# REAL ESTATE VALUATION MODEL ANALYSIS REPORT – ETE II (CAT 2)

R Abhijit Srivathsan
2448044

## Summary

This report analyzes a machine learning pipeline developed for predicting house prices in New Taipei City, Taiwan. The analysis covers data preprocessing, exploratory data analysis, feature selection, model development, and evaluation. The final polynomial regression model achieved a test R² of 0.794, indicating strong predictive capability for real estate valuation in the region.

## 1. Data Preprocessing & Exploratory Data Analysis

### Dataset Overview

The dataset contains 414 real estate transactions from Sindian District, New Taipei City, with 6 features and 1 target variable:

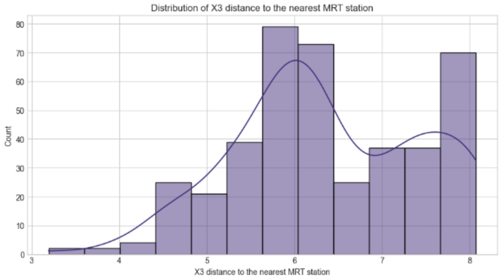
- Features: Transaction date, house age, distance to MRT station, number of convenience stores, latitude, and longitude
- Target: House price per unit area (in 10,000 New Taiwan Dollar/Ping)
  *1 Ping = approximately 3.3 square meters (or about 35.6 square feet)*

### Data Quality Assessment

- **Completeness**: No missing values were detected in any columns
- **Outliers**: Outlier detection using IQR method identified:
  – 37 outliers in distance to MRT station
  – 8 outliers in latitude
  – 35 outliers in longitude
  – 3 outliers in house price
  – All outliers were handled by capping at IQR boundaries

## Distribution Analysis

- **Skewness Assessment**:
  - “Distance to MRT station” showed significant positive skewness (1.216)
  - Log transformation successfully reduced skewness to -0.094
  - Other features had acceptable skewness levels



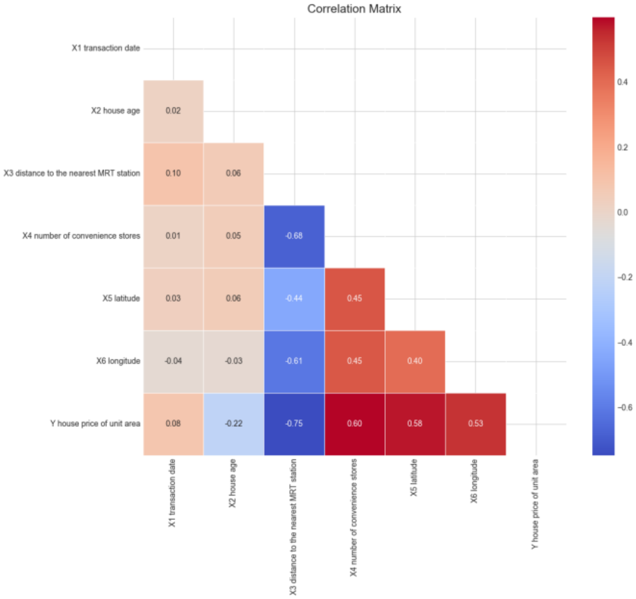
## Feature Scaling

- Features were standardized to ensure comparability and improve model performance

## Correlation Analysis

The correlation with the target variable revealed:

- Strong negative correlation with distance to MRT station (-0.749)
- Strong positive correlation with number of convenience stores (0.599)
- Strong positive correlation with latitude (0.577)
- Moderate positive correlation with longitude (0.532)
- Weak negative correlation with house age (-0.216)
- Minimal correlation with transaction date (0.084)



This analysis provides clear evidence that proximity to MRT stations and convenience stores significantly impacts property values, aligning with urban development theories emphasizing the importance of accessibility to transportation and amenities.

## 2. Feature Selection and Engineering

### Feature Selection Approach

The analysis employed correlation-based feature selection, retaining all features due to their varying degrees of correlation with the target. The model coefficients from linear regression further confirmed the importance of each feature.

