SMU Data Science Review

Volume 8
Number 1 *Spring 2024*

Article 12

Predictive Analysis of Local House Prices: Leveraging Machine Learning for Real Estate Valuation

Joey Hernandez
Southern Methodist University, joeyvhernandez@gmail.com

Danny Chang
Southern Methodist University, changd@smu.edu

Santiago Gutierrez
Southern Methodist University, gutierrezs@smu.edu

Paul Huggins
paul.huggins@lanternstudios.com

Recommended Citation

Hernandez, Joey; Chang, Danny; Gutierrez, Santiago; and Huggins, Paul () "Predictive Analysis of Local House Prices: Leveraging Machine Learning for Real Estate Valuation," *SMU Data Science Review*: Vol. 8: No. 1, Article 12.
Available at: https://scholar.smu.edu/datasciencereview/vol8/iss1/12

Predictive Analysis of Local House Prices: Leveraging Machine Learning for Real Estate Valuation

Santiago Gutiérrez, Daniel Chang, Joey Hernandez, Paul Huggins
Master of Science in Data Science, Southern Methodist University,
Dallas, TX 75205 USA
gutierrezs@smu.edu,
changd@smu.edu,
joeyhernandez@smu.edu

Abstract: This paper presents a comprehensive study examining the real estate market potential in the dynamic urban landscapes of Frisco and Plano, Texas. Combining traditional real estate analysis with cutting-edge machine learning techniques, the study aims to predict home prices and assess investment feasibility. Leveraging these findings, the study proposes a strategic focus on predictive modeling and investment potential identification, emphasizing the continual refinement of machine learning models with updated data to accurately forecast changes in the real estate market. By harnessing the predictive power of these models, investors can identify high-growth areas and optimize their investment decisions, thus capitalizing on emerging trends and investment hotspots in Frisco and Plano. This study highlights the potential of advanced analytical tools in guiding investors toward lucrative real estate opportunities in rapidly developing urban environments.

1 Introduction

Real estate investment, recognized for its potential in wealth generation and economic growth, remains a domain historically perceived as exclusive and often inaccessible (Ullah & Sepasgozar, 2020). Prevailing misconceptions have deterred individuals, particularly those from modest backgrounds, from considering it a viable investment option (Ullah & Sepasgozar, 2020). This exclusivity is compounded by the challenge of obtaining real-time data, an essential component for informed investment decisions, potentially leaving investors navigating the market without clear and comprehensive insights (Ullah & Sepasgozar, 2020).

This research addresses complexities inherent in-home valuation within the residential real estate sector. Its primary aim is to simplify and make accessible the process of property valuation for investment purposes, fostering inclusivity and understanding in this domain. By concentrating on residential real estate valuation, the study presents a focused entry point for aspiring investors, guiding them in the accurate valuation of homes. A significant aspect of this research is the application of machine learning techniques. These methods are employed to equip investors with the tools and insights necessary to accurately assess property values, thereby instilling confidence and enhancing the model's ability to identify and leverage market opportunities effectively.

The significance of this research extends beyond the traditional confines of real estate analysis. Plano and Frisco, Texas, are recognized as growing urban areas of population and development (Demographics Frisco Economic Development Corporation, n.d.; Best Cities for Jobs in 2024. (n.d.), presenting a variety of investment opportunities owing to strategic locales and rich amenities. In alignment with the paper's focus, this study employs machine learning and data science techniques to predict house prices and potential rental yields in these areas. Such predictive analysis aims to provide investors with sophisticated tools to assess the future of investments effectively. By focusing on Plano and Frisco, the research contributes to a deeper understanding of local market dynamics and offers a model for leveraging technological advancements in real estate valuation.

The democratization of real estate investments, facilitated by technology and AI, underscores the relevance of this study. Modern platforms that facilitate fractional investments on a global scale have markedly lowered the barriers to entry in the real estate market (Guest Contributors, 2023). This shift promotes urban development and economic growth while introducing greater liquidity into the market. Illustrative of this trend is the burgeoning real estate crowdfunding sector, which, with a current valuation of $19.5 Billion, is expected to experience significant growth in the coming years. (Polaris et al., 2023).

This research occupies a crucial role in deciphering the rapidly changing landscape of real estate investment, particularly in the context of Plano and Frisco, Texas. By harnessing the power of machine learning and data science, it is positioned to provide investors with nuanced, data-driven insights. These insights are imperative for making informed decisions, grounded in thorough analysis and empirical evidence. The study's primary focus is on the predictive analysis of local house prices in these areas, aiming to offer a robust and reliable framework for real estate valuation. It endeavors to bridge the gap between traditional valuation methods and modern technological capabilities, ensuring that investors can navigate this evolving market with confidence and precision.

2 Literature Review

2.1 Real Estate Investment Barriers

Ullah and Sepasgozar (2020) shed light on some of the significant challenges prospective real estate investors face when entering the market (Ullah & Sepasgozar, 2020). A primary challenge in this domain is the presence of financial constraints. These constraints can hinder newcomers from obtaining essential funding or securing loans with favorable terms. This situation frequently leads to investors entering property purchases without a complete understanding, often resulting in difficulties in affording the acquisition. (Ullah & Sepasgozar, 2020). Time constraints can also pose a problem, especially for investors exploring the real estate market while managing other commitments or for individuals who want quick returns (quick buy-ins and buy-outs).

Additionally, locating suitable assets matching an investor's criteria can be daunting. Potential investors need to find properties that align with an individual's investment goals and ensure that investments are priced reasonably, leading to another challenge: the complexity of property valuation. Valuing real estate accurately requires a deep understanding of the market, local conditions, current trends, and costs associated with buying and selling (Ullah & Sepasgozar, 2020). For the real estate market to become more appealing and accessible to a broader range of investors, addressing and alleviating these barriers is crucial.

2.2 Machine Learning in Real Estate

Studies have illuminated the growing usefulness of machine learning in the world of real estate, such as the use of supervised learning and neural networks to predict and validate any real estate value (Hermans, McCord, Davis & Bidanset, 2023, pp. 43+; Varma, Sarma, Doshi & Nair, 2018, pp. 1936-1939) including buildings (Lee, Kim, Choi, Kim and Lee, 2018, pp. 3966+). Its application is particularly notable in areas like price prediction and streamlining costs when dealing with markets that investors might not be familiar with (Jung, Kim, & Jin, 2022, pp. 345+).

By analyzing vast amounts of data and recognizing patterns, especially geospatial inputs, machine learning can provide insights that traditional analytical methods may not provide (Geerts, Vanden Broucke, & De Weerdt, 2023; Schernthanner, 2017). For investors, this means being equipped with predictions backed by a wealth of data points, which can enhance the accuracy of investment decisions. Furthermore, the uncertainties that are often inherent in real estate – such as fluctuating market values, changing neighborhood dynamics, evolving economic trends, or societal values on home characteristics – can be mitigated (Yagmur, Kayakus, & Terzioglu, 2022, pp. 39+). The ability of machine learning to process and analyze data rapidly means that investors can receive timely insights, potentially allowing for proactive rather than reactive decisions.

As technology advances, machine learning integration into the real estate sector holds significant promise for refining investment strategies and reducing potential risks.

2.3 Advanced Techniques and Data Sources

The domain of real estate is experiencing a transformation with the integration of sophisticated computational methods, such as complex machine learning and deep learning models, as seen in the expanding usage of complex regression models (Hong, Choi, & Kim, 2020, pp. 140+; Adeojo, 2021). These advanced analytical approaches go beyond traditional techniques by diving deeper into intricate patterns, correlations, and anomalies that might be overlooked by simpler models (Geerts, Vanden Broucke, & De Weerdt, 2023; Schernthanner, 2017).

Additionally, there is a rising emphasis on harnessing unstructured data – which includes non-traditional sources like social media sentiments, news articles, or even video feeds – and geospatial information to offer a more comprehensive perspective on properties and markets (Geerts, Vanden Broucke, & De Weerdt, 2023). Geospatial data brings in locational insights that can have profound implications for real estate, such as understanding proximity to amenities or assessing the desirability of a location (Geerts, Vanden Broucke, & De Weerdt, 2023; Schernthanner, 2017).

By combining these state-of-the-art models with diverse data sources, there is an enhanced possibility of pinpointing investment opportunities that would have remained obscured in the past (Jung, Kim, & Jin, 2022, pp. 345+). As technology and data science evolve, machine learning will play a pivotal role in redefining the landscape of real estate predictions and insights.

