INSTITUTO
UNIVERSITÁRIO
DE LISBOA

Hunting for Bubbles: A Predictive Model of New York City's Real Estate Market

Nuno Rodrigo Basílio Soares

Master in Data Science

Supervisor:
PhD. Diana Elisabeta Aldea Mendes, Associate Professor,
ISCTE Business School, Department of Quantitative Methods for Business and Economics

Co-supervisor:
PhD. Vivaldo Manuel Pereira Mendes, Associate Professor,
ISCTE Business School, Department of Economics

October, 2023

Department of Quantitative Methods for Management and Economics

Department of Information Science and Technology

Hunting for Bubbles: A Predictive Model of New York City's Real Estate Market

Nuno Rodrigo Basílio Soares

Master in Data Science

Supervisor:
PhD. Diana Elisabeta Aldea Mendes, Associate Professor,
ISCTE Business School, Department of Quantitative Methods for Business and Economics

Co-supervisor:
PhD. Vivaldo Manuel Pereira Mendes, Associate Professor,
ISCTE Business School, Department of Economics

October, 2023

Dedico este trabalho a todos os que tanto me apoiaram.

Acknowledgements

I would like to commence my acknowledgment section by expressing my heartfelt gratitude to my thesis supervisor, Professor Diana Mendes. Her unwavering patience and invaluable feedback have been instrumental in shaping this thesis. Without her guidance and support, I would not have been able to deliver this work. Furthermore, this Master's program marked my introduction to the field of Data Science, and having Professor Diana as my teacher was a pivotal step in taking my first strides in this domain. In addition, I extend my sincere thanks to my thesis co-supervisor, Professor Vivaldo Mendes, for his contributions to the development of this Master's thesis.

I am also deeply indebted to the entire ISCTE community, including the dedicated professors who imparted their knowledge, the diligent librarians who facilitated my research, and the committed security and building maintenance teams who ensured a conducive environment for learning. Their collective efforts have contributed to creating an extraordinary and memorable educational experience for every student at this university.

I would like to extend my heartfelt gratitude to my family for their unwavering moral support throughout these years and their encouragement in pursuing my academic aspirations. A special thanks goes to my parents, whose relentless efforts and sacrifices ensured that I had everything I needed to pursue my education. Their dedication has been the cornerstone of my journey.

I want to express my profound appreciation to my sister, who has been one of my greatest motivators. Knowing that my achievements bring pride to her has been a driving force for me to always strive for excellence. I am deeply grateful to my grandmother for her steadfast trust in my abilities, which has been a constant source of confidence and inspiration.

I also want to acknowledge my friends for the countless laughs and shared moments during the process of writing this thesis. These moments of leisure and relaxation were instrumental in maintaining a healthy work-life balance and provided the endurance needed to persevere.

To those who can no longer share in this achievement with me, I am certain they would be immensely proud and filled with happiness for my accomplishments. Throughout the process of writing, I held them close in my thoughts and carried their memory as a constant source of motivation. I would like to offer a special tribute to my grandfather, who ignited in me the insatiable desire to continuously learn and explore the world around me. His influence was pivotal in shaping my approach to all aspects of life and essentially, who I am.

In closing, I want to extend a heartfelt thank you to my girlfriend, Marta. Her unwavering belief in me and her ability to inspire self-confidence during moments of self-doubt have been invaluable. Marta has been by my side throughout this entire journey, experiencing the highs and lows alongside me. It has been my greatest pleasure to share the moments of joy and enthusiasm with her, knowing that her devoted support was always there when I needed it. As with all the significant milestones we've achieved together, I am delighted to add this one to our shared list of accomplishments.

I look forward to a future filled with many more wonderful moments and shared achievements with those I love by my side.

Resumo

O que define uma bolha imobiliária? Quais são os principais fatores que podem desencadear este fenómeno? Estes fatores são relevantes no contexto de Nova Iorque? As bolhas imobiliárias não são uma novidade, uma delas esteve inclusive ligada à conhecida crise financeira de 2008. Compreender os fatores causais por trás de uma bolha imobiliária, bem como as dinâmicas únicas que caracterizam o famoso mercado imobiliário da cidade de Nova Iorque, são assuntos frequentemente discutidos tanto pela população geral como pelos media.

O objetivo desta dissertação é analisar os principais fatores económicos associados a bolhas imobiliárias, utilizando insights de pesquisas realizadas sobre o tópico em diversas regiões geográficas. Foram inicialmente coletados dados para monitorar estes indicadores económicos referentes aos EUA e, em particular, à cidade de Nova Iorque. Posteriormente, foram construídos modelos de classificação para identificar períodos de bolhas imobiliárias utilizando esses indicadores como input.

O estudo concluiu que a configuração de modelo mais bem-sucedida foi alcançada através da utilização de XGBoost com a lista de features do Teste 1. Tendo o par, com e sem feature selection, alcançado taxas de accuracy de 0,89 e 0,86, respetivamente. As features com um papel significativo nestes modelos de classificação estão alinhadas com aquelas destacadas por outros autores como cruciais para a deteção de bolhas imobiliárias. O contributo para a performance dos modelos do rácio price-to-rent, inflação e taxas de juros demonstraram que estes indicadores são aplicáveis ao mercado imobiliário da cidade de Nova Iorque, corroborando com conclusões feitas por diversos autores para outros mercados.

Palavras-chave: Bolhas imobiliárias, Mercado imobiliário, *Machine Learning*, Modelos de Classificação, Impulsionadores económicos, XGBoost

Abstract

What defines a housing bubble precisely? What forms the consensus on its principal triggering factors? Are these factors relevant to New York City? Housing bubbles are not novel occurrences, one was even associated with the major financial crisis of 2008. Understanding the causal factors behind a housing bubble and the unique dynamics characterizing the renowned New York City real estate market, are subjects frequently discussed by both the public and the media.

The aim of this dissertation is to examine the primary economic factors typically linked to housing bubbles, drawing insights from research conducted in various geographical regions. This research involves gathering data to monitor these economic drivers in the US and New York City markets. Subsequently, classification models were constructed using these variables as inputs to identify periods of housing bubbles. The dissertation also includes an analysis of model performance and a comparison between different models to determine which feature set yields superior results.

This study concluded that the most successful model configuration was achieved by using XGBoost with the feature list of Test 1. This configuration was tested both with and without feature selection, resulting in accuracy rates of 0.89 and 0.86, respectively. Notably, the features that played a significant role in our classification align with those highlighted by other researchers as crucial for housing bubble detection. Features such as the price-to-rent ratio, inflation, and interest rates demonstrated their applicability to New York City, substantiating findings from diverse geographical regions.

Keywords: Housing Bubbles, Housing Market, Machine Learning, Classification Models, Economic Drivers, XGBoost

Table of Contents

ACKNOWLEDGEMENTS **III**

RESUMO **V**

ABSTRACT **VII**

TABLE OF CONTENTS **1**

INTRODUCTION **3**

LITERATURE REVIEW **5**

2.1. Housing Bubbles 5

METHODOLOGY **15**

3.1. The problem visualized from data 15

3.2. Data Extraction and Preparation 16

3.3. First insights about Data 22

3.4. Data Preparation Revisited 23

3.5. Modeling 29

RESULTS AND DISCUSSION **33**

4.1. Features chosen without feature selection 33

4.2. Advanced Feature Selection 33

4.3. XGBoost with and without Advanced Feature Selection 35

4.4. Random Forest with and without Advanced Feature Selection 38

4.5. Neural Network with and without Advanced Feature Selection 41

CONCLUDING REMARKS **45**

REFERENCES **47**

APPENDIX A **51**

APPENDIX B **51**

APPENDIX C **53**

CAPÍTULO 1

Introduction

The ability to identify housing bubbles is of significant importance, not only for individuals purchasing homes but also for investors within the market. The decision of which property to purchase is pivotal for buyers seeking a place to live. The price of the house influences various aspects of the buyer's living situation. If a suitable place is financially out of reach, it might necessitate relocation, potentially impacting one's career and access to infrastructure, among other social factors. From an investor's standpoint, the presence of housing bubbles presents both challenges and opportunities. Identifying and divesting from overpriced real estate can serve as a risk management strategy. It helps prevent acquiring properties at a significantly inflated cost compared to their actual value, which would lead to financial losses upon resale. This aspect is also pertinent to general consumers, as avoiding overpayment for a home increases the likelihood of mitigating financial loss in case relocation becomes necessary.

However, detecting housing bubbles is significant beyond safeguarding home buyers and investors; it also involves the potential for profoundly severe consequences at a macroeconomic level. An illustration of these potential repercussions is the global financial crisis of 2008. The origins of this crisis can be traced to the early 21st Century, characterized by the liberal monetary policy of the Federal Reserve (FED), which fueled the exponential growth of the USA Real Estate market and other financial instruments. The bursting of this housing bubble not only spread across the American market but swiftly extended to other regions, magnifying the crisis into a global phenomenon. Consequently, it yielded far-reaching economic, political, and social impacts on a worldwide scale.

To initiate this discussion, it is important to establish a definition of what constitutes a housing bubble. A housing bubble is said to exist when significant disparities arise between the price and the intrinsic value of an asset. The market enters a bubble state when the market price of an asset surpasses its fundamental value, with the fundamental value being defined as the asset price under conditions of market equilibrium. The existence of a bubble implies a deviation between market value and the equilibrium market value, often driven by speculative behavior. Kindleberger (1987) provides a comprehensive description of this phenomenon, defining it as a rapid surge in the price of an asset or a range of assets. This initial surge generates expectations of further increases and attracts new buyers, typically speculators with no vested interest in the asset's utility or earning potential but rather motivated by the potential profits from trading it. Subsequently, this ascent is followed by a reversal in expectations and a sharp decline in price, often culminating in severe financial or economic crises.

This thesis aimed to enhance the understanding of which economic indicators significantly contribute to detecting a housing bubble. Different combinations of features, as indicated in various studies on the subject, were utilized as inputs. The literature review encompassed studies conducted in various geographical territories, and the objective was to ascertain if the findings from these studies were transferrable to New York City's real estate market. This dissertation addresses several key questions: What features hold the highest significance in housing bubble detection? Can the features employed in other geographical regions be applied to New York City? Among the available economic variables, which set of inputs yields superior performance? Lastly, which models employed in this classification model perform best?

This thesis was organized as follows: Chapter 2 presented a literature review of articles related to housing bubbles and real estate markets, discussing the economic features most influential in causing abnormal behavior in the housing market, potentially leading to housing bubbles. Chapter 3 outlined the data extraction and initial cleaning processes, the transformation of features, dataset merging, and the definition of the target feature. This chapter introduced the three input sets to be used, the feature selection process was explained, and the algorithms – XGBoost, Random Forest, and Neural Network – were briefly presented. Chapter 4 delved into the results, analyzing the model performance and its relationship with the literature reviewed in Chapter 2. Finally, in Chapter 5, concluding remarks were presented, along with some additional alternative tests not covered in this dissertation, which might be interesting to explore in future work based on the literature review.

CHAPTER 2

Literature Review

2.1. Housing Bubbles

The concept of housing bubbles is a subject that is frequently debated, and there is no definitive answer, as various authors hold different perspectives on what constitutes a housing bubble, the criteria for defining one, and the economic indicators that best signal their presence. In this literature review, we will analyze the work of other scholars in this field.

While the causes and contributing factors to housing bubbles can vary, one economic variable and one ratio are frequently examined in this context: interest rates and price-to-income ratios. In the research conducted by Tsai and Lin (2022), it is evident that when property prices significantly outpace rental prices, it can indicate a housing boom, although not necessarily a housing bubble. This study also highlights the critical role of monetary policy in the real estate market. When interest rates are exceptionally low, and financing terms are highly favorable, there is an increased likelihood of speculative behavior, potentially leading to the formation of a housing bubble by delaying necessary price adjustments. Another notable finding is that stable changes in interest rates correspond to stable price-to-rent ratios. Further evidence of the significant role of interest rates in housing bubble formation is presented in Taipalus's (2006) research. The study concludes that one of the primary drivers behind the surge in housing prices in Europe during the early 2000s was the exceptionally low interest rates on housing loans.

