A Machine Learning Approach to the Analysis of Real Estate Prices in the New York Metropolitan Area

Jacky He, Zhiduo Xie, Shou-Kai Cheng, Yibei Li, Yuchuan Zhang
Cornell University
Department of Information Science and Cornell Tech
ph474@cornell.edu, zx324@cornell.edu, sc2745@cornell.edu, yl3692@cornell.edu, yz2947@cornell.edu

Abstract

This project addresses the critical challenge of accurately predicting real estate prices in the New York Metropolitan Area (NYC), a crucial task for market transparency and aiding decisions in the dynamic NYC market. Employing a dataset with detailed property features from Kaggle, we apply and compare three state-of-the-art machine learning models—Linear Regression, Random Forest, and Gradient Boosting—to identify the most effective approach. The system's architecture includes a user-friendly front-end for parameter input and a robust back-end for model processing. Models are evaluated using RMSE, MAE and R², with the best-performing model deployed for tailored price estimations. This comparative analysis not only provides insights into the most effective predictive models but also advances machine learning applications in real estate, offering methodologies that could influence future predictive modeling.

1. Introduction

1.1. Motivation

The task of predicting real estate prices is crucial in markets like New York City (NYC), where complexity and dynamism can challenge both buyers and sellers. Accurate price predictions enhance market transparency and facilitate informed decision-making. This project develops an application aimed at predicting housing prices using advanced machine learning techniques, leveraging a comprehensive dataset with detailed property features.

1.2. Technical Focus

We employ three sophisticated machine learning algorithms—Linear Regression, Random Forest Regression, and Gradient Boosting Regression. These were chosen for their efficacy in handling large and complex datasets typical of NYC real estate data.

1.3. Prior Work

Previous research has often focused on individual machine learning models without extensive comparative analyses. Studies have explored combining traditional and non-traditional data to enhance prediction accuracy. Our work builds upon these foundations by comparing multiple models to determine the most effective predictions in the NYC real estate market.

1.4. Machine Learning Pipeline

The project pipeline includes initial data exploration to identify and visualize key features influencing property prices, followed by data preprocessing, model training, and evaluation using RMSE and R². The final model is deployed in a user-friendly application, enabling customized property price estimations.

1.5. Impact

The social impacts of this application are significant, as it aims to democratize access to real estate data and pricing knowledge, fostering a more transparent and fair market. Ethical considerations include data privacy and avoiding the reinforcement of existing biases in predictive models.

2. Background

Prior research in the realm of housing price prediction has predominantly utilized simpler statistical models such as linear regression to determine the impact of various property-related features on market values. While effective, these models often lack the complexity needed to fully capture the multifaceted nature of real estate markets. Recent advancements have expanded the scope of variables considered, incorporating not only physical attributes of properties but also environmental and socioeconomic factors. More sophisticated algorithms, including Ridge Regression, Support Vector Machines (SVM), and ensemble methods like Gradient Boosting, have been adopted to better account for these complexities.

For instance, Park and Bae's study [2] assesses the efficacy of several machine learning algorithms—including C4.5, RIPPER, Naïve Bayesian, and AdaBoost—for predicting housing prices with a comprehensive dataset from Fairfax County, Virginia. Their findings underscore the heightened accuracy of RIPPER over the other models tested.

Another significant contribution is the hybrid approach that combines Lasso regression with Gradient Boosting models, enhancing predictive accuracy in real estate price modeling,as discussed by Lu al. (2017) [1].

In our work, we aim to build upon these foundations by employing a comparative analysis of advanced machine learning models, specifically Linear Regression, Random Forest, and Gradient Boosting. This approach is designed to leverage the distinct strengths of each model to more accurately predict housing prices in the highly dynamic and complex New York City real estate market.

Our anticipated approach contrasts with earlier methods by focusing on a direct comparison and evaluation of multiple models to identify which most effectively captures the nuances of NYC's market.

3. End-to-End ML Pipeline

3.1. Back-End

3.1.1 Data Collection, Exploration & Processing

The dataset for this study [3], sourced from Kaggle, encompasses a total of 4,801 entries. It includes diverse features pivotal for housing price prediction:

- **Price:** Numerical value indicating the sale price of the property.
- **Postal (Zip code):** Categorical data serving as location identifiers.
- **Square Feet (area):** Numerical value representing the size of the property.
- **Room:** Numerical count of rooms, including bedrooms and bathrooms.
- **Type:** Categorical data indicating the type of property, such as House, Condo, Co-op, Multi-Family, Town House, and Other.

