Housing prices prediction model

Raymond Ali  
GENG-4500 - 60  
University of WindsorWindsor, Canada  
ali5t@uwindsor.caPat Maynard  
GENG-4500 - 60  
University of WindsorWindsor, Ontario  
maynardp@uwindsor.ca

*Abstract*—The housing market is one filled with variables, uncertainty, and unpredictability. There are many factors at play that can cause this instability with major ones being things such as recessions, socio-political events, and even pandemics. Despite all of this, there is a general pattern to be deciphered. Random forest regression has been among the top choice of various Artificial Intelligence models used in attempts to predict housing prices. For this project, we have adopted a random forest regression model using collected data from reputable resources to predict housing prices here in Canada. Data was sourced from the Consumer Price Index, the House Price Index, the Multiple Listing Service, and Canadian Interest Rates in order to train the model. After implementing some adjustments and hyperparameter tuning, the model successfully outputted a graph showing monthly data points illustrating price changes over ten years, albeit with some scaling issues. After revisiting the training data, and adjusting the scaling there, we were able to correct this issue.

Keywords—Machine learning, Housing market, Random Forest Regression

# Introduction

The housing market in Canada is something that carries a lot of weight in the decision-making process of many, if not all Canadians looking to buy property. It has many variables to consider including such things as affordability, supply, investments, interest rates and/or mortgaging, various housing quality issues, and inequality issues. It can be quite an undertaking for first-time homeowners, or anyone who is looking to buy or sell. This project will take a deeper look at the Canadian housing market and formulate a strategic approach to buying or selling properties using machine learning implemented into an artificial intelligence model. A timeline between 2008 and 2023 will be used as the source of data to try and identify key points throughout the individual years, and the greater decade. Variables such as the CPI (Consumer Price Index), the HPI (House Price Index), the MLS (Multiple Listing Service), Interest Rates, and socio-political events such as the 2008 collapse, various recessions, and even COVID-19 are taken into consideration. The model selected is called Random Forest Regression. This model is a predictive technique that combines multiple decision trees to make accurate predictions by taking averages from each tree to form its conclusions. This model will be trained off of the previously mentioned variables to predict the next ten years of housing prices. Though not quite as in-depth as other, more professional studies done across the world, this project will look for patterns and clear up some of the guesswork when it comes to property pricing, hopefully allowing those in the housing market to make better-informed decisions. Random Forest was selected for its ability to learn from non-linear decisions, its high accuracy, its flexibility, and the importance it places on the features used. This decision was also informed by similar studies done in the past attempting to identify the best artificial intelligence model for house price predictions. As a part of this project’s proposal, a literature review was conducted to see how others may have attempted to use machine learning to solve similar problems. Included below is an excerpt from our proposal [7], showing our literature review.

# Literature Review

The problem that this project attempts to solve is the fact that the housing market is rather unpredictable. Many factors can affect this market including general economic conditions, interest rates, supply and demand, government regulations, and even global geopolitical influences [1]. Predicting the housing market can be quite challenging due to its complexity and the influence of numerous interconnected factors. Economic conditions, such as gross domestic product growth, employment rates, and interest rates, constantly fluctuate and can significantly influence housing demand and affordability [2]. Moreover, demographic trends, including population growth, household formation, and migration patterns, vary across regions and impact housing preferences and spatial distribution. Supply and demand dynamics further complicate predictions, as mismatches between housing supply and demand can lead to unpredictable market fluctuations. Government policies and regulations also play a crucial role, introducing additional uncertainties through changes in zoning laws, taxation, and mortgage lending standards [2]. Furthermore, the housing market is influenced by consumer sentiment, investor behavior, and external shocks such as geopolitical events or natural disasters, adding further layers of unpredictability. The intricate interplay of these factors creates a complex and dynamic housing market environment, making it difficult to accurately forecast future trends and outcomes.