Another article examined in this context is the study by Li et al. (2021), which concludes that there exists a correlation between the emergence and subsequent bursting of housing bubbles and the inherent risks within the real estate market. The cycle, as described, typically commences with a rising number of housing loans offered under appealing terms, leading to an increase in speculative activity affecting property prices. The author also attributes part of the blame to the financialization of the housing market, where its rapid and prosperous value appreciation, coupled with its relative stability, renders it highly attractive for allocating surplus funds compared to more traditional assets. Furthermore, the author identifies a research gap in this field, specifically the scarcity of studies adopting a multidisciplinary approach, integrating theory with intelligent algorithms and qualitative analysis—an area that this paper aims to address.

According to Hung and Tzang (2021), the price of a property comprises two components: investment and consumption. The consumption component should typically account for the majority of a property's price. A housing bubble begins to form when the investment component starts growing to the extent that it becomes nearly as prominent as or even more prominent than the consumption share. The authors' model concludes that buyers consider the investment value when acquiring a property, and they are willing to pay varying amounts for the same property depending on economic factors such as interest rates, Loan-to-Value (LTV) ratios, rent values, market prices, volatility, among others. During periods of economic strength, the real estate market tends to exhibit volatility, consequently reducing the consumption share of the total price, particularly in properties with high rental potential. According to the author, this is one of the red flags indicating the potential existence of a housing bubble.

Returning to Kindleberger's (1987) definition of a bubble, it involves a rapid increase in the price of an asset or a group of assets in a continuous process. The initial price surge captures the attention of new buyers, disrupting the existing market equilibrium and leading to further price increases due to the economic forces of supply and demand. The crux of the issue primarily lies in the behavior of market participants who persistently trade overpriced assets. Masiukiewicz and Dec (2015) conclude that multiple indicators can assist in identifying real estate bubbles, with some of the most crucial ones being:

1. **The price-to-income ratio:** This metric reveals how many years of work would be necessary to purchase a property;
2. **Future rent prices:** A positive value indicates that expectations of price increases are factored into the market price;
3. **The price-to-rent ratio**: This ratio helps determine whether it is more advantageous to purchase or rent a property;
4. **Analysis of changes in house price indexes**.

However, as demonstrated in Rentala et al. (2021) work, relying solely on these indicators may prove insufficient. When rental prices surge above the average or when a property is rapidly revalued at a higher price in the short term, it becomes possible to generate significant profits by reselling the property within a year. If this cycle repeats over several years, it may signal the presence of a housing bubble.

To enhance the reliability of such analyses, an interdisciplinary approach, as emphasized by Lepenioti et al. (2020), is essential. Predictive analysis offers a way to overcome some of the limitations associated with descriptive and diagnostic analyses by leveraging historical data to forecast the future. For classification purposes, models like decision trees, logistic regression, and neural networks can be employed. In cases involving segmentation, unsupervised machine learning algorithms such as K-means clustering, hierarchical clustering, or Gaussian mixture models could be utilized. It is important to note, however, that while these predictive models may yield improved results, they are still constrained by their reliance on historical data and prior research.

As a confirmation of other authors research, Dec et al. (2022) conclude that not every disequilibrium in the supply and demand function leads to a housing bubble; the reduction of construction leads to a decreased supply that can't match the existent demand, which leads to a sharp increase in price. The detection and measuring of housing bubbles should be the base of a system of alert for financial crisis, since, a lot of times, housing bubbles and their burst, are at the epicenter of high instability in all economic spheres. According to the author this system should be in the hands of a central bank or other government institution.

While examining the post-pandemic era we are currently experiencing, Afxentiou et al. (2022) study, grounded in both the health and financial crises, investigates the potential for an economic crisis similar to the one in 2008, driven by the COVID-19 pandemic. During the pandemic, the United States real estate market surged to record highs. A significant contributing factor to this phenomenon was the lockdowns and other restrictions, prompting many Americans to reconsider their living arrangements, with a preference for open spaces and homes suitable for remote work, which had become increasingly popular. This surge in demand, coupled with historically low interest rates, led to a rapid and substantial increase in prices, raising concerns about a situation reminiscent of the 2008 crisis.

The American real estate market has been experiencing price growth since 2011, following the burst of the housing bubble, and by the end of 2021, property values had already exceeded the pre-crash peak in 2006 by 4.5%. What distinguishes the current paradigm from the pre-pandemic context is that price pressures in communities with low population density are now at least as pronounced as those in metropolitan areas. This is a shift from the past and can be primarily attributed to the appeal of housing characteristics in sparsely populated areas, which became particularly attractive during the pandemic.

Several factors distinguish the current situation from the 2008 crisis. Despite historically low interest rates close to 0% (although these have recently changed due to high inflation levels), financial institutions are showing greater diligence in managing their portfolios, prioritizing healthy credit quality. Moreover, households have more economic power today, and despite the impact of high inflation, the decline in real household income has not been as sharp as during the 2008 crisis. In fact, real household income, even after adjusting for the effects of high inflation, remains higher than the income levels observed following the 2008 crisis.

The urban exodus triggered by the pandemic in the American market has resulted in people from metropolitan areas, typically with higher incomes, purchasing homes in less populated regions, driving prices in those areas to unprecedented levels. Another contributing factor to increased prices is the influx of Millennials into the housing market for the first time during the pandemic. Additionally, there is a growing presence of institutional investors, such as hedge funds and investment firms, in the real estate market. These investors have an expanding stake in the market with the aim of converting properties into rentals, thereby ensuring a stable source of income for their portfolios. As depicted in Figure 2.1 below, the proportion of properties acquired by investors increased during the pandemic, although it has experienced some retracement since then.

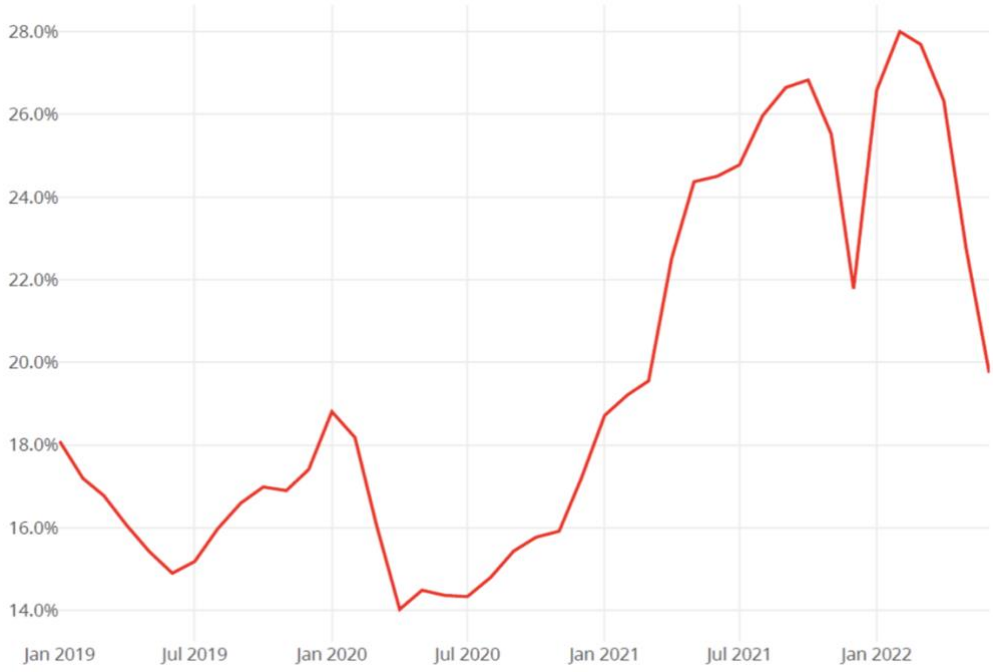

Figure 2.1 - Percentage of Houses bought by investors in the American Real Estate market. Source: Corelogic

As previously mentioned in the study conducted by Rantala et al. (2021), the pace at which these investors are selling their properties could also serve as an indicator of a real estate bubble. As illustrated in Figure 2, more than 10% of the homes purchased by investors in the US market between early 2019 and mid-2021 were resold within the first six months after acquisition.

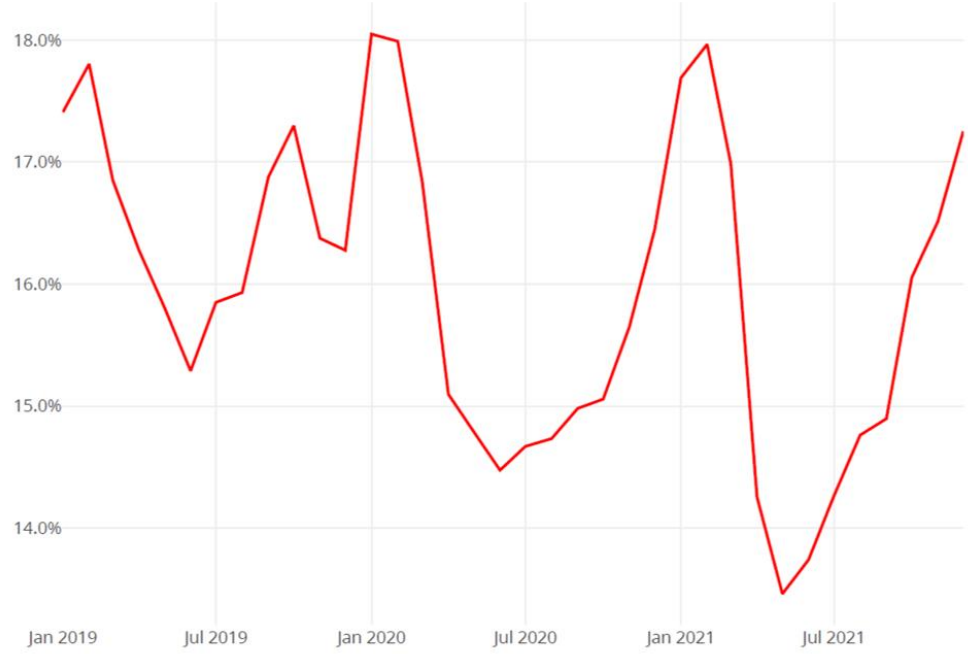

Figure 2.2 - Percentage of properties sold by investors in a period of up to six months after purchase. Source: Corelogic

As previously discussed, a monetary policy featuring very low interest rates, which facilitates easy access to capital, can contribute to price increases. However, this variable may not always apply to investors, as they often make cash offers on properties. Despite some potential trigger points, as argued by Afxentiou et al. (2022), the likelihood of a scenario similar to the 2008 housing crisis is considered low. This is primarily attributed to the record-high prices being driven by supply and demand dynamics, as explained earlier. The authors argue that supply failed to meet demand, partly due to the phenomenon of urban exodus and a reduction in new property construction compared to pre-2008 levels. These factors, coupled with more robust regulations, led Afxentiou et al. (2022) to dismiss the possibility of a housing bubble.

In the course of the research, Leung (2004) arrived at a critical conclusion in the context of his literature review on Macroeconomics and Housing. His findings indicate that standard macroeconomics textbooks often treat housing as a typical consumption good, similar to others, or even overlook its importance altogether. In contrast, traditional housing economics frequently fails to consider the interplay with the broader macroeconomy, except for some theoretical and empirical analyses in urban and housing economics that introduce macroeconomic variables such as inflation, economic growth, GDP, and the unemployment rate as 'control features'. Nevertheless, Leung emphasizes that housing holds significance beyond being a standard consumption good, given its substantial share in household expenditure and total family wealth.

Davis and Heathcote (2001) concluded that the market value of the US residential property stock approximately matches the annual average GDP. The same authors, in 2003, discuss the observations that non-residential investment lags behind GDP, while residential investment precedes GDP. Furthermore, they establish a connection between the construction of new units and housing cycles, highlighting the significance of construction-intensity for volatility. This is attributable to the highly volatile nature of productivity in the construction sector and the labor-intensive nature of construction. Additionally, they note that residential structures depreciate at a slow rate, which significantly contributes to the incentive for concentrating new structure production during periods of high relative productivity.

Englund and Ioannides (1997) analyze the dynamics of housing prices in 15 OECD countries and find that these countries exhibit similar patterns in their housing price changes. Their research highlights a consistent pattern of price changes over time, where the first-year price change strongly correlates with the following year's change. This suggests that if housing prices increase in one year, it is likely that they will continue to rise in the subsequent year. Furthermore, the study identifies factors such as the current rate of economic growth and fluctuations in interest rates as having a substantial impact on housing prices.

