LMM House Price: Linear Mixed Model for Buyer Biased Advanced House Price Prediction

Yifan Xu
University of Illinois at Urbana-Champaign
yx52@illinois.edu

Nicole Chen
University of Illinois at Urbana-Champaign
yingder2@illinois.edu

Guan-Hong Lin
University of Illinois at Urbana-Champaign
gl48@illinois.edu

Abstract

Accurately predicting house prices is a critical challenge in real estate analytics, as prices are influenced by not only many real estate factors but buyers' preferences. Traditional regression models, such as linear regression, LASSO, and ridge regression, often fail to capture buyer-specific variations, which significantly impact final transaction prices. To address this limitation, we propose a novel **Linear Mixed Model (LMM)-based house price prediction framework** that explicitly incorporates buyer preferences as random effects. Our approach extends the House Prices - Advanced Regression Techniques dataset by simulating buyer-specific attributes, allowing the model to better reflect real-world purchasing behaviors. This study highlights the advantages of LMMs in real estate price modeling and provides a more interpretable, buyer-aware prediction framework. Our findings believe that incorporating buyer preferences into house price prediction models can significantly improve forecasting accuracy, offering valuable insights for both home buyers, sellers and real estate professionals.

1 Introduction

When purchasing a house, a buyer may have only a few specific requirements in mind. However, the actual condition of a house is influenced by numerous factors, and its price is often determined by a complex interplay of these variables. Accurately estimating the value of a house is a critical challenge. A precise price prediction can help buyers secure properties at a fair market price while providing data-driven justifications to sellers regarding their property's value.

House price prediction presents several challenges. One major difficulty is identifying the most influential features among a vast number of variables. The dataset used in this study contains over 80 features, leading to a highly sparse feature matrix, which increases the risk of overfitting. Additionally, different buyers may have varying expectations regarding house prices, introducing noise into the dataset and further complicating the prediction task.

The central problem in this study is how to predict house prices based on a large set of property attributes while effectively selecting the most relevant and impactful predictors from a multitude of irrelevant variables.

Existing methods for house price prediction include linear regression, LASSO regression, and ridge regression. While these methods can partially address the challenges posed by sparse feature matrices, they struggle to capture buyer-specific preferences, which significantly influence final transaction prices. As a result, traditional models fail to fully align with the individualized decision-making process of home buyers.

To address this limitation, this study introduces a Linear Mixed Model (LMM)-based house price prediction framework that explicitly incorporates buyer preferences. We propose a Linear House Price Prediction Model, which leverages data augmentation techniques to further optimize prediction accuracy. Experimental results demonstrate that our approach achieves a higher prediction accuracy compared to conventional methods, effectively improving house price estimation.

2 Related Work

House price prediction has been a widely studied problem in real estate analytics, with various approaches proposed to model the relationship between property attributes and market prices. Traditional statistical models, such as linear regression, have been extensively used for this task. The methods provide interpretable models by assigning weights to different property features. However, their performance is often limited by the high-dimensional and sparse nature of real estate datasets, leading to overfitting and suboptimal predictions.

To mitigate issues arising from sparse feature matrices, regularization techniques such as LASSO[5] and ridge regression[1] have been explored . LASSO regression enforces sparsity by penalizing the absolute sum of coefficients, effectively selecting a subset of relevant features, while ridge regression applies an L2 penalty to reduce overfitting. Despite their advantages, these methods primarily focus on the intrinsic attributes of houses and fail to account for external factors such as buyer-specific preferences and market dynamics.

Recent research has acknowledged the importance of buyer preferences in house price prediction. Traditional models treat house prices as a function of structural and locational attributes, overlooking the fact that final transaction prices are highly dependent on individual buyer decisions. Some studies have attempted to incorporate external economic indicators or behavioral data, but these approaches remain limited in modeling buyer-specific variations[3].

