House Price Change Predictions with Multiple Variables

Xiaoyan Zhang, Michelle Kim, Taylor Crockett EPPS 6323, Dr. Karl Ho

In recent years, the real estate market in Frisco, TX has experienced substantial growth and fluctuation. Accurate house price predictions are critical for both homebuyers and investors in this burgeoning area. This study aims to develop a house price prediction model trained on data from zip codes in and around Frisco using linear regression, random forest, and gradient boosting tree models. Data was collected from Redfin for selected zip codes and integrated with Housing Price Index (HPI) data from Federal Reserve Economic Data (FRED). Substantial data preprocessing, including cleaning and handling missing values, was performed to ensure data quality. The dataset was organized by month and merged with HPI data, accounting for HPI lag. Several predictor variables were employed, such as the number of beds and baths, square footage, lagged HPI data, house age and log-transformed lot size. Compared to other tested models, the gradient boosting tree model emerged as the best predictor based on the least square distance metric. The study explores the potential of machine learning techniques in the real estate market for predicting house prices, with implications for homebuyers and investors in decision-making processes. In rapidly growing areas characterized by significant increases in property values, these techniques may prove particularly valuable for forecasting trends and identifying investment opportunities. Further research and validation of the results may enhance the model's performance and generalizability to other high-growth geographic areas.

1 Introduction

As the real estate market in the United States continues to experience unprecedented growth and volatility, accurate house price prediction has become an increasingly crucial tool for various stakeholders, including homebuyers, investors, policymakers, real estate developers, and financial institutions. Frisco, Texas, a rapidly growing city in the Dallas-Fort Worth metroplex, exemplifies this trend. The city's remarkable growth is driven by several factors, such as the relocation of numerous corporate headquarters, including the Dallas Cowboys' World Headquarters, Toyota, and PGA of America, a strong local economy, and an outstanding public school system that consistently ranks among the best in the nation. Additionally, the city boasts world-class sports venues, retail and entertainment options, and a low cost of living relative to other major metropolitan areas.

Frisco's emergence as a secondary business district within the Dallas-Fort Worth metroplex has further contributed to its growth and prominence in the region. The city's strategic location and pro-business environment have attracted a diverse range of industries, reinforcing its role as a thriving economic center that complements the primary business hubs in the metroplex. This rapid expansion has led to an increased demand for housing, which has driven up property values, making accurate house price predictions all the more essential for various stakeholders.

Statewide, Texas has seen a surge in attractiveness for real estate investments, thanks to its business-friendly environment, low taxes, and affordable housing. These factors, coupled with a strong economy and job market, have made Texas an appealing destination for both domestic and international migrants, many of whom have made a home in Frisco. However, the recent pandemic-induced recession and the subsequent post-pandemic boom have created a volatile environment for the United States economy. This has led to a period of high inflation and the potential for increased interest

rates, which could have significant implications for the real estate market in the near future.

In the face of rapidly evolving real estate landscapes, such as the one in Frisco, and the broader economic context in the United States, it is crucial to develop accurate house price prediction models that can help guide decision-making for a wide range of stakeholders. The primary goal of this study is to address the challenge of predicting house prices in Frisco, TX, using machine learning techniques based on Housing Price Index (HPI) data. By focusing on Frisco, this research will contribute to the understanding of how machine learning can be employed to predict real estate prices in high-growth areas and under varying economic conditions. The results of this study will not only provide valuable insights for the Frisco real estate market but also contribute to the broader knowledge of applying machine learning techniques in real estate market analysis, particularly in the context of rapidly changing economic environments.

The primary objectives of this study are to develop house price prediction models using three different machine learning techniques and evaluate their levels of performance. By doing so, the study aims to answer the following questions: What are the key features influencing real estate value in high-growth suburban areas? How well do different machine learning techniques perform when predicting house prices? And finally, how can the selected model be used to forecast future house price trends and inform decision-making processes for various stakeholders?

The scope of this study is limited to the Frisco, TX real estate market and some adjacent zip codes. While the study utilizes a comprehensive set of predictor variables, the selection of variables may still not cover all possible factors influencing house prices in the study area.