In the research conducted by Tsatsaronis and Zhu (2004), it is concluded that a robust and enduring relationship exists between inflation, interest rates, and fluctuations in housing prices. This connection implies that prolonged periods of high inflation, followed by abrupt decelerations in price growth, may lead to a misalignment between housing prices and the fundamental factors that determine their long-term value. Such circumstances underscore the importance of vigilance on the part of monetary authorities.

Lecat and Mésonnier (2005) discovered that, regardless of the analytical approach employed, it becomes evident that financial factors exert a substantial influence on house prices. Their findings indicate that interest rates and the availability of loans impact housing prices, and this influence is further driven by financial deregulation. Over time, the increased accessibility of credit, coupled with shifts in household income and population trends, has significantly reshaped the housing market. However, these transformations have also given rise to potential economic and financial challenges, prompting debates regarding monetary policy. As it stands, the conclusions drawn by Lecat and Mésonnier (2005) have played a noteworthy role in the context of the subprime crisis.

In the 2011 housing bubble survey conducted by Mayer, many of the macroeconomic factors presented align with the findings of other scholars on the subject. Notably, the survey introduces an intriguing definition of a housing bubble worth highlighting, which shares similarities with Kindleberger's (1987) description. According to this definition, a housing bubble occurs when there are extreme fluctuations in house prices, characterized by a growth rate of 20% or more annually for two to three years, followed by a comparable decline over the subsequent three years. Applying this criterion, the author concludes that housing bubbles are relatively frequent phenomena.

Demonstrating that bubble events are relatively common, Rebelo et al. (2011) reveal that boom-and-bust cycles are widespread phenomena in the housing market, occurring in various countries and across different time periods. Furthermore, they establish a connection between the influx of new entrants into the market, underscoring the significant influence of supply and demand dynamics in housing booms. This corroborates the findings of other researchers on this topic.

In Bourassa et al. (2019) study, the primary objective was to identify the most effective method for monitoring and detecting price bubbles across six metropolitan housing markets in three countries using 30 years of quarterly data. Their findings established that the price-to-rent ratio emerged as the most dependable approach for detecting price bubbles. This aligns with the insights of other researchers previously discussed in this literature review, who also emphasized the significance of this indicator. The price-to-rent ratio offers a critical measure of housing affordability, with far-reaching implications for individuals and the broader economy. Furthermore, even in the context of investment in residential real estate, the price-to-rent ratio is often employed to assess the attractiveness of real estate markets. Elevated ratios can discourage investors, potentially reducing speculative buying and fostering a more stable market.

Lu et al. (2015) analyzed the Asian market, specifically focusing on Penang. Their study concentrated on various indicators, including the housing price index, consumer price index, GDP, interest rates, and housing supply factors. The least squares regression method was the chosen analytical technique, applied over the period from 2000 to 2012. While the authors did not discover evidence of a housing price bubble during this timeframe, their analysis confirmed that inflation and borrowing costs (inferred from interest rates) played a substantial role in elucidating fluctuations in housing prices in the region. This finding concurs with previous research that also underscored the importance of these factors in detecting housing bubbles.

Shen et al. (2005) examined the potential existence of a price bubble in Beijing and Shanghai in 2003. Their analytical approach encompassed the Granger causality test and generalized impulse response analysis. Utilizing features including income, stock market indexes, housing vacancy rates, and housing price indexes as inputs, the authors arrived at the conclusion that properties in Shanghai were overvalued. This underscores the utility of these indicators in predicting housing bubbles.

Dispasquale and Wheaton (1996) concur that the primary determinant of how the housing market transforms and evolves over time is the size and growth of the economy within a given country or region. In this context, Gwartney et al. (2004) establish a direct correlation between economic growth and income growth, with the former being a prerequisite for the latter. GDP is frequently regarded as the principal indicator of economic growth, particularly when assessing GDP on an individual level through metrics like per capita GDP. An increase in per capita GDP is typically associated with income growth, enabling individuals to allocate more resources toward goods and services. This heightened economic activity and enhanced spending capacity profoundly impact the housing market, often driving up property prices. Consequently, GDP and per capita GDP are considered integral metrics for scrutiny when investigating this subject, a consensus shared by the other works reviewed in this literature analysis.

Finally, Glaeser et al. (2008) deduce that real estate bubbles are significantly linked to the supply elasticity within the real estate market. This underscores the imperative nature of incorporating models with supply-related features as inputs.

The articles examined in this section elucidate the multifaceted economic variables that influence housing prices and can be employed in identifying housing bubbles. While some of these variables are readily discernible and quantifiable, others present more intricate challenges for verification and application. The precise definition of a housing bubble, as well as the pivotal variables for its prediction, remain topics of ongoing discussion. Nevertheless, there exists a substantial common ground, as evidenced by the reviewed literature, that can be harnessed for this purpose.

CHAPTER 3

Methodology

3.1. The problem visualized from data

As evidenced in the Literature Review, numerous economic indicators are frequently associated with the housing market. These indicators offer insights into overall economic conditions, financial stability, and the dynamics of housing supply and demand. The following information was compiled about the United States of America market and the New York City market:

3.1.1. *US Gross Domestic Product (GDP)* – As defined by the Organization for Economic Co-operation and Development (OECD), GDP is the standard measure of value added through the production of goods and services in a country during a specific period. This data was selected due to its ability to indicate overall economic activity and growth, factors that can influence the housing market.

3.1.2. *NYC Employment and unemployment rates* – This data assumes significance as it reflects the labor market conditions of New York City, potentially impacting housing affordability and demand. Typically, high employment and low unemployment rates signify a robust housing market.

3.1.3. *US Interest Rates* – This dataset comprises the FED interest rates, which represent the rate at which the central bank borrows money from commercial banks. This rate usually serves as the base rate indexed to mortgage rates. As discussed earlier and in the introduction, lower interest rates can make borrowing an attractive option, facilitate access to capital, and enhance housing demand.

3.1.4. *US Consumer Price Index (CPI)* – The Consumer Price Index gauges the overall change in consumer prices based on a representative basket of goods and services over time. It is used for Calculating inflation or deflation rates, given their connection to housing affordability. Being that inflation is a rise in prices, which can be translated as a decline of purchasing power over time, this is an important variable to consider as high inflation equals a reduced purchasing power of potential buyers, directly affecting housing affordability.

3.1.5. *NYC Household Income* – Higher incomes generally support a healthier housing market. This data is integral to calculating a price-to-income ratio, a pivotal element in analyzing the existence of a housing price bubble.

3.1.6. *NYC Building permits* – We included this dataset to assess the issuance of building permits, offering insights into the supply side of the housing market. Increased levels of issued permits may suggest potential growth or market oversupply.

3.1.7. *NYC Rental vacancy rates* – Rental vacancy rates measure the proportion of unoccupied rental units. Lower vacancy rates often indicate strong rental demand and a potentially robust housing market.

3.1.8. *US Consumer Confidence Index (CCI*) – This indicator provides insights into future household consumption and savings based on responses about expected financial situations, employment, and savings capability. It reflects consumer optimism about the economy and their willingness to make significant purchases, including housing.

3.1.9. *NYC Housing Price Index* – This index measures property price fluctuations for single-family properties in New York City. It functions as an indicator of housing price trends and also operates as an analytical tool to estimate changes in mortgage default rates, prepayments, and housing affordability.

3.1.10. *NYC Median Rent* – This dataset encompasses median rental rates for housing units ranging from studios to four-bedroom residences. It can be utilized to analyze price trends and also as a supplementary tool for calculating price-to-rent ratios.

These indicators *per se* would hold limited significance without a dataset, including the sale prices of real estate in New York City. For this purpose, there is also a dataset encompassing the real estate officially sold in the city from 2003 to 2022.

3.2. Data Extraction and Preparation

The data extracted, as mentioned earlier, originated from diverse sources and arrived in different formats. Due to this variability, certain preparatory tasks were necessary before the commencement of analysis, which will be outlined in this section. These tasks encompass various steps, including merging datasets to consolidate all category-related information into a single file, renaming, dropping, and reordering columns, converting and updating file types, among other tasks.

3.2.1. US Gross Domestic Product (GDP)

The source of data on the US Gross Domestic Product (GDP) is attributed to the US Bureau of Economic Analysis. The extraction process of this dataset entailed the utilization of FRED (Federal Reserve Economic Data), an online repository encompassing an extensive collection

of economic data time series. FRED is managed by the esteemed Federal Reserve Bank of St. Louis, serving the fundamental purpose of furnishing monetary data to amplify comprehension of the Federal Reserve's policy deliberations.

	DATE	GDP	A939RX0Q048SBEA	GDPC1	GDPC1_PC1	A939RX0Q048SBEA_PC1
0	1947-01-01	243.164	14213.0	2034.450	.	.
1	1947-04-01	245.968	14111.0	2029.024	.	.
2	1947-07-01	249.585	14018.0	2024.834	.	.
3	1947-10-01	259.745	14171.0	2056.508	.	.
4	1948-01-01	265.742	14326.0	2087.442	2.60473	0.79505

Figure 3.1 - GDP dataset before executing Jupyter Notebook A.1..

The data was retrieved in .csv format and subsequently incorporated into Jupyter Notebook A.1[1]. The script embedded within this notebook was crafted to bestow fresh nomenclature upon the dataset's columns, thus fostering clarity and intelligibility. Concomitantly, this script orchestrated the strategic reordering of columns, enhancing the dataset's coherence and analytical utility. The final dataset was saved in .csv format.

	date	nominal_gdp	real_gdp	real_gdp_change	real_gdp_per_capita	real_gdp_per_capita_change
0	01/01/1947	243.164	2034.450	NaN	14213.0	NaN
1	04/01/1947	245.968	2029.024	NaN	14111.0	NaN
2	07/01/1947	249.585	2024.834	NaN	14018.0	NaN
3	10/01/1947	259.745	2056.508	NaN	14171.0	NaN
4	01/01/1948	265.742	2087.442	2.60473	14326.0	0.79505

Figure 3.2 - GDP dataset after executing Jupyter Notebook A.1..

3.2.2. *NYC Employment and unemployment rates*

The dataset under consideration originated from the US Department of Labor and was publicly accessible through direct extraction from the New York State Department of Labor (DOL). The dataset was initially provided in .xlsx format, necessitating a conversion process for optimal compatibility and usability. To achieve this conversion, Jupyter Notebook A.2. was employed, transforming the dataset into the widely accepted .csv format, a more suitable format for subsequent analysis.

[1] Jupyter notebooks are available upon request. Please contact us at nrbss@iscte-iul.pt for more information. The complete file list can be found in Appendix A.

	YEAR	Labor Force	Employment	Emp/Pop	Unemployed	Unemp Rate	LFPART	Unnamed: 7	Unnamed: 8	Unnamed: 9
0	1976-01-01 00:00:00	3066.605	2723.016	47.8	343.589	11.2	53.856779	NaN	NaN	NaN
1	1976-02-01 00:00:00	3065.430	2722.421	47.8	343.009	11.2	53.855060	NaN	0.0	NaN
2	1976-03-01 00:00:00	3064.867	2722.931	47.9	341.936	11.2	53.864095	NaN	0.0	NaN
3	1976-04-01 00:00:00	3067.313	2726.299	47.9	341.014	11.1	53.916558	NaN	-0.1	NaN
4	1976-05-01 00:00:00	3071.973	2730.681	48.0	341.292	11.1	54.017461	NaN	0.0	NaN

Figure 3.3 - NYC Employment and unemployment dataset before executing Jupyter notebook A.3..

The dataset, once obtained, exhibited formatting inadequacies that required rectification. Addressing these issues, a separate Jupyter Notebook script, A.3., was executed. This script aimed to enhance the overall quality of the dataset by rendering column names more intelligibly and improving data integrity. Moreover, this script included procedures to handle extraneous data entries, effectively bypassing irrelevant rows. Additionally, the script facilitated the elimination of two vacant columns, followed by strategically renaming and reordering columns to foster structural coherence. The results of these enhancements were then captured in a new .csv file.

	date	labor_force	employed	employment_vs_population	unemployed	unemployment_rate	change_unemployment_rate	labor_force_participation
0	1976-01-01	3066.605	2723.016	47.8	343.589	11.2	NaN	53.856779
1	1976-02-01	3065.430	2722.421	47.8	343.009	11.2	0.0	53.855060
2	1976-03-01	3064.867	2722.931	47.9	341.936	11.2	0.0	53.864095
3	1976-04-01	3067.313	2726.299	47.9	341.014	11.1	-0.1	53.916558
4	1976-05-01	3071.973	2730.681	48.0	341.292	11.1	0.0	54.017461

Figure 3.4 - NYC Employment and unemployment dataset after executing Jupyter notebook A.3..

