GENG-4500 – 60 Project Proposal: 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, but there is a general pattern to be deciphered. Random forest regression has been the top choice of AI model to attempt to predict housing prices. For this project, we will attempt to adapt a random forest regression model using collected data from reputable resources to predict housing prices here in Canada.

Keywords—machine learning, the housing market, random forest regression

# Introduction

# Literature Review

# 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, 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.

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).

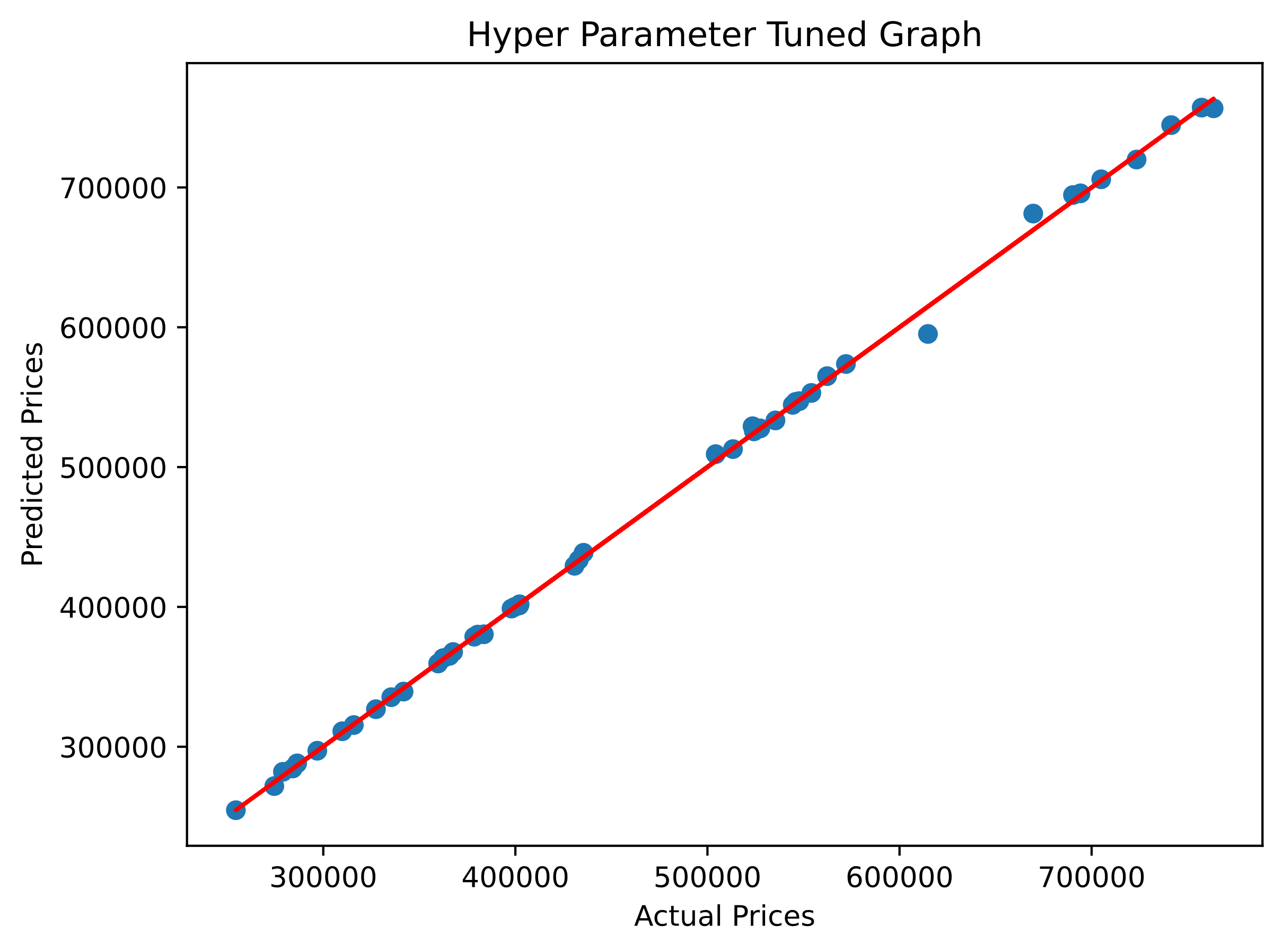


Fig. 1

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. The scale of the values of the data still seems to be off the mark, but it 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 further analyze the outputs 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]). The only issue noted is that the price point value is likely scaled to the initial values from the data set and not the jump-off points of the final data point values, thus giving us an accurate prediction of change in values while not showing the proper dollar amounts for the time frame presented.