To address these limitations, Linear Mixed Models (LMMs) have been introduced in various domains, offering a powerful approach to handle hierarchical and grouped data[2]. However, their application in house price prediction remains underexplored. By incorporating buyer preferences as random effects, LMMs can better model the variability in final sale prices. Furthermore, data augmentation techniques have been successfully applied in various machine learning tasks to enhance generalization and mitigate overfitting[4]. The integration of LMMs with data augmentation provides a promising direction for improving house price prediction models.

This study builds upon these existing methodologies by introducing a Linear House Price Prediction Model, which integrates Linear Mixed Models to account for buyer-specific preferences and uses data augmentation to enhance robustness and predictive performance. Bridging the gap between traditional regression models and real-world buyer behavior, our approach aims to provide a more accurate and interpretable framework for house price prediction.

3 Method

In this study, we apply Linear Mixed Models (LMMs) to predict house prices using the House Prices - Advanced Regression Techniques dataset. Traditional regression models often struggle with sparse feature matrices and do not account for buyer-specific variations, which significantly influence final transaction prices. By leveraging LMMs, a key concept learned in this course, we incorporate both fixed effects (house attributes) and random effects (buyer preferences and neighborhood influences) to improve prediction accuracy. Additionally, we extend the dataset by simulating buyer preference data, allowing us to capture heterogeneous decision-making patterns in real estate transactions.

3.1 Model Formulation

A standard linear regression model assumes a deterministic relationship between input features and the target variable:

$$y = X\beta + \epsilon$$

where y denotes house price and X is feature matrix. ϵ is random noise. In order to find a regression coefficient β with

$$\hat{\beta} = \underset{\beta}{\operatorname{argmin}} \|\mathbf{y} - \mathbf{X}\beta\|_2^2$$

However, house prices are influenced not only by property attributes but also by buyer preferences and neighborhood-specific factors. To account for these effects, we incorporate random effects into the model:

$$y = X\beta + Zu + \epsilon, \quad u \sim N(0, I_z\sigma_u^2), \quad \epsilon \sim N(0, I\sigma_\epsilon^2)$$

where y denotes house price and X is feature matrix. ϵ is random noise. Z represents design matrix for buyer's preference and u is buyer's preference and neighborhood influences. By incorporating random effects, LMM allows us to model group-dependent variations that traditional regression models fail to capture.

3.2 Dataset Expansion: Incorporating Buyer Preferences

The House Prices - Advanced Regression Techniques dataset consists of 1460 observations and 79 features, including structural attributes (e.g., OverallQual, GrLivArea, GarageCars, TotalBsmtSF), location attributes (e.g., Neighborhood, MSZoning), market factors (e.g., YearBuilt, SaleCondition). However, the dataset does not contain buyer-specific information, which is a crucial factor in real-world transactions.

To extend the dataset, we simulate buyer preference data, assigning each transaction to a hypothetical buyer ID. Buyer preferences are modeled using size preference and New vs. Old Preference. For Size Preference, Buyers may prefer large or small houses. For New vs. Old Preference, some buyers favor newer homes, while others seek vintage properties. We introduce a categorical variable BuyerID, where each buyer ID represents a unique preference pattern. This allows us to capture individual variations in valuation through the random effect u_Buyer in our LMM.

References

[1] Arthur E Hoerl and Robert W Kennard. Ridge regression: Biased estimation for nonorthogonal problems. *Technometrics*, 12(2):55–67, 1970.

[2] José C Pinheiro and Douglas M Bates. Linear mixed-effects models: basic concepts and examples. *Mixed-effects models in S and S-Plus*, pages 3–56, 2000.

[3] R Aswin Rahadi, Sudarso Kaderi Wiryono, Deddy P Koesrindartoto, and Indra Budiman Syamwil. Relationship between consumer preferences and value propositions: a study of residential product. *Procedia-Social and Behavioral Sciences*, 50:865–874, 2012.

[4] Connor Shorten and Taghi M Khoshgoftaar. A survey on image data augmentation for deep learning. *Journal of big data*, 6(1):1–48, 2019.

[5] Robert Tibshirani. Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 58(1):267–288, 1996.