Nepal House Price Prediction

Aryan Malla

24 Feb 2025

Abstract

This report outlines a machine learning approach to predict house prices in Nepal using a dataset comprising features like district, land area, and road access. Due to the dataset's messiness, small size, and age (two years old), prediction accuracy was challenging, further complicated by annual house price fluctuations. An averaging ensemble of multiple models was employed, outperforming hyperparameter tuning, with metrics of Mean Squared Error (MSE) 0.1048, Mean Absolute Error (MAE) 0.2255, Root Mean Squared Error (RMSE) 0.3237, and R^2 Score 0.4188 on a log-transformed target ('Price'). SHAP and LIME visualizations, supplemented by additional EDA plots, provide interpretability and insights into feature importance and data characteristics. This document details the methodology, results, challenges, and deployment instructions.

1 Introduction

The real estate market in Nepal is shaped by factors such as location, property size, and infrastructure access. This project predicts house prices using a dataset from two years ago, featuring variables like district, land area, road access, and bedrooms. However, the data's messiness (e.g., missing values, inconsistencies), small size, and age posed significant hurdles. House prices change annually due to economic shifts and regional developments, rendering the dataset partially outdated. Despite these challenges, a simple averaging ensemble of regression models was developed, with the target variable ('Price') log-transformed to mitigate skewness. Performance metrics were evaluated on this log scale, achieving moderate success. This report elaborates on the methodology, results, and insights, supported by SHAP, LIME, and expanded EDA visualizations.

2 Methodology

2.1 Data Preprocessing

The dataset required substantial preprocessing due to its messy and limited nature:

- **Column Removal**: Dropped 'PARKING', 'BUILDUP AREA', 'AMENITIES', and 'TITLE' due to excessive missing values or low relevance.
- **Unit Conversions**: Converted 'Land Area' to Aana and 'Road Access' to feet.
- **Location Splitting**: Split 'Location' into 'Street' and 'District'.

- **Outlier Handling**: Removed outliers to enhance model stability.
- **Missing Values**: Filled numerical gaps with medians and categorical gaps with "missing."
- **Target Transformation**: Log-transformed ‘Price‘ to normalize its skewed distribution.

2.2 Model Development

The dataset’s constraints, initially explored eight regression models—Random Forest, CatBoost, Linear Regression, XGBoost, Ridge Regression, LightGBM, Decision Tree, and Lasso Regression—but focused on ensemble performance due to poor individual model results on this noisy, small dataset.

Instead, I adopted a simple averaging ensemble, combining predictions from Random Forest, CatBoost, Linear Regression, XGBoost, Ridge Regression, and LightGBM. This approach involved:

- Training each model on the preprocessed dataset using 5-fold cross-validation to ensure robustness across folds, given the small dataset size. - Predicting on the test set and averaging the log-transformed predictions:

$$\text{ensemble}_{\text{pred}} = \frac{\text{rf}_{\text{pred}} + \text{cat}_{\text{pred}} + \text{lr}_{\text{pred}} + \text{xgb}_{\text{pred}} + \text{ridge}_{\text{pred}} + \text{lgb}_{\text{pred}}}{6.0}$$

- Evaluating performance using MSE, MAE, RMSE, and R^2 on the log scale, aligning with the target transformation.

The ensemble outperformed hyperparameter-tuned individual models by reducing variance and mitigating overfitting, leveraging the diversity of tree-based (Random Forest, CatBoost, XGBoost, LightGBM) and linear (Linear Regression, Ridge Regression) models. Tree-based models captured non-linear patterns (e.g., interactions between ‘Land Area (Aana)‘ and ‘Street‘), while linear models stabilized predictions for simpler relationships, but their combined averaging reduced variance more effectively than tuning, which amplified overfitting on the noisy, small dataset.