Additionally, the seasonality index may be overrepresented due to the rapid increase in house prices, potentially leading to biased estimates of the seasonality effects given that it was often the highest-weighed variable in variable importance. The use of decomposition methods, such as seasonal decomposition of time series (STL), could be worth evaluating in the future to separate the seasonal component from the trend component and more accurately estimate the seasonality effects. However when the seasonality index was removed as a variable in our models, the out-of-bag (OOB) and mean squared error (MSE) changed to a nearly negligible degree, suggesting that further investigation is needed.

Furthermore, the HPI as sourced from FRED is for the greater West South Central census division and does not specifically represent Frisco. Potential omitted variables include neighborhood characteristics, proximity to amenities, local employment, local school quality, crime rates, and transportation infrastructure. These variables were omitted due to lack of time and data, however, they should be included in future studies given more resources.

Additionally, the machine learning techniques chosenlinear regression, random forests, and gradient boosting treesdo not encompass all possible modeling approaches suitable for house price prediction, but rather represent a selection of popular and widely used methods in the field. Other potential models that were not considered in this study include support vector machines, neural networks, and Bayesian regressions.

The study is organized after the introduction in the following order: a literature review, methods of data collection, methods of data analysis, modeling results, and a conclusion of the study, also followed by an appendix of code and references.

2 Literature Review

The prediction of house prices in booming cities has increasingly become an area of interest in both academic research and practical applications, as rapid urbanization and unaffordability present challenges for potential homeowners and investors Brown (2022) and Connelly (2022). The rising demand for housing, coupled with uncertain economic conditions and escalating prices,

has led to the need for more accurate and reliable prediction models that can inform decision-making processes Wile (2023).

Existing studies have explored the application of machine learning techniques for house price prediction, taking into account various factors such as location, size, amenities, and seasonality, among others Vaddi et al. (2022) and Adetunji et al. (2022). For instance, Vaddi et al. (2022) investigated the use of machine learning in predicting house prices based on estate attributes and visual images, employing a "a multi-kernel regression approach...to predict the house price from both visual cues and estate attributes" that demonstrated superior performance compared to baseline methods. Adetunji et al. (2022) explored the use of the random forest techniques on Boston housing data from the UCI Machine Learning repository to evaluate the performance of their proposed model.

Despite the growing body of literature on machine learning and house price prediction, there is still a need for research that focuses specifically on rapidly booming cities where unaffordability and uncertainty are prevalent. Collin County, for example, has emerged as the most expensive housing market in North Texas, with the median home sales price reaching $460,000, a staggering 26.4% increase from the previous year Brown (2022). In addition, the affordability of homes in the area has been on a steady decline, with middle-income families finding it increasingly difficult to purchase homes Connelly (2022). This underscores the importance of developing accurate prediction models that can help potential homebuyers and investors navigate the complex real estate landscape in booming cities. In light of these challenges, our study aims to contribute to the existing literature by offering a performance evaluation of not just one, but various machine learning techniques, including linear regression, random forests, and gradient boosting trees, in predicting house prices.

3 Data Collection

To predict sale prices in Frisco, we conducted a comprehensive data collection process that encompassed two main sources. Our primary per-sale listing dataset was obtained from Redfin, a leading online multiple listing service (MLS), which provided city house prices for Frisco from March 2021 through March 2023. This dataset consists of 3,048 entries and includes the following variables: sold month, property type, address, city, zip code, price, beds, baths, location, square footage, lot sizes, year built, price per square foot, latitude, longitude, and MLS identifier. We defined the house price as our dependent variable.

It is important to note that the sold month variable in the dataset is recorded as the first date of each month. Additionally, all properties in the dataset are single-family residential homes. Due to consequences of casting a wide net during web scraping, the city designation for some houses may be listed as either Frisco or its adjacent cities Hackberry, The Colony, and Little Elm. During the data cleaning process, we removed any entries with missing values in the price variable to ensure the accuracy and reliability of our analysis.

In addition to the Redfin dataset, we obtained housing price index (HPI) data from Federal Reserve Economic Data (FRED), an online compilation of data publications by the U.S. Federal Housing Finance Agency. The HPI is a measure of the change in single-family house prices over time, providing valuable context for our study. Unfortunately, we were unable to access the HPI data specific to Frisco, so we used the available data for the West South Central census division, which encompasses the region that includes Frisco. The HPI data spans from January 2000 through September 2022, with dates recorded as the first day of each month.

