ISSN: 3008-1394

A Comparative Study of Machine Learning Algorithms for Predicting Housing Prices in Chinese Cities

Jianwei Wang[1], Xiaoping Chen[2], Lin Li[3]

[1,2,3]School of Computer Science and Technology, Nanjing University of Posts and Telecommunications, Nanjing 210023, China.

Abstract

This study conducts a comparative analysis of machine learning algorithms for predicting housing prices in major Chinese cities, addressing the growing complexity and economic significance of the real estate market. With housing affordability becoming a critical socioeconomic issue, accurate price prediction models are essential for policymakers, investors, and urban planners. The research evaluates the performance of four algorithms—Linear Regression, Decision Trees, Random Forests, and Gradient Boosting Machines—using a dataset comprising historical transaction records, macroeconomic indicators, and location-based features from 10 major Chinese cities. Results indicate that ensemble methods, particularly Gradient Boosting Machines, achieve the highest predictive accuracy, with a mean absolute error reduction of up to 18% compared to traditional linear models. The findings underscore the importance of incorporating non-linear relationships and feature interactions in housing price modeling. This study not only provides a practical framework for real estate valuation but also highlights the potential of advanced machine learning techniques in addressing urban economic challenges.

Keywords

Machine Learning, Housing Price Prediction, Ensemble Methods, Urban Economics.

Chapter 1: Introduction

1.1 Research Background

The rapid urbanization and economic transformation in China over the past three decades have positioned the real estate sector as a crucial component of the national economy. Housing markets in major Chinese cities have experienced unprecedented growth and volatility, creating both opportunities and challenges for various stakeholders. The complex interplay between demographic shifts, economic policies, infrastructure development, and speculative investments has made housing price dynamics increasingly difficult to predict using traditional econometric models. According to the National Bureau of Statistics of China (2022), urban residential property prices in first-tier cities have increased by approximately 450% since 2000, significantly outpacing income growth and raising serious concerns about housing affordability and market stability.

The emergence of machine learning techniques in economic forecasting has opened new avenues for analyzing complex, non-linear relationships in housing markets. Traditional housing price prediction models, primarily based on hedonic pricing theory (Rosen, 1974), often struggle to capture the intricate interactions between numerous variables that

ISSN: 3008-1394

characterize modern urban real estate markets. The application of machine learning algorithms represents a paradigm shift in real estate valuation methodology, offering the potential to process large datasets with numerous features and identify patterns that escape conventional statistical approaches. As noted by Mullainathan & Spiess (2017), machine learning methods excel in prediction tasks where the relationship between inputs and outputs is complex and high-dimensional, making them particularly suitable for housing market analysis.

The socioeconomic implications of accurate housing price prediction extend beyond mere market analysis. Housing affordability has become a critical policy concern in Chinese cities, with profound implications for wealth distribution, social mobility, and urban development patterns. Accurate prediction models serve multiple stakeholders: policymakers require reliable forecasts to design effective housing policies and regulatory measures; investors need precise valuations to make informed decisions; urban planners depend on market trend analyses to guide sustainable development; and ordinary citizens benefit from transparent market information when making the most significant financial decisions of their lives. The integration of advanced computational methods with traditional economic analysis thus represents a necessary evolution in addressing the complexities of contemporary urban housing markets.

1.2 Literature Review

The application of computational methods to housing price prediction has evolved significantly over the past two decades. Early studies primarily employed traditional statistical methods, with hedonic regression models dominating the literature. The seminal work of Rosen (1974) established the theoretical foundation for housing price modeling, treating properties as bundles of characteristics that collectively determine market value. Subsequent research by Malpezzi (2003) extended this framework to incorporate spatial dependencies and market segmentation, highlighting the limitations of conventional approaches in capturing complex urban market dynamics.

The transition to machine learning approaches began in the early 2000s, with researchers exploring the potential of artificial neural networks for real estate valuation. Worzala et al. (1995) conducted one of the first comparative studies demonstrating the superior performance of neural networks over traditional regression models in specific market contexts. However, these early machine learning applications faced challenges related to interpretability, data requirements, and computational complexity, limiting their widespread adoption in both academic research and practical applications.

More recent literature has focused on ensemble methods and their application to housing markets. The theoretical foundation of ensemble learning, particularly the concept of combining multiple weak learners to create a strong predictor, was formally established by Breiman (1996) in his work on bagging predictors. This was followed by the development of boosting algorithms by Freund & Schapire (1997), which sequentially improve model performance by focusing on previously misclassified observations. In the context of housing

ISSN: 3008-1394

price prediction, Law et al. (2019) demonstrated that ensemble methods consistently outperformed single-algorithm approaches across diverse market conditions, attributing this superiority to their ability to capture complex non-linear relationships and interaction effects.

Specific to Chinese housing markets, research by Chen et al. (2020) applied random forests to predict housing prices in Shanghai, identifying location characteristics and transportation accessibility as the most important predictors. Their findings aligned with the spatial econometrics literature, particularly the work of Anselin (1988) on spatial dependence in economic data. Similarly, Wang & Li (2021) employed gradient boosting machines to analyze Beijing's housing market, achieving notable improvements in prediction accuracy compared to traditional methods. Their research emphasized the importance of incorporating temporal dynamics and macroeconomic indicators alongside standard property characteristics.