2.4 Inclusivity in Real Estate Investment

The world of real estate holds immense potential for wealth generation and economic growth. Nevertheless, despite its vast opportunities, some misconceptions and barriers have historically limited its accessibility (Ullah & Sepasgozar, 2020). One such prevalent notion is that real estate investing is an exclusive domain for the affluent, often discouraging individuals from modest backgrounds from exploring it as a viable investment option (Ullah & Sepasgozar, 2020). Coupled with this is the challenge of obtaining real-time data, which is crucial in making informed investment decisions. Potential investors might navigate the market unthinkingly without timely and accurate data, adding to investment hesitations.

This study aims to dismantle these barriers, creating a more accessible environment for all aspiring real estate investors, with a particular emphasis on residential real estate. By concentrating exclusively on the residential sector, the study offers an in-depth and specialized analysis for both newcomers and seasoned investors. This focus allows for a comprehensive understanding of the intricacies and trends within the residential market, equipping investors with the knowledge and tools necessary for informed decision-making.

3 Methods

3.1 Data Collection

This research aimed to garner comprehensive information that would provide deep insights into the real estate market for the geographical areas of Frisco and Plano, Texas (Figure 1 and 2, respectively). These areas were chosen due to potential and signs of significant growth and the influx of new businesses, making them particularly interesting for real estate market analysis.

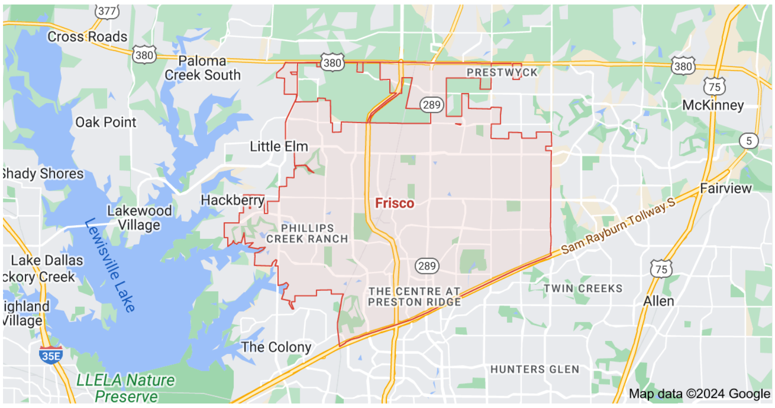


Fig. 1. This is a geographical map of Frisco, Texas, depicting the locations of real estate listings that were sourced.

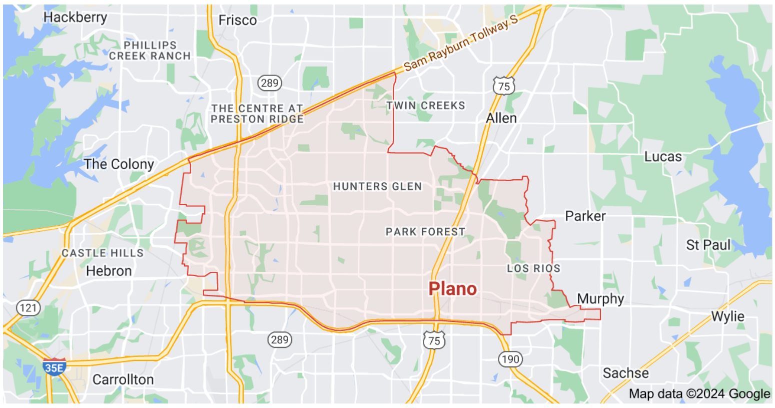


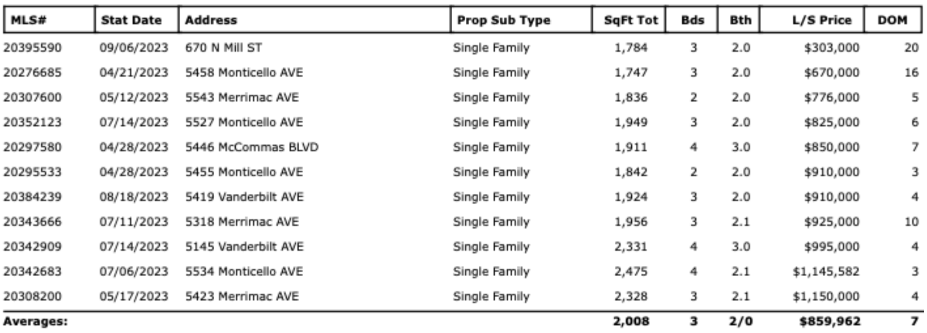
Fig. 2. This is a map of Plano, Texas, showing the locations of real estate listings sourced.

3.1.1 Housing Data

The primary source of our data was the “Sold” listings of the Redfin. This platform was chosen due to its extensive listings and up-to-date information on recent property sales.

Finally for further insights on housing prices and relevant market factors, “Multiple Listing Service (MLS) data will be leveraged: The MLS is a trusted source of real estate data, offering a vast repository of property listings, sales, and pricing information. The research utilized MLS data to understand the broader market trends and gather specific property details.

Table 1. Dataset of MLS Listings in the Dallas-Fort Worth Area



MLS#	Stat Date	Address	Prop Sub Type	SqFt Tot	Bds	Bth	L/S Price	DOM
20395590	09/06/2023	670 N Mill ST	Single Family	1,784	3	2.0	$303,000	20
20276685	04/21/2023	5458 Monticello AVE	Single Family	1,747	3	2.0	$670,000	16
20307600	05/12/2023	5543 Merrimac AVE	Single Family	1,836	2	2.0	$776,000	5
20352123	07/14/2023	5527 Monticello AVE	Single Family	1,949	3	2.0	$825,000	6
20297580	04/28/2023	5446 McCommas BLVD	Single Family	1,911	4	3.0	$850,000	7
20295533	04/28/2023	5455 Monticello AVE	Single Family	1,842	2	2.0	$910,000	3
20384239	08/18/2023	5419 Vanderbilt AVE	Single Family	1,924	3	2.0	$910,000	4
20343666	07/11/2023	5318 Merrimac AVE	Single Family	1,956	3	2.1	$925,000	10
20342909	07/14/2023	5145 Vanderbilt AVE	Single Family	2,331	4	3.0	$995,000	4
20342683	07/06/2023	5534 Monticello AVE	Single Family	2,475	4	2.1	$1,145,582	3
20308200	05/17/2023	5423 Merrimac AVE	Single Family	2,328	3	2.1	$1,150,000	4
Averages:				**2,008**	**3**	**2/0**	**$859,962**	**7**

Description: The image displays Table 1, which provides a snapshot of Multiple Listing Service (MLS) entries for single-family homes in the Dallas-Fort Worth area (Holmes, 2023, pp. 4).

3.1.2 Time Frame

The study focuses on homes sold within the three months preceding January 2024. This recent and specific time frame was selected to ensure the data reflects the current market conditions, ensuring that our findings are both relevant and timely.

3.1.2.1 Rationale for Snapshot Approach

1. **Relevance:** By focusing on a short and recent time frame, the research could be ensured to reflect the data that displayed the most up-to-date market dynamics, making the findings particularly pertinent for stakeholders looking for the current situation.
2. **Manageability:** A snapshot approach allowed us to delve deep into the data without being overwhelmed by the sheer volume that a more extended time frame might have introduced. This meant more detailed and focused analyses.

3.1.2.2 Potential Limitations and Consideration

While the snapshot approach has its advantages, it's essential to acknowledge potential limitations:

1. **Limited Temporal Scope:** By focusing solely on our 3-month snapshot before January 2024, there will likely be a cost of missing out on broader trends or cyclic patterns that a longer time frame might reveal, limiting our modeling to predictions of current prices rather than future prices.

2. **External Events:** Any significant events or disruptions specific to January 2024 or even 2023-2024 could influence the data, potentially making it less generalizable to other months or years.

3.2 Data Preprocessing, Cleaning, and Integration

Once the data is obtained, the next stage in any data-driven study is the preprocessing and cleaning of the raw data. The integrity and reliability of the results heavily depend on the quality of the input data, particularly when leveraging machine learning models that thrive on well-structured, consistent data.

