Beijing Housing Price Forecast

Objective

Beijing's real estate market is a hub of high demand. There's speculation that the value of community features—especially educational access due to *school district housing*[1] policies—could significantly influence property prices. Using a dataset of Beijing housing prices from 2011 to 2017 sourced from Lianjia.com, a prominent real estate platform in China, the objective of this project is to develop a predictive model for housing prices in Beijing and identify the key factors in the determination of real estate market prices.

Methodology

This project[2] employs XGBoost, a gradient-boosted decision tree algorithm renowned for its proficiency with time series data. By utilizing Recursive Feature Elimination (RFE), the model identifies the top eight features that most significantly predict the 'squareMeterPrice.' Through ensemble learning and careful parameter optimization, the model's accuracy is heightened while preventing overfitting.

The dataset was divided for model training and testing at the threshold of January 1, 2016. Data up to December 31, 2015, is used for training, and data from January 2016 onwards is reserved for testing.

Let $\boldsymbol{X}$ be the matrix of predictors and y be the vector of housing prices. The XGBoost model aims to learn the function $f: \boldsymbol{X} \rightarrow y$ by minimizing the objective function: $\mathcal{L}(\theta) = \sum_{i=1}^{n} l\big(y_i, f(x_i; \theta)\big) + \lambda \sum_{i=1}^{T} {\omega_j}^2$, where l is the loss function (i.e. root mean squared error), ω_j represents the weight of the j-th tree, T is the total number of trees, λ is the L2 regularization term. Predicators x includes:

- totalPrice: The total price of the house.
- square: The area of the house in square meters.
- buildingType: The type of building (encoded as categorical variable).
- district: The district where the house is located (e.g. Haidian District, Fengtai District).
- communityAverage: The average price in the community.
- daysSince20110101: Number of days since January 1, 2011 (i.e. the earliest data).
- tradeYear: The year when the trade occurred.
- districtAverage: The average price in the district.

Results

1) *Top features (Graph 1-3)*
   a) tradeYear: as in graph 1, the median of squareMeterPrice steadily increased over the years.
      - Importance score of 51.8% means tradeYear accounts for over 50% of the model's predictive power and likely captures temporal market trends, economic condition, and inflation effect.
   b) communityAverage: as in graph 2, communityAverage has a moderate positive relationship with squareMeterPrice.
      - Importance score of 23.5% means a relatively strong predictive power, which implies local conditions, including education and amenities access and showcases strong positive correlation with target data.
2) *Model accuracy (Graph 4-5)*
   a) *AtoP (Actual to Prediction ratio)* of 1.029, meaning the model's average predictions are only about 2.9% lower than the actual values, which is close to the perfect highly accurate.
   b) The R^2 score of 0.9686 is considered very high. It means that approximately 96.86% of the variance in target variable, *square meter price*, can be explained by this model.
   c) As in graph 5, predictions closely follow the actual data trends and model performance improves over time.

Conclusion

As the common real estate quotes said "location, location, location" and "time is more valuable than money", both location-specific factors and temporal market trends should be carefully considered when assessing property values.

---

[1] School district housing: free admissions for household-registered students within certain school districts.

[2] GitHub Repository: https://github.com/sni13/HousingPrice_Beijing

Beijing Housing Price Forecast – Presentation Slides

Beijing Housing Price Forecast

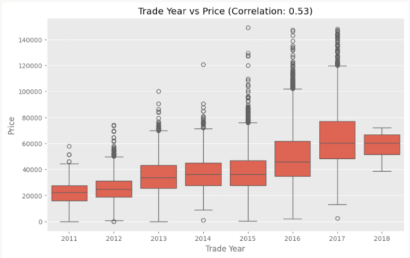
Author: Sailing NI
Code: GitHub Link

- Motivation: what are the key drives for Beijing housing price hikes?
- Ensemble Learning Model:
  - XGBoost Regression with L2 regularization -> effective with time-series data
  - RFE (Recursive Feature Elimination) -> pick best 8 features & avoid overfitting
- Predicators x:

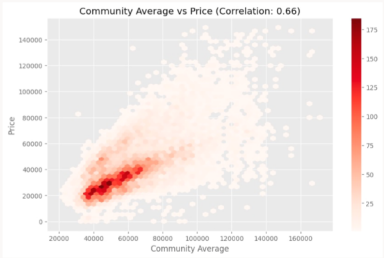
Property Attributes	Location Specific	Temporal Factors
totalPrice	district	daysSince20110101
square	communityAverage	tradeYear
buildingType	districtAverage	

Important Features

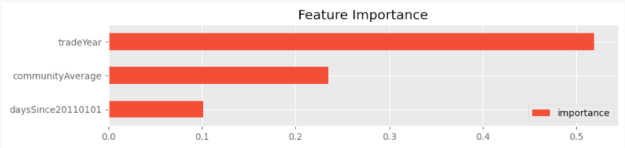
Author: Sailing NI
Code: GitHub Link



Graph 1: Year of Trade Time vs Square Feet Prices



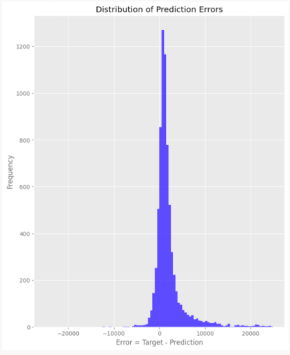
Graph 2: Square Feet Prices: Community vs City Level



Graph 3: Top 3 features in the XGBoost Model

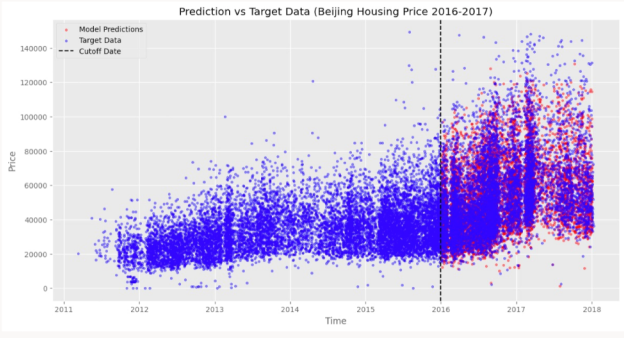
Evaluation Results

Author: Sailing NI
Code: GitHub Link



Graph 4: Distribution of Prediction Error

- Most of the prediction Errors are small values around 0, with an ***AtoP*** score of 1.029, showing high prediction accuracy.



Graph 5: Square Feet Prices: Prediction vs Target (2016-01-01 ~ 2017-12-31)

- The target prices are plotted **on top of** model predicted ones.
- Both red(predicted) points covered by blue(target) and a high R^2 score of 0.9686