Before creating our regression models, we performed further data cleaning by removing variables that were not relevant to our analysis, such as longitude, latitude, house association, and MLS identifier. This streamlined dataset allowed us to focus on the most pertinent factors for predicting house prices in Frisco.

Month sold	City	ZIP	Price ($)	Beds	Baths	Sq. ft.	Lot size	Year blt.	Pr./sq. ft.	MLS	Lat.	Long.
11/1/2022	Carr.	75007	375,000	3	2	1821	8059	1974	206	201…	33…	-96…
4/1/2022	Carr.	75007	389,900	4	2	1706	7841	1979	229	200…	33…	-96…
9/1/2022	Carr.	75007	608,000	3	3	2721	9191	1986	223	201…	32…	-96…
9/1/2022	Carr.	75007	410,000	3	2	1822	7492	1984	225	201…	32…	-96…
8/1/2022	Carr.	75007	598,000	5	3.5	3141	9845	1989	190	200…	33…	-96…
5/1/2022	Carr.	75007	400,000	3	2	1818	8756	1974	220	200…	33…	-96…
5/1/2022	Carr.	75007	325,000	3	2	1506	7928	1974	216	200…	33…	-96…
12/1/2022	Carr.	75007	420,000	3	2.5	2028	10454	1985	207	202…	33…	-96…
3/1/2022	Carr.	75007	395,000	3	2.5	1996	4704	2004	198	200…	33…	-96…
5/1/2022	Carr.	75007	375,000	3	2	1883	9191	1984	199	200…	33…	-96…

Table 1: An example of collected Redfin data

Date	HPI
1/1/2021	210.81409
2/1/2021	214.37410
3/1/2021	220.39817
4/1/2021	226.77821
5/1/2021	233.07686
6/1/2021	240.13970

Table 2: An example of collected HPI data

4 Data Analysis

In order to develop an accurate and reliable model, it is crucial to thoroughly understand the data and its variables. To ensure the validity of our analysis, we examined the quality and relationships between variables through a series of data analysis techniques as follows: frequency analysis, cross tables, histograms, trend and scatter plots, summary statistics, handling of missing values, imputation, and outlier identification. Following this, we carefully derived new variables to promote better modeling while ensuring that it would not promote collinearity to an excessive degree.

4.1 Frequency analysis

Frequency analysis is an essential starting point for understanding variables, particularly categorical ones. We assessed the frequency distribution of houses by zip code, number of baths, and number of bedrooms.

Table 3 indicates that the majority of houses in the dataset are located within zip codes 75033, 75034, 75035, and 76036, with zip code 75035 having the highest number of houses. Houses not in those zip codes are likely products of the wide net cast during the data collection: we chose not to exclude those houses because they will not negatively affect the analysis and modeling as a whole. Table 4 shows that most houses have 2 to 4¡ bathrooms, while very few have more than 6 or less than 2.

4.2 Cross table

Cross tables examine the relationships between paired variables, particularly categorical or binned variables. This analysis determines the correlation between variables and their reasonableness in relation to one another, which in turn facilitates variable selection during the modeling process. Exemplified below are cross tables for number of beds by zip code, number of baths by zip code, and number of beds by number of baths.

4.3 Visualizations

Visualizations, such as histograms, trend plots, and scatter plots, allow us to observe the distribution and relationships between variables. For instance, histograms reveal the distribution of variables like listing price, living square feet, and lot size. Trend plots illustrate variable values over time, such as price by month or by bedroom. Scatter plots, on the other hand, depict the correlation between two variables, such as price and living square feet.

75007	75033	75034	75035	75036	75056	75068	75071	75072	75078	78373
313	727	450	1,149	559	9	90	4	2	5	1

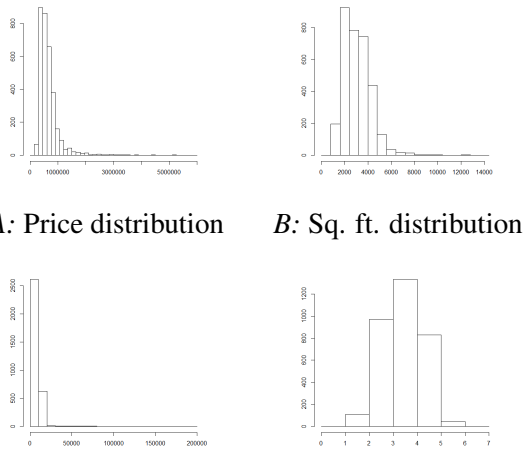
Table 3: Count of houses by ZIP

1	1.5	2	2.5	3	3.5	4	4.5	5	5.5	6	6.5	7.5	9	10
5	4	921	491	364	579	428	299	67	95	40	12	2	1	1