3.3.1 Addressing Missing Values

One common challenge in datasets is the presence of missing values. Ignoring or mishandling these can lead to biased or inaccurate results. To combat this, imputation methods were implemented and tailored to the nature and distribution of each variable. The features were cross-referenced with the listings for categorical data to understand value missingness better.

An example can be found when addressing missing values in “Garage Spaces.” These observations tended to still have a value in “Parking Spaces,” but it was correlated with other types of parking, such as “Carport Spaces.” When these instances were encountered, the imputation of the missing value for “Garage Spaces” was filled with 0 because of the assumption that there was no specific “Garage Spaces.” Regarding numerical data, mean and median imputations were considered based on the data distributions for each feature being investigated. For a feature that did not fall into these categories, such as High School Name, we utilized a K Nearest Neighbors imputation approach rather than simply imputing with a median.

3.3.2 Outlier Detection and Treatment

Outliers, or extreme values that deviate significantly from other observations, can disproportionately influence the outcomes of some machine learning models. Using statistical methods, outliers were identified and assessed.

In some cases, outliers were retained if observations represented genuine, significant observations. In others, outliers were either transformed or removed to prevent them from skewing the model's performance. Because there are varying submarkets within the geographical area, specific neighborhoods stand out much more than others in terms of price, desirability, etc. Variables such as these were considered in determining how to approach dating with outliers and aided decisions in adjusting the scope of the research to capture the real estate market being investigated best.

3.3.3 Rectifying Inconsistencies

Datasets, especially those extracted from diverse sources like Redfin, can sometimes need to be more consistent. These could be in the form of varied naming conventions, different measurement units, or mislabeled categories. The data sets were meticulously reviewed to identify and rectify these discrepancies, ensuring uniformity and consistency across all data points.

3.3.4 Feature Engineering and Selection Process

Feature engineering entailed developing new variables or modifying existing ones to capture pertinent data better. Extra characteristics were engineered, such as the creation of city distance-based features, to improve the models' capacity for price prediction.

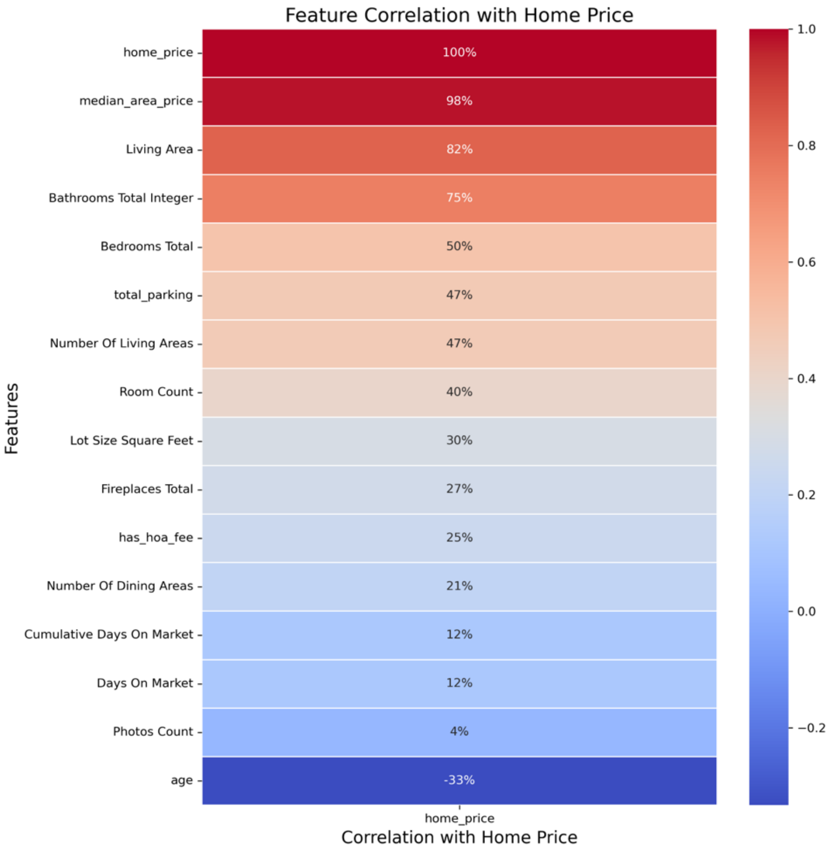


Fig. 3. This correlation heatmap illustrates the percentage of correlation of various housing features with home price, highlighting median area price and living area as the most positively correlated factors. At the same time, age shows a notable negative correlation.

3.3.5 Data Challenges

One of the primary challenges encountered was scraping data from Redfin. Ensuring the correct and consistent data formatting required significant effort due to the diverse nature of listings and the varying levels of detail provided. Additionally, structuring this data into a usable format for analysis posed its own challenges, requiring meticulous attention to detail and validation of data integrity.

3.4 Prediction Modeling and Selection Criteria

Selecting the right machine learning model is pivotal for ensuring accurate and actionable insights. The selection criteria revolved around:

1. **Data Type:** The nature of the data significantly influences algorithm selection for instance, numerical data on house prices over time may benefit from time-series analysis models such as ARIMA to capture trends and seasonality. Categorical data, such as property types or amenities, might lean toward ensemble methods like random forests.
2. **Prediction Task:** The objective of the prediction task also plays a part in the choice of methods we proceed with. For classification, such as assigning investment grades to properties, we consider logistic regression or decision trees. For regression tasks aimed at forecasting home prices, linear regression, and gradient boosting machines can be considered for the balance of accuracy and performance.
3. **Model Interpretability:** Given the substantial financial stakes in real estate investments, preference is given to models that offer higher interpretability. This enables stakeholders to understand the rationale behind predictions, ensuring transparency. However, the ultimate selection is biased towards models demonstrating superior predictive accuracy, acknowledging the trade-off between complexity and interpretability.

3.4.1 Regression Models

Traditional regression models, such as Lasso Regression and Linear Regression, were implemented for the foundational framework of the predictions. These models are instrumental in elucidating the linear relationships between various variables and real estate values, offering insights and interpretations that shed light on the factors influencing home prices.

Lasso Regression is effective in identifying significant predictors of house prices and plays a crucial role in feature selection and dimensionality reduction. By applying a penalty to the absolute size of the coefficients, Lasso Regression can drive some coefficients to zero, effectively eliminating those variables from the model. This capability is invaluable to simplify the model and focus on the most impactful features, thereby enhancing model interpretability and potentially improving prediction performance by reducing the complexity of the feature space.

3.4.2 Ensemble Models

Ensemble approaches, specifically Random Forest and XGBoost (Extreme Gradient Boosting), were used to detect nonlinear relationships and complexities in the data. These models are renowned for their abilities to increase prediction precision by combining several weak learners into a cohesive, strong predictive model.

Random Forest employs a collection of decision trees to create a more generalized model that reduces overfitting and enhances performance on unseen data by averaging the predictions of numerous trees. While XGBoost, on the other hand, stands as a refined implementation of gradient boosting designed for speed and performance. It systematically corrects the mistakes of prior models in the sequence, optimizing both computational speed and predictive accuracy.

3.4.3 Clustering Techniques

Agglomerative clustering was employed to uncover inherent groupings within the dataset (Figure 4). This hierarchical clustering technique is particularly useful for identifying and understanding the data structure without the need for pre-specifying the numbers of clusters. By iteratively merging the closest pairs of clusters, this method built up a hierarchical tree of clusters known as a dendrogram. The dendrogram serves as a visual representation, offering an intuitive understanding of the data's hierarchical structure and the relationship between its components.

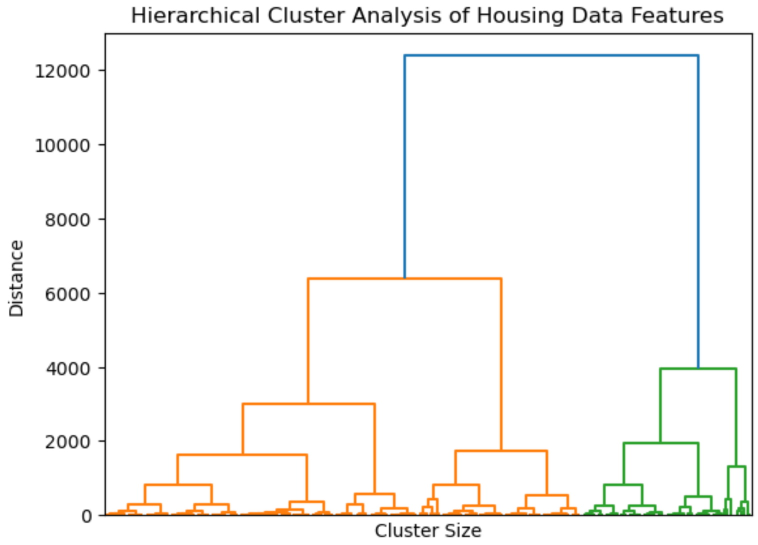


Fig. 4. This is a Hierarchical Clustering Dendrogram depicting the groupings of housing market data features based on similarities. Figure 4 illustrates the results of applying Agglomerative Clustering to housing market data, with the dendrogram revealing natural groupings among features that influence local house prices.

