Housing Price Prediction Model

Jesus Perez, Georgios Filippos Bogdos, Gideon Koech

Abstract

This research focuses on utilizing Artificial Intelligence (AI) techniques to develop a model for accurately estimating the median value of homes in California. The study addresses the gap in research related to AI-driven approaches and user engagement in home price estimation. The findings have implications for various stakeholders, including banks, brokers, and individuals seeking informed decision-making in the real estate market. The results of our analysis indicate that the Extreme Gradient Boosting model achieved the highest R-squared value of 0.84, followed by the Random Forest Regression model with an R-squared value of 0.82. The Linear Regression model and Lasso regression model achieved R-squared values of 0.68 and 0.67, respectively. These findings demonstrate the effectiveness of AI techniques in improving the accuracy of housing price predictions.

Problem Statement

The primary purpose of this project was to evaluate how well Artificial Intelligence methods perform in comparison to one another. The key research question however can be summarized as follows:

By using the dataset supplied, construct a model of housing prices that can accurately estimate the median value of a home in the state of California.

Introduction

The housing market has drawn a lot of academic research during the last few decades. According to Shiller (2005), the reason real estate is so well-liked is because it offers both a place to live and a means of income. The housing business, which is inextricably entwined with the financial sector, is very important to the economy. A fall in the housing market frequently has significant effects and has the ability to start recessions and economic crises. Several countries have seen significant price changes throughout history, which were often preceded by an all-time high and then followed by a sharp decrease. Price changes have a significant impact on a household's well-being, business cycles, and financial stability. Due to these results, price movements may be used to produce indicators for financial regulatory organizations, central banks, and other economic stakeholders (Rosen, 1974).

Artificial Intelligence (AI) uses a variety of technological tools and mathematical procedures to extract information from data. These tools are always applicable in scenarios where it would be very challenging to manually review massive amounts of data. Depending on their location, houses can have a wide range of features and price points. For instance, if a huge property is located in a wealthy, attractive neighborhood as opposed to a one in a lower income neighborhood one, the price may be greater. To increase the precision of the predictions generated, a variety of pre-processing techniques will be used on the data gathered from the experiment.

Background

Estimating market prices for homes is challenging due to various influencing factors such as market power, quality, advertising, and brand recognition. The "hedonic model," proposed by Lancaster in 1966, suggests that homes consist of different features that consumers consider when making purchases, and altering these attributes does not significantly impact the product or price prediction algorithms. Indicators of supply and demand, along with affordability, further contribute to price estimation (Rosen, 1974).

Following the global financial crisis of 2008, the real estate market experienced a prolonged decline, particularly in major cities, until the end of 2011. Since 2012, the housing market has shown an upward trend attributed to a decreasing supply of available homes, increased demand, and subsequent price escalation (Soni & Goyal, 2018). This has prompted economists and market analysts to prioritize the development of more accurate forecasting models to safeguard against future economic downturns (Park & Kwon Bae, 2015).

Research Relevance

In the real estate industry, predicting a market's price is not a novel concept. For many people who are involved in or impacted by the real estate market, determining the value of a piece of property is crucial. The bank must ascertain the cost of the property before a customer purchase it. In order for the seller to receive the maximum amount of money, real estate agents must set the appropriate price. Finally, private individuals want to know the value of their home or a potential new home so they may make more informed decisions about when to sell or buy. Businesses and organizations can employ Artificial Intelligence to derive value and information from data. The utilization of AI may help these actors perform better. AI techniques can also assist banks, brokers, and individuals in making better decisions when it comes to anticipating real estate values (Park & Kwon Bae, 2015).