Table 4: Count of houses by num. baths

	Beds						
ZIP	1	2	3	4	5	6	7
75007	0	9	195	94	15	0	0
75033	0	5	125	338	250	9	0
75034	0	5	123	150	158	13	1
75035	0	1	355	472	301	20	0
75036	1	86	125	239	107	1	0
75056	0	1	2	6	0	0	0
75068	0	0	47	36	6	1	0
75071	0	0	1	3	0	0	0
75072	0	0	1	1	0	0	0
75078	0	0	0	3	2	0	0
78373	0	0	0	0	1	0	0

Table 5: Cross table of beds against ZIP



A: Price distribution *B:* Sq. ft. distribution

C: Lot size distribution *D:* Beds distribution

Figure 1: Histograms of variable distribution. Both the house price and living square feet histograms are right-skewed, suggesting the need for log transformations, particularly for linear regression modeling results.

4.4 Summary statistics

Similar to a histogram, summary statistics provide exact values of the variable distribution, particularly for numeric variables. In this case, quantile statistics determine percentiles at 0%, 1%, 10%, 25%, 50%, 75%, 90%, 95%, 99%, and 100%, as well as the minimum (0% percentile) and maximum (100% percentile) values, mean, and median (50% percentile) of the variable. The percentile tables are shown below.

4.5 Missing values and outliers

Addressing missing values and outliers is crucial for accurate modeling. Different models require distinct treatments for missing values and outliers. For instance, linear regression cannot accommodate missing values and is highly sensitive to outliers. Conversely, gradient boosting trees can handle missing values without additional treatment. In the most recent dataset used, the target variable price has one missing value, which necessitates the removal of that observation. Additionally, when identified, outliers that are clear symptoms of human input error are adjusted by elimination or imputation, for instance imputing the 290,806,560 square-foot entry as 30,000 square feet (approximately the area of a large mansion).

4.6 Deriving new variables

Derived variables can offer valuable insights not readily available from the original variables in a dataset, such as hidden relationships and trends that may significantly impact the studys results. We introduced several derived variables to facilitate better modeling, some of which we deemed crucial: age of house, total beds and baths, beds-to-baths ratio, living square footage-

Table 6: Quantiles of house price variables

Price									
quantile((house_df$PRICE), c(0, 0.01, 0.1...0.95, 0.99, 1))									
0%	1%	10%	25%	50%	75%	90%	95%	99%	100%
189900	285000	365000	435000	575000	750000	975000	1229250	2099153	5250000
Beds									
quantile((house_df$BEDS), c(0, 0.01, 0.1...0.95, 0.99, 1))									
0%	1%	10%	25%	50%	75%	90%	95%	99%	100%
2	2	3	3	4	5	5	5	6	7
Baths									
quantile((house_df$BATHS), c(0, 0.01, 0.1...0.95, 0.99, 1))									
0%	1%	10%	25%	50%	75%	90%	95%	99%	100%
1	2	2	2	3	4	4.5	5	6	10
Lot Size									
quantile((house_df$LOTSIZE), c(0, 0.01, 0.1...0.95, 0.99, 1))									
0%	1%	10%	25%	50%	75%	90%	95%	99%	100%
1	3290	5489	6534	7841	9409	11848	13852	20502	17246

to-beds ratio, and a seasonality index by month. For instance, the age of the house is nearly always a significant driver of the price, but only years built and sold are represented on Redfin. Furthermore, the month sold can capture seasonality patterns in the housing market, which often affect supply and demand dynamics of house prices.

Another vital derived variable not listed above is the lagged HPI measure. It is essential when using HPI values in a dataset to consider that HPI measures are known to lag behind real market trends. To account for this lagging effect, we introduced 1-, 2-, 3-, and 6-month lagged HPI variables, as well as HPI percent change for the same lag intervals.

5 Modeling

To prepare the modeling process, we partitioned the dataset using a stratified sampling approach, allocating 70% of entries for training and the remaining 30% for testing. The following code illustrates the data splitting procedure.