For this project, we will attempt to use a random forest regression model to possibly predict housing prices based on past collected data. By utilizing an ensemble of decision trees trained on random subsets of the data, random forest regression can effectively model the nonlinearities and interactions between various factors influencing housing prices [3]. Features such as economic indicators, demographic trends, housing supply, and demand dynamics can be incorporated into the model, allowing it to capture the multidimensional nature of the housing market depending on the level of complexity needed. Additionally, the random forest algorithm's ability to handle large datasets and mitigate overfitting enhances its predictive performance [4]. By leveraging the collective results of multiple decision trees and aggregating their predictions, random forest regression provides robust and accurate estimates of housing prices, enabling people to make informed decisions in buying, selling, or investing in real estate.

There are several papers already published that attempt to use machine learning as a tool to help predict pricing outcomes. In the first article researched, by Rawool et al [3], the authors take note of the advancing nature of machine learning and it potential to be applied to the housing market. They apply various models such as linear regression, decision tree regression, k-means regression, and random forest regression [3]. They ultimately find that random forest regression provides the most accurate results. This research was incredibly thorough and detailed as they investigated multiple models to determine the best one. Their methods were well thought out and carefully considered. This paper convinced us that the random forest regression model is the best way to go for an introductory attempt at solving the unpredictability of the market. In the next article, Adetunji et al [4], the authors use an in-depth random forest regression model to predict better results than using just a house price index. This is because the HPI is a repeat sale index that tracks average price shifts in repeat transactions [4]. This makes the HPI unreliable as it is a rough predictor based on averages. Their random forest regression model used 506 data entries with 14 features for evaluation. Though their model is far more in-depth than ours, they proved that this is the ideal choice for predicting prices. For the article, Ake [5], the author takes on a rather study in what can be considered a combination of the two previous articles. He sets out to investigate a multitude of AI models to evaluate their application in predicting housing prices. He uses simple linear regression, decision tree regression, random forest regression, and ridge and lasso regression [5]. This is a very large and detailed research with many considerations taken. He ultimately finds that random forest regression produces the highest r-squared value, and the smallest root-mean-squared error value, and is, therefore, the most appropriate model to use for this type of application, proving the previous article correct. Finally, in the article, Truong et al [6], the authors once again research a multitude of AI models to determine which is most effective. They use three different types of machine learning methods and two techniques in machine learning [6]. They are random forest, XGBoost, LightGBM, hybrid regression, and stacked generalization regression. This is a highly detailed and very extensively researched paper that produced results that show the pros and cons of all methods analyzed. Random forest had the lowest error, while prone to overfitting. XGBoost and LightGBM both had decent accuracy, but the best time complexities. Hybrid regression performed extremely well due to its generalization. Stacked generalization though complicated in nature produced the most accurate results by far. This extensive research went far beyond the methods previously analyzed but showed that going farther can often change the perspective.

# Methodology

The methodology for this project roughly follows the standard seven principles for machine learning.

1. New Input data

After extensive iterations with a consolidated data set from various sources and only coming out with a thirty percent prediction accuracy, it was clear our input data needed to change. A new data set that outlined many more property types was available from the MLS database.

1. Preparing the data

The first step in preparing the data for processing is to convert the date strings into datetime objects that Python can recognize and work with. Another issue was the data set needed to be limited to the minimum length of all columns of data available.

After that, the rest of the data was separated based on whether it was an HPI or a raw dollar value. Beyond that, the features of each were broken down by the property type for further distinction.

The most important part of preparing the data for our final iteration of the process was to apply the scaler to it. We went with the robust scaler class offered in the SkLearn libraries. It lets our model adapt the incoming data to project a future value, thus solving our dollar value scale problems.

1. Training the model

During the training phase of the project, we took a different approach from previous iterations. We left the non-tuned model outputting what would have been produced alongside the hyperparameter-tuned model for further in-depth analysis of how the problem is viewed (Fig. 1) and how much the tuning had an impact on the results.