3.2.3. *US Interest Rates*

For the US Interest Rates, the dataset sourced its information from the Board of Governors of the Federal Reserve System, with FED rates serving as the dataset content. This data was extracted from FRED and originally included daily frequency data from 1954 to 2023. The extracted file was in .csv format, and script A.4., was executed to undertake fundamental data refinement tasks, including column renaming and reordering. The resulting dataframe was saved in a separate .csv file.

3.2.4. *US Consumer Price Index (CPI)*

This data originated from the International Monetary Fund (IMF) and was accessible to the public via their data website. For the extraction of the requisite data, two distinct search queries were executed. The initial query yielded a .xlsx file encompassing Consumer Price Index (CPI) values, with a monthly frequency spanning from January 2000 to June 2023. The second query

provided an additional .xlsx file detailing the percentage change in CPI from the preceding year, covering the same timeframe.

In order to harmonize the data for analysis, a script named A.5. was employed. This script facilitated the conversion of both .xlsx files into the .csv format, rendering them more amenable to analytical processes. Subsequently, another script, A.6., was executed. This script, functioning as a data integration tool, merged the two distinct .csv files. The date was employed as an indexing mechanism. Furthermore, it addressed the issue of missing values by implementing the Pandas .ffill() function, ensuring a consistent value for each date within a given month.

The data transformation steps culminated in the creation of an enriched dataframe. This resulting dataframe was saved as a new .csv file, poised for further analysis.

3.2.5. *NYC Household Income*

The data source for New York City's household income was the US Census Bureau, and the data was obtained from FRED in .csv format. To enhance data coherence and compatibility, script A.7. was executed. This script undertook the task of renaming columns and adjusting their data types within the dataset. Following these alterations, the dataframe was saved into a new .csv file.

3.2.6. *NYC Building permits*

The data source for this dataset is likewise the US Census Bureau, with the data being acquired from FRED in the .csv format. To facilitate data alignment and harmonization, script A.8. was executed, employing the same procedure outlined in Module 3.2.5.

3.2.7. *NYC Rental vacancy rates*

The data source for the Rental Vacancy Rates is also the US Census Bureau. The data was extracted from FRED in the .csv format. To achieve uniformity and coherence within the dataset, script A.9. was executed. This script performed the identical procedure elucidated in Module 3.2.5.

3.2.8. *US Consumer Confidence Index (CCI*)

The data source for this dataset was the OECD (Organisation for Economic Co-operation and Development), it was obtained from their data website in the .csv format. Subsequently, script A.10. was executed. This script's purpose was to exclude unnecessary columns, rename the

retained columns, and adjust their order within the dataset. The resultant updated dataframe was then preserved as a new .csv file.

3.2.9. *NYC House Price Index*

The data source for this dataset is the US Federal Housing Finance Agency, and the dataset was obtained from FRED in the .csv format. Script A.11. was executed to perform two operations: adjusting the data types of specific columns and renaming those columns for clarity. The resulting adjusted dataframe was then saved in a new .csv file.

3.2.10. *NYC Median Rent*

The data source for this dataset is the US Department of Housing and Urban Development's Office of Policy Development and Research, and this information was made available through their website. For each year ranging from 2003 to 2023, a .xlsx file containing median rent information for all US counties was downloaded. Script A.12. was employed to convert these files into .csv format.

Since the datasets initially encompassed information for all US counties, a data filtering process was enacted - to present data pertaining to New York City exclusively. Subsequently, column names were adjusted, and the dataset was reordered for consistency. All individual dataframes were merged into a single dataset, consolidating information on New York City's median rent.

Within this combined dataframe, an additional column was introduced to provide a single median rent value without division based on residence configuration. The script A.13. was responsible for implementing these changes.

3.2.11. *NYC Sale Data*

This dataset was obtained from the New York City Department of Finance via their website. The initial extraction process involved downloading one .xls or .xlsx file for each year between 2003 and 2022, corresponding to the five New York City Boroughs: Bronx, Brooklyn, Manhattan, Queens, and Staten Island. To facilitate data transformation, five scripts, A.14.i. to A.14.v., were executed. These scripts were designed to convert the datasets for each borough into the .csv format. It's noteworthy that the format of the .xls file changed over the years, and this converter script took that into account when performing the transformation from .xls/.xlsx to .csv.

After this initial transformation, five additional scripts, A.15.i. to A.15.v., were executed. Their purpose was to consolidate information spanning all years for each borough into a single .csv file, while ensuring consistent column names throughout the datasets. With this data now organized into five .csv files, one for each borough, each adhering to a compatible format, script A.16. was implemented. This script compiled all the information related to property sales in New York City into a single comprehensive file.

As is the case with all other datasets mentioned here, these were the initial steps applied to these datasets, and additional transformations were later performed as necessary when any of the applied steps required further refinement, following the principles of the CRISP-DM methodology.

3.3. First insights about Data

In this section, we conducted a brief examination of all eleven datasets previously detailed in the first section of this chapter. This analysis was performed within a new script, named 'House Bubble Prediction', which will serve as the platform for executing the remaining parts of the project. Our primary objective was to develop a thorough understanding of the structure of each dataset. This encompassed various aspects, including determining the number of rows and columns, identifying data types, exploring basic statistical characteristics, and assessing the extent of missing data. Table 3.1 provides a concise summary of this gathered information:

Dataset	Number of Columns	Number of Rows	Max number of empty rows
US Gross Domestic Product (GDP)	6	306	4
NYC Employment statistics	8	570	1
FED Interest Rates	2	25261	0
US Consumer Price Index (CPI)	3	8553	0
NYC Household Income	3	39	1
NYC Building Permits	3	426	12
NYC Rental Vacancy Rates	3	37	1
US Consumer Confidence Index (CCI)	3	562	0
NYC House Price Index	4	193	4
NYC Median Rent	8	21	0
NYC Sales	21	1861418	428918

Table 3.1 - Initial datasets structure

3.4. Data Preparation Revisited

In this section of the dissertation, both data cleaning steps and data understanding steps were performed. These steps include changing variable types, graphic visualizations of datasets, handling missing values, normality tests, examining linear correlations between variables, and filtering datasets. The subsection below will explain the most critical changes to each dataset.

Except for the NYC Sales dataset, which necessitated more intricate filtering procedures, the methods applied to each dataset exhibited considerable similarity. This section outlines the common practices followed. Most of these datasets showed missing values within the column responsible for representing the percentage change of the economic feature concerning the preceding available period. This deficiency arose due to the absence of data about the previous period, rendering it impossible to calculate the percentage change for that particular timeframe. Consequently, these instances of missing data were imputed with a value of zero, thereby eliminating null entries.

Furthermore, the Shapiro-Wilk test was systematically conducted for each dataset to assess the normal distribution of respective columns. The following steps involved dimensionality reduction for these datasets by removing highly correlated features. The Pearson correlation coefficient was computed, and features displaying significant correlations were subsequently eliminated.

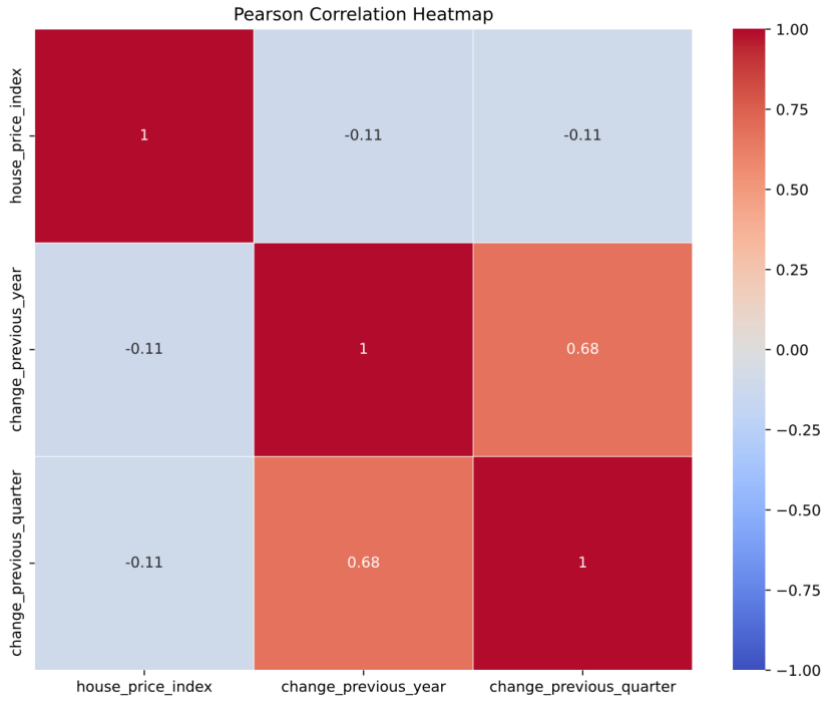

Figure 3.6 - Pearson Correlation Matrix for the House Price Index dataset

Subsequently, since most of these datasets contained information from periods preceding our sale data, a decision was made to filter the datasets to display only the data from 2002 onwards. This process significantly reduced the dimensionality of the datasets. The final step applied to each dataset was the transformation to a monthly frequency. These datasets had varying original frequencies, such as daily, monthly, quarterly, or annual. For features with a daily frequency, the monthly values were computed as the median of the daily values for that month. For features with quarterly or annual frequencies, additional monthly rows were generated for the corresponding time periods, and these values were filled using interpolation. Table 3.2 presents a summary of the operations performed for each dataset.

Dataset	Missing values	Shapiro-Wilk Test, which columns had a Normal distribution?	Columns eliminated due to high correlation	Data filters	Conversion to a monthly format
US Gross Domestic Product (GDP)	Imputed with 0	None	'nominal_gdp', 'real_gdp', 'real_gdp_change'	From 2002 onwards	Interpolation from quarterly
NYC Employment and unemployment rates	Imputed with 0	None	'labor_force', 'employed', 'employment_vs_population', 'unemployed'	From 2002 onwards	None
US Interest Rates	No missing data	None	None	From 2002 onwards	Monthly median from daily values
US Consumer Price Index (CPI)	No missing data	None	None	From 2002 onwards	Monthly median from daily values
NYC Household Income	Imputed with 0	'real_median_household_income_change'	None	From 2002 onwards	Interpolation from annual
NYC Building Permits	Imputed with 0	None	None	From 2002 onwards	None
NYC Rental Vacancy Rate	Imputed with 0	None	None	From 2002 onwards	Interpolation from annual
US Consumer Confidence Index (CCI)	No missing data	None	None	From 2002 onwards	None
NYC House Price Index	Imputed with 0	None	'change_previous_quarter'	From 2002 onwards	Interpolation from quarterly
NYC Median Rent	No missing data	'median_rent_1bdr', 'median_rent_2bdr', 'median_rent_3bdr', 'median_rent_4bdr', 'median_rent'	'city', 'median_rent_studio', 'median_rent_1bdr', 'median_rent_2bdr', 'median_rent_3bdr', 'median_rent_4bdr'	None	Interporlation from annual

Table 3.2 – Data cleaning and understating tasks undergone in each dataset

3.4.1. NYC Sale Data

For the primary dataset, the initial step involved creating a column named 'quarter,' which supported grouping sales by quarter, enabling the plotting of charts with average and median sale prices. Subsequently, as done with previous datasets, the Shapiro-Wilk normality test was applied to the sale price. This test concluded that the 'sale_price' variable did not adhere to a normal distribution.

Since this dataset encompassed all property sales in New York City, including non-residential properties, it required thorough cleaning. To initiate the cleaning process (and exclude non-residential properties), an analysis of the 'building_class_sale' variable was conducted. To gauge the extent of this task, the unique entries within each code were initially counted. Subsequently, an auxiliary column was created, with values representing the first letter of the building class code. However, certain codes denoted mixed usage, necessitating their separation into regular groups and the residential subsection of said group. This division was particularly required for class codes 'R' and 'V.' With this categorization completed, the number of non-residential entries was calculated using the NYC Building classification[2] document provided by the NYC Department of Finance to identify residential codes. Following this filtration, it was determined that the dataset contained 185,678 rows corresponding to non-residential properties that needed to be removed. This represented approximately 9.98% of the original dataset.