The dataset provides a solid foundation for applying machine learning techniques due to its comprehensive set of ground truth labels for property prices.

Data visualization techniques employed include scatterplots to assess relationships between price and square footage, and histograms and boxplots to evaluate the price distribution across different zip codes.

Our preprocessing approach includes:

- **Handling Missing Data:** Imputation or removal of entries ensures completeness of the dataset.
- **Encoding Categorical Features:** One-hot encoding transforms categorical data into a numerical format required by ML algorithms.
- **Normalizing Continuous Features:** Standardization of features like square footage normalizes data to zero mean and unit variance, aiding in algorithm performance.
- **Removing Price Outliers:** Outliers are removed to prevent model skew and improve accuracy.
- **Reducing Dimensionality:** PCA is applied when necessary to reduce feature space and prevent overfitting.

These preprocessing steps are designed to optimize data quality and relevance, ensuring robust model training and evaluation.

3.1.2 Methods and Model Training

We apply Linear Regression, Random Forest, and Gradient Boosting due to their effectiveness in handling the complex, non-linear relationships typical in housing data. We anticipate that ensemble methods like Random Forest and Gradient Boosting will outperform Linear Regression due to their robustness against overfitting and ability to handle heterogeneous data.

The model inputs will include features such as type, size, and location of properties, while the output will be the estimated market price of each property.

The data will be divided into an 80% training set and a 20% testing set. We will employ k-fold cross-validation within the training set to optimize model parameters and prevent data leakage.

Parameters for Random Forest will include the number of trees and maximum tree depth. For Gradient Boosting, we will tune the learning rate and the number of estimators. Regularization techniques will be used in Linear Regression to avoid overfitting.

3.1.3 Model Evaluation

The models will be evaluated using the following metrics:

- **Root Mean Squared Error (RMSE)**: Measures the average magnitude of the errors, providing a clear indication of model accuracy.
- **Mean Absolute Error (MAE)**: Offers an average of the absolute errors, advantageous for its robustness to outliers.

- **R-squared (R²)**: Represents the proportion of variance in the dependent variable predictable from the independent variables, useful for assessing the explanatory power of the model.

These metrics are standard in regression analysis and will help in assessing the models comprehensively, including their tendencies toward over- or underfitting.

Experiments will include hyperparameter tuning, variations in feature sets, and different training configurations. Detailed documentation of performance on training and test datasets, learning curves, and error analysis will ensure reproducibility and provide clear insights into model behavior.

3.1.4 Model Deployment

The web application will allow users to input property details and receive price estimates. This interface will be designed for ease of use to cater to a broad user base, including non-technical individuals.

The application aims to enhance transparency in the real estate market and support informed decision-making. We will address ethical concerns by implementing robust data privacy measures and regularly updating the model to handle biases and ensure accuracy.

3.2. Front-End (Streamlit)

3.2.1 Home Page

Quick Start Guide: Instructions on how to use the application to get housing price estimates. **User Inputs:** Features a file uploader where users can upload their dataset. Once a file is uploaded, a preview of the data is displayed.

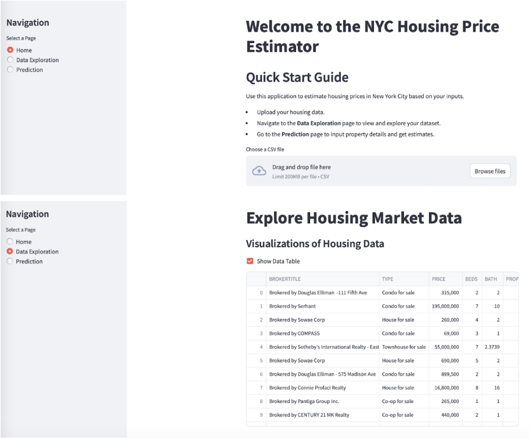
3.2.2 Data Exploration Page

Interactive Charts: Users can see the preview of the housing data as shown in Figure 1
Map Views: Users can see the map views of the housing price in NYC. Demo in Figure 2

3.2.3 Prediction Page

Input Form: Users input property details through dropdown menus and sliders.
Submit Button: Triggers the prediction model.
Results Display: Shows the predicted housing price.