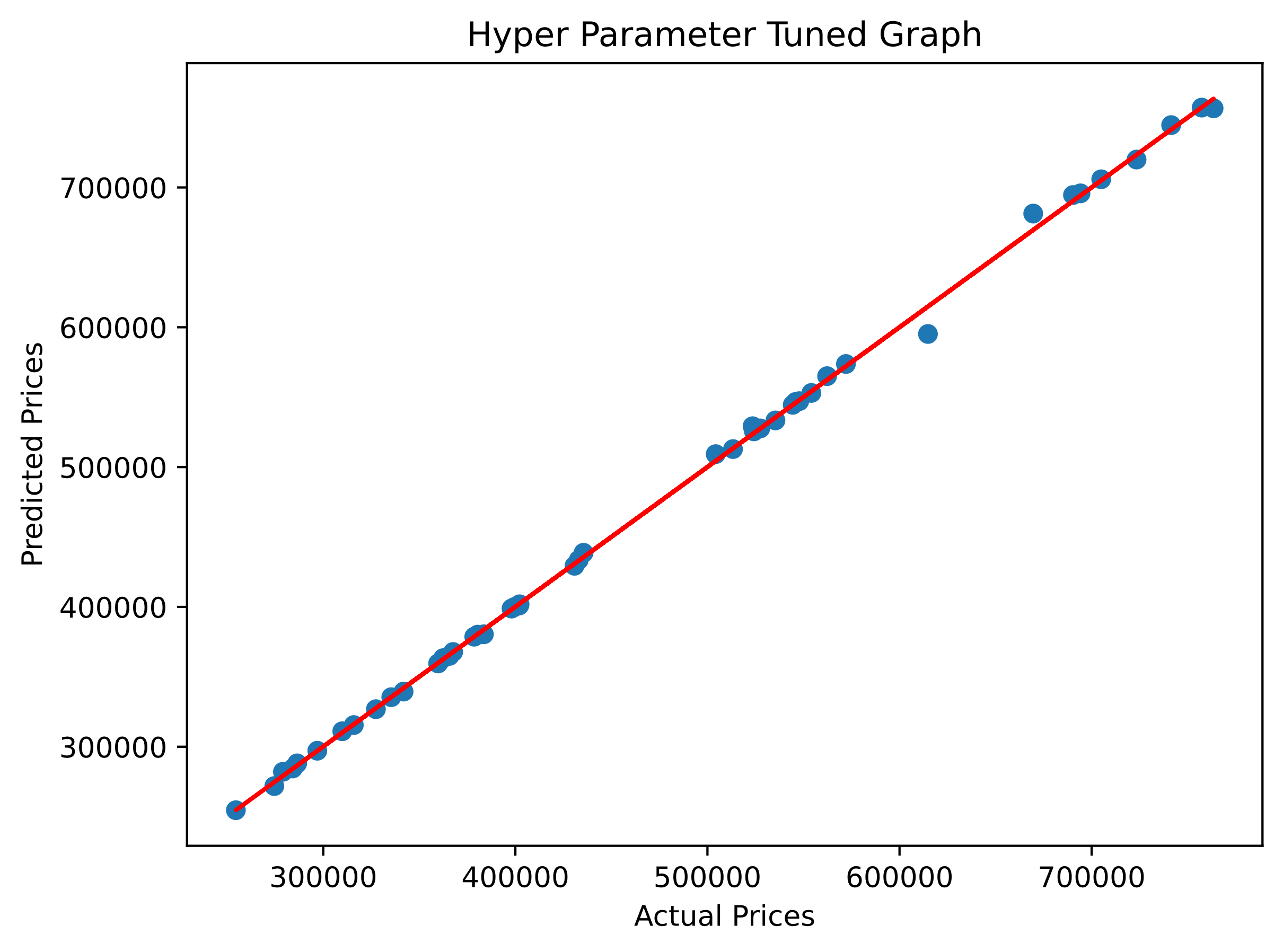


Fig. 1

1. Evaluation

Once the model was trained there was a noticeable difference in how the data looked from previous attempts. Most notably the model knew we were trying to find dollar values now and not just HPI points. With the dollar amounts scaling correctly, we noticed the accuracy has improved substantially from the initial attempts.

1. Hyperparameter Tuning

The biggest change in the model training was the use of hyperparameter tuning. The decision to use random search was made early on to try to increase the randomness of the inputs for the model to work with, thus giving us a more diverse approach to training.

Once we had our model fit originally, we instantiated a random search object to fit as well. From then on, we could see the two models making their predictions in parallel to one another for a clear picture of what was being achieved.

1. Prediction

The final predictions of the models would then be used in a newly instantiated random forest model to analyze the outputs further and hopefully give us a more accurate view of the problem. This new instance of the model would use the number of estimators selected by the previous random search model and the data set from the first random forest model.