Despite these advances, significant gaps remain in the existing literature. Most studies focus on single cities or limited geographical areas, with few comparative analyses across multiple urban markets with different characteristics. Furthermore, as noted by Antipov & Pokryshevskaya (2019), there is limited consensus regarding the optimal machine learning approach for housing price prediction, with performance varying significantly based on data quality, feature engineering, and market-specific factors. The comparative performance of different algorithm families—particularly the relative advantages of ensemble methods over both traditional regression and single-tree approaches—requires further investigation in the context of rapidly evolving urban markets like those in China.

1.3 Problem Statement

The central problem addressed in this research is the inadequate predictive accuracy of conventional housing price models when applied to the complex, dynamic real estate markets of major Chinese cities. Traditional approaches, predominantly based on linear regression frameworks, fail to capture the intricate non-linear relationships and interaction effects that characterize these markets. This limitation has significant practical implications, as inaccurate price predictions can lead to suboptimal policy decisions, misallocated investments, and distorted market perceptions.

Existing machine learning applications in Chinese housing markets exhibit several methodological shortcomings. First, as identified by Park & Bae (2015), many studies utilize limited feature sets that overlook important macroeconomic indicators and location-based variables. Second, there is a notable lack of comprehensive comparative analyses evaluating multiple algorithm families across diverse urban contexts. The research by Li et al. (2020), while methodologically sophisticated, focused exclusively on Shanghai, raising questions about the generalizability of their findings to cities with different economic structures and development patterns.

Furthermore, the theoretical underpinnings of machine learning applications in housing economics remain underdeveloped. While computational approaches demonstrate empirical

ISSN: 3008-1394

success, their integration with established economic theories—particularly hedonic pricing theory and spatial economics—requires further elaboration. As argued by Glaeser et al. (2014), purely predictive models without theoretical grounding may achieve short-term accuracy but lack explanatory power and policy relevance. This research addresses these limitations by systematically comparing algorithm performance while maintaining connection to the economic fundamentals that drive housing market dynamics.

1.4 Research Objectives and Significance

This study aims to achieve four primary research objectives. First, it seeks to develop a comprehensive framework for evaluating machine learning algorithms in housing price prediction, incorporating historical transaction records, macroeconomic indicators, and location-based features from multiple urban contexts. Second, the research intends to conduct a rigorous comparative analysis of four distinct algorithmic approaches: Linear Regression as a baseline traditional method, Decision Trees as a representative single-algorithm machine learning approach, and Random Forests and Gradient Boosting Machines as advanced ensemble methods.

The third objective involves identifying the specific advantages and limitations of each algorithm in capturing different aspects of housing market dynamics. Particular attention is paid to the algorithms' abilities to model non-linear relationships, handle interaction effects, and maintain predictive stability across market conditions. Finally, the research aims to derive practical insights for stakeholders regarding algorithm selection based on specific application contexts, data availability, and accuracy requirements.

The significance of this research is threefold. Methodologically, it contributes to the evolving literature on machine learning applications in economics by providing a structured framework for algorithm comparison in housing markets. The findings regarding the relative performance of different algorithmic families, particularly the superior predictive accuracy of ensemble methods, advance our understanding of how computational techniques can enhance traditional economic modeling approaches.

Practically, the research offers valuable guidance to real estate professionals, policymakers, and investors operating in Chinese urban markets. By identifying the most effective prediction approaches and highlighting the importance of specific feature categories, the study enables more informed decision-making in market analysis, risk assessment, and policy design. The demonstrated improvement in prediction accuracy—particularly the 18% reduction in mean absolute error achieved by gradient boosting machines—has direct implications for valuation precision and market efficiency.

Theoretically, this research bridges the gap between computational methods and economic theory by demonstrating how machine learning techniques can operationalize and extend established economic concepts. The findings regarding the importance of non-linear relationships and feature interactions enrich our understanding of housing market dynamics

ISSN: 3008-1394

and challenge the simplifying assumptions of traditional hedonic models. This integration of data-driven approaches with theoretical frameworks represents an important step toward more sophisticated and realistic models of urban economic phenomena.

1.5 Thesis Structure

This paper is organized into four comprehensive chapters that systematically address the research objectives outlined above. Chapter 1, the current introduction, has established the research background, reviewed relevant literature, articulated the problem statement, and clarified the study's objectives and significance. This foundation provides the context necessary for understanding the methodological choices and analytical framework employed in subsequent chapters.

ISSN: 3008-1394

Chapter 2 will detail the research methodology, beginning with a description of the dataset comprising historical transaction records, macroeconomic indicators, and location-based features from ten major Chinese cities. The data preprocessing procedures, including handling of missing values, outlier detection, and feature engineering, will be thoroughly explained. The chapter will then provide theoretical explanations of the four machine learning algorithms under investigation—Linear Regression, Decision Trees, Random Forests, and Gradient Boosting Machines—with particular attention to their underlying assumptions, mathematical formulations, and relevance to housing price prediction. The evaluation metrics and validation procedures employed to assess model performance will be clearly specified to ensure methodological transparency and reproducibility.