In addition to the hierarchical clustering, the silhouette score was utilized as a metric to assess the quality of the clusters formed. The silhouette scores provide a measure of how similar an object is to its cluster (cohesion) compared to other clusters (separation). High silhouette scores indicate well-fitted objects within clusters and a clear distinction between clusters, which is essential for validating the consistency within clusters of data.

3.5 Model Evaluation

To thoroughly evaluate the performance of the machine learning models, a suite of evaluation metrics, specifically Mean Absolute Error (MAE), Root Mean Squared Error (RMSE), R-squared (R^2), and Mean Absolute Percentage Error (MAPE) was utilized. These metrics provide a quantifiable measure of the models' accuracy and give insight into the reliability and consistency of the predictions. Using these metrics could ensure that the models are both precise and dependable in forecasting real estate trends.

3.6 Prediction Generalization and Overfitting

The robustness of machine learning models often depends on the ability to generalize to new, unseen data. One of the significant challenges in machine learning is overfitting, where a model might perform exceptionally well on the training data but needs to generalize effectively on new data.

In addition to standard cross-validation techniques, with k-fold cross-validation being the primary method, we specifically set aside a hold-out set comprising data newly gathered from a separate but comparable geographic area not represented in the training set. This hold-out set serves as a critical test for our models' ability to generalize beyond the training environment.

3.6.1 Cross-Validation Explained

Cross-validation is a resampling technique used in machine learning to assess the performance of algorithms on independent datasets. It provides a holistic view of the model's performance, reducing the risk of unintentionally tailoring it too closely to a specific subset of data.

3.6.2 K-Fold Cross-Validation Explained

In k-fold cross-validation:

1. **Partitioning:** The entire dataset is divided into k equal (or nearly equal) subsets or "folds". Typically, values of k such as 5 or 10 are chosen based on the dataset's size and the specific problem domain.
2. **Training and Testing:** In this process, the model undergoes training on k-1 of the folds and is evaluated on the one-fold that is left out. This cycle is

executed k times, ensuring that each k fold serve as a validation set exactly once.

3. **Aggregation:** The k results from the folds can then be averaged (or otherwise combined) to produce a single estimation of model performance.

3.6.3 Benefits of K-Fold Cross-Validation

- **Bias Reduction:** Training and validating the model on different data subsets minimizes the risk of model bias towards a specific data arrangement.
- **Variance Reduction:** Averaging the results over multiple rounds of testing reduces variance, leading to a more stable and reliable performance metric.
- **Optimal Utilization of Data:** Since every data point gets to be in both the training and testing dataset, it ensures that the model is exposed to the entire data set, optimizing the data's use.

3.7 Ethical Considerations

In the realm of research, especially one that delves into sectors as sensitive as real estate, ethical considerations are vital. This study was designed and conducted with a firm commitment to upholding the highest ethical standards, ensuring that all stakeholders' rights, privacy, and interests were safeguarded.

3.7.1 Data Privacy and Anonymization

- **Data Collection Ethics:** When gathering data from platforms like Redfin and MLS listings, it's crucial to ensure that no personally identifiable information (PII) is captured, stored, or analyzed. The data collection methodologies adhered strictly to this principle.
- **Anonymization Techniques:** Post data collection, any information that could potentially be traced back to an individual or a specific property was anonymized. This included obfuscating details such as exact addresses, owner names, and other unique identifiers.

3.7.2 Democratizing Real Estate Investment

- **Promoting Inclusivity:** Historically, the real estate sector has had high entry barriers, often sidelining potential investors who lack hefty capital or intricate market knowledge. This research's underlying ethos is the democratization of real estate investment, ensuring that a broader segment of the population can participate and benefit.
- **Lowering Entrance Barriers:** By leveraging cutting-edge technology and in-depth analysis, the goal was to provide insights that make the investment process more transparent and approachable. This transparency can empower

individuals, irrespective of experience in real estate investments, to make informed decisions and potentially enter the market.

3.8 Software and Tools

All data preprocessing, model development and analyses were performed using the Python programming language within the Visual Studio Code IDE. For the data manipulation, analysis, and visualization tasks, a range of Python libraries were utilized, including pandas, seaborn, sklearn, statsmodels, NumPy, and matplotlib. These tools and libraries are widely accepted and prevalent in data science and machine learning tasks.

3.9 Prediction Generalization and Overfitting

One significant limitation is the availability and quality of data. The datasets used may suffer from errors, omissions, or biases, which can inherently affect the reliability and validity of our predictions. Furthermore, while our machine learning models demonstrate promising accuracy and reliability in forecasting real estate values, performance is intrinsically tied to the quality, accuracy, and representativeness of the data employed.

Moreover, the ever-shifting nature of real estate markets means that external factors not captured in historical data, such as sudden economic shifts, changes in zoning laws, or unforeseen natural disasters, can impact the predictive accuracy of our models. Future work aimed at enhancing data collection methods, refining model algorithms, and expanding the geographic diversity of the datasets may help to mitigate these limitations.

3.10 Algorithms and Data

To glean meaningful insights from the dynamic real estate markets of Central and North Dallas, it is crucial to employ a combination of sophisticated machine learning algorithms and robust data collection techniques. This section delves into the algorithms chosen for this study and the data sources and methodologies underpinning the research.

3.10.1 Algorithms Utilized

- **Linear Regression:** One of the most foundational algorithms in predictive modeling, linear regression will be used to establish relationships between various independent variables (such as property features and location metrics) and dependent variables (like property prices). Its simplicity and interpretability make it an ideal choice for understanding the linear relationships within the dataset.

- **Random Forest:** A more sophisticated ensemble learning method, the Random Forest algorithm creates a 'forest' of decision trees during training and outputs the mode of the classes (classification) or mean prediction (regression) of individual trees for unseen data. This approach is adept at handling complex datasets with potential nonlinear relationships and offers a more holistic view of the data's intricacies.
- **Preprocessing with Scikit-Learn:** Before feeding data into prediction models, it's essential to preprocess it to ensure optimal model performance. Scikit-Learn's preprocessing packages will be used to handle tasks such as data normalization, encoding categorical variables, and handling missing values.

3.10.2 Data Sources and Collection Methodology

1. **Redfin:** Redfin offers an extensive array of property listings, transaction prices, and historical market data, with a unique emphasis on user-friendly technology and transparency in real estate transactions. Leveraging Redfin's innovative location-based search capabilities, market analysis tools, and comprehensive listings download for this study will pave the way for our data collection process and general insights.
2. **MLS (Multiple Listing Service):** Serving as a trusted and extensive database for real estate professionals, MLS provides detailed property listings, sales data, and other crucial real estate metrics. A Real Estate Agent supplied access to and the ability to extract relevant data from MLS for the targeted regions of Central and North Dallas.

4 Results

4.1 Overview and Scope

The following section presents the outcomes of our predictive analysis on local house prices, using advanced machine learning techniques to forecast real estate values. Central to the integrity of our evaluation is the employment of hold-out data sets- distinct subsets of data that the models have yet to encounter during the training phase. This approach ensures that our performance metrics reflect the models' capabilities to generalize to new, unseen data, thereby providing a robust assessment of predictive power.

To quantify the accuracy and reliability of our models, metrics including R-Squared Adjusted (R^2 Adjusted), Root Mean Squared Error (RMSE), Mean Absolute Percentage Error (MAPE) and Mean Absolute Error (MAE) were selected.

Acknowledging the crucial influence of geographical location on real estate valuation, our analysis has been stratified by city. This segmentation allows for a more nuanced understanding of the models' performance, recognizing that market dynamics and property values can vary significantly across different locales. Such an approach enables a more granular interpretation of the data, reflecting localized patterns and trends that could be obscured in a broader, non-segmented analysis.

