Project Proposal

Benjamin Kodi, Dafne Arreola, Jada Duggins, Alan Cedeno

Bias in AI-Driven Housing Price Predictions: A Case Study of the New York Metropolitan Area

1 Introduction

Real estate pricing is one of the many data analysis and prediction tasks that machine learning algorithms are employed for. However, these algorithms frequently provide biased predictions because of biased training data, algorithmic design, and historical socioeconomic inequities. If the training data contains biases such as gentrification, income segregation, or redlining, the model will perpetuate them, impacting real estate patterns and strengthening systemic disparities.

Zillow's Zestimate is a well-known example of AI-driven real estate pricing since it estimates house prices based on past sales data, geographical trends, and property qualities. While such models provide valuable insights, they may exacerbate housing inequalities by overvaluing or undervaluing properties in specific neighborhoods, particularly those with historically marginalized populations. This study investigates how AI-based housing price models inflate or deflate values in different New York Metropolitan Area demographic areas.

This study examines Zillow real estate data, U.S. Census demographics, and Airbnb rental market trends to identify bias tendencies in AI-generated pricing projections. By pointing out differences between AI-generated property valuations and real market patterns, we hope to contribute to the larger conversation on algorithmic bias in housing markets.

2 Research Questions

This study explores the following key research questions:

1. Due to AI-driven pricing models, are housing prices inflated more in some demographic regions?

2. How do algorithmic predictions compare to actual market trends in diverse socioeconomic neighborhoods?

3. What demographic factors (e.g., income, race, rental patterns) correlate most strongly with discrepancies in AI-predicted versus actual home values?

4. How does short-term rental activity (Airbnb) influence AI-driven New York Metro Area price estimates?

3 Data Description

We collected information from many sources, each providing a distinct perspective on NYC real estate, demographics, and short-term rentals, in order to investigate these concerns. We may investigate biases in AI pricing forecasts and their effects on society with the use of these datasets.

3.1 Zillow Real Estate Data

- **Source:** Zillow Research Data (Zillow.com)
- **Description:** Provides historical and current home value and rental price estimates from Zillow's Home Value Index (ZHVI) and Observed Rent Index (ZORI).
- **Key Variables:**
  - Median Home Price (ZHVI)
  - Zillow Observed Rent Index (ZORI)
  - Market Temperature Index (measuring demand vs. supply)
  - Pending Sales & Inventory Levels
- **Timeframe:** Monthly data from 2015–2025.
- **Purpose:** Analyze how AI-driven price estimates compare across NYC boroughs and neighborhoods.

3.2 U.S. Census Data (American Community Survey - ACS)

- **Source:** U.S. Census Bureau (census.gov)
- **Description:** Provides demographic, economic, and housing-related statistics for the New York Metropolitan Area.
- **Key Variables:**
  - Median Household Income
  - Racial and Ethnic Composition (%)
  - Owner-Occupied vs. Renter-Occupied Housing (%)

  - Educational Attainment Levels

- **Timeframe:** Latest available data from 2023–2024.
- **Purpose:** Examine the relationship between neighborhood demographics and AI-predicted housing prices.

3.3 Inside Airbnb Data

- **Source:** Inside Airbnb (insideairbnb.com)
- **Description:** Details Airbnb short-term rental activity in NYC, including pricing, availability, and host characteristics.
- **Key Variables:**
  - Average Listing Price
  - Neighborhood and Zip Code
  - Host Type (Individual vs. Multi-Listing Host)
  - Occupancy Rate (%)
- **Timeframe:** 2023–2024 Airbnb listings.
- **Purpose:** Evaluate the role of Airbnb-driven gentrification in shaping AI-driven housing prices.

4 Methodology and Analysis

We will combine descriptive statistics, inferential analysis, and machine learning techniques to analyze the impact of AI-driven pricing models on housing prices. Our approach consists of many steps. This will include data cleaning and preprocessing, which will handle missing values using imputation techniques. It will also standardize and normalize numerical data to improve comparability.

Another stage is to perform exploratory data analysis (EDA). This will create summary data such as mean, median, and standard deviation to help understand how home prices vary among demographics. We will also conduct correlation analysis to identify relationships between housing prices and demographic factors such as income, race, and education level.

Another step is bias detection in AI-driven pricing models. We will apply fairness metrics and compare model predictions across different demographic groups to identify pricing discrepancies. We will also conduct a comparative analysis of different datasets. We will evaluate how pricing trends differ between Zillow's Zestimate, U.S. Census Housing Data, and the Inside Airbnb Dataset. We will identify any inconsistencies between real estate market prices and AI-generated estimates.

Additionally, we will use predictive modeling and scenario analysis. This will simulate alternative scenarios where biased data is adjusted to observe potential price corrections. We will also implement regression models to estimate the impact of demographic variables on housing prices.

After these analyses, we will summarize findings on biases in AI-driven pricing models and discuss potential regulatory or algorithmic adjustments to correct unfair pricing. We will provide insights on how stakeholders can interpret and address bias in housing price predictions.

5 Expected Findings and Implications

We anticipate that our examination of multiple datasets will reveal that AI-driven pricing models do not consistently raise housing costs in every population. Discrepancies are expected due to variations in AI pricing models, training data, and historical pricing patterns within communities. Our research will highlight trends emphasizing the complex relationship between AI pricing and housing affordability.

6 Conclusion

The increasing use of AI-driven real estate pricing models necessitates a thorough examination of biases embedded in these algorithms. This study will provide empirical evidence of how these biases manifest in the New York Metropolitan Area and offer data-driven recommendations to promote greater accountability and transparency in AI-generated price estimates.

7 Group Agreement and Work Contract

Through much discussion, we have collectively decided that each group member will make time to meet regularly to discuss individual progress. We will delegate a fair amount of work to each group member and set deadlines for each task so that we can maintain punctuality and efficiency. If a team member cannot keep up with their portion of delegated work due to valid reasons, other group members will assist in any way that they can. Lastly, although work only needs to be finished before deadlines, we will encourage an environment of getting tasks done sooner rather than later.