3 Results

3.1 Ensemble Performance

The ensemble’s performance on the log-transformed target was:

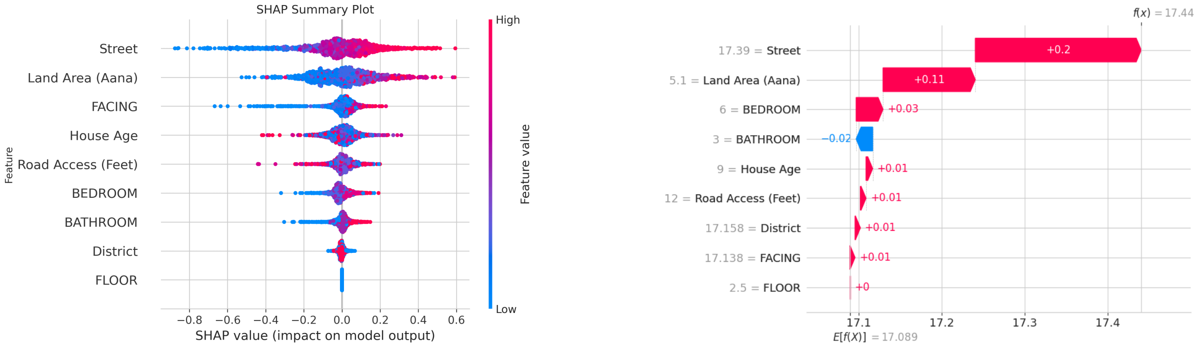
- **MSE**: 0.1048
- **MAE**: 0.2255
- **RMSE**: 0.3237
- **R^2 Score**: 0.4188

An R^2 of 0.4188 indicates that 41.88% of the variance in log-transformed prices is explained, a moderate outcome given the dataset’s limitations.

3.2 Interpretability and EDA

To address the dataset’s messiness and enhance understanding, SHAP, LIME, and additional EDA visualizations were utilized:

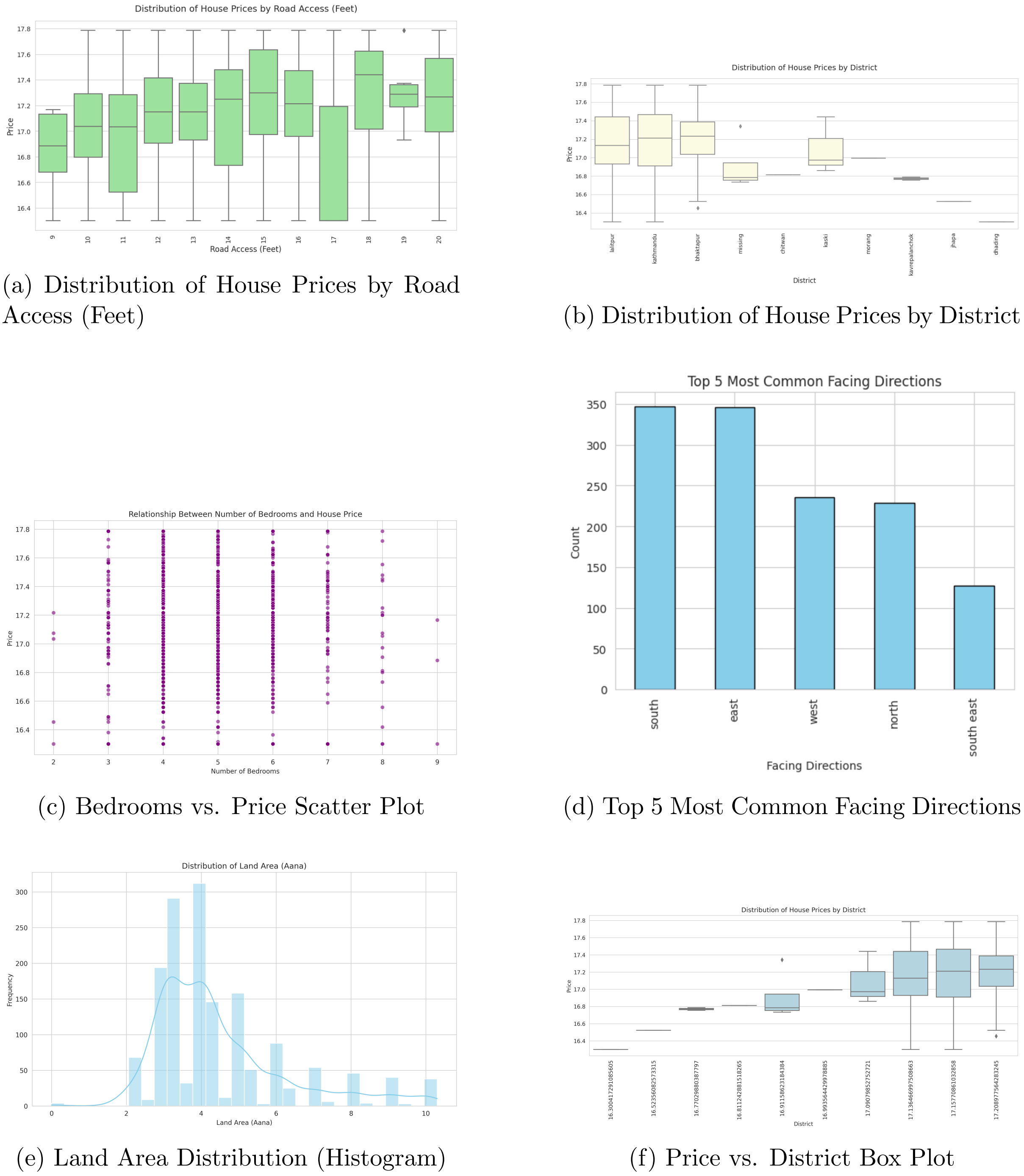
- **SHAP Summary Plot** (Figure 1a): Highlights global feature impacts, e.g., ‘Land Area (Aana)‘ and ‘Street‘.
- **LIME Explanation** (Figure 1b): Offers local explanations for specific predictions.
- **EDA Plots** (Figures 2a to 2f): Include scatter plots, histograms, and box plots to explore data distributions and relationships.