Lastly, the results' scope maintains that the target feature "Sold Price" represents the final transaction price of homes. It is important to note that this price is based on listings marked as "sold," thereby assuming the listed price equates to the actual transaction value, not including taxes, fees, and other miscellaneous closing costs. The model, therefore, provides insights into the determinants of final selling prices within the housing market under these conditions.

4.2 Linear Regression

Linear Regression is the starting point in the range of predictive tools utilized for this research. First, it helps us set a benchmark by clearly showing how different factors (independent variables) influence the sales price of homes across a given market. This insight is vital for comparing the performance of more complex models we use in our analysis. By identifying key features and the presence of a significant effect on the target feature, a deeper understanding can be gained of what drives home prices in each market, guiding both the analysis and potential recommendations.

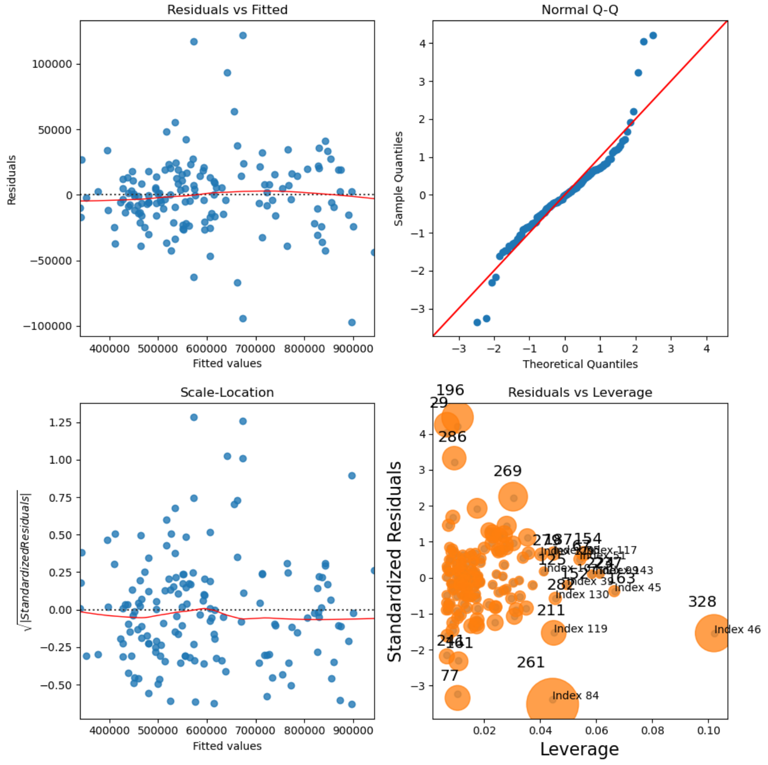


Fig. 5. This is the Linear Regression Diagnostics plot for the Frisco model. The image displays a set of linear regression diagnostic plots for the Frisco Model, including residuals versus fitted values, a normal Q-Q plot, a scale-location plot, and a residual versus leverage plot, which are used to assess the model's assumptions and identify potential outliers or influential points.

The summary output of the linear regression model was generated to assess the fit of the model.

- **Residuals Versus Fitted:** This plot checks the assumptions of linearity and homoscedasticity of residuals. Ideally, the residuals should appear as a random cloud of points disbursed around the horizontal axis. In this case, there is a deviation from a constant variance in the middle of the plot near the 675,000 area.
- **Normal Q-Q:** This Quantile-Quantile plot checks the assumption that the residuals are normally distributed. Points following the red line indicate normality. The plot above shows the deviation of this line in the tails, which indicates potential issues with outliers.
- **Scale-Location (or Spread–Location):** This plot provides information on whether residuals are spread equally along the range of predictors. Similar to the Residuals Versus Fitted plot, a small uptick near the center of the plot can be seen.
- **Residuals Versus Leverage:** This plot illustrates the outliers which have an overbearing influence on the regression line. The points in this plot are labeled by observation index, and those with high leverage are further away from the center of the plot horizontally. There appears to be evidence of several observations with high leverage, which aligns with evidence seen earlier in this residual analysis.

Table 2. Linear Regression Summary Results - Frisco

Feature	Coefficient	Standard Error	T-Statistic	P-Value	Lower Bound 0.025	Upper Bound 0.975
Age	-5612.87	2296.13	-2.444	0.016	-1.02e+4	-1075.69
Total Number of Bathroom s	3274.76	3166.22	1.034	0.303	-2981.73	9531.25
Number of Fireplaces	-1.018e+04	2435.34	-4.180	0.000	-1.5e+04	-5367.95
Median Estimated Price Based on	1.415e+5	3306.01	42.791	0.000	1.35e+05	1.48e+5

Living Area						

Table 3. Linear Regression Summary Results – Plano

Feature	Coefficient	Standard Error	T-Statistic	P-Value	Lower Bound 0.025	Upper Bound 0.975
Age	-6311.49	3679.09	-1.707	0.090	-1.36e+04	984.83
Living Area	8650.11	7969.12	1.085	0.279	-7077.22	2.44e+04
Median Estimated Price Based on Living Area	1.267e+05	8421.86	15.03	0.000	1.1e+05	1.43e+05

Table 4. Linear Regression Results – Frisco

Performance Metric	Hold Out Test Score
RMSE	30,454.59
MAE	18,716.04
MAPE	3.46%
R^2- Adjusted	.94

Table 5. Linear Regression Results – Plano

Performance Metric	Hold Out Test Score
RMSE	35,734.73
MAE	23,301.38
MAPE	4.94%
R^2- Adjusted	.94

4.3 Random Forest

Building upon the insights from the simple linear regression model, this research now shifts to employing Random Forest as a more complex predictive tool to enhance the accuracy and robustness of our housing market analysis. Random forest is an ensemble method that combines multiple decision trees to improve prediction accuracy and offers an advantage in dealing with non-linearity and interaction effects between variables that are present in housing data.

Table 6. Random Forest Results – Frisco

Performance Metric	Validation Set Score	Hold Out Set Score
RMSE	27,093.56	28,628.44
MAE	18,734.23	18,646.20
MAPE	3.13%	3.59%

Table 7. Random Forest Results – Plano

Performance Metric	Validation Set Score	Hold Out Set Score
RMSE	39,372.59	40,314.84
MAE	25,761.93	25,780.46
MAPE	5.5%	5.5%

4.4 Dimensionality Reduction

Several techniques were employed to reduce the dimensionality of the features in the data so that the models could utilize and retain only the most informative features.

Pearson's R Correlation
We started with Pearson's R correlation analysis to assess the linear relationship between independent and target variables. This technique allowed us to quantify the strength and direction of the linear correlations between features and home prices. By setting a threshold for correlation coefficients, we could eliminate features with negligible linear relationships to the target, thus reducing dimensionality while preserving those variables most likely to influence the model's predictions (Ramsey & Schafer, 2013, pp. 194).

ANOVA Testing
Further, we utilized ANOVA (Analysis of Variance) testing to examine the categorical variables in our dataset. ANOVA helps identify whether there are statistically significant differences in the means of the target variable across different categories of each feature (Ramsey & Schafer, 2013, pp. 217-218). By comparing the variance between groups to the variance within groups, ANOVA testing enabled us to discern which categorical features have a meaningful impact on the target variable. Features that showed significant variance affecting the target variable were retained, whereas those with minimal impact were considered for removal.

LASSO Regression

Lastly, we applied LASSO (Least Absolute Shrinkage and Selection Operator) regression, a powerful regularization and feature selection technique. LASSO imposes a penalty on the absolute size of the coefficients, effectively shrinking some of them to zero (James et al., 2023, pp. 485). This property of LASSO makes it exceptionally useful for dimensionality reduction, as it automatically selects a subset of the most predictive features while discarding the rest. Through LASSO regression, we were able to refine our feature set further, keeping only those variables that contribute significantly to the model's predictive accuracy.