Another variable requiring value removal was 'sale_price.' The NYC Department of Finance noted that all transactions with a price of zero dollars denoted transactions conducted without a cash consideration, such as donations to institutions or transfers from parents to their children. Given that these values could potentially distort the perception of New York City's housing market, the decision was made to eliminate them. There were 474,810 transactions recorded without cash consideration, amounting to 25.51% of the dataset. After removing these rows, 64.51% of the original dataset remained, comprising 1,200,915 rows.

[2] A link to the New York City building class code is provided in Appendix B.

In discussions that arose during the analysis of variables for input selection, a decision was reached to retain only the sale price. This decision was based on the dynamic nature of the sale price, directly relevant to housing bubbles. All other columns were excluded because they were deemed static and did not exhibit a clear business sense connection to the study of housing bubbles. Subsequently, the dataset consisted of 1,200,915 sale prices spanning from 2003 to 2022.

To ensure uniformity with the other variables - containing economic information in monthly frequency - a new dataframe was created. Within this dataframe, each row corresponds to a specific month within the specified timeframe. These rows were populated with the median sale price relevant to their respective month.

3.4.2. Join Datasets

All datasets are merged in this stage, utilizing the 'sale_date' column within the NYC Sale dataset as the basis for indexing. An operational function named 'clean_and_merge_datasets' has been constructed to streamline this procedure. Within this function, the core dataset is designated, accompanied by an enumeration of all ancillary datasets slated for integration. A 'left' join operation is executed between the primary dataset and its supplementary counterparts. There is also an option to designate the name of the date column, although the default 'date' is retained.

In creating the 'merged_dataset,' two essential attributes commonly utilized in housing market analysis required inclusion. These attributes consist of the price-to-income ratio and the price-to-rent ratio. To calculate the former, a new column denoted 'price_to_income_ratio' was introduced. This column was formed by dividing values from the 'sale_price' column by the 'real_household_income.' For the latter, a column named 'price_to_rent_ratio' was established. Its values were computed by multiplying the 'median_rent' column by 12 and then dividing it into the 'sale_price' column.

The final step in this phase is the definition of our target variable, 'is_bubble.' To accomplish this, a plot was generated to identify growth and explosive growth periods. The specified periods led to the creation of four distinct labels:

1. **Explosive Growth**: These represent periods characterized by significant and swift increases in median sale prices. In the code, "explosive growth" is identified as months where the difference (diff) in median sale prices exceeds the 'growth_threshold,' and this upward trajectory persists over three consecutive months. Alternatively, a single month is also classified as experiencing explosive growth if the difference is

exceptionally high (greater than or equal to 30,000). These periods of explosive growth are indicated with green dashed lines on the plot.

2. **Explosive Decrease**: These signify periods marked by significant and rapid declines in median sale prices. In the code, "extreme decrease" is defined as months where the difference (diff) in median sale prices falls below the 'decrease_threshold,' and this downward trend persists over three consecutive months. Alternatively, a single month is designated as an extreme decrease if the difference is extraordinarily low (less than or equal to -30,000). These periods of extreme decrease are marked with red dashed lines on the plot.
3. **Positive Extreme**: These denote periods with substantial increases in median sale prices over a year, following a rolling 12-month approach. In the code, "positive extreme" is ascribed to intervals where the sale price increases by 80,000 or more for a year. These positive extreme periods are distinguished with yellow dashed lines on the plot.
4. **Negative Extreme**: These indicate periods characterized by substantial decreases in median sale prices over a year, following a rolling 12-month approach. In the code, "negative extreme" is attributed to intervals where the sale price decreases by 80,000 or more annually. These negative extreme periods are identified with purple dashed lines on the plot.

Subsequently, intervals between episodes of explosive growth and explosive decrease, or between periods of positive and negative extremes, were also emphasized on the graph.

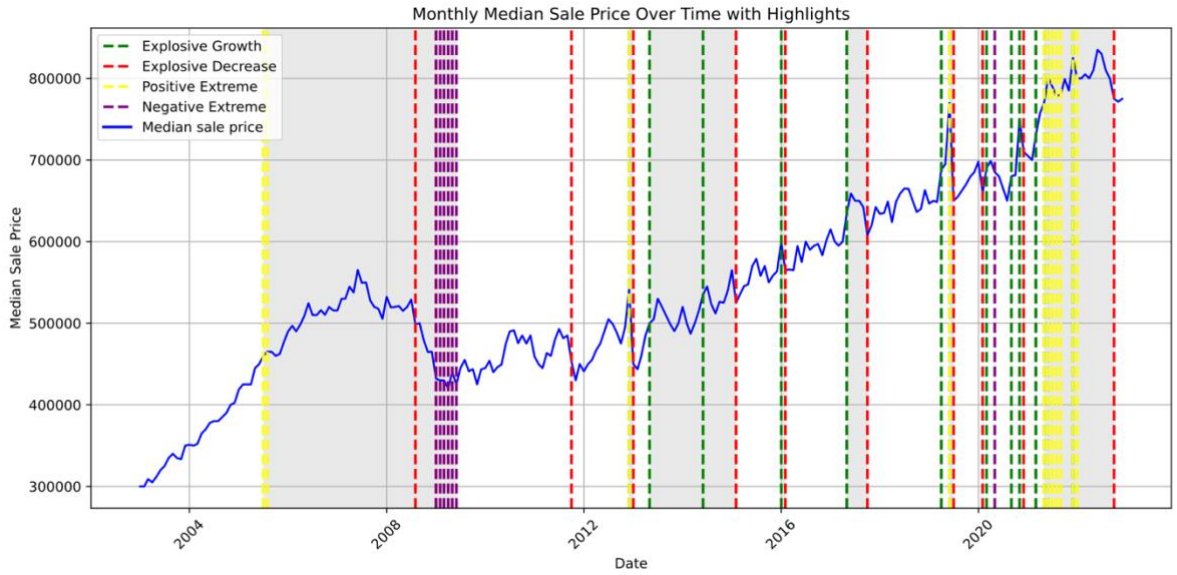

Figure 3.5 - Median sale price with periods of extreme changes in price highlighted

The highlighted intervals occurred between July 2005 and June 2009, from December 2012 to January 2013, May 2013 to February 2015, January 2016 to February 2016, May 2017 to October 2017, June 2019 to July 2019, and May 2021 to October 2022. These intervals were then employed to categorize the periods in between as a housing bubble within our newly established column 'is_bubble'.

After augmenting the columns detailed in this section to the merged dataset, the resulting dataset consisted of 240 rows (one for each month) and 24 columns[3]. Among these rows, 140 are not classified as part of a housing bubble period, while 100 are.

3.5. Modeling

This section overviews the classification models adopted and the prerequisite steps that precede model deployment. The classification task involved the application of three distinct algorithms to build the models: XGBoost, Random Forest, and a Neural Network. Each of these models underwent dual training phases. Initially, they were trained using the complete set of features designated for the specific test. Subsequently, they were retrained utilizing the subset of features identified through advanced feature selection. Three distinct tests were conducted, outlined as follows:

1. **Test 1:** The list of features employed in this test includes 'sale_price', 'real_gdp_per_capita', 'unemployment_rate', 'fed_rate', 'inflation', 'real_median_household_income', 'authorized_housing_units', 'cci', 'house_price_index', 'rental_vacancy_rate', 'median_rent', 'population', 'price_to_income_ratio', 'price_to_rent_ratio', and 'is_bubble';
2. **Test 2:** This test utilized this list of features: 'sale_price', 'real_gdp_per_capita_change', 'change_unemployment_rate', 'fed_rate', 'inflation', 'real_median_household_income_change', 'authorized_housing_units_change', 'cci', 'house_price_index_change', 'rental_vacancy_rate_change','median_rent', 'population', 'price_to_income_ratio', 'price_to_rent_ratio', 'is_bubble';
3. **Test 3:** All features within 'merged_dataset'.

Prior to model execution using the complete set of selected features, a seed number was established to ensure the reproducibility of results. Subsequently, the dataset was divided into two subsets: a training set and a test set, following a 70% to 30% ratio. This division was further delineated into target and input sets, with the target variable being 'is_bubble'.

[3] A list of the final feature names and their descriptions is provided in Appendix C.

In the test involving advanced feature selection, two distinct methods were applied: ANOVA and Lasso regularization. Both approaches followed a 70% - 30% split for data partitioning. The data was further categorized within each split into a feature matrix (x) and a target variable (y).

ANOVA, or Analysis of Variance, is a statistical technique for exploring differences among group means within a sample. In this analysis, ANOVA plays a pivotal role by identifying the top 5 features based on their p-values concerning the target variable. We deploy 'SelectKBest' from scikit-learn, utilizing the 'f_classif' score function to assess feature significance as determined by ANOVA. The resulting set, 'anova_selected_features', comprises the chosen features strongly associated with the target variable.

Lasso, short for Least Absolute Shrinkage and Selection Operator, constitutes a variant of linear regression that leverages L1 regularization. It supplements the linear regression cost function with a penalty term, fostering sparsity in feature coefficients. In our feature selection strategy, Lasso regularization is harnessed through logistic regression (LogisticRegression) featuring an L1 penalty term. The outcome, 'lasso_selected_features', encompasses the feature names that Lasso regularization has identified as significant.

Subsequently, a plot was generated, showcasing the features selected by both ANOVA and Lasso regularization. These mutually selected features were retained to retrain the model, utilizing them as inputs in the modeling process.

3.5.1. XGBoost

As mentioned earlier, XGBoost is an algorithm included in this paper's model selection. It falls under the gradient boosting algorithm family, which utilizes ensemble techniques to amalgamate predictions from multiple weak learners, often decision trees, resulting in a robust predictive model. XGBoost is known for its effectiveness in tackling various machine learning challenges, including regression, classification, ranking, and more. Its adaptability, robustness, and efficient implementation make it a compelling choice for our current classification problem.

3.5.2. Random Forest

We further provide a concise introduction to the Random Forest algorithm. They are categorized within the ensemble learning family, specifically the bagging algorithms. The primary objective of bagging algorithms is to elevate predictive accuracy by amalgamating the predictions from multiple base learners, often represented by decision trees. The selection of Random Forest is based on its reputation for simplicity, effectiveness, and robust performance

in handling classification problems. Notably, it offers the advantage of minimizing overfitting, a common issue in complex machine learning models, and provides a valuable tool for feature selection, making it an interesting tool to use as a model.

3.5.3. Neural Network

A neural network represents a computational model inspired by the architecture and functionality of the human brain. It consists of layers comprising interconnected artificial neurons, commonly known as nodes or units. These neurons are systematically arranged into an input layer, one or more hidden layers, and an output layer. The decision to employ Keras is rooted in its attributes as a high-level deep learning library, streamlining neural network design and training processes.

CAPÍTULO 4

Results and Discussion

4.1. Features chosen without feature selection

In the previous section, 3.5., three distinct Tests were presented to address various hypotheses regarding prediction of housing bubbles. The rationale behind the feature selection in Test 1 aimed to retain economic variables in their original state to evaluate their effectiveness in predicting the target value. In Test 2, the hypothesis tested was the connection between these economic variables and the presence of housing bubbles, while considering that changes in these variables might provide improved predictive inputs. Finally, Test 3 initially employed all dataset variables as inputs to assess the model's performance without potential bias.

4.2. Advanced Feature Selection

When implementing feature selection with the feature set listed in Test 1, the features selected by both ANOVA and Lasso regularization include interest rates, inflation, consumer confidence index, and the price-to-rent ratio. These features effectively differentiate between the two groups in our target. This outcome aligns with the expectations based on previous research, which consistently emphasized the significance of these features. The selection of the top five features, with four of them being the same in both methods, reflects a consensus on feature importance. This suggests that the selected features are not arbitrary but exhibit a robust relationship with the target variable.

