Hierarchical Bayesian Modeling for Real Estate Valuation

AN INTELLIGENT APPROACH TO AUTOMATED VALUATION MODELS (AVMS) USING THE AMES HOUSING DATASET

Faheem Arif

MA23BTECH11010

Table of Contents

3
The Problem

4
The Solution

5
Project Goals

6
The Dataset: Ames

7
Intermediate Results: Data Preparation

8
Steps Ahead: EDA & Modeling Phase

9
Summary & Impact

MA23BTECH11010

The Problem: Standard Automated Valuation Models are Brittle

1

Challenge: How do we predict house prices reliably, especially in neighborhoods with sparse data?

2

Approach 1: Complete Pooling (e.g., one big Lasso/Ridge model)

- How it works: Treats all neighborhoods the same. Learns a single, city-wide effect for every feature.
- Problem: Ignores that a fireplace might be worth more in an upscale neighborhood than a starter-home area. It's too simplistic.

3

Approach 2: No Pooling (e.g., one model per neighborhood)

- How it works: Builds a completely independent model for each of the 25 neighborhoods.
- Problem: If a neighborhood has only 5 sales, the model's estimates will be wildly unstable and untrustworthy. It fails on sparse data.

The Solution: A Hierarchical Bayesian Model

Core Idea: Partial Pooling

- The model learns a city-wide average for price effects.
- It treats each neighborhood's effect as a variation from that average.
- It "borrows statistical strength" from data-rich neighborhoods to improve estimates for data-sparse ones.

The Outcome:

- Stable: Estimates for sparse-data neighborhoods are "shrunk" towards the city-wide mean, preventing extreme predictions.
- Honest: Provides a full probability distribution for each prediction, directly quantifying uncertainty.

Project Goals

1. To construct a **hierarchical regression model** for house valuations where individual neighborhood and global information is used jointly.
2. To generate not just point-estimates, but full **predicted posterior distributions** giving a more honest perspective of the valuation uncertainty.
3. To compare insights and interpretability of insights using standard techniques (like **Lasso / Ridge regression** and foundational **Bayesian Regression** techniques).

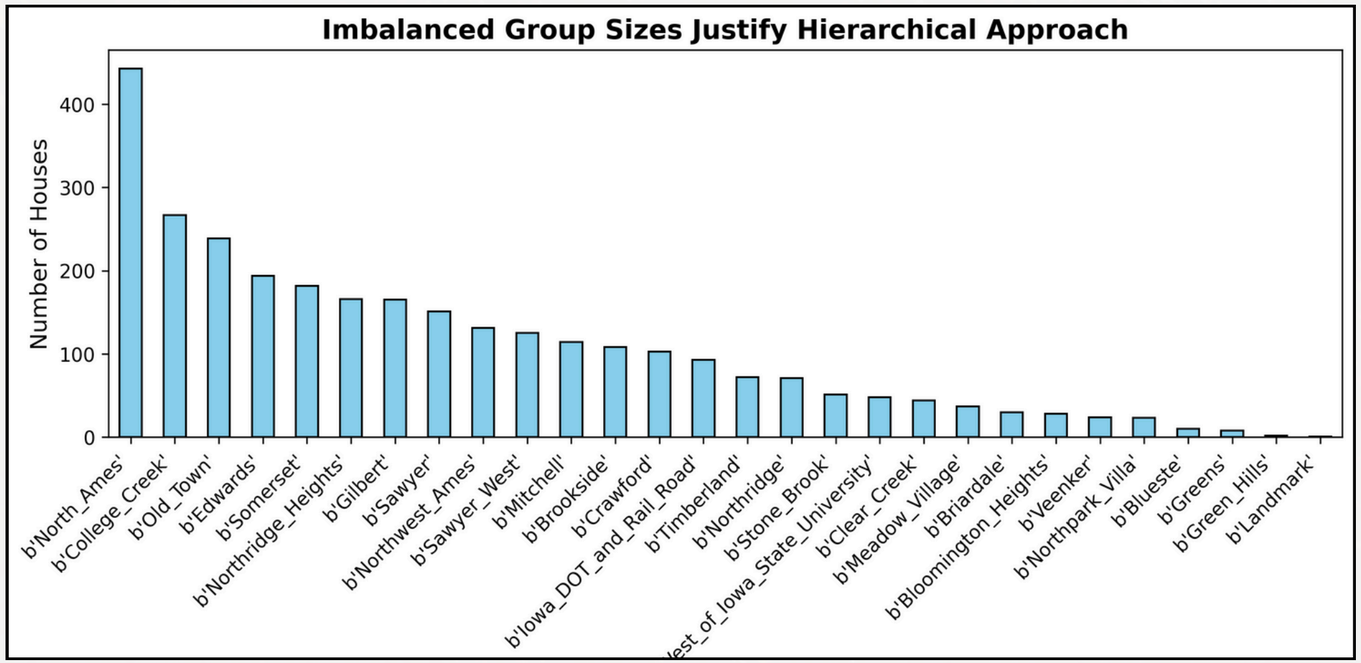
The Dataset: Ames, Iowa (2006-2010)

Size: 2930 property sales.

Features: 79 explanatory variables (e.g., Pool Count).

Key Hierarchical Variable:
Neighborhood

- Contains 25 distinct neighborhood groups.
- **Crucially, the number of sales per neighborhood is highly imbalanced.**



Intermediate Results: Data Cleaning & Preparation

The first major phase was a rigorous data cleaning process. The raw data contained 'hidden' missing values as placeholder strings, which required a systematic approach to standardize and clean

Handled "Meaningful Absence"

- For some columns (e.g., Alley, Pool_QC), NaN means "No Feature." These were correctly filled with a 'None' category.

Standardized Missingness:

- Replaced all placeholders with numpy.nan to enable standard processing.

Imputed True Missing Data:

- For columns like Lot_Frontage, NaN means unknown data. These were imputed using the neighborhood-level median.

Corrected Data Types:

- For columns like Lot_Frontage, NaN means unknown data. These were imputed using the neighborhood-level median.

Steps Ahead: EDA & Modeling Phase

Exploratory Data Analysis:

- On a cursory look at the data, we noticed some skewness in the house price data. I am considering some transformations (taking logarithms of all prices).

- Further, we need to find powerful predictive features like *Overall_Qual*, *Gr_Liv_Area*, *Total_Bsmt_SF*, and *Garage_Cars* which show the strongest linear relationship with log-price.

Modeling Phase:

- **Baselines**:
  - Model 1 (Complete Pooling): Ridge/Lasso Regression on the entire dataset.
  - Model 2 (No Pooling): A separate Ridge model for each neighborhood.
- **Simple Bayesian:**
  - Replicate the pooling and no-pooling models in PyMC to build a Bayesian baseline.
- **The Hierarchical Model:**
  - Model 3 (Partial Pooling): Construct the final hierarchical model.
  - Will handle posterior correlation with the NUTS sampler.