### Feature Transformation

- Log transformation was applied to "Distance to MRT station" to address skewness
- Feature scaling was implemented to standardize the range of variables

### Feature Importance Analysis

Based on linear regression coefficients:

1. Distance to MRT station (-7.177): Most influential feature with strong negative impact
2. Latitude (3.815): Second most influential with positive impact
3. House age (-2.799): Third most influential with negative impact
4. Transaction date (1.661): Moderate positive impact
5. Number of convenience stores (1.186): Moderate positive impact
6. Longitude (0.385): Least influential with slight positive impact

### Impact on Model Performance

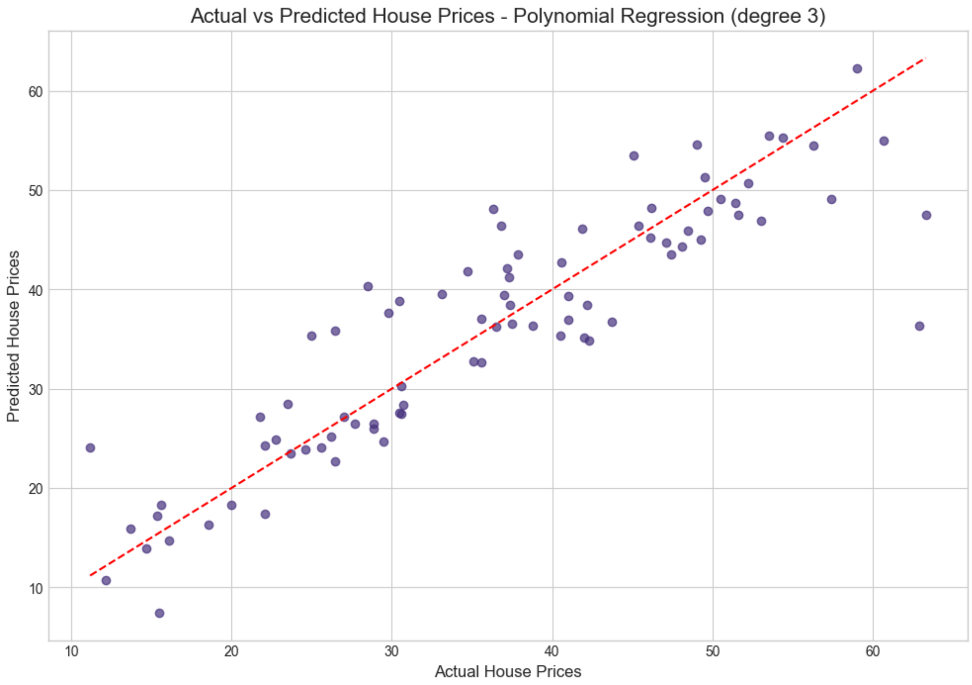
- The transformation of the "Distance to MRT station" feature likely improved model performance by normalizing its distribution
- Retaining all features was justified as regularization methods (Ridge and Lasso) did not eliminate any features, suggesting all variables contribute meaningful information
- The polynomial feature engineering significantly improved model performance, increasing $R^2$ from 0.689 (linear) to 0.794 (polynomial degree 3)

## 3. Model Development and Hyperparameter Tuning

### Model Selection Process

The analysis evaluated four types of regression models:

1. **Multiple Linear Regression (Baseline)**
   - Training R²: 0.690
   - Testing R²: 0.731
   - Testing RMSE: 6.720
2. **Ridge Regression**
   - Hyperparameter tuning via cross-validation
   - Best alpha: 1
   - Training R²: 0.690
   - Testing R²: 0.731
   - Testing RMSE: 6.714
   - Minimal improvement over baseline
3. **Lasso Regression**
   - Hyperparameter tuning via cross-validation
   - Best alpha: 0.0001
   - Training R²: 0.690
   - Testing R²: 0.731
   - Testing RMSE: 6.720
   - No feature elimination (all coefficients non-zero)
   - Performance similar to baseline
4. **Polynomial Regression**
   - Degrees tested: 2 and 3
   - Degree 2: Testing R² of 0.776, RMSE of 6.134
   - Degree 3: Testing R² of 0.794, RMSE of 5.872
   - Significant improvement over linear models



## Hyperparameter Tuning Approach

- **Ridge and Lasso**: Cross-validation to determine optimal regularization strength (alpha)
- **Polynomial Regression**: Testing different polynomial degrees (2 and 3)

## Justification of Final Model Selection

Polynomial Regression (degree 3) was selected as the final model due to:

- Highest testing R² (0.794)
- Lowest testing RMSE (5.872)
- Ability to capture non-linear relationships between features and target
- Reasonable complexity without severe overfitting (difference between training R² of 0.816 and testing R² of 0.794 is acceptable)

# 4. Model Evaluation

## Performance Metrics

The final Polynomial Regression (degree 3) model achieved:

- **Training metrics**:
  - $R^2$: 0.816
  - MSE: 31.663
  - RMSE: 5.627
  - MAE: 4.085
- **Testing metrics**:
  - $R^2$: 0.794
  - MSE: 34.475
  - RMSE: 5.872
  - MAE: 4.260
- **Cross-validation**:
  - Mean CV $R^2$: 0.669
  - Standard Deviation of CV $R^2$: 0.077

## Interpretation of Model Performance

- **Explained Variance**: The $R^2$ of 0.794 indicates the model explains approximately 79.4% of the variance in house prices, which is strong for real estate prediction
- **Prediction Error**: RMSE of 5.872 (in 10,000 New Taiwan Dollar/Ping) represents the average prediction error
- **Cross-validation Stability**: The model shows some variability across folds (SD of 0.077), suggesting moderate dependence on the specific data split
- **Generalization**: The relatively small gap between training and testing $R^2$ (0.022) suggests good generalization without severe overfitting

## Comparative Analysis

- Polynomial Regression (degree 3) outperformed all other models:
  - 6.3% improvement in $R^2$ over linear regression
  - 12.6% reduction in RMSE compared to linear regression
  - Consistently better performance in both training and testing sets

# 5. Data-Driven Insights and Recommendations

## Key Findings

1. **Location Factors Dominate Price Determination**
    - Proximity to MRT stations is the strongest predictor of property value
    - Each standard deviation decrease in distance to MRT station increases property value by approximately 7.18 units (after accounting for polynomial terms)
    - Geographic coordinates (especially latitude) significantly influence pricing, indicating clear neighborhood value differentiation
2. **Accessibility to Amenities Drives Value**
    - Number of convenience stores shows strong positive correlation with property prices
    - Properties with more nearby convenience stores command higher prices, indicating the premium for convenience
3. **Property Characteristics Impact**
    - Newer properties tend to command higher prices (negative coefficient for house age)
    - The non-linear relationship captured by polynomial terms suggests complex interactions between property age and other factors
4. **Market Timing Influence**
    - Small positive coefficient for transaction date suggests slight market appreciation over the study period
    - The effect is relatively minor compared to location and property-specific factors

## Recommendations for Stakeholders

1. **For Investors and Buyers**:
    - Prioritize properties near MRT stations for better value retention and appreciation
    - Consider the number of convenience stores as a key indicator of potential property value
    - Balance the trade-off between property age and location advantages
2. **For Property Developers**:
    - Focus development in areas with good MRT access or planned future stations
    - Incorporate convenience store spaces in development plans to enhance property value
    - Consider the non-linear relationships between features when assessing potential development sites

3. **For Real Estate Professionals**:

   - Use the polynomial model for more accurate property valuation
   - Emphasize proximity to transportation and amenities in property marketing
   - Recognize that properties farther from MRT stations need other significant compensating factors to achieve comparable values

4. **For Policy Makers**:

   - Invest in expanding MRT network to increase property values in developing areas
   - Encourage mixed-use development that incorporates convenience stores and amenities
   - Consider the significant impact of transportation infrastructure on property tax base

### Model Improvement Opportunities

1. **Additional Features**:

   - Include school district information
   - Add property-specific features (floor level, view, building quality)
   - Incorporate neighborhood safety statistics

2. **Advanced Modeling Techniques**:

   - Explore ensemble methods (Random Forest, Gradient Boosting)
   - Consider spatial regression models to better capture geographic effects
   - Implement time series components to better model market trends

3. **More Sophisticated Feature Engineering**:

   - Create interaction terms between key features
   - Develop accessibility indices combining MRT distance and convenience stores
   - Generate neighborhood clusters based on geographic coordinates

## Conclusion

The developed polynomial regression model demonstrates strong predictive capability for real estate valuation in New Taipei City. The analysis confirms the critical importance of location factors, particularly proximity to public transportation and conveniences, in determining property values. The model’s performance ($R^2$ of 0.794) provides stakeholders with a reliable tool for property valuation while highlighting the complex, non-linear relationships that exist in real estate markets.