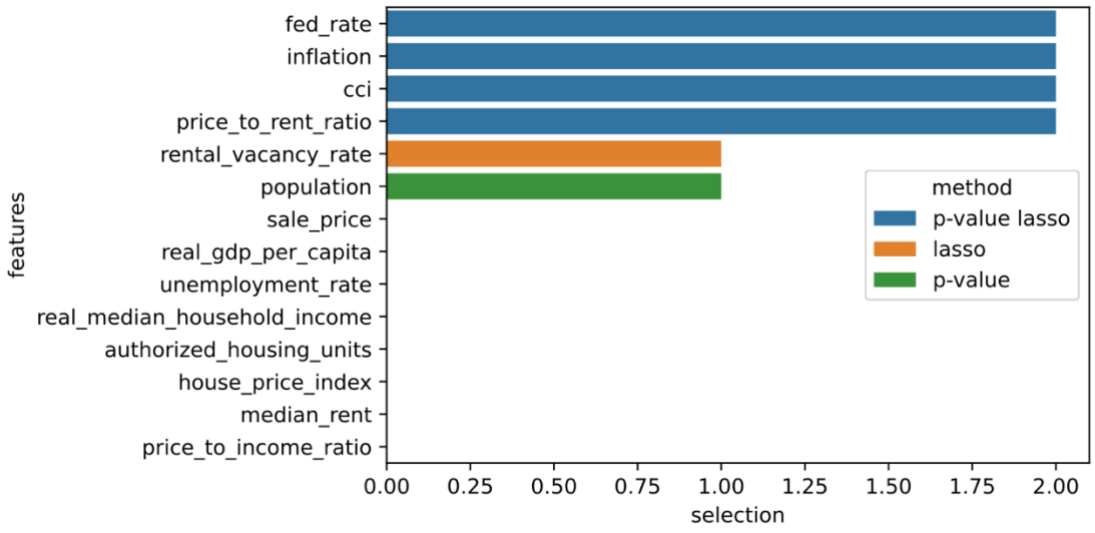

Figure 4.1 - Feature selection with ANOVA and Lasso regularization for test 1

Transitioning to Test 2, we observe that the degree of feature importance agreement and consistency is not as pronounced between the two methods as with the features in Test 1. Nevertheless, we note that two of the features selected by both ANOVA and Lasso regularization in Test 2 align with two features chosen in the previous test. Notably, the price-to-rent ratio and inflation are among the features that recurrently appeared in the literature review for this paper.

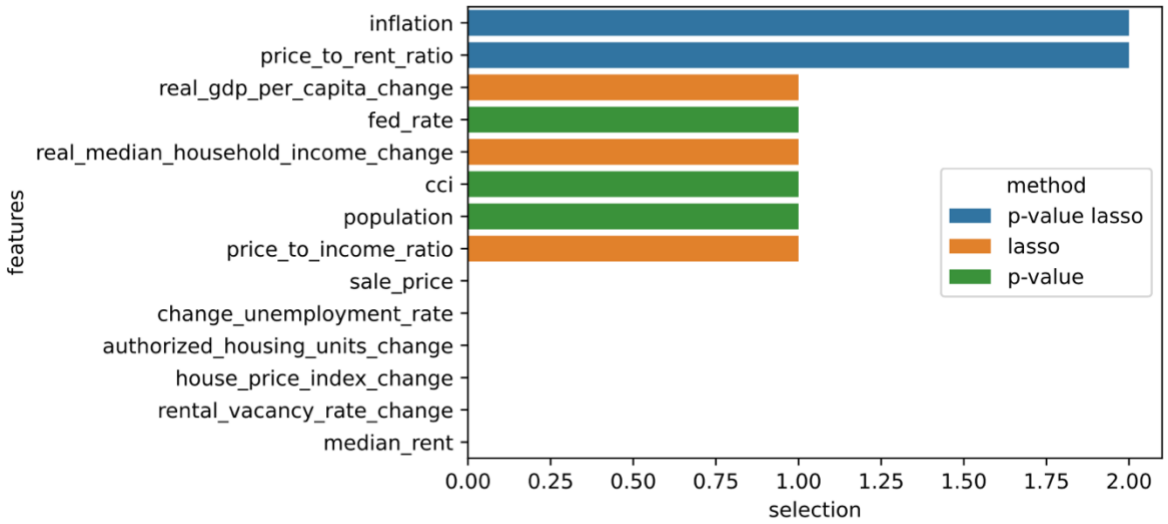

Figure 4.2 - Feature selection with ANOVA and Lasso regularization for test 2

In conclusion, Test 3 yields results quite similar to those of Test 2 in terms of feature importance agreement and consistency. It's noteworthy that both features selected in Test 3 were also chosen in Test 1. Of particular interest is the price-to-rent ratio, the only feature consistently selected by both methods in all tests. This feature holds significant importance in housing bubble detection, with Bourassa et al. (2019) even describing it as the most reliable indicator for identifying housing bubbles.

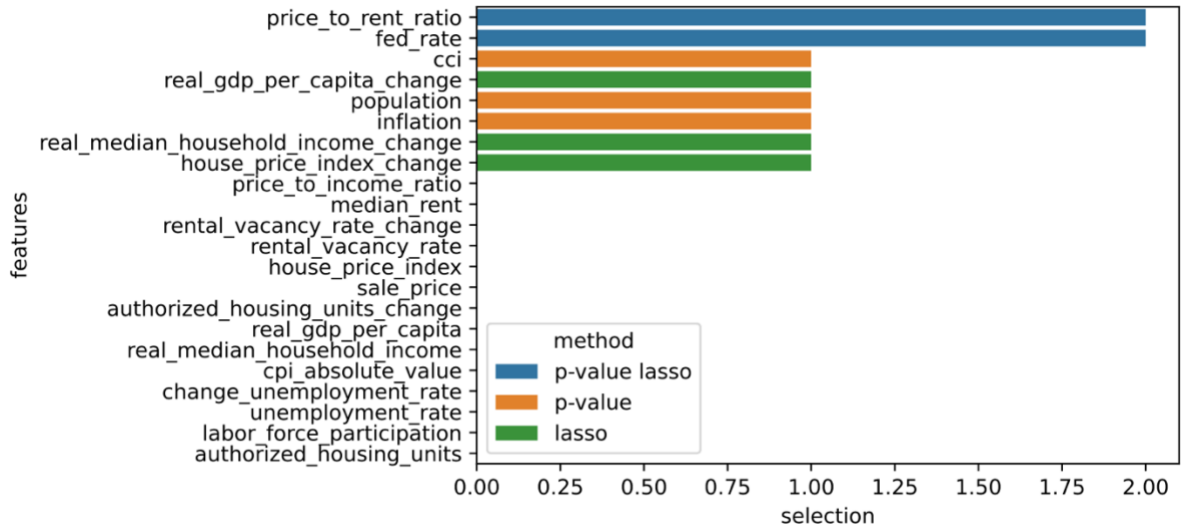

Figure 4.3 - Feature selection with ANOVA and Lasso regularization for test 3

With the features identified for scenarios with and without feature selection, the next step is to proceed to the results of the models constructed using XGBoost, Random Forest, and the Neural Network.

4.3. XGBoost with and without Advanced Feature Selection

4.3.1. XGBoost with and without Advanced Feature Selection – Test 1

The results displayed considerable promise in the initial XGBoost test conducted without advanced feature selection. The model achieved a noteworthy accuracy score of 0.89. Furthermore, the performance was characterized by balanced classes, as indicated by the f1-scores.

Upon evaluating feature importance, it became evident that the most influential determinants within the ensemble of trees were real median household income, the price-to-rent ratio, and inflation. This observation is consistent with findings from various researchers, including Lecat and Mésonnier (2005) and Gwartney et al. (2004) regarding household income, as well as studies by Masiukiewicz and Dec (2015) and Bourassa et al. (2019) concerning the price-to-rent ratio, along with research by Leung (2004) and Lu et al. (2015) on inflation.

For this same test set, XGBoost was implemented with advanced feature selection, and a slight decline in accuracy to 0.86 was observed. Nonetheless, the f1-scores demonstrated that the model consistently maintained balanced predictions among the classes. This marginal reduction in accuracy might be linked to the loss of information resulting from removing specific features compared to the test without feature selection. However, this trade-off led to the creating of a more straightforward and interpretable model that focused on the most influential features.

When examining feature importance, there was almost a tie among three key features. These features, ranked in order of significance, included inflation, the consumer confidence index, and the price-to-rent ratio. Prior studies have also pinpointed inflation and the price-to-rent ratio as vital factors. What distinguished this analysis was the notable importance assigned to the consumer confidence index. This discovery resonated with the research by Kindleberger (1987), which implied that individuals might be willing to pay more for a house when they have a high level of confidence in the asset's future appreciation. While this particular factor was not extensively explored in existing literature, it emerged as an intriguing feature for analysis due to its potential influence on consumer spending, including in the housing market, possibly leading to property valuations surpassing fundamental values.

4.3.2. XGBoost with and without Advanced Feature Selection – Test 2

In the second test set employing XGBoost without advanced feature selection, the model's performance, while slightly inferior to that in Test 1, still delivered favorable results. An analysis of feature importance unveiled the three most influential features: sale price, the price-to-rent ratio, median rent, and the change in real household income. Notably, the last two features shared the same level of importance. It's worth mentioning that the sale price, which had also been employed in Test 1, had not previously exhibited such high feature importance. This variation in importance was intriguing, considering that sale price is a fundamental indicator in housing bubble prediction, as affirmed by the research of Kindleberger (1987) and Mayer (2011).

Furthermore, the change in real household income was of particular interest due to its noteworthy importance, as supported by studies conducted by Afxentiou et al. (2022), Shen et al. (2005), Lecat and Mésonnier (2005), and Gwartney et al. (2004).

Later, in the second test, when applying XGBoost with advanced feature selection, the model's performance experienced a significant decline compared to the test without feature selection. Nonetheless, it is important to highlight that the model still maintained a relatively high level of accuracy, especially considering it relied on just two input features. This outcome underscores the exceptional significance associated with these specific features.

Comparing the feature importance, it was observed that inflation narrowly surpassed the price-to-rent ratio, albeit by a small margin. As previously mentioned, both of these features had been identified as critical in housing bubble detection and prediction, and this test reaffirmed the significance of these findings.

Despite the evident performance decrease with feature selection, it's noteworthy that the model maintained a relatively high accuracy rate, even with just two input features. This outcome further highlights the paramount importance of these specific features within the context of the research problem.

4.3.3. XGBoost with and without Advanced Feature Selection – Test 3

The third test, utilizing XGBoost without advanced feature selection, emerged as the most robust model in the analysis. It achieved the highest accuracy and demonstrated exceptionally high f1-scores.

Upon exploring the features with the highest feature importance, three key factors came to the forefront: the percental change in the house price index, the change in real median household income, and the price-to-rent ratio. Notably, house price indexes have been recognized as one of the most critical features for identifying housing bubbles, a point underscored by the work of Masiukiewicz and Dec (2015). It's worth mentioning that the labor force participation rate emerged as the fourth most important feature. This is significant, as it served as one of the control macroeconomic features employed by Leung (2004) in their research.

In this same test, a substantial drop in performance became evident when XGBoost was applied with advanced feature selection. This decline was primarily attributed to reducing the model's input variables to just two. It's important to note that additional tests, although not included in this report, indicated that performance could be significantly improved by introducing one more feature selected by either the ANOVA or Lasso regularization methods. For instance, if, in addition to the price-to-rent ratio and the federal interest rate, inflation was also included, the accuracy would rise to 0.86, a performance that matches one of the first tests.

However, despite the performance drop, it can be concluded once again that the price-to-rent ratio and interest rates remain of great significance in detecting housing bubbles. In terms of their feature importance, the model indicated that the price-to-rent ratio held significantly more importance than the interest rate, although the latter still retained a high level of significance. This reaffirmed the critical role of these variables in the context of housing bubble detection.

Test	Without advanced feature selection	With advanced feature selection
1	F1-score false: 0.91 F1-score true: 0.86 Accuracy: 0.89	F1-score false: 0.88 F1-score true: 0.83 Accuracy: 0.86
2	F1-score false: 0.90 F1-score true: 0.84 Accuracy: 0.88	F1-score false: 0.79 F1-score true: 0.69 Accuracy: 0.75
3	F1-score false: 0.92 F1-score true: 0.87 Accuracy: 0.90	F1-score false: 0.76 F1-score true: 0.62 Accuracy: 0.71

Table 4.1 - Test results with and without feature selection using XGBoost

4.4. Random Forest with and without Advanced Feature Selection

4.4.1. Random Forest with and without Advanced Feature Selection – Test 1

In the first test, Random Forest was employed without advanced feature selection, yielding results in terms of accuracy and f1-scores similar to those achieved with XGBoost.