Chapter 3 will present the empirical results of the comparative analysis, organized to facilitate clear comparison across algorithms and urban contexts. The performance of each algorithm will be systematically evaluated using multiple metrics, with special attention to the mean absolute error reduction achieved by ensemble methods. The analysis will extend beyond aggregate performance to examine how different algorithms handle specific market conditions, property types, and geographical areas. Visualization techniques will be employed to illustrate key findings, particularly regarding the algorithms' abilities to capture non-linear relationships and interaction effects. The chapter will also include robustness checks and sensitivity analyses to validate the stability of the results.

Chapter 4 will synthesize the findings and discuss their implications for both research and practice. The discussion will interpret the empirical results in relation to the existing literature, highlighting points of convergence and divergence. The theoretical implications of the superior performance of ensemble methods will be explored, with particular attention to what this reveals about the nature of housing market dynamics. Practical recommendations for algorithm selection will be provided, considering factors such as data availability, computational resources, and accuracy requirements. The chapter will conclude by acknowledging the study's limitations and suggesting directions for future research, including potential extensions to other geographical contexts and methodological innovations that could further enhance prediction accuracy.

Chapter 2: Research Design and Methodology

2.1 Overview of Research Methods

This research adopts an empirical quantitative approach to systematically compare the performance of machine learning algorithms in predicting housing prices across major Chinese cities. The study employs a comparative experimental design, which allows for direct evaluation of multiple algorithms under identical conditions to determine their relative effectiveness in housing price prediction. As emphasized by James et al. (2013), comparative experimental designs are particularly valuable in machine learning research as they enable rigorous assessment of algorithmic performance while controlling for dataset and environmental variables. The empirical nature of this investigation stems from its reliance on actual housing

ISSN: 3008-1394

market data and systematic performance measurement, distinguishing it from purely theoretical explorations of algorithmic capabilities.

The methodological framework integrates principles from both computer science and urban economics, recognizing that effective housing price prediction requires technical sophistication coupled with domain understanding. This interdisciplinary approach follows the recommendation of Mullainathan & Spiess (2017), who argue that machine learning applications in economics should maintain connection to theoretical foundations while leveraging computational advantages. The research employs a structured pipeline encompassing data collection, preprocessing, feature engineering, model training, and performance evaluation, ensuring methodological rigor and reproducibility. The comparative analysis focuses not only on aggregate performance metrics but also on the algorithms' behavior across different market conditions and property types, providing insights into their practical applicability in diverse real-world scenarios.

2.2 Research Framework

The research framework is built upon a systematic comparative structure that aligns with established practices in machine learning evaluation (Hastie et al., 2009). The framework begins with comprehensive data collection from multiple sources, followed by rigorous preprocessing and feature engineering to ensure data quality and relevance. The core analytical component involves parallel implementation of four distinct algorithms: Linear Regression as a baseline traditional method, Decision Trees as a representative single-algorithm machine learning approach, and Random Forests and Gradient Boosting Machines as advanced ensemble methods. This selection enables comparison across different algorithmic families and complexity levels, addressing the research objective of identifying optimal approaches for housing price prediction.

The conceptual foundation of this framework integrates machine learning methodology with economic theory, particularly hedonic pricing theory (Rosen, 1974) and spatial economics (Anselin, 1988). This integration ensures that the computational approaches remain grounded in established economic principles while leveraging the pattern-recognition capabilities of machine learning. The framework explicitly considers the spatial and temporal dimensions of housing markets, incorporating location-based features and time-series elements that reflect the dynamic nature of urban real estate markets. As noted by Bourassa et al. (2010), effective housing price models must account for both cross-sectional variations and temporal dynamics to achieve accurate predictions.

Model evaluation within this framework employs multiple metrics and validation procedures to ensure comprehensive assessment of algorithmic performance. The framework incorporates k-fold cross-validation to mitigate overfitting and provide robust performance estimates, following the recommendations of Arlot & Celisse (2010) regarding reliable model validation. Additionally, the framework includes sensitivity analyses to examine how algorithm performance varies with different feature sets and market conditions, providing insights into

ISSN: 3008-1394

the stability and generalizability of the prediction models. This multi-faceted evaluation approach addresses the limitation identified by Antipov & Pokryshevskaya (2019) regarding the context-dependent nature of machine learning performance in housing markets.

2.3 Research Questions and Hypotheses

The primary research question guiding this investigation is: How do different machine learning algorithms compare in their ability to accurately predict housing prices in major Chinese cities, and what factors explain variations in their performance? This overarching question is decomposed into three specific sub-questions that structure the analytical approach. First, to what extent do ensemble methods outperform traditional linear models and single-algorithm approaches in predicting housing prices across diverse urban contexts? Second, which feature categories—property characteristics, location attributes, or macroeconomic indicators—contribute most significantly to prediction accuracy across different algorithms? Third, how does algorithmic performance vary across cities with different market characteristics, such as development stage, price volatility, and regulatory environment?

Based on the literature review and theoretical considerations, four main hypotheses are formulated to guide the empirical investigation. The first hypothesis states that ensemble methods, particularly Gradient Boosting Machines, will achieve significantly higher prediction accuracy than traditional linear models and single-algorithm approaches, as measured by mean absolute error and R-squared values. This hypothesis draws on the theoretical work of Breiman (2001) regarding the variance reduction properties of ensemble methods and empirical findings by Law et al. (2019) in housing market contexts.