Further, we would need to make a set of datetime objects for each month of the next ten years and then apply our data set to it for the model to finally give us a prediction of what it thinks the housing market will look like in the future.

1. Output of model

The final output of the culmination of the models was a graph showing monthly data points over the next ten years (Fig. 2 [2024-2034]).

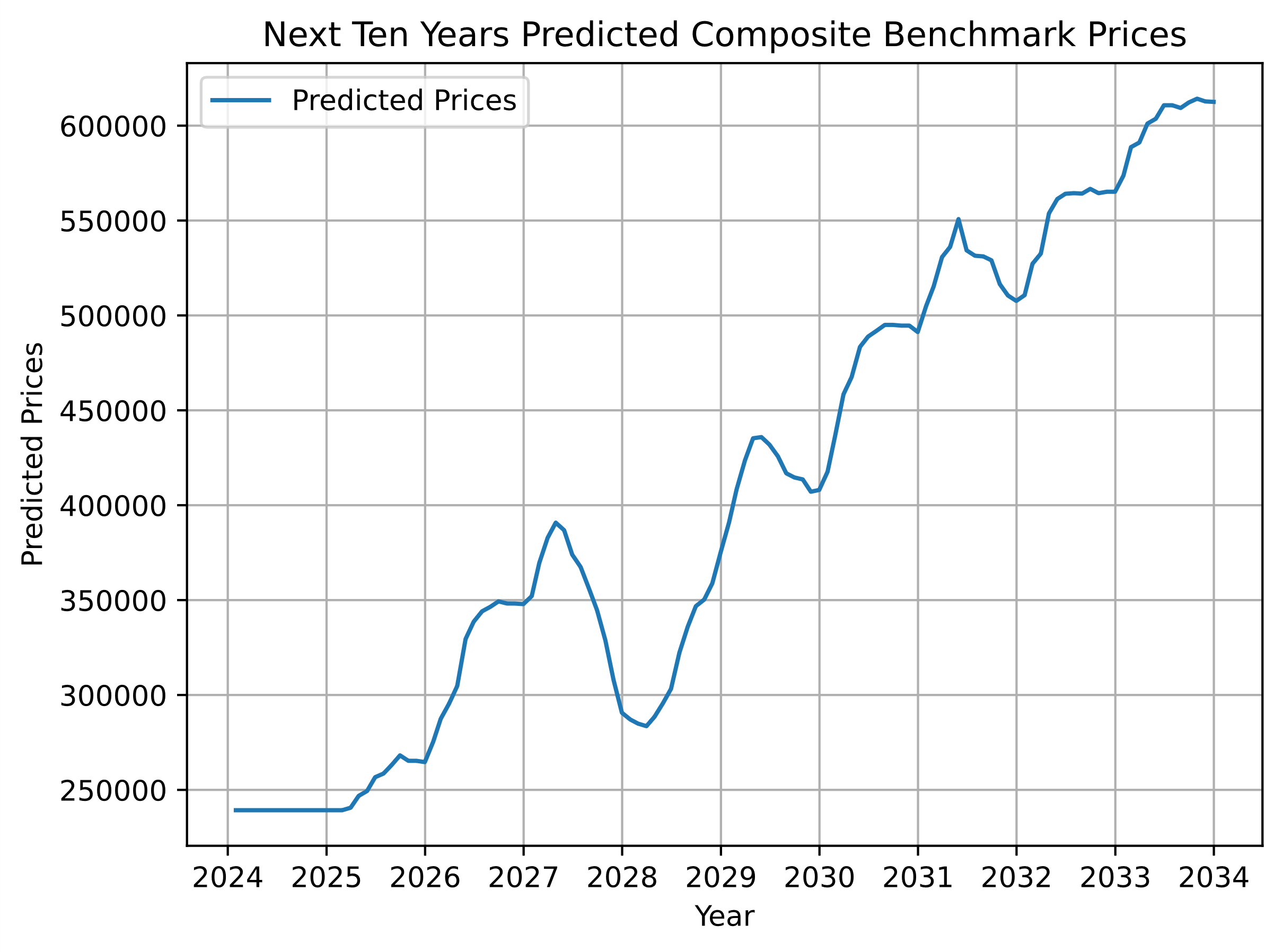


Fig. 2

# Results & Discussion

While disappointed in the scaling issues we had been faced with we pursued many alternatives to the functions, libraries, and tools taught in class. One of the most notable is SkTime, this is a library built almost exclusively for forecasting future data when you only have historical data sets and a time frame you want to explore. With more time and research into SkTime, we feel that the problems we have been facing with our model would be solvable promptly. However, this approach is outside the scope of this class and thus it would be like cheating in our opinion.