4.5 Model Result Analysis – Linear Regression

In analyzing the Linear Regression model results for the real estate market of Plano and Frisco, TX, several key insights emerge. For Frisco, as indicated in Table 3, the model yielded an RMSE of 30,454.59, MAE of 18,716.04, and a MAPE of 3.46%, with an R^2- Adjusted of .94. These figures suggest a high level of accuracy in the model's predictive capability, with an average error margin of less than 5% of the actual home values. The high R^2- Adjusted value underscores the model's strength in explaining the variability of the response data around its mean. Conversely, for Plano, Table 4 shows the model with a slightly less accurate RMSE of 35,734.73, MAE of 23,301.38, and a MAPE of 4.94%, though maintaining an R^2- Adjusted of .94. This indicates a consistent ability to account for the variability in home prices, despite the higher error rates in predictions compared to Frisco.

In Table 2, the OLS Regression summary for Frisco reveals that the Median Estimated Price Based on Living Area is a highly significant predictor ($p < 0.001$), with a large positive coefficient, suggesting that as the median price per square foot increases, so does the predicted selling price of the home. The negative coefficient for Age indicates that newer homes tend to fetch higher prices. While the Total Number of Bathrooms and the Number of Fireplaces show statistical significance in their contribution to the selling price, their coefficients suggest differing impacts on the price.

The living_area_price_estimate feature stands out in Table 3 with a significant positive coefficient, reinforcing its role as a crucial determinant of house pricing in the area. The F-statistic probability nearing zero indicates the model's robustness, confirming that the model is statistically significant. However, the Durbin-Watson statistic of approximately 2 suggests no autocorrelation in the residuals, which supports the independence of the observations. The Omnibus test's probability indicates that the residuals are normally distributed. In contrast, the Jarque-Bera test significantly rejects the null hypothesis, indicating that the residuals might not be normally distributed, which is a point that may need further investigation or model refinement.

4.5 Model Result Analysis – Random Forest

The Random Forest results for Frisco and Plano provide a comprehensive view of the model's performance in the real estate valuation context. For Frisco, as seen in Table 6, the Random Forest model achieves an RMSE of 27,093.56 on the validation set and

28,628.44 on the hold-out set, which suggests a relatively low spread of errors in the predictions. The MAE scores, 18,734.23 for the validation set and 18,646.20 for the hold-out set, indicate a strong central tendency in the model's predictive accuracy, with the average error being within a reasonable range of the actual home values. The MAPE values of 3.13% for the validation set and 3.59% for the hold-out set reflect a proportional accuracy that aligns well with the market dynamics, considering the complexity of housing price predictions.

For Plano, as depicted in Table 7, the Random Forest model results indicate an RMSE of 39,372.59 on the validation set and 40,314.84 on the hold-out set, which are higher than those for Frisco, pointing to a greater variance in the predictive errors for the Plano housing market. The MAE is also higher, at 25,761.93 for the validation set and 25,780.46 for the hold-out set, suggesting that the predictions in Plano may be less central to the actual values than in Frisco. The consistency in MAPE, at 5.5% for both validation and hold-out sets, denotes that the model's percentage errors are uniform across different subsets of data. However, they are higher than those in Frisco, which might be due to specific local market factors affecting Plano.

The Random Forest model's performance in both cities indicates a robust predictive capability, though it's evident that the model is more accurate in Frisco than in Plano. The higher errors in Plano could be attributed to a more diverse and complex real estate market or the model not capturing some locality-specific nuances that influence housing prices. Despite these differences, the Random Forest model stands out for its ability to handle non-linear relationships and interaction effects between features without needing transformation. It is particularly useful in the heterogeneous and complex domain of real estate valuation.

5 Discussion

5.1. Research Interpretations

Linear Regression Model

The analysis of the linear regression model results for the real estate markets in Plano and Frisco, Texas, provides significant insights into their respective housing price dynamics. For Frisco, the model showcases a high level of predictive accuracy, with an RMSE of 30,454.59 MAE of 18,716.04, and a MAPE of 3.46%. These metrics indicate that the models' predictions are on average within less than 5% of the actual home values underlining a strong ability to explain the variability of home prices around their mean.

The performance in Plano, although slightly less accurate (RMSE of 35,734,73, MAE of 23,301.38, and MAPE of 4.94%), still maintains a high R^2– Adjusted value of .94. This consistency in explaining home price variability suggests that the model is robust across different market conditions, even though the error margins are slightly higher for Plano compared to Frisco.

Further, the OLS Regression summary for Frisco points to the Median Estimated price based on living area as a significant predictor of housing prices, indicating a direct correlation between the price per square foot and the overall selling price. The significance of the total number of bathrooms and the number of fireplaces, despite having different impacts on the prices, also highlights the nuanced nature of housing valuations.

Random Forest Model

The Random Forest model extends our understanding of the real estate valuation in Frisco and Plano. In Frisco, the model demonstrates a superior performance with lower RMSE and MAE values on both validation and hold-out sets compared to Linear Regression, suggesting a more consistent prediction with actual home values. This is attributed to the model's ability to handle complex, non-linear relationships and interactions between features without requiring transformation, an advantage in the diverse and intricate real estate market.

Conversely, the performance in Plano, while robust, indicates a higher variance in prediction errors. This could reflect a more complex and diverse housing market in Plano, or potentially, the model's limitations in capturing certain local market nuances. Despite this, the uniform MAPE values across different data subsets denote a consistent level of predictive accuracy, showcasing the model's adaptability.

5.2. Research Implications

The comparison between Linear Regression and Random Forest models across both cities underscores the complexity of real estate market prediction. The differences in model performance can be attributed to the inherent characteristics of the models themselves—Linear Regression's strength in capturing linear relationships and Random Forest's ability to manage non-linear dynamics and interaction effects.

The higher predictive accuracy in Frisco across both models suggests a less volatile market or a more uniform distribution of housing features influencing prices. In contrast, the greater prediction errors in Plano hint at a more heterogeneous market with potentially unique local factors at play.

The discrepancies in the residuals' distribution, as indicated by the Durbin-Watson statistic and the Jarque-Bera test, suggest areas for further investigation. While the independence of observations is confirmed, the question of normal distribution of residuals, especially in the Linear Regression model, highlights the need for potential model refinement or the exploration of alternative models to better capture the underlying market dynamics.

This analysis not only sheds light on the predictive capabilities of both models in the context of real estate valuation in Frisco and Plano but also highlights the importance of model selection and refinement in accurately capturing market complexities.

5.3. Notable Findings

One of the key takeaways of this study is the paramount importance of leveraging the price of recently sold homes within the same geographical area as a feature in predicting real estate values. This approach provided a dynamic and highly relevant measure of market conditions, capturing the immediate effects of supply and demand fluctuations on property prices. This feature proved to be a critical determinant in enhancing the model's predictive accuracy, reflecting the real-time market valuation more accurately than traditional static metrics.

The inclusion of this geographical price feature underscores the significance of spatial factors in real estate valuation. It highlights how properties in proximity not only share similar characteristics but also how their values are interdependent and are highly related to nearby recent transactions. This finding suggests a more granular layer of market dynamics, where local sales activities significantly impact the valuation of nearby properties, reinforcing the need for models to incorporate localized market trends for precision in predictions.

Furthermore, the utilization of the random forest model stands out as a crucial methodological choice in this research. Random Forest's ability to manage complex, non-linear relationships and its inherent mechanism for handling overfitting through ensemble learning made it particularly effective in digesting the nuanced and multi-dimensional data inherent to real estate markets. The model's performance in both Frisco and Plano – notably its superior predictive accuracy in Frisco – demonstrates its robustness and adaptability to diverse market conditions.

These findings collectively illustrate the transformative impact of integrating innovative data features and sophisticated modeling techniques on the predictive analysis of real estate markets. The creation of a feature based on the price of recently sold homes within the same geographical area, paired with the strategic utilization of the Random Forest model, not only provided a deeper understanding of market dynamics, but also set a new benchmark for accuracy and reliability in real estate valuation.

5.4. Limitations

While this study provides important insights into the real estate markets of Plano and Frisco, TX, it is not without limitations. The following are key constraints that should be acknowledged:

- **Data Collection Limitations:** The primary data source for this study involved scraping websites such as Redfin or Zillow. While these platforms offer extensive listings and transaction data, they may not capture the entire market or all relevant property attributes. Additionally, the data extracted from these sources may be subject to reporting delays, inaccuracies, or biases.
- **Exclusion of Municipality and Tax Zones:** The study did not incorporate information on municipality and tax zones, which can significantly influence property values. These factors affect property taxes, access to services, and investment in infrastructure, all of which play crucial roles in determining real estate prices.
- **Reliance on Publicly Available Data:** The reliance on publicly available data sources may limit the comprehensiveness and depth of the analysis. Official records and proprietary datasets could provide more accurate and detailed insights into property characteristics, transaction histories, and market trends.
- **Generalizability of Findings:** The findings from Plano and Frisco may not be directly applicable to other regions due to geographical, economic, and

regulatory differences. The real estate market dynamics can vary significantly across different locations, affecting the generalizability of the results.