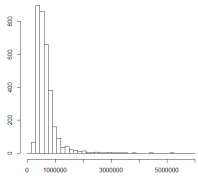
```
library(rsample)
set.seed(123)
initial_split_df <- initial_split(df,
    ↪ prop = 0.7)
train <- training(initial_split_df)
test <- testing(initial_split_df)
```

This stratification ensures that the distribution of the target variable, house prices, was similar across both the training and testing sets, providing a representative sample of the data. The training set was used to develop and calibrate the three models used, namely linear regression, random forest, and gradient boosting tree models, while the testing set served as an unbiased evaluation of each models performance.

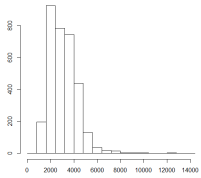
Each model considered log-transformed price as its dependent variable because log transformation helps stabilize the variance, reduce the impact of outliers, and better satisfy the assumption of normality in the distribution of errors. This, in turn, leads to more accurate and reliable predictions, especially when dealing with highly skewed data such as house prices.

5.1 Models used

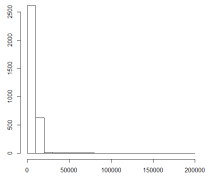
The linear regression model is a popular statistical technique that assumes a linear relationship between predictor and response variables. Its advantages are simplic-



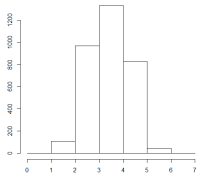
A: Price distribution



B: Sq. ft. distribution



C: Lot size distribution



D: Beds distribution

Figure 2: The relationship between price and various variables. Trend plots *A* and *B* reveal outliers within the data at the extremes, and *C* indicates an upward trend with seasonal fluctuations. Scatter plot *D* demonstrates a positive correlation between the size of the house and its price.

ity, computational efficiency, and interpretability of coefficients. However, it is limited by its inability to handle complex non-linear relationships, interactions between features, and outliers.

The random forest model combines decision trees to make a more accurate model. It is less disrupted by outliers and noise, and models non-linear relationships and feature interactions. Estimates of feature importance, also called variable importance, can identify the most relevant variables for predicting house prices. However, compared to linear regression, the random forest model is more complex and computationally expensive. There is also a risk of overfitting, especially with small datasets or when hyperparameters are not properly tuned; to train a properly tuned random forest model can prove extremely time-consuming.

Lastly, the gradient boosting tree model is a machine learning algorithm that adds decision trees in an iterative manner to minimize residual errors. It is flexible in accommodating different loss functions and has a reputation of high predictivity and accuracy, modeling non-linear relationships and interactions between features. However, similar to random forest modeling, gradient boosting trees can prove complex, computationally expensive, and prone to overfitting if hyperparameters like the learning rate or number of boosting iterations are not carefully chosen.

5.1.1 Linear regression

The code below demonstrates the process of fitting the linear regression model on the training data in R:

```
lreg_model <- lm(log_price ~
                 BATHS +
                 SQUAREFEET +
                 hpi_lag3 +
                 hpi_3m_pct +
                 houseage +
                 LOTSIZE_imputed_ind +
                 log_lotsize,
                 data = train)
```

However, during the model development process, some predictor variables were excluded from the final model due to concerns about multicollinearity, overfitting, or negative contribution to the models predictive performance. The necessary exclusion of these variables reflects the limitation of linear regression models to capture complex relationships between predictors and target variables, especially as the number of relationships increase.

To assess the performance of the final linear regression model, we calculated the mean squared error (MSE) on the testing strata resulting in an MSE of 0.0286. MSE and model comparison will be discussed further in the model comparison subsection.

5.1.2 Random forest

To implement the random forest model in R, we utilized the "randomForest" package. First, we optimized the model using the "tuneRF" function, which tunes the number of variables at each split and minimizes the out-of-bag (OOB) error rate. The OOB error rate is the rate at which the model is incorrect when predicting new data points that were not used in the training process.

Specifically, tuneRF focuses on the "mtry" hyperparameter, which determines the number of variables tested at each branch. Our analysis indicated that an "mtry" value of 7 resulted in the lowest OOB error rate.