There was a significant amount of time spent trying to predict the data of each column in the future data time frame using various approaches. Unfortunately, the outputs were not accurate enough to make predictions (Fig. 3)

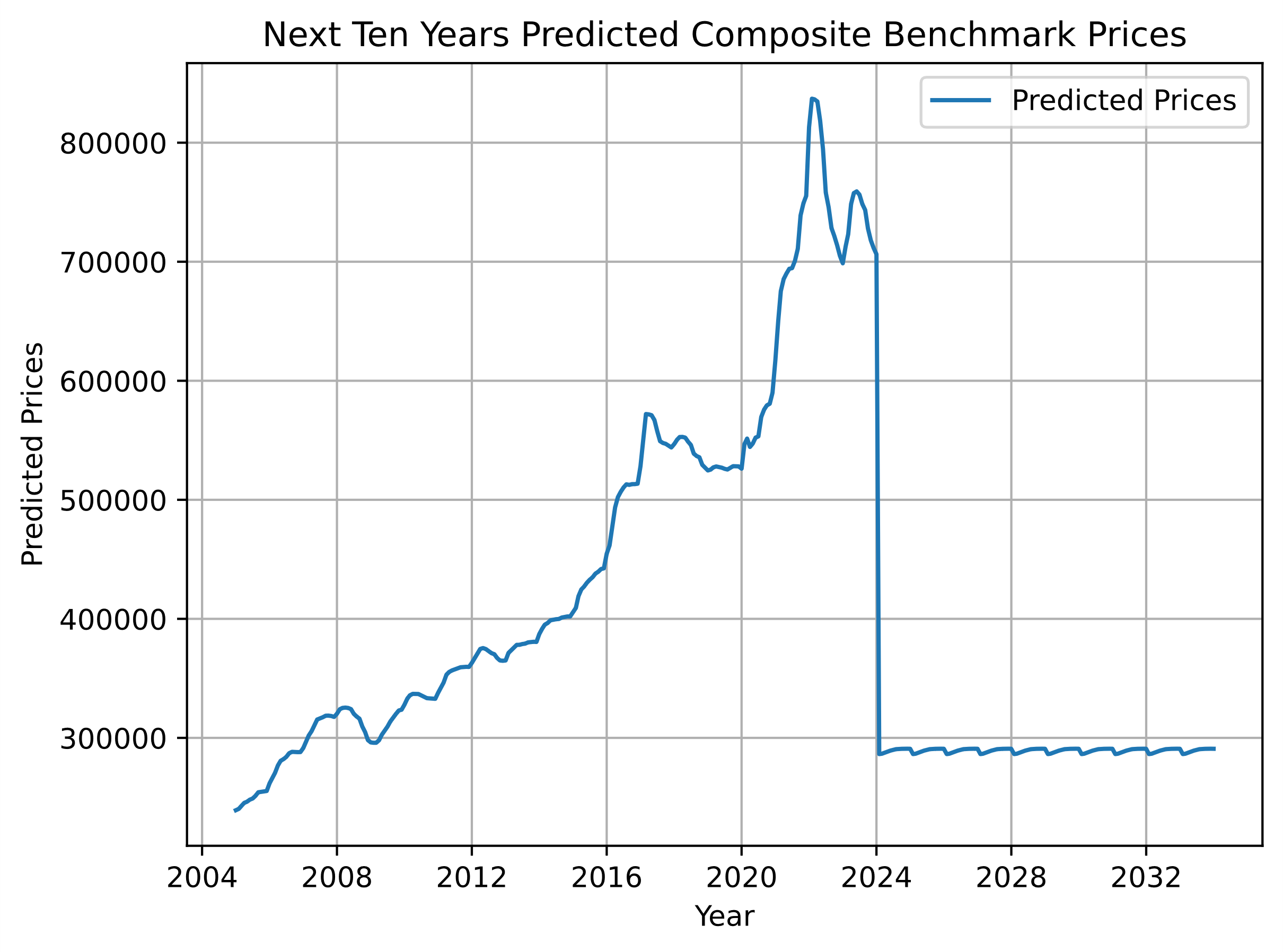


Fig. 3

Eventually, we reached out to Mohammad Hedayati, and he directed us to revisit scaling the initial data for training. After applying that change, we were able to get properly scaled numbers and trends for our final output.