The top three features identified were the price-to-rent ratio, rental vacancy rate, and median rent. The price-to-rent ratio, a widely utilized feature in housing bubble detection, continued demonstrating its significance as the feature with the highest importance in the model.

Additionally, the rental vacancy rate, as employed by Shen et al. (2005), played a notable role in this analysis. It serves as a valuable indicator for assessing housing market conditions. A surplus of rental properties, as indicated by a high vacancy rate, could signify excessive construction or insufficient demand, factors that could contribute to the formation of a housing bubble.

The importance of median rent was also highlighted. Research by Hung and Tzang (2021) established a connection between rent values and the prices consumers are willing to pay for housing. This feature's high importance underscores its impact on sale prices and, consequently, its role in shaping or bursting a housing bubble.

Random Forest was applied with advanced feature selection for the same first test set, and a more significant drop in results was observed. Although the accuracy remained at 0.81, it's worth noting that the f1-score for the True class decreased to 0.78.

Despite this decrease in performance, the model with advanced feature selection still yielded respectable results. Its simplicity and interpretability make it a viable choice, even with reduced accuracy. Concerning feature importance, the model assigned nearly equal significance to all features, a pattern akin to what was observed in the XGBoost model with advanced feature selection.

Notably, despite the closely ranked features, inflation maintained its position as the most significant feature, aligning with the conclusions reached by Tsatsaronis and Zhu (2004) and Lu et al. (2015). This consistency reaffirms the importance of inflation in housing bubble detection.

4.4.2. Random Forest with and without Advanced Feature Selection – Test 2

In the second test, Random Forest was utilized without advanced feature selection, resulting in performance very similar to that achieved with XGBoost. Interestingly, even the feature importance rankings were identical, a notable contrast from what was observed in Test 1.

Additionally, for this second test set, Random Forest was applied with advanced feature selection, the results significantly deteriorated, falling below the performance levels observed for XGBoost with advanced feature selection. When exploring feature importance in this context, both features were found to have very similar levels of significance. However, in the case of Random Forest, the price-to-rent ratio held a slightly higher level of importance, in contrast to the feature importance rankings observed in Test 2 using XGBoost. This divergence suggests that the choice of the algorithm can influence the relative importance of features, underscoring the importance of carefully considering the selection of algorithms in housing bubble detection.

4.4.3. Random Forest with and without Advanced Feature Selection – Test 3

In the third test, Random Forest was employed without advanced feature selection, yielding the highest model performance observed thus far. The performance matched that of test 3 using XGBoost without feature selection.

Regarding feature importance, the price-to-rent ratio emerged as the most critical feature, holding a substantial lead over the others. Following this key feature were inflation and rental vacancy rates in the second and third positions, which exhibited similar levels of importance.

Again, in Test 3, Random Forest was employed with advanced feature selection, and a significant drop in performance was observed. This decrease in performance was attributed to the strict use of features selected by both ANOVA and Lasso regularization methods. However, this approach ensured that the selected input features substantially impacted the model's predictive capabilities.

The price-to-rent ratio held the highest importance between the two selected features, although interest rates still carried considerable influence. This observation highlights the importance of these features in the context of housing bubble prediction, even in a reduced-feature model.

Test	Without advanced feature selection	With advanced feature selection
1	F1-score false: 0.91 F1-score true: 0.86 Accuracy: 0.89	F1-score false: 0.87 F1-score true: 0.78 Accuracy: 0.83
2	F1-score false: 0.90 F1-score true: 0.84 Accuracy: 0.88	F1-score false: 0.75 F1-score true: 0.64 Accuracy: 0.71
3	F1-score false: 0.92 F1-score true: 0.87 Accuracy: 0.90	F1-score false: 0.78 F1-score true: 0.62 Accuracy: 0.72

Table 4.2 - Test results with and without feature selection using Random Forest

4.5. Neural Network with and without Advanced Feature Selection

4.5.1. Neural Network with and without Advanced Feature Selection – Test 1

For the first test set, a neural network was employed without advanced feature selection, resulting in suboptimal performance. The model struggled to achieve meaningful results, with an F1-score of 0.00 for the false category. This disappointing outcome can be attributed to the relatively low number of training set rows compared to the number of features. The model faced challenges in identifying significant patterns in the data, and there was a high risk of noise playing a substantial role in the model's predictions.

Later, when a neural network was employed with advanced feature selection, in the first test set, the performance experienced a significant improvement. The model achieved an accuracy of 0.79 and displayed much more balanced classes, as evident in the f1-scores.

In terms of feature importance, inflation emerged as the most critical feature, followed by the price-to-rent ratio and the Consumer Confidence Index (CCI). Notably, interest rates exhibited relatively low feature importance in this model, which deviates from the findings of most other authors in the field. However, it's important to acknowledge that this particular model had the lowest performance among those considered in the study. This observation highlights the intricate relationship between feature importance and model performance, emphasizing the necessity for meticulous feature selection in neural network-based housing bubble prediction.

4.5.2. Neural Network with and without Advanced Feature Selection – Test 2

In the second test, a neural network was employed without advanced feature selection, resulting in better performance than the first test. However, the overall performance was still relatively modest. There was a significant increase in accuracy, particularly noteworthy for the improved class balance, as evidenced by the substantial difference in the f1-score for the False class.

Regarding feature importance, sale price and median rent emerged as the most pivotal features, with exceptionally low importance values assigned to the other features.

A neural network was utilized now with advanced feature selection, in this same second test set, the performance further improved, yielding satisfactory results, especially considering that only two input features were employed.

Remarkably, the model assigned the majority of feature importance to inflation. However, it's crucial to acknowledge that due to the limited size of the training set, this model remained less robust compared to the one using Random Forest with advanced feature selection. Although the neural network had a slightly better accuracy score, the f1-scores suggested that the Random Forest model with advanced feature selection provided a closer representation of the real-world dynamics of the problem under study. This comparison highlights the trade-offs and considerations involved in selecting an appropriate model for housing bubble prediction, especially when dealing with limited training data.

4.5.3. Neural Network with and without Advanced Feature Selection – Test 3

In the third test, a neural network was employed without advanced feature selection, achieving the highest accuracy among the tests that used this approach. However, despite the marginally better accuracy, the model exhibited an inferior class balance in the f1-scores compared to Test 2. This observation suggests that, overall, the model's performance was less satisfactory.

The features with greater importance in this scenario were sale price, median household income, and median rent.

Afterwards, as done with previous tests, the third test set had its neural network used with advanced feature selection; there was a significant increase in accuracy compared to the model without advanced feature selection. However, it's important to note that the classification success rate for each class dropped significantly, as evidenced by the reduced f1-score for the True class.

Regarding feature importance, almost all the significance was attributed to the interest rates, with the price-to-rent ratio being considered less important in this test. It's worth noting that the relatively small training set for this neural network could significantly influence this outcome, emphasizing the importance of dataset size in the performance and feature importance of neural network models.

Test	Without advanced feature selection	With advanced feature selection
1	F1-score false: 0.00 F1-score true: 0.59 Accuracy: 0.42	F1-score false: 0.84 F1-score true: 0.69 Accuracy: 0.79
2	F1-score false: 0.62 F1-score true: 0.51 Accuracy: 0.57	F1-score false: 0.80 F1-score true: 0.57 Accuracy: 0.72
3	F1-score false: 0.67 F1-score true: 0.44 Accuracy: 0.58	F1-score false: 0.75 F1-score true: 0.35 Accuracy: 0.64

Table 4.3 - Test results with and without feature selection using a Neural Network

In summary, most results provide valuable insights into the key features for predicting housing bubbles. These findings align with the literature review, highlighting the importance of these features in preventing or controlling housing bubble situations. Notably, while the number of authorized housing units for construction had some feature importance in the models, it was not among the top most important features. Additionally, ANOVA or Lasso regularization did not select it as a feature to retain. This observation is interesting, as it differs from the findings of Dec et al. (2022), Afxentiou et al. (2022), Davis and Heathcote (2001), and Shen et al. (2005), who identified significant relationships between the number of construction units and the existence of a housing bubble.

CHAPTER 5

Concluding Remarks

The prediction of the existence of housing bubbles is a complex topic that commences with defining what constitutes a housing bubble. In this study, we primarily adopted the definitions provided by Kindleberger (1987) and Mayer (2011), emphasising the significant and rapid increase in housing prices.

Our research concluded that numerous macroeconomic features are associated with housing bubbles, with some being more pertinent than others. Based on our findings, it can be asserted that certain features commonly used to detect the existence of a housing bubble include the price-to-rent ratio, as highlighted by Masiukiewicz and Dec (2015), Bourassa et al. (2019); interest rates, as mentioned by Tsai and Lin (2022), Taipalus (2006), Hung and Tzang (2021), Afxentiou et al. (2022), Ioannides and Englund (1997), Tsatsaronis and Zhu (2004), Lecat and Mésonnier (2005), and Lu et al. (2015); and inflation, as indicated by Leung (2004), Tsatsaronis and Zhu (2004), and Lu et al. (2005).

Altogether, 18 models were developed during this study, employing XGBoost, Random Forest, and Neural Networks, both with and without advanced feature selection, using the ANOVA and Lasso regularization methods.

Among these models, the pair that yielded the most favorable results was XGBoost in Test 1, which predominantly incorporated economic variables in their original form using absolute values. In this configuration, the three most critical features for housing bubble classification were the real median household income, the price-to-rent ratio, and inflation without advanced feature selection. Conversely, with advanced feature selection, the significant features were inflation, the Consumer Confidence Index (CCI), and the price-to-rent ratio.

It is noteworthy to mention that the role of income and consumer confidence in the context of housing bubbles was also examined by Lecat and Mésonnier (2005), Gwartney et al. (2004), and Kindleberger (1987).

In contrast, the models that exhibited the poorest performance were those developed using neural networks, which can be attributed to the limited size of the training set.

It is noteworthy that the price-to-rent ratio consistently emerged as highly significant in all models, followed by interest rates and inflation, which featured prominently in most models. These findings lead us to conclude that research conducted in other housing markets appears to apply to the New York City housing market, as the most important features in our models align with those identified in other studies.

As future work in this field, it would be intriguing to explore a model that exclusively utilizes features with a daily frequency. A larger number of data points could potentially enhance the performance of a neural network model, making it more feasible. Additionally, a model incorporating stock indexes, similar to the approach employed by Shen et al. (2005), could provide valuable insights into whether the dynamics of stock bubbles extend to real estate bubbles.

This research aims to contribute to a more comprehensive understanding of the topic and provide valuable insights into the macroeconomic indicators that warrant close monitoring. This knowledge can help stakeholders take proactive measures to prevent the emergence of housing bubbles. In cases where such bubbles already exist, this research equips stakeholders with the information needed to know which features to address in order to manage the situation effectively and strategically, minimizing potential externalities and ensuring a more stable housing market.

References

Kindleberger, C. (1987), "Bubbles, Eatwell, J., Milgate, M. and Newman", P. (Eds.), The New Palgrave: A Dictionary of Economics, Stockton Press, New York, New York, p. 281

H. Akaike (1973), "Information Theory as an Extension of the Maximum Likelihood Principle", in B. N. Petrov, and F. Csaki, (Eds.), Second International Symposium on Information Theory, Akademiai Kiado, Budapest, pp. 267-281.

D.T. Anderson, J.C. Bezdek, M. Popescu, and J.M. Keller (2010), "Comparing Fuzzy, Probabilistic, and Possibilistic Partitions", IEEE Transactions on Fuzzy Systems, 18(5), 906-918.

Tsai, I. -., & Lin, C. -. (2022). A re-examination of housing bubbles: Evidence from european countries. Economic Systems, 46(2)

Li, S., Liu, J., Dong, J., & Li, X. (2021). 20 years of research on real estate bubbles, risk and exuberance: A bibliometric analysis. Sustainability (Switzerland), 13(17)

Hung, C. -., & Tzang, S. -. (2021). Consumption and investment values in housing price: A real options approach. International Journal of Strategic Property Management, 25(4), 278-290.