We conducted a dry run to build an initial random forest model using $mtry = 7$ and starting with 1,000 random trees. A larger number of trees usually results in improved model accuracy and reduced overfitting, but it also increases computational time and resource consumption. The 1,000-tree model took approximately 30 seconds to produce on the local system. A plot of the OOB error rate against the number of trees constructed showed that it reached a general minimum around 500 trees, which was the value ultimately chosen for the final random forest model. After tuning and training, this model produced an MSE of 0.0204 from its predictions. However, a variable importance measure showed that the seasonality index was most important, indicating that the random forest model may be subject to overrepresentation from that variable without trend decomposition.

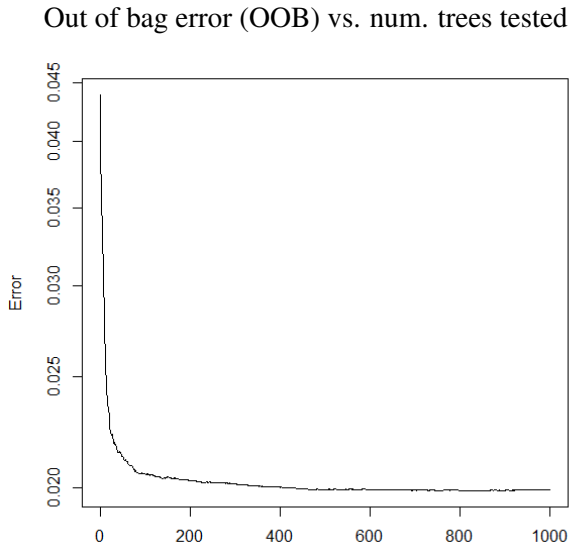


Figure 3: A random forest model with $mtry = 7$ and $nTree_0 = 1000$ shows stabilization at $nTree = 500$

5.1.3 Gradient boosting tree

In order to implement the gradient boosting tree model, we employed the gbm package in R. To optimize the models performance, we conducted a grid search for the hyperparameters, including number of trees, learning rate, and maximum tree depth. That initial analysis concluded that a combination of 200 trees, a learning rate of 0.05, and a maximum tree depth of 10 provided the best balance between model accuracy and computational efficiency.

Using the optimal hyperparameters, the gradient boosting tree model clearly showed that HPI lagged by 6 months was the most important feature, followed distantly by the number of beds and HPI lagged by 3 months. Subsequently, we further explored different hyperparameter combinations, evaluating their performance graphically. Based on this analysis, we ultimately chose a final model with 60 trees ($nTree = 60$), a maximum tree depth of 10 ($max_depth = 10$), and a learning rate of 0.1. This refined model demonstrated even better performance while maintaining computational efficiency.

Based on the tuning graph above we choose $nTree = 60$, $max_depth = 10$, and a learning rate of 0.1. The final training code with parameters is below:

```
best_model <- gbm(log_price ~ ., data =
    ↪ train_data, n.trees = 60,
                interaction.depth = 10,
                    ↪ shrinkage = 0.1,n.
                    ↪ minobsinnode = 5,
                distribution = "gaussian")
best_pred <- predict(best_model, newdata
    ↪  = test_data)
MSE.gbm = mean((test_data$log_price -
    ↪ best_pred)^2)
```

Upon completion of the model training, the gradient boosting tree model yielded a mean squared error (MSE) of 0.019315 on the test dataset.

5.2 Model comparison

Each of these models has its own advantages and limitations, which were considered before and after evaluating their performance. Outside of qualitative pros and cons, predictive accuracy of each model was quantified using the mean squared error (MSE) computed on the

testing dataset.

MSE is a reliable measure of model performance, as it calculates the average squared difference between the predicted and actual values, essentially representing the distance between predicted and actual values. Lower MSE values indicate better predictive accuracy, and the metric is particularly useful for comparing different models in terms of their ability to minimize prediction errors.

The linear regression model, with its simplicity, computational efficiency, and interpretability of coefficients, yielded an MSE of 0.0286. However, this model was limited in its ability to capture complex relationships between predictor variables and the target variable, as demonstrated by the exclusion of some variables during the model development process.

The random forest model, which is more robust and can handle non-linear relationships, produced an MSE of 0.0204 after tuning and training. Despite its improved accuracy over the linear regression model, the variable importance measure indicated that the seasonality index was the most important feature, suggesting potential overrepresentation of this variable without trend decomposition.