(a) SHAP Summary Plot

(b) LIME Waterfall Plot

Figure 1: Interpretability Visualizations



(a) Distribution of House Prices by Road Access (Feet)

(b) Distribution of House Prices by District

(c) Bedrooms vs. Price Scatter Plot

(d) Top 5 Most Common Facing Directions

(e) Land Area Distribution (Histogram)

(f) Price vs. District Box Plot

Figure 2: Exploratory Data Analysis Visualizations

4 Insights and Discussion

4.1 Key Insights

- **Feature Importance**: SHAP plots identify 'Land Area (Aana)' and 'Street' as dominant predictors, consistent with real estate trends. - **Dataset Challenges**: The data's messiness (e.g., missing values, inconsistencies), small size, and two-year age hindered accuracy. Annual house price changes further reduced its relevance. - **Ensemble Superiority**: Average ensembling outperformed hyperparameter tuning, likely due to its ability to mitigate overfitting in a limited dataset.

The dataset's small size and messiness—evident in scattered EDA plots (e.g., Figures 2a, 2c)—made pattern detection difficult. Its two-year age introduced discrepancies, as house prices fluctuate annually, impacting prediction reliability.

4.2 Trade-offs, Advantages, and Disadvantages

- **Advantages**:

- *Robustness*: Ensembling reduces individual model errors.
- *Interpretability*: SHAP and LIME enhance trust in predictions.
- *Stability*: Log transformation minimizes extreme value effects.

- **Disadvantages**:

- *Moderate Accuracy*: R^2 of 0.4188 suggests underfitting or missing predictors.
- *Data Age*: Two-year-old data misaligns with current prices.
- *Complexity*: Log-scale predictions need back-transformation for practical use.

- **Trade-offs**:

- *Simplicity vs. Precision*: Ensembling is efficient but less precise than tuned models on cleaner data.
- *Data Cleaning vs. Loss*: Removing outliers and columns simplifies analysis but risks losing insights.

5 Implementation

Deploy the model via:

- **Local**: Download the project files, install dependencies with pip install -r requirements.txt, and run streamlit run app.py.
- **Online**: Access the model at `https://nepalhousepriceprediction.streamlit.app/` for instant predictions through a web interface.

The application accepts inputs like district, land area (Aana), road access (feet), bedrooms, and facing direction, returning log-transformed price predictions that require back-transformation for actual values.

6 Conclusion

This project addresses house price prediction in Nepal using a messy, small, two-year-old dataset, challenged by annual price fluctuations. Preprocessing (e.g., unit conversions, outlier removal, log transformation) and an averaging ensemble achieved moderate performance ($R^2 = 0.4188$), with SHAP, LIME, and EDA visualizations offering insights into Land Area (Aana) and Street as key predictors. Despite limitations—dataset size, messiness, and obsolescence—the ensemble outperformed hyperparameter tuning, reflecting its robustness. Future work could incorporate fresher data, advanced ensemble techniques (e.g., stacking), or additional features (e.g., economic indicators) to enhance accuracy and relevance in Nepal's dynamic real estate market.