# Conclusion

The conclusions drawn from this endeavor are that we needed to pay close attention to the data set used as the input to get proper predictions, the end-user would still need to interpret the data presented for errors.

Other things to note would be, if this data presented is correct a homeowner could sell their property for maximum profit in the spring of 2027 and buy again for most value added in the early months of 2028. And we see this trend again on a smaller scale in the spring of 2031 and the early winter of 2032. Thus, the prediction could be made that selling in the spring and buying in the early winter months are the best times. And finally, we proposed that at least once a decade there was a major event that impacted pricing. It looks like this event is set to take place in 2027-2028 for this decade.

# References

1. B. Mérő, A. Borsos, Z. Hosszú, Z. Oláh, N. Vágó, A high-resolution, data-driven agent-based model of the housing market, Journal of Economic Dynamics and Control, Volume 155, 2023, 104738, ISSN 0165-1889, https://doi.org/10.1016/j.jedc.2023.104738.
2. C. Zhao, F. Liu, Impact of housing policies on the real estate market - Systematic literature review, Heliyon, Volume 9, Issue 10, 2023, e20704, ISSN 2405-8440,
3. A.G. Rawool, D. V. Rogye, S. G. Rane, V. A. Bharadi, House price prediction using machine learning, IRE Journals, Volume 4, Issue 11, 2021, ISSN 2456-8880
4. A.B. Adetunji, O. N. Akande, F. A. Ajala, O. Oyewo, Y. F. Akande, G. Oluwadara, House price prediction using random forest machine learning technique, Science Direct, Procedia Computer Science 199, pages 806 – 813, 2022
5. I. Ake, Combining machine learning models to predict house prices, Southhampton Solent University, Faculty of Business, Law, and Digital Technologies, 9 September 2022
6. Q. Truong, M. Nguyen, H. Dang, B. Mei, Housing price prediction via improved machine learning techniques, Science Direct, Procedia Computer Science 174, pages 433-442, 2020
7. R. Ali, P. Maynard, Project Proposal: Housing prices prediction model, GENG 4500-60, University of Windsor, 14 February 2024

##### Appendices

Appendix A: Initial Data Set Example

Appendix B: New Data Set Example

Appendix C: Comparison of Outputs

Appendix A

Initial Data Set Example

The table object below (Table 1) displays the initial data set and how it was a concatenation of different tables into one table. The unfortunate outcome was that there were not enough features and data in this set to produce a useful outcome. Thus, it needed to be replaced.



Table 1

Appendix B

New Data Set Example

The new data set below (Table 2) shows a more robust set to work with. It has the HPI we initially used to predict along with real dollar values of different types of properties. This proved to be the most important asset as it gave us a link between the index value and monetary values. Along with the robust set of features, this was our final data set.



Table 2

Appendix C

Comparison of Outputs

Looking at the difference in outputs between hyper-tuned (Fig. 4) and non-tuned (Fig. 3) helped us to understand the approach to the problem and how accurate our predicted prices would be. Below is a comparison of some of the output data between the models.

Non-Tuned:

[705388. 335518. 254518. 595919. 271312. 695601. 297031. 681523. 544252.

434100. 719617. 288173. 365141. 527562. 757081. 509602. 564637. 315523.

744558. 547313. 380374. 546812. 756147. 533693. 430256. 284273. 572798.

378761. 380179. 363323. 694114. 326685. 401831. 311097. 282660. 438490.

553013. 401057. 359562. 527512. 367784. 525590. 512719. 339292. 399752.

398923.]