Finally, the gradient boosting tree model, which is known for its flexibility and high performance, was initially trained using 200 trees, a learning rate of 0.05, and a maximum tree depth of 10. Further hyperparameter tuning led to the selection of a final model with 60 trees, a maximum tree depth of 10, and a learning rate of 0.1. This refined model achieved the lowest MSE of 0.019315, indicating superior predictive accuracy among the three models.

In summary, the gradient boosting tree model is chosen as the best model for predicting house prices because it demonstrates the highest predictive accuracy without sharing the same limitations as the other models, such as inability to capture complex relationships (linear regression) or overrepresentation of certain variables (random forest). While the random forest and linear regression models may be more appropriate in situations where interpretability or computational efficiency are of higher importance, the gradient boosting tree model offers a more reliable prediction with its superior performance in terms of minimizing MSE on the test dataset.

5.3 Market trends

The historical analysis of house prices, as depicted in the figure below, reveals a general upward trend over time. Notably, a pronounced increase in prices is observed between November 2021 and March 2022. The final month's forecast from our gradient boosting tree model, which exhibited the best predictive performance among the evaluated models, suggests that house prices will continue to rise despite the concurrent increase in interest rates and persistent high inflation.

In the figure, the green line represents the monthly percentage change in house prices. A striking peak of nearly 20% is evident around November 2021. The majority of the monthly price percentage changes are positive, indicating a general appreciation in house values over time. This observation aligns with the well-established economic principle that housing prices tend to increase in the long run, influenced by factors such as population growth, economic development, and limited land supply in urban areas.

6 Conclusion

In conclusion, this study aimed to investigate the performance of three predictive modeling techniqueslinear regression, random forest, and gradient boosting treein forecasting house prices in booming, quickly growing middle-class areas such as Frisco, TX. The models were developed using a dataset incorporating a variety of relevant features and derived variables, such as the number of bedrooms and bathrooms, square footage, house age, historical price indices, bath-to-bed ratio, and log-transformed price and lot size. The gradient boosting tree model demonstrated the best performance among the evaluated models, as evidenced by the lowest mean

squared error (MSE). This superior performance can be attributed to the model's ability to capture complex nonlinear relationships and interactions between features, which is not possible with simpler models like linear regression.

The findings of this study have important implications for homebuyers, investors, and policymakers, as they provide valuable insights into the dynamics of the housing market in rapidly growing areas. The predictive performance of the gradient boosting tree model suggests that this technique can be effectively utilized to forecast house prices and inform decision-making processes for various stakeholders. However, it is important to note that the choice of the best model depends on the specific problem, dataset characteristics, and the desired balance between interpretability, complexity, and predictive performance. Moreover, limitations related to zip code selection, predictor variables, and lack of result validation may impact the generalizability of the findings. Future research should aim to validate these findings, anticipate the incorporation of future data to expand the modeling over time, and explore additional features, alternative modeling techniques, and larger datasets to further enhance the accuracy and generalizability of the predictive models. Furthermore, studies investigating the underlying factors driving housing market trends can help in understanding the market dynamics and devising effective policies to promote housing affordability and market stability in high-growth geographic areas.

7 Synergy Report

This project consisted of three segments: ideation, proposal, and final presentation with a written paper. Each team member worked hard in all three segments, contributing to the theoretical background, coding, analysis, and writing.

Michelle contributed to the literature review, description of data collection, and linear regression model and analysis.

Taylor contributed to the abstract, introduction, conclusion, random forest model and analysis, and editing of the project paper.

Xiaoyan contributed to the data collection, data analysis, gradient boosting tree model, model comparison, and market trend analysis.

References

Adetunji, Abigail Bola et al. (2022). "House price prediction using random forest machine learning technique". en. In: *Procedia Comput. Sci.* 199, pp. 806–813.

Brown, Steve (Jan. 2022). "Collin County home prices outpace North Texas with gains". en. In: *The Dallas Morning News*.

Connelly, Christopher (Aug. 2022). "Collin County is DFW's least affordable area for homebuyers. Denton County isn't far behind". en. In: *KERA*.

Vaddi, Sai Surya et al. (2022). "House price prediction via visual cues and estate attributes". In: *Advances in Visual Computing*. Cham: Springer Nature Switzerland, pp. 91–103.

Wile, Rob (Jan. 2023). "Inflation cooled in December to 6.5%, but the Fed is likely to keep interest rates high". en. In: *NBC News*.

Appendicized Code

Appendix content