Literature Review

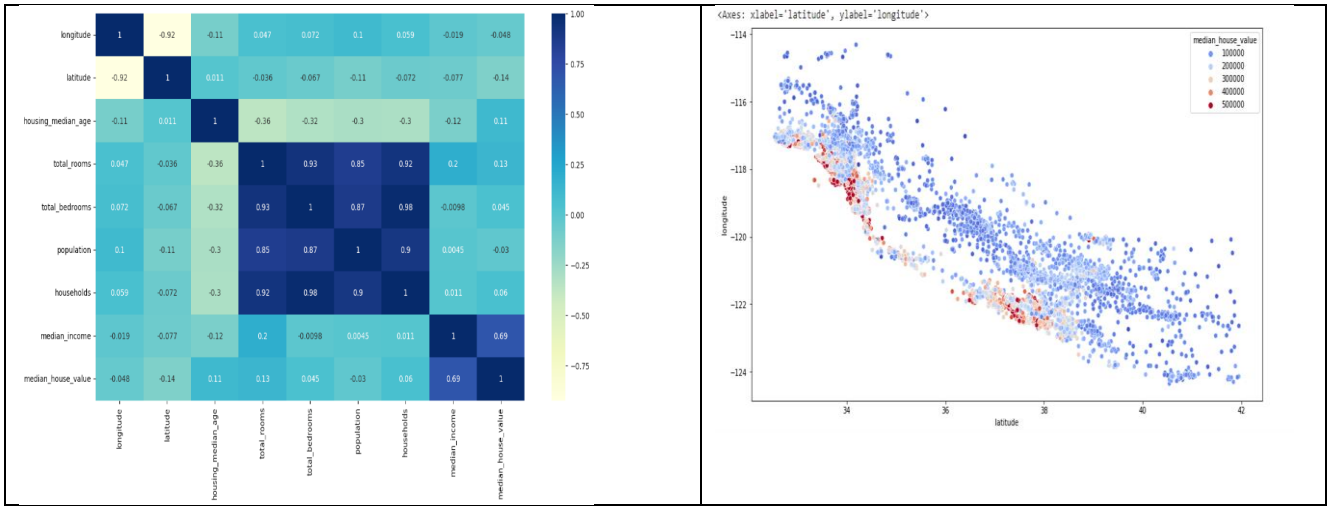
Several related works that are important for fitting appropriate models to Housing Data have been discussed in this section. The report adds to a limited but rising body of research on Artificial Intelligence in the housing market. The body of research on predicting house prices is extensive and includes both traditional classification, regression, and autoregressive models. Only relevant machine learning literature is presented in this area.

The research study conducted by (Martin, 2011; Baker, 2008; Acharya & Richardson, 2009) helps to understand the relevancy of anticipating changes in the house prices considering the fall of the home price bubble, that started the financial crisis in the United States in 2007. These findings emphasize the significance of identifying the early warning signs of large economic changes developing over time and demonstrate the significant effects of volatility or shocks in the housing market on actual economic activity. Other important economic factors have an impact on and have a direct impact on house price predictions. Other significant economic phenomena and economic systems are directly impacted by the capacity to forecast home prices.

Dataset

The California house price dataset was used to validate our methods in this study. The US Census Bureau has released Census Data for California, which has 20640 records. The sample dataset contains 10 distinct metrics for each Californian block group, such as population, median income, and median housing price. The median house value attribute of the dataset will be predicted utilizing the various features as independent variables. The dataset contains the following fields or variables: Longitude, Latitude, Housing Age Median, Total Rooms, Count of Bedrooms, Population, Households, Ocean proximity, Median Income and Median House Value.

Dataset Visualization



Methods of Prediction

- Linear Regression: Describes the relationship between variables by fitting a line to the data that has been gathered.
- Random Forest: Supervised learning technique that utilizes the ensemble learning approach to perform regression.
- LSTM: A type of recurrent neural network capable of learning order dependence in sequence prediction problems.
- XGBoost: XGBoost is an implementation of gradient boosted decision trees designed for speed and performance.
- XAi: Creates a system that is more transparent, interpretable, and accountable.

Methodology

- Data Exploration: Load data into our code, view and analyze the data set loaded.
- Data Preprocessing: Check for duplicates, remove missing values, remove missing values and scale data.
- Feature Engineering: Transform and generate new features.
- Model Training: Model selection and hyperparameter training.

- Model Evaluation: R squared performance evaluation matrices used to evaluate models

Experimental Setup and Implementation Tools

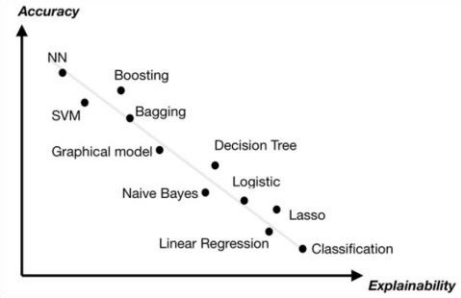
The experiments were carried out on a Lenovo ThinkPad laptop equipped with an Intel(R) Core (TM) i7-6600U CPU running at a clock speed of 2.60GHz, 2.81 GHz, and 8.00 GB of RAM. The computational environment utilized for the experiments was Google Colab, along with several Python libraries, including Pandas, Scikit-Learn, Seaborn, Keras, and Xgboost. These libraries provided the necessary functionality for data manipulation, machine learning algorithms, visualization, and boosting techniques, enabling a comprehensive analysis of the research objectives.