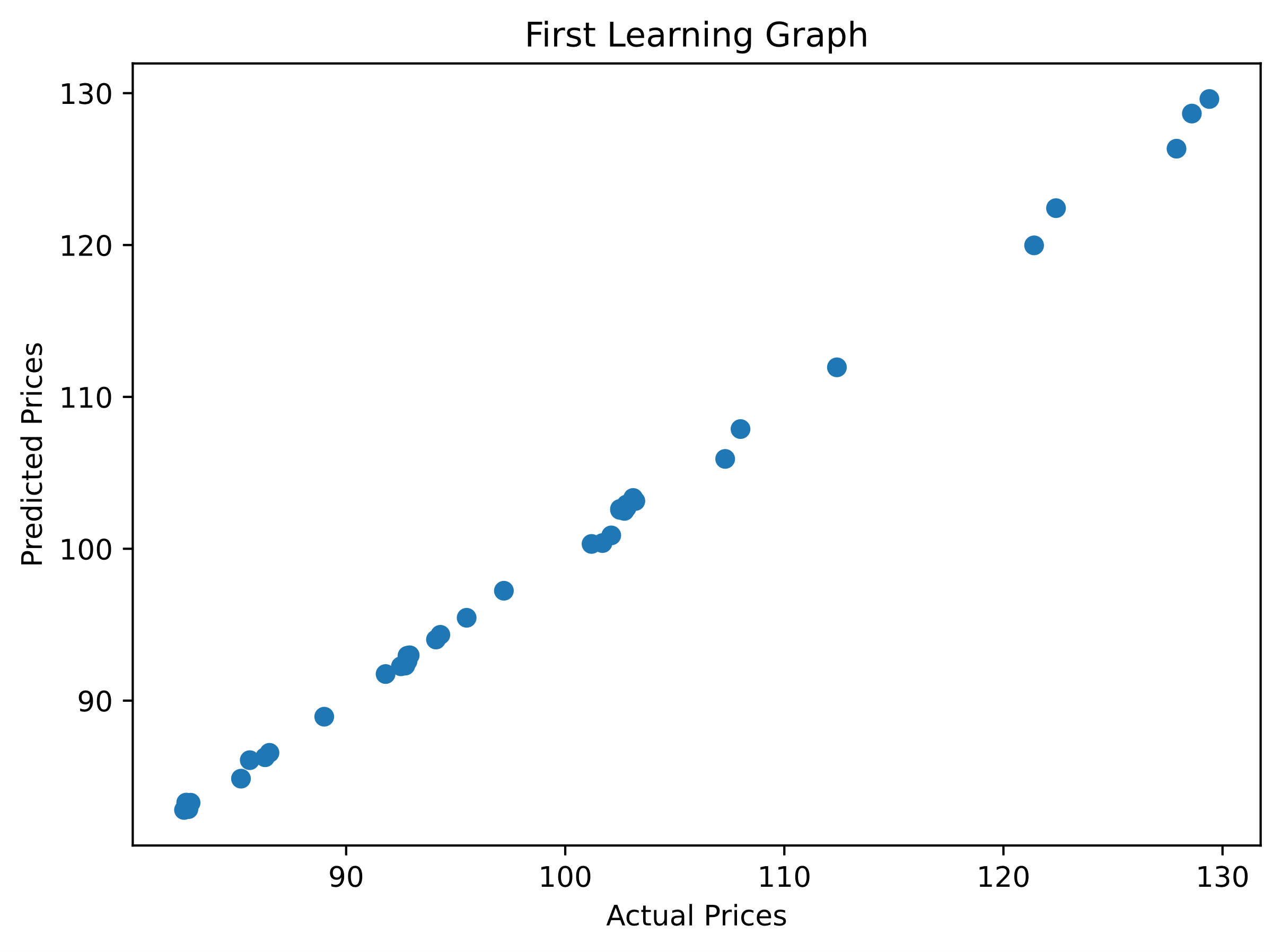


Fig. 4

Tuned:

[705835.5 335436.5 254601.5 595204. 271866.5 695769.5 297143.5 681312.

544519. 433602.5 719956. 288098. 364984.5 527658.5 757091.5 509260.

565034. 315514.5 744571.5 547158. 380433. 546587.5 756596.5 533371.5

429416. 284452.5 573729.5 378699.5 380191.5 363352.5 694567.5 326858.