Menus and Input Widgets include a Navigation Bar at the top of page,and various input widgets such as dropdowns and sliders. Data Visualizations are powered by libraries like Plotly, integrated into Streamlit. Demo in Figure 3



Welcome to the NYC Housing Price Estimator

Quick Start Guide

Use this application to estimate housing prices in New York City based on your inputs.

- Upload your housing data.
- Navigate to the **Data Exploration** page to view and explore your dataset.
- Go to the **Prediction** page to input property details and get estimates.

Choose a CSV file

Drag and drop file here
Limit 200MB per file • CSV

Explore Housing Market Data

Visualizations of Housing Data

☑ Show Data Table

	BROKERTITLE	TYPE	PRICE	BEDS	BATH	PROP
0	Brokered by Douglas Elliman -111 Fifth Ave	Condo for sale	315,000	2	2	
1	Brokered by Serhant	Condo for sale	195,000,000	7	10	
2	Brokered by Sowae Corp	House for sale	260,000	4	2	
3	Brokered by COMPASS	Condo for sale	69,000	3	1	
4	Brokered by Sotheby's International Realty - East	Townhouse for sale	55,000,000	7	2.3739	
5	Brokered by Sowae Corp	House for sale	690,000	5	2	
6	Brokered by Douglas Elliman - 575 Madison Ave	Condo for sale	899,500	2	2	
7	Brokered by Connie Profaci Realty	House for sale	16,800,000	8	16	
8	Brokered by Pantiga Group Inc.	Co-op for sale	265,000	1	1	
9	Brokered by CENTURY 21 MK Realty	Co-op for sale	440,000	2	1	

Figure 1: Schematic diagram of the web application's user interface layout.

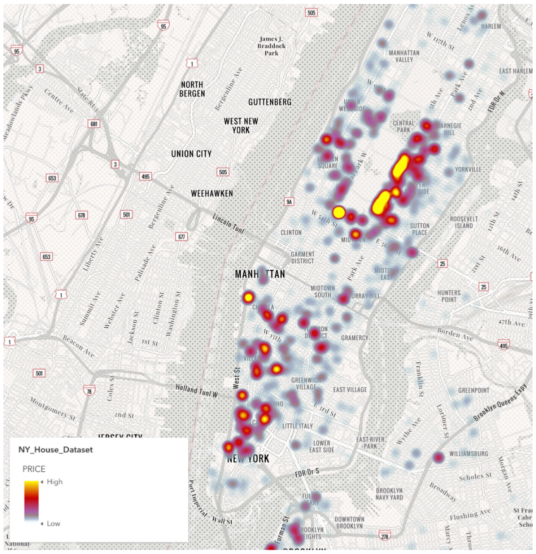


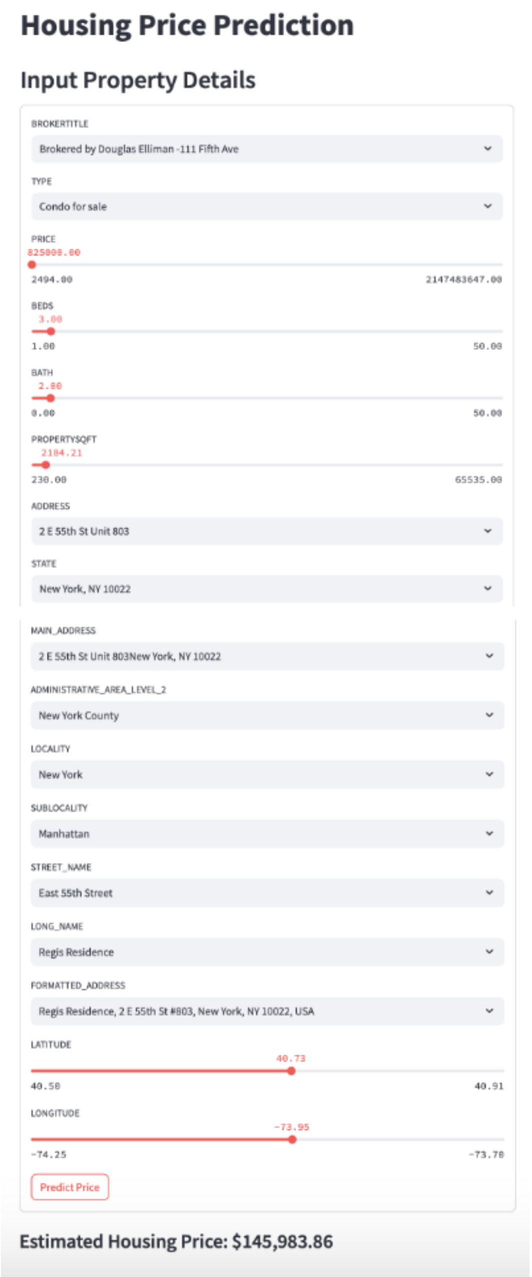
Figure 2: Map view of the housing price distribution across New York City.