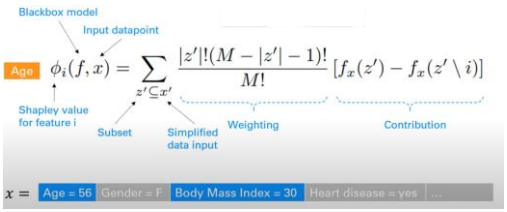
Results

Model	R-Squared Value
Linear Regression	0.68
Random Forest Regression	0.82
Extreme Gradient Boosting	0.84
Lasso regression	0.67

XAi (Novel Idea Attempt)

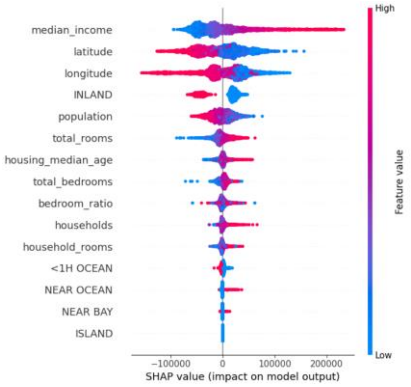
Explainable AI (XAI) is a prominent field within artificial intelligence (AI) research, dedicated to developing AI systems capable of providing explanations for their decisions and behaviors. The primary objective of XAI is to enhance the transparency, interpretability, and accountability of AI systems. This pursuit is essential to ensure the responsible and ethical utilization of AI technologies. Various XAI techniques, including feature importance analysis, partial dependence plots, and counterfactual explanations, play a pivotal role in elucidating the decision-making processes of AI systems and offering valuable insights into their inner workings. By leveraging these XAI methodologies, researchers can shed light on the factors influencing AI decisions and contribute to the advancement of trustworthy and explainable AI systems.





SHAP (Shapley Additive Explanations) is a mathematical approach that provides explanations for the predictions made by machine learning models. Grounded in game theory principles, SHAP can be employed to elucidate the predictions of any machine learning model by quantifying the contribution of each feature towards the overall prediction. By calculating the SHAP values, the contribution of each feature in explaining the disparity between the average prediction of the model

and the specific prediction of an instance can be comprehended. Moreover, SHAP facilitates an understanding of how different combinations of features contribute to the overall prediction.



Challenges

- Despite our expectation that LSTM would provide the best prediction, we encountered issues with our algorithm functioning properly.
- The prediction of prices is influenced by various factors, particularly dependent on the location.
- We were also constrained by limited computing power.

Conclusion

In this research study, the California dataset was utilized to explore various regression models, including the Simple Linear Regression Model, Lasso Regression Model, and the Random Forest Model. Through our analysis, it has been determined that the XGBoost model exhibits the highest values of the R-square, making it the most suitable model for this dataset. Therefore, we conclude that the XGBoost model is the optimal choice for predicting house prices based on this specific dataset. Furthermore, our research demonstrates the viability of advanced machine learning algorithms such as LR, RF, and Lasso as valuable tools for real estate researchers in the field. However, it is important to acknowledge that these machine learning techniques do possess their own limitations, which should be taken into consideration when interpreting the results.

Individual Contributions

Georgios – Filippos Bogdos	- Compiling Report, Code, Colab Debugging, Presenting
Jesus Perez	- Report, Documentation, Code Review, Code, Presenting
Gideon Koech	- Compiling Report, Data Analysis, Code, Presenting

Other Groups Grades

Group	Grade
1	100
2	100
3	100

4	100
5	100
6	100
7	100
8	100

References

K. E. Case, J. M. Quigley, and R. J. Shiller, "Comparing Wealth Effects: The Stock Market vs. the Housing Market," Southern Economic Journal, Vol. 72, no. 2, pp. 269-283, 2005.

Rosen, S. (1974) Hedonic Prices and Implicit Markets: Product Differentiation in Pure Competition. Journal of Political Economy, 82, 34-55. https://doi.org/10.1086/260169

Prasad, K. Venkata, and K. Ravindra. "A comparative study of regression models for prediction of house prices in Ames, Iowa." Journal of Construction Engineering and Management. Volume 138, no. 4 (2012): 499-508.

Park, B., and Bae, J.K. Using machine learning algorithms for housing price prediction: The case of Fairfax County, Virginia housing data, Expert Systems with Applications, Volume 42, Issue 6, Pages 2928-2934, 2015.

Singh, R. K., and S. Singh. "Predicting housing prices with machine learning techniques." International Journal of Computer Applications 173, no. 7 (2017): 1-7.

Soni, N. K., and N. P. Goyal. "Housing price prediction using machine learning techniques: A review." arXiv preprint arXiv:1809.06354 (2018).