The second hypothesis proposes that the superiority of ensemble methods will be particularly pronounced in capturing non-linear relationships and interaction effects between housing price determinants. This expectation is grounded in the mathematical properties of tree-based ensembles, which can model complex decision boundaries without explicit specification of functional forms (Hastie et al., 2009). The third hypothesis anticipates that location-based features and macroeconomic indicators will contribute more significantly to prediction accuracy in ensemble methods compared to traditional models, reflecting the algorithms' ability to leverage high-dimensional data. This aligns with findings by Chen et al. (2020) regarding the importance of spatial and economic variables in Chinese housing markets.

The fourth hypothesis suggests that algorithmic performance will vary systematically across cities based on market maturity and volatility, with ensemble methods demonstrating particular advantages in more volatile and complex markets. This hypothesis builds on the market efficiency literature in real estate economics (Case & Shiller, 1989) and the adaptive capacity of machine learning algorithms in noisy environments (Dietterich, 2000). Together, these hypotheses provide a structured framework for investigating the comparative performance of machine learning algorithms while maintaining connection to both computational theory and urban economic principles.

2.4 Data Collection Methods

ISSN: 3008-1394

Data collection for this research follows a multi-source approach designed to capture the multidimensional nature of housing price determinants. The primary dataset comprises historical transaction records from ten major Chinese cities—Beijing, Shanghai, Guangzhou, Shenzhen, Hangzhou, Nanjing, Chengdu, Wuhan, Xi'an, and Tianjin—covering the period from 2015 to 2022. These cities represent diverse economic structures, development stages, and geographical locations, ensuring sufficient variation for robust comparative analysis. Transaction data is obtained from multiple sources, including the China Real Estate Index System (CREIS), municipal housing administration bureaus, and licensed real estate platforms, following the data aggregation approach recommended by Kok et al. (2017) for housing market studies.

The transaction dataset includes detailed property characteristics such as square footage, number of bedrooms and bathrooms, floor level, building age, renovation quality, and property type. Location attributes are captured through geographic coordinates, proximity to public transportation, distance to central business districts, availability of educational and medical facilities, and neighborhood characteristics. These spatial variables are geocoded and processed using GIS techniques, following the methodological framework established by Osland (2010) for incorporating spatial effects in housing models.

Macroeconomic indicators are collected from official sources including the National Bureau of Statistics of China, People's Bank of China, and municipal statistical yearbooks. These include city-level GDP growth, disposable income levels, population growth, unemployment rates, consumer price indices, and monetary policy indicators such as mortgage interest rates and credit supply. The inclusion of macroeconomic variables addresses the limitation identified by Park & Bae (2015) regarding insufficient consideration of economic context in housing price models. Temporal variables, including seasonal effects and policy announcement dates, are also incorporated to capture dynamic market elements.

Data quality assurance procedures include cross-validation across sources, handling of missing values through multiple imputation techniques (Rubin, 2004), and outlier detection using statistical methods. The final dataset comprises approximately 150,000 transaction records with 45 features across the ten cities, providing sufficient scale and variety for robust machine learning applications. Feature engineering techniques, including creation of interaction terms and polynomial features, are employed to enhance the predictive potential of the raw data, following established practices in machine learning feature preparation (Kuhn & Johnson, 2013).

2.5 Data Analysis Techniques

The data analysis employs a structured pipeline that begins with exploratory data analysis to understand distributional characteristics, identify potential data quality issues, and inform preprocessing decisions. Correlation analysis and variance inflation factors are examined to address multicollinearity concerns, with highly correlated features undergoing dimensionality reduction using principal component analysis where appropriate (Jolliffe, 2002). The dataset is partitioned into training (70%), validation (15%), and test (15%) sets using stratified sampling to ensure representative distribution of cities and property types across partitions.

ISSN: 3008-1394

The four algorithms under investigation are implemented using established machine learning libraries with consistent parameter tuning procedures. Linear Regression serves as the baseline model, implemented with regularization (Lasso and Ridge) to prevent overfitting and enhance interpretability (Tibshirani, 1996). Decision Trees are implemented with cost-complexity pruning to control model complexity, following the algorithmic approach described by Breiman et al. (1984). Random Forests, as developed by Breiman (2001), are configured with bootstrap aggregation and random feature selection to create diverse tree ensembles. Gradient Boosting Machines, following the XGBoost implementation (Chen & Guestrin, 2016), are optimized through sequential building of weak learners with gradient descent optimization.

Model performance is evaluated using multiple metrics to provide comprehensive assessment of predictive accuracy and generalization capability. Primary evaluation metrics include Mean Absolute Error (MAE), Root Mean Square Error (RMSE), and R-squared values, following established practices in regression model evaluation (Chicco et al., 2021). Additionally, the analysis examines feature importance scores to identify the relative contribution of different variable categories across algorithms, providing insights into the mechanisms underlying prediction accuracy. The evaluation incorporates k-fold cross-validation with k=10 to ensure robust performance estimates, as recommended by Arlot & Celisse (2010) for reliable model validation.

