales price



Residual plot



Housing price app predictor

- The agents can use the app predictor which will take the account the different factors such as location, flat size, and proximity to amenities
- The app can be used to estimate and list housing prices

Limitations of app predictor

- The app does not account for government policies and market cooling measures
- There could be global and economic events that affect the housing market that is not accounted for by the model

Learning Points

- Delegating tasks to be more efficient and effective
- Have to try many models and feature combinations to obtain a good prediction result
- Hypothesise to have a goal, open-mind to be accepting
- To balance and manage project time well

Task management: Trello