Masiukiewicz, P., Dec, P. (2015). Behawioralne aspekty baniek ceowych i sposoby ich dezaktywacji (Behavioral aspects of price bubbles and ways to deactivate them). Ekonomista, 4.

Rantala, J., Rantanen, A., Yllikäinen, M., & Holopainen, T. (2021). Weakness of real estate collateral valuation policy in changed financial world

Lepenioti, K., Bousdekis, A., Apostolou, D., & Mentzas, G. (2020). Prescriptive analytics: Literature review and research challenges. International Journal of Information Management, 50, 57-70.

Afxentiou, D., Harris, P., & Kutasovic, P. (2022). The COVID-19 Housing Boom: Is a 2007–2009-Type Crisis on the Horizon? Journal of Risk and Financial Management, 15(8), 371.

Dec, P., Główka, G., & Masiukiewicz, P. (2022). Price Bubbles in the Real Estate Markets - Analysis and Prediction. WSEAS Transactions on Business and Economics, 19, 292-303.

Leung, C. (2004). Macroeconomics and Housing: A Review of the Literature. Chinese University of Hong Kong.

Davis, M., and Heathcote, J. (2001). Housing and the Business Cycle. International Economic Review, Department of Economics, University of Pennsylvania and Osaka University Institute of Social and Economic Research Association, Vol. 46, (3): 751- 784.

Englund, P., and Ioannides, Y. M. (1997). House Price Dynamics: An International Empirical Perspective. Journal of Housing Economics, Vol. 6, (2): 119-136.

Tsatsaronis, K. and Zhu, H. (2004). What Drives Housing Price Dynamics: Cross- Country Evidence. BIS Quarterly Review.

Lecat, R., & Mésonnier, J.-S. (2005). What role do financial factors play in house price dynamics? Banque de France Bulletin Digest, No. 134, 27, 31-33.

National Aboriginal Capital Corporation Association, NACCA, (2005). The Role of Housing in an Economy.

Phillips, P. C. B., Shi, S. P., and Yu, J. (2011). Testing for Multiple Bubbles. Singapore Management University, Working Paper Series No. 09-2011.

Bourassa S. C., Hoesli M., Oikarinen E .(2019) Measuring house price bubbles. Real Estate Econ 47(2):534–563

Taipalus K. (2006) A global house price bubble? Evaluation based on a new rent-price approach. Bank of Finland research discussion papers

Lu LY, Lee JYM, Al-Mulali U, Ahmad NA, Mohammad IS (2015) Housing bubble in Penang prediction and determinants. J Teknol 73(5)
Shen Y, Hui EC, Liu H (2005) Housing price bubbles in Beijing and Shanghai. Manag Decis 43(4):611–627
Dipasquale, D. and Wheaton, W.C. (1994), ‘‘Housing market dynamics and the future of housing prices’’, Journal of Urban Economics, Vol. 35, pp. 1-27.
Gwartney, J.D. et al. (2004), Economics: Private and Public Choice (Chinese Edition), CITIC Publishing House, Beijing, p. 369.
Adelino, M., Schoar, A., & Severino, F. (2018). The role of housing and mortgage markets in the financial crisis. Annual Review of Financial Economics, 10, 25-41.
Malone, T. (2022, August 24). The Beginning of the End? Single-Family Investor Activity Takes a Step Back in Q2.
Glaeser, E. L., Gyourko, J., & Saiz, A. (2008). Housing supply and housing bubbles. Journal of Urban Economics, 64(2), 198-217.
Mayer, C. (2011). Housing Bubbles: A Survey. Annual Review of Economics. 3.
Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., & Duchesnay, É. (2011). Scikit-learn: Machine Learning in Python. Journal of Machine Learning Research, 12(85), 2825-2830.
Yiu, M. S., Yu, J., & Jin, L. (2013). Detecting bubbles in Hong Kong residential property market. Journal of Asian Economics, 28, 115-124.
Cevik, S., & Naik, S. (2023). Bubble Detective: City-Level Analysis of House Price Cycles. IMF Working Paper, European Department
Hou, Y. (2010). Housing price bubbles in Beijing and Shanghai? International Journal of Housing Markets and Analysis, 3(1), 17-37.
Ayan, E., & Eken, S. (2021). Detection of price bubbles in Istanbul housing market using LSTM autoencoders: a district-based approach. Soft Computing, Advance online publication. https://doi.org/10.1007/s00500-021-05677-6
Rouwendal, A., & Longhi, S. (2008). The Effect of Consumers' Expectations in a Booming Housing Market: Space-time Patterns in the Netherlands, 1999–2000. Housing Studies, 23(2), 291-317. https://doi.org/10.1080/02673030801893107.
NYC Department of Finance. (2022). Property Annualized Sales Update. https://www.nyc.gov/site/finance/taxes/property-annualized-sales-update.page
U.S. Census Bureau. (2023). Rental Vacancy Rate for New York [NYRVAC], retrieved from FRED, Federal Reserve Bank of St. Louis; https://fred.stlouisfed.org/series/NYRVAC
U.S. Census Bureau. (2023). New Private Housing Structures Authorized by Building Permits for New York-Newark-Jersey City, NY-NJ-PA (MSA) [NEWY636BPPRIV], retrieved from FRED, Federal Reserve Bank of St. Louis; https://fred.stlouisfed.org/series/NEWY636BPPRIV
OECD. (2023). Consumer confidence index (CCI) (indicator). doi: 10.1787/46434d78-en.
U.S. Federal Housing Finance Agency. (2023). All-Transactions House Price Index for New York [NYSTHPI], retrieved from FRED, Federal Reserve Bank of St. Louis; https://fred.stlouisfed.org/series/NYSTHPI
U.S. Bureau of Economic Analysis. (2023). Gross Domestic Product [GDP], retrieved from FRED, Federal Reserve Bank of St. Louis; https://fred.stlouisfed.org/series/GDP
U.S. Census Bureau. (2023). Real Median Household Income in New York [MEHOINUSNYA672N], retrieved from FRED, Federal Reserve Bank of St. Louis; https://fred.stlouisfed.org/series/MEHOINUSNYA672
Malone, T. (2022, August 24). The Beginning of the End? Single-Family Investor Activity Takes a Step Back in Q2. Corelogic. Retrieved from

https://www.corelogic.com/intelligence/the-beginning-of-the-end-single-family-investor-activity-takes-a-step-back-in-q2/
US Department of Labor. (2023). Developments in the New York City Labor Market. https://dol.ny.gov/labor-statistics-new-york-city-region
Board of Governors of the Federal Reserve System (US). (2023). Federal Funds Effective Rate Selected Interest Rates (Daily) - H.15. https://www.federalreserve.gov/releases/h15/
US Department of Housing and Urban Development's Office of Policy Development and Research. (2023). 50th Percentile Rent Estimates. https://www.huduser.gov/portal/datasets/50per.html
International Monetary Fund. (2023). Consumer Price Index (CPI). https://data.imf.org/?sk=4ffb52b2-3653-409a-b471-d47b46d904b5

Appendix A

A. Scripts used in initial data transformation:

1. 'US GDP – Column Name Change.ipynb'
2. ‘'xlsx to csv converter - NYC Employment Statistics.ipynb'
3. 'NYC Employment Statistics - Column Name Change.ipynb'
4. 'FED Rates – Column Name Change'
5. 'xlsx to csv converter - CPI.ipynb'
6. 'Combine CPI into Single CSV (includes cleaning of missing values).ipynb'
7. 'NYC Household Income - Column Name Change.ipynb'
8. 'NYC Building Permits - Column Name Change.ipynb'
9. 'NYC Rental Vacancy - Column Name Change.ipynb'
10. 'CCI - Column Name Change.ipynb'
11. ‘NYC House Price Index - Column Name Change.ipynb'
12. ‘xlsx to csv converter - Median NYC Rent.ipynb'
13. ‘NYC Median Rent - Column Name Change and Merger.ipynb'
14. 'xlsx to csv converter for NYC sales – borough.ipynb'
    i. ‘'xlsx to csv converter for NYC sales – Bronx.ipynb'
    ii. ‘'xlsx to csv converter for NYC sales – Brooklyn.ipynb'
    iii. ‘'xlsx to csv converter for NYC sales – Manhattan.ipynb'
    iv. ‘'xlsx to csv converter for NYC sales – Queens.ipynb'
    v. ‘'xlsx to csv converter for NYC sales – Staten Island.ipynb'
15. 'Combine borough data into Single CSV.ipynb'
    i. 'Combine Bronx data into Single CSV.ipynb'
    ii. 'Combine Brooklyn data into Single CSV.ipynb'
    iii. 'Combine Manhattan data into Single CSV.ipynb'
    iv. 'Combine Queens data into Single CSV.ipynb'
    v. 'Combine Staten Island data into Single CSV.ipynb'
16. ‘'Combine all boroughs data into Single CSV.ipynb'

Appendix B

B. NYC Sales Dataset auxiliary itens:

1. New York City’s building code classification
2. Glossary of Terms for Property Sales Files

Appendix C

C. Model details

C.1. Data Details – Merged dataset

Frequency: Monthly

Range: January 2003 to December 2022, 240 observations

Individual Series details:

Sale_price: Median sale price of housing units in NYC in the time period of that row.

real_gdp_per_capita: Real gross domestic product per capita. Real Gross Domestic Product (GDP) per capita is a measure of the economic performance and standard of living in a country. It represents the total economic output of a country, adjusted for inflation, divided by its population. This metric is often used to assess and compare the relative prosperity and economic well-being of different countries or regions.

real_gdp_per_capita_change: Percentual change in real GDP per capita compared to the previous year available.

labor_force_participation: Proportion of New York City's working-age population that is either employed or actively seeking employment.

unemployment_rate: The unemployment rate is the percentage of the total New York City labor force that is currently unemployed and actively seeking employment. It is a key economic indicator that reflects the health of an economy, with a higher unemployment rate indicating a greater level of economic distress.

change_unemployment_rate: Percentual change in the unemployment rate in New York City when compared to the previous year available.

fed_rate: The Fed rate, is the interest rate at which depository institutions lend reserve balances to other depository institutions overnight. It is one of the most important tools used by the U.S. Federal Reserve to implement monetary policy and influence the overall economic and financial conditions in the United States.

inflation: Inflation, as the percentage change in the Consumer Price Index (CPI), represents the increase in the overall price level of a defined basket of consumer goods and services over a specific period. It quantifies the erosion of the purchasing power of a currency.

cpi_absolute_value: The Consumer Price Index (CPI) is a measure of the average change over time in the prices paid by urban consumers for a market basket of consumer goods and services, with the base period value set at 100. It is used to track and compare changes in the price level of this basket of goods and services relative to the prices in the base period.

real_median_household_income: The Real Median Household Income for New York City. It is the income that divides the household income distribution into two equal parts, with half the households earning more and half earning less. It is adjusted for inflation to account for changes in the real value of income over time, providing a more accurate picture of the purchasing power and economic well-being of the median household.

real_median_household_income_change: Percentual change in New York City's real median income when compared to the previous year.

authorized_housing_units: This series represents the total number of building permits for all structure types in New York City.

authorized_housing_units_change: Percentual change in New York City's number of building permits.

cci: The Consumer Confidence Index (CCI) is a numerical measure that reflects the degree of optimism or pessimism among consumers regarding the current and future economic conditions within a specific country.

house_price_index: This House Price Index is a statistical measure designed to track and assess changes in the prices of residential properties over time within New York City.

house_price_index_change: Percentual change in the New York City's house price index when comparaed to the previous available year.

rental_vacancy_rate: The Rental Vacancy Rate for New York City, it is a measure that indicates the percentage of available rental housing units that are vacant and not currently occupied by tenants.

rental_vacancy_rate_change: Percentual change in New York City's rental vacancy rate when compared to the information available for the previous year.

median_rent: The median rent paid by tenants for residential properties in New York City.

Population: Number of people living in New York City in the designated time period.

price_to_income_ratio: The Price-to-Income Ratio is a measure that quantifies the relationship between residential property prices and household income in a specific location. It was calculated by dividing the median home price by the median household income in New York City.

price_to_rent_ratio: The Price-to-Rent Ratio is a measure that evaluates the cost-effectiveness of buying a home compared to renting one in a specific location. It was calculated by dividing the median home price by the median rent in New York City.

is_bubble: Defines if there is a housing bubble for the time period specified in that row.