401875.5 311074. 282068. 438711. 553024. 401020. 359569.5 529227.

367746.5 525554. 512868. 339437.5 399973. 398814.5]

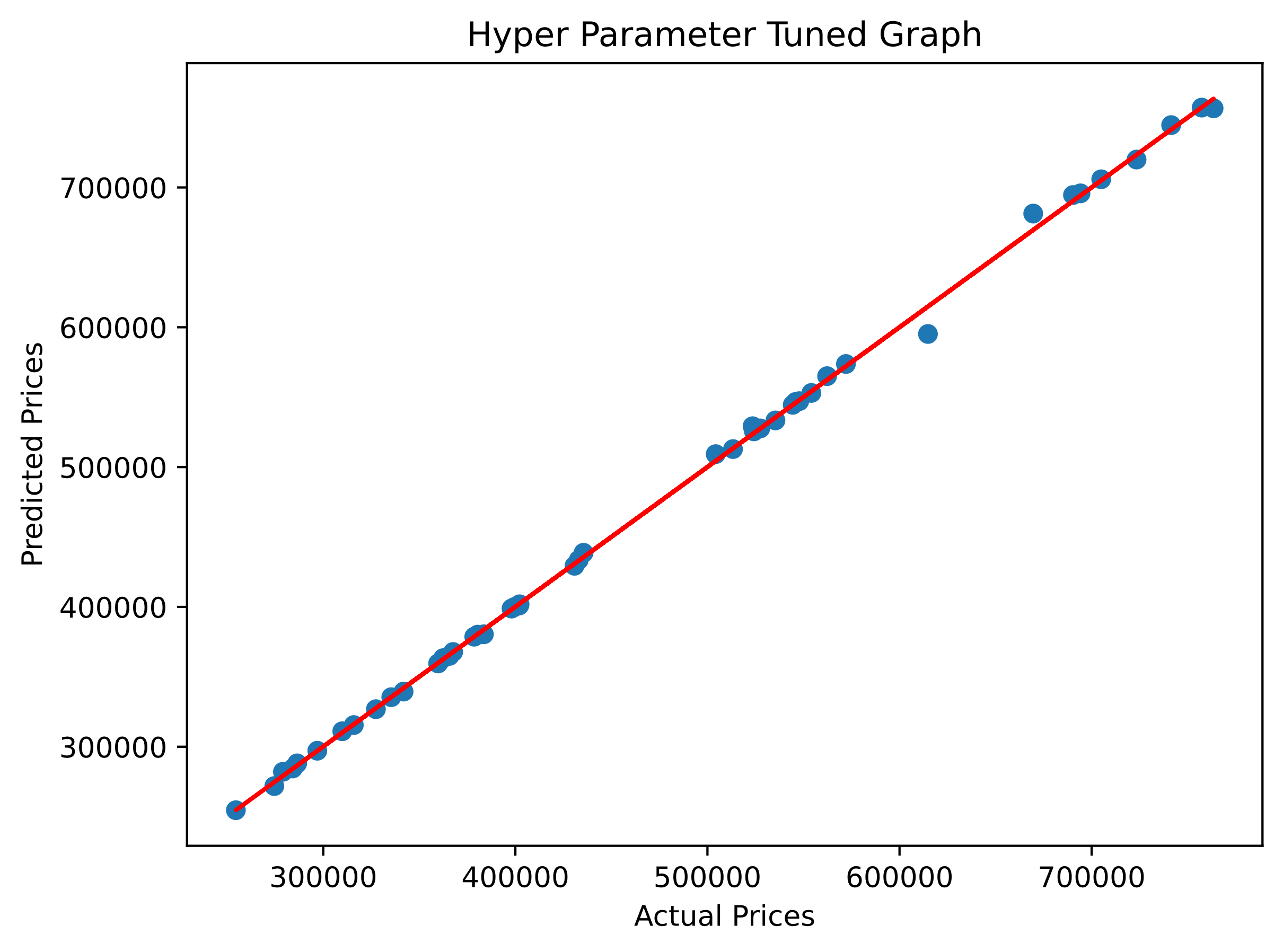


Fig. 5