- **Emerging Market Conditions:** The rapidly changing nature of real estate markets means that the findings may become dated as new economic conditions, policies, and trends emerge. Continuous updates and refinements of the models are necessary to maintain their relevance and accuracy.

5.5. Ethics

In our research, sensitive details such as neighborhood names, school districts, and precise locations of properties are present. On the one hand, these details provide invaluable insights into the complex dynamics of housing markets, enabling more granular analysis and targeted interventions. For example, understanding how schools are distributed across districts can shed light on educational disparities and inform policies to improve access to quality education. However, while enhancing the accuracy and depth of our analysis, such information raises significant privacy and discrimination concerns.

The specificity of our study's data reveals the intricate fabric of socioeconomic disparities and increases the risk of perpetuating entrenched biases in housing markets. By shedding light on the nuanced characteristics of neighborhoods, school districts, and individual properties, our dataset unintentionally exposes the structural inequities that underpin housing accessibility and affordability. In many cases, identifying socioeconomic disparities can unintentionally reinforce existing biases, exacerbating disparities in access to housing and resources. While necessary for comprehensive analysis, the detailed nature of our dataset presents a slew of privacy concerns that must not be overlooked. The risk of inadvertently identifying individual properties and exposing homeowners to unwanted attention or privacy breaches is significant, emphasizing the importance of strict privacy safeguards.

Considering these concerns, researchers must take a principled approach to data handling and analysis. This includes establishing strong privacy protocols to anonymize or aggregate sensitive information, thereby protecting individuals' and communities' identities while allowing for meaningful analysis. Furthermore, we must be vigilant against using data to perpetuate biases or reinforce systemic inequalities.

A two-pronged approach addressed these multifaceted concerns: anonymize data and selectively remove features. We use anonymization to obscure identifying information while retaining the dataset's analytical integrity. By removing features that pose the greatest privacy risks, such as precise property locations, we reduce the possibility of inadvertent disclosure while protecting homeowners' and residents' privacy rights. Furthermore, using such detailed information raises ethical concerns about consent and transparency. Homeowners and residents may be unaware that their information is being used for research purposes, nor do they fully understand the potential consequences of its disclosure. Without adequate safeguards, there is a risk of inadvertently violating people's privacy rights and undermining trust in the research process.

Our ethical framework prioritizes the principles of fairness, transparency, and inclusion. We support analyses that empower marginalized communities, highlight diverse voices, and promote equitable outcomes for all people. By prioritizing ethics in our research endeavors, we maintain the scientific process's integrity and foster trust and accountability in the larger community. Our approach underscores the importance of responsible data use, advocating for fair, transparent, and beneficial analyses for all individuals.

5.6 Future Research

The exploration of machine learning applications in assessing housing prices in Plano and Frisco, TX, has unveiled substantial opportunities for deeper understanding and innovation within the real estate sector. Integrating technology and data analytics holds the promise of transforming market analysis, valuation accuracy, and investment strategies.

Moving forward, several strategic avenues emerge for expanding the scope and efficacy of this research:

1. **Partnering with Real Estate Data Sources:** Collaborating with comprehensive real estate data providers such as Zillow or Redfin could significantly enhance the quality and breadth of data accessible for analysis. These partnerships could offer more granular, up-to-date, and wide-ranging datasets, including historical sales data, listing details, and user-generated content, which are invaluable for refining predictive models. Access to such rich data repositories would improve the accuracy of current valuation models and enable the exploration of new dimensions in market dynamics.
2. **Incorporating Socio-economic Factors:** Future research should consider integrating socioeconomic indicators such as income levels, employment rates, educational attainment, and demographic trends. These factors play a crucial role in shaping housing demand and prices. Understanding the interplay between socioeconomic conditions and real estate valuations can provide more holistic insights, aiding in developing models that reflect the complexities of human behavior and economic conditions in housing markets.
3. **Accessing MLS Data for Closing Prices**: Gaining access to Multiple Listing Service (MLS) data could offer a treasure trove of information on closing prices, which are pivotal for accurate market valuations. MLS databases are rich sources of real-time, comprehensive data on property transactions, including sale prices, listing times, and property features. Leveraging MLS data can enhance model precision by providing the most current and accurate reflection of market transactions.
4. **Assessing the Impact of Fractional Investments**: With the rise of fractional investments in real estate, examining impacts on market liquidity, accessibility, and long-term market dynamics is essential. Research could explore how democratizing access to real estate investment affects pricing, ownership patterns, and investment strategies, potentially reshaping the market landscape.
5. **Identifying and Mitigating Inherent Biases**: It's crucial to address the potential biases and pitfalls that may arise from relying on machine learning

models in real estate investment. Future studies should focus on identifying these biases, understanding implications, and developing strategies to mitigate related effects, ensuring that models are accurate and equitable.

By pursuing these recommendations, future research can build upon the foundations laid by this study, pushing the boundaries of what's possible in real estate market analysis and investment strategy development. The journey towards a more data-informed, technologically advanced real estate sector is challenging and exhilarating, promising a future where decisions are more nuanced, informed, and effective.

6 Conclusion

This research has meticulously analyzed real estate variables to explain the factors that determine the final selling prices of homes within the housing market of Plano and Frisco, TX. Central to our findings is the pivotal role played by the living_area_price_estimate feature, which emerged as the most significant predictor within our models. This feature, engineered to capture the median sold price of comparably sized homes in recent transactions, has allowed for an enhanced predictive accuracy that is intimately aligned with the dynamics of the local real estate market.

The Random Forest model exhibits a strong predictive capability for both cities, with Frisco's validation and hold-out test MAE scores being 18,734.23 and 18,646.20,respectively, and Plano's scores slightly higher at 25,761.93 and 25,780.46. The MAPE scores also suggest a good fit, with Frisco at 3.13% for the validation set and 3.59% for the hold-out set, whereas Plano shows a higher error with 5.5% for both sets. These metrics indicate that the Random Forest model is particularly effective in predicting home sale prices in Frisco.

Our models, especially the Random Forest, show robustness and reliability in predicting home sale prices in Frisco and Plano, with Frisco's model slightly outperforming Plano's. The strength of these models is evident in their performance metrics, highlighting their practical applicability in the real estate market. Stakeholders can leverage these insights for a more informed evaluation and prediction of property valuations. Moving forward, continuous refinement of our models and the integration of new data sources will enhance the precision and depth of our understanding of the real estate market's complexities.

It is imperative to reiterate that the scope of our model's applicability hinges on the assumption that the 'Sold Price' reflects the final transaction value, exclusive of additional financial obligations such as taxes and fees. This assumption anchors the model in a realistic context, ensuring the derived insights are practical and actionable. The research conducted herein is not merely academic; it holds substantial real-world importance by offering a nuanced lens through which stakeholders can evaluate and predict property valuations. Refining our methodologies and embracing new data sources opens the door to even more precise modeling and a deeper understanding of the real estate market's complex tapestry.