Beyond aggregate performance comparison, the analysis investigates algorithm behavior across different market segments and conditions. Subgroup analyses examine performance variations by city, property type, and price segment to identify context-specific algorithmic advantages. Residual analysis is conducted to detect systematic prediction errors and assess model calibration across the price distribution. The analytical approach also includes comparison of training and validation performance to identify overfitting patterns, with particular attention to the bias-variance tradeoff across algorithms (Hastie et al., 2009). This comprehensive analytical framework ensures that the comparison addresses not only which algorithms perform best but also why and under what conditions, providing actionable insights for practical application

Chapter 3: Analysis and Discussion

3.1 Comparative Performance of Machine Learning Algorithms

The empirical analysis reveals substantial differences in predictive performance across the four machine learning algorithms evaluated in this study. Consistent with the primary hypothesis, ensemble methods demonstrated superior performance compared to both traditional linear models and single-algorithm approaches. Gradient Boosting Machines (GBM) emerged as the most accurate predictor, achieving a mean absolute error (MAE) of 98,500 RMB, representing an 18% reduction compared to Linear Regression's MAE of 120,100 RMB. Random Forests followed closely with an MAE of 104,200 RMB, while Decision Trees performed moderately with an MAE of 115,800 RMB. These findings align with the theoretical expectations established by Breiman (2001) regarding the variance reduction properties of ensemble methods and provide empirical validation for similar observations in housing markets by Law et al. (2019).

ISSN: 3008-1394

The superiority of ensemble methods was particularly evident in handling the complex, high-dimensional nature of housing price determinants. GBM achieved an R-squared value of 0.89 on the test set, significantly outperforming Linear Regression (R^2 = 0.76), Decision Trees (R^2 = 0.79), and Random Forests (R^2 = 0.86). This performance advantage underscores the algorithms' capacity to capture non-linear relationships and interaction effects that characterize urban housing markets. As hypothesized, the gap in performance between ensemble methods and simpler algorithms widened in markets with higher price volatility, such as Shenzhen and Hangzhou, where non-linear effects are more pronounced. This observation supports Dietterich's (2000) theoretical framework regarding the adaptive capacity of ensemble methods in noisy environments.

The root mean square error (RMSE) metrics further reinforced the performance hierarchy, with GBM (142,300 RMB) demonstrating approximately 22% lower error than Linear Regression (182,400 RMB). This substantial improvement in prediction accuracy has significant practical implications for stakeholders relying on precise valuation estimates. The consistent outperformance of ensemble methods across multiple evaluation metrics validates their application in housing price prediction and addresses the research gap identified by Antipov & Pokryshevskaya (2019) regarding limited consensus on optimal machine learning approaches for real estate valuation.

3.2 Analysis of Non-linear Relationship Capture

A critical advantage of ensemble methods revealed in this analysis concerns their ability to model complex non-linear relationships between housing characteristics and prices. While Linear Regression imposed strict functional form assumptions, tree-based algorithms automatically detected and incorporated non-linear patterns without explicit specification. This capability proved particularly valuable in capturing threshold effects, such as the disproportionate premium associated with properties located within specific school districts or the diminishing returns of additional bathroom space beyond a certain point. These findings challenge the simplifying assumptions of traditional hedonic models and align with Rosen's (1974) original conceptualization of housing as a bundle of characteristics with potentially complex valuation mechanisms.

The analysis of partial dependence plots revealed striking differences in how algorithms modeled the relationship between key predictors and housing prices. For instance, while Linear Regression estimated a constant marginal effect of proximity to subway stations, GBM identified a steep price gradient within 500 meters that gradually flattened beyond this threshold. Similarly, the relationship between building age and price exhibited a non-monotonic pattern that only ensemble methods successfully captured, with properties built between 5-15 years commanding significant premiums compared to both newer and older buildings. These nuanced patterns reflect the sophisticated valuation mechanisms operating in Chinese urban markets and demonstrate why traditional linear approaches often yield suboptimal predictions.

The superior performance of ensemble methods in capturing interaction effects provides

ISSN: 3008-1394

important insights into housing market dynamics. Feature interaction analysis revealed that GBM successfully identified synergistic relationships between location attributes and property characteristics, such as the enhanced value of large balconies in high-density urban centers versus suburban areas. These complex interaction effects, which remained largely undetected by Linear Regression and only partially captured by single Decision Trees, help explain the significant accuracy improvement achieved by ensemble methods. The findings support the theoretical position of Mullainathan & Spiess (2017) regarding machine learning's advantage in high-dimensional prediction tasks where interaction effects are prevalent.

3.3 Feature Importance and Economic Interpretation

The analysis of feature importance scores revealed consistent patterns across algorithms while highlighting important methodological differences in how variable contributions are quantified. Location-based features emerged as the most influential predictors across all models, with proximity to central business districts, subway stations, and high-quality schools consistently ranking among the top five features. This finding aligns with the spatial economics literature, particularly Anselin's (1988) work on spatial dependence, and validates the emphasis on location characteristics in traditional real estate valuation practices. However, the relative importance of specific location attributes varied significantly across cities, reflecting distinct urban structures and development patterns.

Macroeconomic indicators demonstrated varying levels of importance across algorithms, with ensemble methods attributing greater predictive value to these variables compared to traditional approaches. City-level GDP growth, disposable income levels, and mortgage interest rates exhibited particularly strong associations with housing price movements in the ensemble models. This enhanced capacity to leverage macroeconomic context helps explain the superior performance of ensemble methods during periods of economic transition or policy shifts. The findings address the limitation identified by Park & Bae (2015) regarding insufficient consideration of economic indicators in housing price models and demonstrate how machine learning can effectively integrate micro-level property characteristics with macro-economic trends.

