Prediction Interval Competition II: House Price

Week 2 Report - Experimental Design and Implementation Plan

孫亞瑄
Institute of Data Science
National Cheng Kung University
Tainan, Taiwan
RE6134014@gs.ncku.edu.tw

許漢權
Department of Computer Science
and Information Engineering
National Cheng Kung University
Tainan, Taiwan
Q56135019@gs.ncku.edu.tw

ABSTRACT

This report presents our experimental design for the Kaggle Prediction Interval Competition II: House Price challenge. We propose a three-stage modeling approach progressing from baseline quantile regression to advanced Conformalized Quantile Regression (CQR) methods. Our preprocessing pipeline handles 200,000 property records with comprehensive feature engineering, addressing high cardinality categorical variables and target variable skewness. The experimental framework prioritizes Conformalized Quantile Regression to achieve theoretical coverage guarantees, complemented by hyperparameter optimization and ensemble methods. Our validation strategy employs nested cross-validation with multi-objective optimization balancing Winkler Interval Score minimization and 90% coverage accuracy. Initial baseline implementation demonstrates feasibility, with LightGBM quantile regression serving as the foundation for advanced uncertainty quantification techniques.

1 PROBLEM FORMULATION AND OBJECTIVES

1.1 Task Definition

We aim to predict 90% confidence intervals for house prices, optimizing the Winkler Interval Score:

$$W\_\alpha = \begin{cases} (u-l) + (2/\alpha)(l-y), & if\ y < l \\ (u-l), & if\ l \le y \le u \\ (u-l) + (2/\alpha)(y-u), & if\ y > u \end{cases}$$

where α = 0.1 for 90% intervals, requiring balance between interval width and coverage accuracy.

1.2 Success Criteria

- Primary: Minimize Winkler Interval Score on test set

- Secondary: Achieve 88-92% coverage rate (close to nominal 90%)

- Tertiary: Maintain competitive ranking in top 50 on leaderboard

2 DATA PREPROCESSING STRATEGY

2.1 Feature Analysis and Engineering Plan

Identified Feature Categories:

- Numerical features: 38 (including price-related, spatial, property characteristics)

- Low cardinality categorical: 3 (join_status, city, submarket)

- High cardinality categorical: 4 (subdivision, zoning, sale_date, sale_warning)

Planned Feature Engineering Pipeline:

1. Domain-specific features:
   - house_age = 2024 - year_built
   - total_rooms = beds + bath_full + bath_3qtr + bath_half
   - land_usage_ratio = sqft / sqft_lot
   - imp_ratio = imp_val / (imp_val + land_val)
   - total_view_score = sum of all view_* features
2. Interaction features (Week 3):
   - sqft × grade (size-quality interaction)
   - latitude × longitude (precise location)
   - land_val × imp_val (total property investment)
3. Temporal features:
   - Extract year, month, season from sale_date
   - Market trend indicators based on sale timing

2.2 Missing Value Strategy

- sale_nbr (21% missing): Median imputation + missing indicator

- subdivision (8.8% missing): "Unknown" category + frequency encoding

- submarket (0.9% missing): "Unknown" category

2.3 Outlier Handling Strategy

Multi-tier approach:

1. Extreme outliers: 1%-99% percentile clipping

2. Target variable: Log transformation to handle right skewness (2.09)

3. Property-specific: Domain knowledge filters (e.g., sqft > 100)

3 MODEL ARCHITECTURE AND EXPERIMENTAL DESIGN

3.1 Three-Stage Modeling Approach

Stage 1: Baseline Models (Completed)

- Random Forest with uncertainty quantification

- LightGBM Quantile Regression (5%, 95% percentiles)

Stage 2: Advanced Models (Week 3)

- Conformalized Quantile Regression (CQR)

- Heteroscedastic Neural Networks

Stage 3: Optimization (Week 3)

- Hyperparameter tuning

- Model selection and ensembling

3.2 Planned Model Implementations

Conformalized Quantile Regression (Priority 1)

Algorithm:

1. Split data: Train (60%), Calibrate (20%), Validation (20%)

2. Train quantile regressors $\hat{q}_{\alpha/2}$, $\hat{q}_{1-\alpha/2}$ on Train

3. Compute conformity scores on Calibrate:

$$E_i = \max(\hat{q}_{\alpha/2}(X_i) - Y_i, Y_i - \hat{q}_{1-\alpha/2}(X_i))$$

4. Find $(1-\alpha)(1+1/n)$-quantile of $\{E_i\}$: $\hat{Q}$

5. Prediction intervals: $[\hat{q}_{\alpha/2}(X) - \hat{Q}, \hat{q}_{1-\alpha/2}(X) + \hat{Q}]$

3.3 Hyperparameter Optimization Plan

LightGBM Quantile Tuning Search space:

- learning_rate: [0.01, 0.03, 0.05, 0.07, 0.1]

- num_leaves: [31, 63, 127, 255]

- feature_fraction: [0.7, 0.8, 0.9, 1.0]

- bagging_fraction: [0.7, 0.8, 0.9, 1.0]

- min_child_samples: [10, 20, 30, 50]

- reg_alpha: [0, 0.1, 0.5, 1.0]

- reg_lambda: [0, 0.1, 0.5, 1.0]

Optimization method: Optuna with 100 trials per quantile

4 VALIDATION FRAMEWORK

4.1 Cross-Validation Strategy

Nested CV approach:

- Outer loop: 5-fold CV for model selection

- Inner loop: 3-fold CV for hyperparameter tuning

- Final validation: 20% holdout for unbiased estimation

4.2 Evaluation Metrics Hierarchy

1. Primary: Winkler Interval Score (competition metric)

2. Coverage diagnostics: Empirical coverage rate

3. Width analysis: Average interval width

4. Calibration plots: Coverage vs. confidence level

5. Quantile diagnostics: Pinball loss for individual quantiles

4.3 Model Selection Criteria

Multi-objective optimization:

- Minimize Winkler score (weight: 0.7)

- Achieve target coverage |coverage - 0.9| (weight: 0.3)

- Pareto frontier analysis for trade-offs