References

1. Ullah F, Sepasgozar SME. Key Factors Influencing Purchase or Rent Decisions in Smart Real Estate Investments: A System Dynamics Approach Using Online Forum Thread Data. *Sustainability*. 2020; 12(11):4382. https://doi.org/10.3390/su12114382
2. Choy, L. H. T., & Ho, W. K. O. (2023). The Use of Machine Learning in Real Estate Research. *Land*, *12*(4), NA. https://link-gale-com.proxy.libraries.smu.edu/apps/doc/A747445725/AONE?u=txshracd2548&sid=bookmark-AONE&xid=ed6a95a2
3. Jung, J., Kim, J., & Jin, C. (2022). DOES MACHINE LEARNING PREDICTION DAMPEN THE INFORMATION ASYMMETRY FOR NON-LOCAL INVESTORS? *International Journal of Strategic Property Management*, *26*(5), 345+. https://link-gale-com.proxy.libraries.smu.edu/apps/doc/A730948405/AONE?u=txshracd2548&sid=bookmark-AONE&xid=1eab7110
4. Lee, W., Kim, N., Choi, Y.-H., Kim, Y. S., & Lee, B.-D. (2018). Machine Learning based Prediction of The Value of Buildings. *KSII Transactions on Internet and Information Systems*, *12*(8), 3966+. https://link-gale-com.proxy.libraries.smu.edu/apps/doc/A558230780/AONE?u=txshracd2548&sid=bookmark-AONE&xid=5e54c3d1
5. Geerts, M., vanden Broucke, S., & De Weerdt, J. (2023). A Survey of Methods and Input Data Types for House Price Prediction. *ISPRS International Journal of Geo-Information*, *12*(5), NA. https://link-gale-com.proxy.libraries.smu.edu/apps/doc/A752424454/AONE?u=txshracd2548&sid=bookmark-AONE&xid=48a6c324
6. Schernthanner, Harald, et al. "Spatial Modeling and Geovisualization of Rental Prices for Real Estate portals." *International Journal of Agricultural and Environmental Information Systems* [IJAEIS], vol. 8, no. 2, Apr. 2017, p. NA. *Gale Academic OneFile*, https://link.springer.com/chapter/10.1007/978-3-319-42111-7_11
7. Hermans, L. D., McCord, M. J., Davis, P. T., & Bidanset, P. E. (2023). An Exploratory Approach to Composite Modelling for Real Estate Assessment and Accuracy. *Journal of Property Tax Assessment & Administration*, *20*(1), 43+. https://link-gale-com.proxy.libraries.smu.edu/apps/doc/A754172157/AONE?u=txshracd2548&sid=bookmark-AONE&xid=b09c3630
8. Yagmur, A., Kayakus, M., & Terzioglu, M. (2022). House price prediction modeling using machine learning techniques: a comparative study. *Aestimum*, (81), 39+. https://link-gale-com.proxy.libraries.smu.edu/apps/doc/A753206901/AONE?u=txshracd2548&sid=bookmark-AONE&xid=69061747
9. Hong, J., Choi, H., & Kim, W.-S. (2020). A HOUSE PRICE VALUATION BASED ON THE RANDOM FOREST APPROACH: THE MASS APPRAISAL OF RESIDENTIAL PROPERTY IN SOUTH KOREA. *International Journal of Strategic Property Management*, *24*(3), 140+. https://link-gale-com.proxy.libraries.smu.edu/apps/doc/A626121942/AONE?u=txshracd2548&sid=bookmark-AONE&xid=36d44d82
10. Carrillo, G. (2019). Predicting Airbnb Prices with Machine Learning: A case study using data from the City of Edinburg, Scotland. Towards Data Science. Retrieved from:https://towardsdatascience.com/predicting-airbnb-prices-with-machine-learning-and-location-data-5c1e033d0a5a
11. Adeojo, J. (2021). Predicting Housing Prices with Machine Learning: An end-to-end project for Kaggle’s advanced regression techniques competition. Towards Data

Science. Retrieved from: https://towardsdatascience.com/predicting-house-prices-with-machine-learning-62d5bcd0d68f

12. Jiao, J., & Bai, S. (2020). An empirical analysis of Airbnb listings in forty American cities. *Cities,* 99(102618). University of Texas at Austin. Retrieved from: https://www.sciencedirect.com/science/article/abs/pii/S0264275119306559
13. Truong, Q. Nguyen, M. Dang, H. Mei, B. (2020). Housing Price Prediction via Improved Machine Learning Techniques. *Procedia Computer Science,* 174(433-442). Texas Christian University. Retrieved from: https://www.sciencedirect.com/science/article/abs/pii/S0264275119306559
14. Qureshi, Adeel; Mushailov, Iosif; Herrera, Patricia; Hale, Phillip; and McDaniel, Reannan (2022) "A Framework for Predicting the Optimal Price and Time to Sell a Home," SMU Data Science Review: Vol. 6: No. 2, Article 16. Retrieved from: https://scholar.smu.edu/datasciencereview/vol6/iss2/16
15. Li X. (2022). Prediction and Analysis of Housing Price Based on the Generalized Linear Regression Model. *Computational intelligence and neuroscience*, *2022*, 3590224. Retrieved from: https://doi.org/10.1155/2022/359022 (Retraction published Comput Intell Neurosci. 2023 Aug 2;2023:9891083)
16. Zapata, J. (2021). Short-Term Rental Data Analysis: An Analysis of the Impact of Short-Term Rental Properties in the City of Dallas. City of Dallas. Retrieved from: https://dallascityhall.com/government/citymanager/Documents/FY%2020-21%20Memos/STR%20Data%20Analysis%2005032021.pdf
17. G. K. Kumar, D. M. Rani, N. Koppula and S. Ashraf, "Prediction of House Price Using Machine Learning Algorithms," 2021 5th International Conference on Trends in Electronics and Informatics (ICOEI), Tirunelveli, India, 2021, pp. 1268-1271. Retrieved From: https://doi.org/10.1109/ICOEI51242.2021.9452820.
18. A. Varma, A. Sarma, S. Doshi and R. Nair, "House Price Prediction Using Machine Learning and Neural Networks," 2018 Second International Conference on Inventive Communication and Computational Technologies (ICICCT), Coimbatore, India, 2018, pp.1936-1939. Retrieved From:https://doi.org/10.1109/ICICCT.2018.8473231.
19. M. Mahyoub, A. A. Ataby, Y. Upadhyay and J. Mustafina, "AIRBNB Price Prediction Using Machine Learning," 2023 15th International Conference on Developments in eSystems Engineering (DeSE), Baghdad & Anbar, Iraq, 2023, pp. 166-171. Retrieved From: https://doi.org/10.1109/DeSE58274.2023.10099909.
20. A. Lektorov, E. Abdelfattah and S. Joshi, "Airbnb Rental Price Prediction Using Machine Learning Models," 2023 IEEE 13th Annual Computing and Communication Workshop and Conference (CCWC), Las Vegas, NV, USA, 2023, pp. 0339-0344. Retrieved From: https://doi.org/10.1109/CCWC57344.2023.10099266.
21. Tsiaperas, T. (2023, August 7). How Dallas-Fort Worth became America's boomtown. Axios. https://www.axios.com/2023/08/07/dallas-fort-worth-america-boomtown#
22. Polaris Market Research. (2023, June). *Real estate crowdfunding market size, Share Global Analysis Report, 2023-2032*. Polaris. https://www.polarismarketresearch.com/industry-analysis/real-estate-crowdfunding-market
23. Guest Contributors, G. (2023, July 27). *Breaking barriers: How fintech is Democratizing Real Estate Investing*. Nasdaq. https://www.nasdaq.com/articles/breaking-barriers:-how-fintech-is-democratizing-real-estate-investing.
24. McGown, J. (2021, July 29). Why the Dallas Moving Surge Is Here to Stay. D Magazine. Retrieved From: https://www.dmagazine.com/commercial-real-estate/2021/07/why-the-dallas-moving-surge-is-here-to-stay/

25. Google. (2024). Geographic map of Plano, Texas [Map]. Retrieved from https://www.google.com/maps/place/Plano,+TX/@33.061262,-96.7366254,12z/data=!3m1!4b1!4m6!3m5!1s0x864c21da13c59513:0x62aa036489cd602b!8m2!3d33.0198431!4d-96.6988856!16zL20vMDEzbTR2?entry=ttu
26. Google. (2024). Geographic map of Frisco, Texas [Map]. Retrieved from https://www.google.com/maps/place/Frisco,+TX/@33.1502712,-96.9929477,11z/data=!3m1!4b1!4m6!3m5!1s0x864c3c1abc09943d:0xcad371b035c8fea2!8m2!3d33.1506744!4d-96.8236116!16zL20vMDEzbTNy?entry=ttu
27. Demographics | Frisco Economic Development Corporation. (n.d.). Friscoedc.com.https://friscoedc.com/data-and-research/demographics
28. Best Cities for Jobs in 2024. (n.d.). WalletHub. Retrieved February 10, 2024, from https://wallethub.com/edu/best-cities-for-jobs/2173
29. James, G., Witten, D., Hastie, T., Tibshirani, R., & Taylor, J. (2023). An Introduction to Statistical Learning. Springer Nature.
30. Ramsey, F. L., & Schafer, D. W. (2013). The statistical sleuth: a course in methods of data analysis. Brooks/Cole, Cengage Learning.
31. Holmes, Jan. 2023. Comparative Market Analysis. Ebby Halliday, Realtors.