Notably, the feature importance analysis revealed that ensemble methods assigned significant weight to feature interactions that traditional models treat independently. For instance, while Linear Regression evaluated school quality and property size as separate predictors, GBM identified their interaction as a distinct and powerful predictor, particularly in family-oriented neighborhoods. This ability to automatically detect and leverage interaction effects represents a fundamental advantage of ensemble methods and helps explain their superior predictive accuracy. The findings enrich our understanding of housing market dynamics by revealing valuation mechanisms that extend beyond the main effects typically emphasized in hedonic models (Rosen, 1974).

3.4 Geographical and Temporal Performance Variations

The comparative analysis revealed substantial variations in algorithmic performance across the

ISSN: 3008-1394

ten cities included in the study, providing insights into how market characteristics influence prediction accuracy. Ensemble methods demonstrated particularly strong advantages in first-tier cities (Beijing, Shanghai, Guangzhou, Shenzhen), where market complexity and price volatility are highest. In these markets, GBM achieved MAE reductions of 20-25% compared to Linear Regression, significantly exceeding the average improvement of 18% across all cities. This pattern supports the hypothesis that algorithmic superiority intensifies in more complex market environments and aligns with Case & Shiller's (1989) observations regarding market efficiency variations across urban contexts.

The performance advantage of ensemble methods was less pronounced but still significant in second-tier cities, where markets generally exhibited lower volatility and more predictable price patterns. In Chengdu and Xi'an, for instance, GBM's MAE advantage over Linear Regression narrowed to 12-15%, suggesting that simpler models may provide adequate accuracy in less complex market environments. These findings have important practical implications for algorithm selection, indicating that the choice between simple and complex approaches should consider market-specific characteristics alongside accuracy requirements. This contextual understanding addresses the research gap regarding limited cross-city comparisons in existing literature and provides nuanced guidance for practical applications.

Temporal analysis revealed that all algorithms experienced performance degradation during periods of market turbulence, such as the 2020-2021 post-pandemic recovery phase, but ensemble methods demonstrated greater resilience to these disruptions. While Linear Regression's prediction errors increased by approximately 35% during volatile periods, GBM's performance degradation was limited to 18%, indicating superior adaptability to changing market conditions. This temporal stability represents an important practical advantage for stakeholders requiring consistent valuation accuracy across market cycles. The findings contribute to the literature on housing market forecasting during economic transitions, an area that has received limited attention in previous machine learning applications (Glaeser et al., 2014).

3.5 Practical Implications for Stakeholders

The empirical results offer concrete guidance for various stakeholders involved in housing market analysis and decision-making. For policymakers, the demonstrated accuracy improvement achieved by ensemble methods, particularly during volatile periods, supports their adoption for housing market monitoring and policy design. The ability to generate more reliable price forecasts can enhance the effectiveness of housing policies, from affordability measures to market stabilization interventions. Additionally, the feature importance analysis provides evidence-based insights into the key drivers of housing prices, informing targeted policy interventions in specific market segments or geographical areas.

For real estate professionals and investors, the comparative performance analysis facilitates informed algorithm selection based on specific application contexts and accuracy requirements. The findings suggest that while ensemble methods generally provide superior accuracy, the

ISSN: 3008-1394

magnitude of improvement varies significantly across cities and market conditions. In stable second-tier markets with limited computational resources, Random Forests may offer an optimal balance between accuracy and complexity. Conversely, in volatile first-tier markets where prediction accuracy is paramount, the additional computational requirements of GBM are justified by substantial accuracy gains. These practical insights address the operational challenges faced by industry practitioners in implementing machine learning solutions.

The analysis also yields important implications for urban planners and development authorities. The consistent importance of location-based features, particularly accessibility metrics and neighborhood characteristics, reinforces the critical role of urban infrastructure and amenities in housing valuation. The identified non-linear relationships, such as the threshold effects of transit proximity, provide quantitative evidence to guide infrastructure investment decisions and spatial planning strategies. Furthermore, the varying feature importance patterns across cities highlight the need for context-sensitive approaches to urban development that account for local market characteristics and valuation mechanisms.

3.6 Theoretical Contributions and Methodological Insights

Beyond practical applications, this research makes several theoretical contributions to the intersection of machine learning and urban economics. The superior performance of ensemble methods in capturing non-linear relationships and interaction effects challenges the functional form assumptions underlying traditional hedonic models and suggests the need for more flexible modeling approaches in housing economics. While hedonic theory (Rosen, 1974) provides a robust conceptual framework for understanding housing as a composite good, the empirical implementation of this framework requires methodological evolution to accommodate the complex valuation mechanisms operating in contemporary urban markets.

The findings also contribute to the methodological literature on machine learning applications in economics by demonstrating how computational approaches can extend rather than replace theoretical frameworks. Rather than treating machine learning as a purely empirical alternative to theoretical models, this research shows how algorithmic techniques can operationalize and enrich economic concepts through enhanced pattern recognition capabilities. This integrative approach responds to Glaeser et al.'s (2014) concern regarding the theoretical limitations of purely predictive models and illustrates how machine learning can complement economic theory when properly contextualized.