3.2.4 Integration of Front- and Back-End

Model training is triggered by user actions, such as submitting the input form. The results will be shown based on user selections, providing an engaging user experience.

3.2.5 Design Considerations

The UI is responsive and accessible, with high contrast and easy navigation options to accommodate users with visual impairments.



Housing Price Prediction

Input Property Details

BROKERTITLE
Brokered by Douglas Elliman -111 Fifth Ave

TYPE
Condo for sale

PRICE
825000.00
2494.00 2147483647.00

BEDS
3.00
1.00 50.00

BATH
2.00
0.00 50.00

PROPERTYSQFT
2184.21
230.00 65535.00

ADDRESS
2 E 55th St Unit 803

STATE
New York, NY 10022

MAIN_ADDRESS
2 E 55th St Unit 803New York, NY 10022

ADMINISTRATIVE_AREA_LEVEL_2
New York County

LOCALITY
New York

SUBLOCALITY
Manhattan

STREET_NAME
East 55th Street

LONG_NAME
Regis Residence

FORMATTED_ADDRESS
Regis Residence, 2 E 55th St #803, New York, NY 10022, USA

LATITUDE
40.73
40.50 40.91

LONGITUDE
-73.95
-74.25 -73.70

Estimated Housing Price: $145,983.86

Figure 3: Schematic diagram of the web application's user interface layout.

4. Risk & Mitigation

Data Quality and Completeness: Reliable, comprehensive data is essential. We'll source credible data and utilize robust preprocessing to ensure accuracy.
Model Overfitting: To avoid overfitting, we'll employ cross-validation and regularization, and set aside part of our dataset for final testing.

5. Expected Outcomes

Well-performed Model: By implementing and fine-tuning the three models, we can achieve well-performing models.
Higher Accuracy: Using advanced machine learning, we anticipate more accurate housing price predictions.
Market Insights: The model will provide deeper insights into market dynamics, improving understanding of price influences.

6. Team Member Contribution

6.1. Technical Components

Front-end: **Yuchuan** Responsible for data visualization. This involves creating interactive graphics and charts to display the data, allowing users to understand trends and insights visually.
Back-end: **Yibei** Handles data processing and Arcgis. This includes cleaning, organizing, and preparing data for analysis, ensuring that the data input into the machine learning models is accurate and structured. **Zhiduo, Jacky and Shou-Kai** Collaboratively work on ML model training. They are tasked with developing, training, and refining machine learning algorithms that the project relies on for data analysis and predictions.

6.2. Writing Components

Introduction and Data Description: Yibei and Shou-Kai
Data Visualization: Yuchuan
Machine Learning Models: Zhiduo and Jacky
Results, Discussion and Future Work: A collaborative effort from all team members

References

[1] Sifei Lu, Zengxiang Li, Zheng Qin, XuLei Yang, and Rick Goh. A hybrid regression technique for house prices prediction. pages 319–323, 12 2017. 2

[2] Bobae Park and Jae Kook Bae. Using machine learning algorithms for housing price prediction: The case of fairfax county, virginia housing data. *Expert Systems with Applications*, 42(6):2928–2934, 2015. 2

[3] Nelgiriye Withana. New york housing market dataset. https://www.kaggle.com/datasets/nelgiriyewithana/new-york-housing-market, 2021. 2