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.

The model aims to leverage a gradient-boosted ensemble of decision trees to forecast Beijing's square meter housing prices[2], recognizing the critical elements that affect price formation processes.

Model Specification

The model[3] utilizes XGBoost, known for its effectiveness with time series data, and employs Recursive Feature Elimination (RFE) to pinpoint the most impactful eight features for predicting the target variable "price per square." This process enhances performance and mitigates overfitting. The model's parameters are carefully fine-tuned to balance bias and variance.

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.

Training

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.

Results

1) *Model accuracy*: the model demonstrates high accuracy with few large errors. Predictions closely follow the actual data trends and show improved model performance over time.
2) *Top features*: tradeYear and communityAverage emerge as the most influential features, indicating strong correlations with housing prices.
   a) tradeYear, as the top feature, likely captures temporal market trends and inflation effect. A general upward trend in housing prices suggests that housing prices have been rising annually.
   b) communityAverage, a strong representation for local conditions, including educational and amenities access, showcases strong positive correlation with squareFeetPrice, supporting the hypothesis that community factors significantly affect housing prices.
3) *Implication*: 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] Public Dataset: Housing prices of Beijing from 2011 to 2017, originally fetched from Lianjia.com.

[3] GitHub Repository:

Beijing Housing Price Forecast

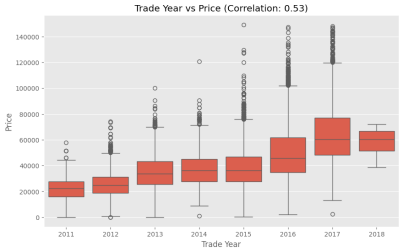
Author: Sailing NI
Code: GitHub Link

- 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:

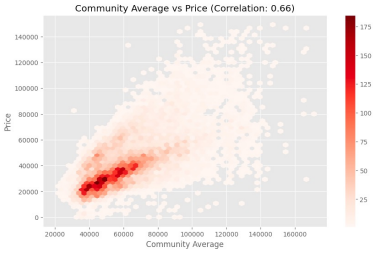
Property Attributes	Location Specific	Temporal Factors
totalPrice	**district**	**daysSince20110101**
square	**communityAverage**	**tradeYear**
buildingType	**districtAverage**	

Important Features

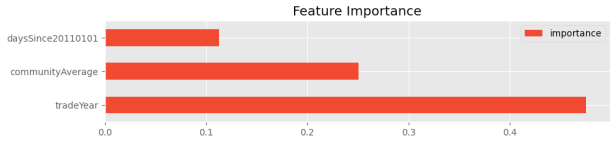
Author: Sailing NI
Code: GitHub Link



Graph 1: Year of Trade Time vs Square Feet Prices



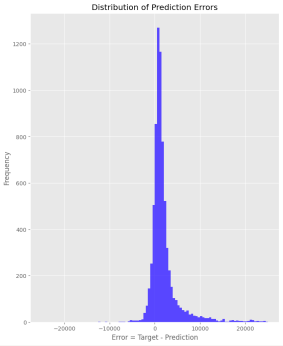
Graph 2: Square Feet Prices: Community vs City Level



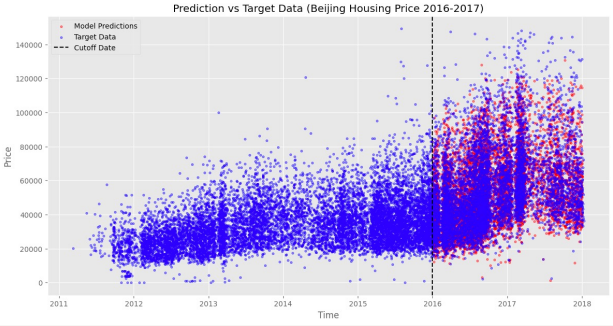
Graph 3: Top 3 features in the XGBoost Model

Evaluation Results

Author: Sailing NI
Code: GitHub Link



Graph 1: Most of the prediction Errors are small values around 0, showing high prediction accuracy.



Graph 2: Square Feet Prices: Prediction vs Target (2016-01-01 ~ 2017-12-31)

- The target prices are plotted **on top of** model predictions.
- Most red(predicted) points are invisible, covered by blue(target) ones, meaning most predictions align with target data.