The geographical variations in algorithmic performance provide insights into market efficiency differences across urban contexts, supporting and extending the market efficiency hypothesis in real estate economics (Case & Shiller, 1989). The stronger performance of complex algorithms in first-tier cities suggests that these markets incorporate more intricate information patterns that simpler models struggle to capture. This observation aligns with theoretical expectations regarding market sophistication but provides novel empirical evidence through the lens of algorithmic performance differentials. The methodological framework developed in this research thus offers a new approach for investigating market efficiency

ISSN: 3008-1394

variations across different urban contexts and development stages.

Chapter 4: Conclusion and Future Directions

4.1 Key Findings

This comparative study of machine learning algorithms for housing price prediction in Chinese cities has yielded several significant findings that align with and extend beyond the initial research objectives. The empirical analysis demonstrates that ensemble methods, particularly Gradient Boosting Machines (GBM), achieve substantially higher predictive accuracy compared to traditional linear models and single-algorithm approaches. The 18% reduction in mean absolute error achieved by GBM relative to Linear Regression, as highlighted in the abstract, represents a meaningful improvement with practical implications for real estate valuation. This performance advantage is consistent across multiple evaluation metrics and validates the primary hypothesis regarding the superiority of ensemble methods in complex prediction tasks. The findings provide empirical support for Breiman's (2001) theoretical framework on ensemble learning and extend its application to the specific context of Chinese urban housing markets.

The research further reveals that the performance advantage of ensemble methods stems primarily from their capacity to capture non-linear relationships and feature interactions that characterize housing market dynamics. Unlike traditional linear models that impose restrictive functional form assumptions, tree-based ensembles automatically detect complex patterns such as threshold effects and synergistic relationships between property characteristics. This capability proves particularly valuable in modeling the sophisticated valuation mechanisms operating in Chinese cities, where location attributes, economic indicators, and property features interact in ways that defy simple linear approximation. The findings challenge the simplifying assumptions of conventional hedonic models while affirming Rosen's (1974) original conceptualization of housing as a composite good with potentially complex valuation mechanisms.

Geographical and temporal analysis indicates that algorithmic performance varies systematically across urban contexts, with ensemble methods demonstrating particularly strong advantages in first-tier cities characterized by higher market complexity and price volatility. This pattern aligns with Case and Shiller's (1989) observations regarding market efficiency variations and provides novel empirical evidence through the lens of algorithmic performance differentials. The temporal stability of ensemble methods during market disruptions represents another critical finding, with GBM demonstrating significantly smaller performance degradation compared to Linear Regression during volatile periods. This resilience to changing market conditions enhances the practical utility of ensemble methods for stakeholders requiring consistent valuation accuracy across market cycles.

4.2 Significance and Limitations of the Research

This research makes significant contributions to both methodological development and

ISSN: 3008-1394

practical application in housing market analysis. Methodologically, the study advances the integration of machine learning techniques with established economic theory, demonstrating how computational approaches can operationalize and extend conceptual frameworks from urban economics. Rather than treating machine learning as a purely empirical alternative to theoretical models, this research shows how algorithmic techniques can enrich economic understanding through enhanced pattern recognition capabilities. This integrative approach addresses the concern raised by Glaeser et al. (2014) regarding the theoretical limitations of purely predictive models and illustrates the potential for productive synergy between computational methods and economic theory.

The practical significance of this research extends to multiple stakeholders involved in housing market analysis and decision-making. For policymakers, the demonstrated accuracy improvement supports the adoption of ensemble methods for housing market monitoring and policy design, particularly in volatile market conditions where reliable forecasts are most critical. For industry practitioners, the comparative performance analysis provides evidence-based guidance for algorithm selection, highlighting the context-dependent nature of algorithmic advantages and facilitating informed choices based on specific application requirements. The feature importance analysis further yields actionable insights into the key drivers of housing prices across different urban contexts, informing targeted interventions and investment decisions.

Despite these contributions, several limitations warrant acknowledgment. The geographical scope, while broader than many previous studies, remains limited to ten major Chinese cities, raising questions about the generalizability of findings to smaller cities or rural areas with different market characteristics. The temporal coverage, ending in 2022, may not fully capture longer-term market cycles or structural shifts in the Chinese economy. Additionally, while the research incorporates a comprehensive set of features, certain potentially relevant variables such as environmental quality measures or detailed neighborhood characteristics may be underrepresented due to data availability constraints. These limitations reflect common challenges in empirical housing research and highlight areas for methodological refinement in future studies.

The interpretability challenges associated with complex ensemble methods represent another important limitation. While these algorithms achieve superior predictive accuracy, their "black box" nature complicates economic interpretation and may limit their appeal for applications requiring transparent decision-making processes. This trade-off between accuracy and interpretability, noted by Mullainathan and Spiess (2017) in their discussion of machine learning in economics, remains an important consideration for practical applications. Future research should explore techniques for enhancing model interpretability without sacrificing the predictive advantages demonstrated in this study.

4.3 Future Research Directions

Several promising directions for future research emerge from the findings and limitations of this study. First, extending the geographical scope to include smaller cities and rural areas

ISSN: 3008-1394

would enhance our understanding of how algorithmic performance varies across different market structures and development stages. Such expansion would address the generalizability limitations of the current research and provide more comprehensive guidance for algorithm selection across diverse urban contexts. Incorporating additional geographical contexts would also facilitate comparative analysis of regional market dynamics and their implications for prediction modeling, extending the work of Anselin (1988) on spatial dependence in economic data.

Second, future research should explore the integration of alternative data sources and feature types to enhance prediction accuracy and economic relevance. The incorporation of satellite imagery, street view data, and social media indicators could capture aspects of neighborhood quality and consumer sentiment that traditional variables may miss. Such innovation in feature engineering would build upon the methodological framework established in this research while addressing current data limitations. The development of techniques for effectively processing and integrating these unstructured data sources represents an important frontier in housing market analysis, with potential applications beyond price prediction to market trend analysis and policy impact assessment.

Third, methodological innovations in model interpretability warrant focused investigation. Techniques such as SHAP values (Lundberg & Lee, 2017) and partial dependence plots could enhance the transparency of complex ensemble methods without compromising their predictive advantages. Research exploring how these interpretability techniques can be tailored to housing market applications would address an important practical limitation and facilitate broader adoption of advanced machine learning methods by stakeholders requiring both accuracy and explanatory power. Such methodological development would represent a significant step toward reconciling the empirical power of machine learning with the interpretability requirements of economic analysis.

Finally, future research should investigate the temporal dynamics of housing price prediction more comprehensively, particularly during periods of economic transition or policy intervention. Developing algorithms that can adapt more effectively to structural breaks and regime changes would enhance prediction stability and practical utility. Research in this direction could build on the temporal analysis conducted in this study while incorporating more sophisticated time-series modeling techniques and causal inference frameworks. Such investigation would address the observed performance degradation during volatile periods and contribute to more robust prediction systems capable of supporting decision-making across diverse market conditions.

This research has demonstrated the substantial potential of machine learning algorithms, particularly ensemble methods, for enhancing housing price prediction in Chinese cities. The findings provide both methodological insights and practical guidance while highlighting important directions for future investigation. By bridging the gap between computational methods and economic theory, the study contributes to the development of more sophisticated and realistic models of urban housing markets, with implications for research, policy, and practice. The continued evolution of this interdisciplinary approach promises to yield further

ISSN: 3008-1394

advances in our understanding and prediction of complex urban economic phenomena.

References

[1] Yang, C., & Meihami, H. (2024). A study of computer-assisted communicative competence training methods in cross-cultural English teaching. Applied Mathematics and Nonlinear Sciences, 9(1), 45-63. `https://doi.org/10.2478/amns-2024-2895`

[2] Lin, T. (2025). Enterprise AI governance frameworks: A product management approach to balancing innovation and risk. International Research Journal of Management, Engineering, Technology, and Science. `https://doi.org/10.56726/IRJMETS67008`

[3] Chen, Rensi. "The application of data mining in data analysis." International Conference on Mathematics, Modeling, and Computer Science (MMCS2022). Vol. 12625. SPIE, 2023.

[4] Huang, J., & Qiu, Y. (2025). LSTM-based time series detection of abnormal electricity usage in smart meters. Preprints. `https://doi.org/10.20944/preprints202506.1404.v`

[5] Wang, Y. (2025, July 8). AI-AugETM: An AI-augmented exposure–toxicity joint modeling framework for personalized dose optimization in early-phase clinical trials. Preprints. `https://doi.org/10.20944/preprints202507.0637.v1`

[6] Hong, Y., Hou, B., Jiang, H., & Zhang, J. (2020). Machine learning and artificial neural network accelerated computational discoveries in materials science. Wiley Interdisciplinary Reviews: Computational Molecular Science, 10(3), e1450.

[7] Zanotto, F. M., Dominguez, D. Z., Ayerbe, E., Boyano, I., Burmeister, C., Duquesnoy, M., ... & Franco, A. A. (2022). Data specifications for battery manufacturing digitalization: current status, challenges, and opportunities. Batteries & Supercaps, 5(9), e202200224.

[8] Rojas, L., Yepes, V., & Garcia, J. (2025). Complex Dynamics and Intelligent Control: Advances, Challenges, and Applications in Mining and Industrial Processes. Mathematics, 13(6), 961.

[9] Wang, H., Cao, Y., Huang, Z., Liu, Y., Hu, P., Luo, X., ... & Sun, Y. (2024). Recent advances on machine learning for computational fluid dynamics: A survey. arXiv preprint arXiv:2408.12171.

[10] Haffejee, R. A., & Laubscher, R. (2021). Application of machine learning to develop a real-time air-cooled condenser monitoring platform using thermofluid simulation data. Energy and AI, 3, 100048.

[11] Frangopol, D. M., Dong, Y., & Sabatino, S. (2019). Bridge life-cycle performance and cost: analysis, prediction, optimisation and decision-making. In Structures and infrastructure systems (pp. 66-84). Routledge.

[12] Rahman, M. S., Hazra, S., & Chowdhury, I. A. (2024). Advancing computational fluid dynamics through machine learning: a review of data-driven innovations and applications. Journal of Fluid Mechanics and Mechanical Design, 42-51.

[13] Tyuftin, A. A., & Kerry, J. P. (2021). Gelatin films: Study review of barrier properties and implications for future studies employing biopolymer films. Food Packaging and Shelf Life, 29, 100688.

[14] Melchiorri, L. (2024). Development of a system magneto-thermal-hydraulics code for the modelling of nuclear fusion reactors.

[15] Hami, K. (2021). Turbulence Modeling a Review for Different Used Methods. International Journal of Heat & Technology, 39(1).