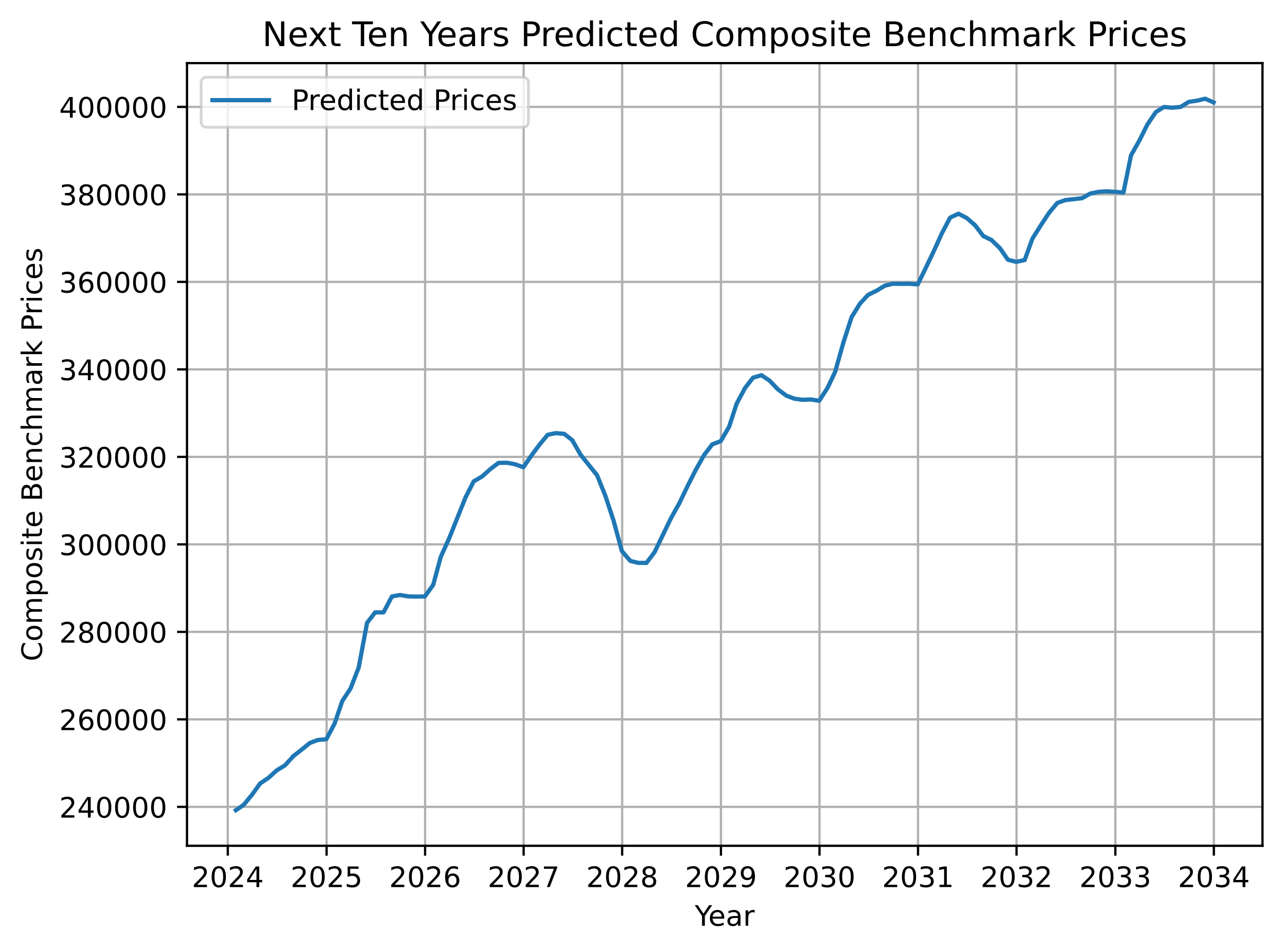


Fig. 2

# 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, and we would need to take a much more in-depth course on machine learning to get better results from a project taking on a problem of this scale.

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.

# 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

##### 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 compares 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.]

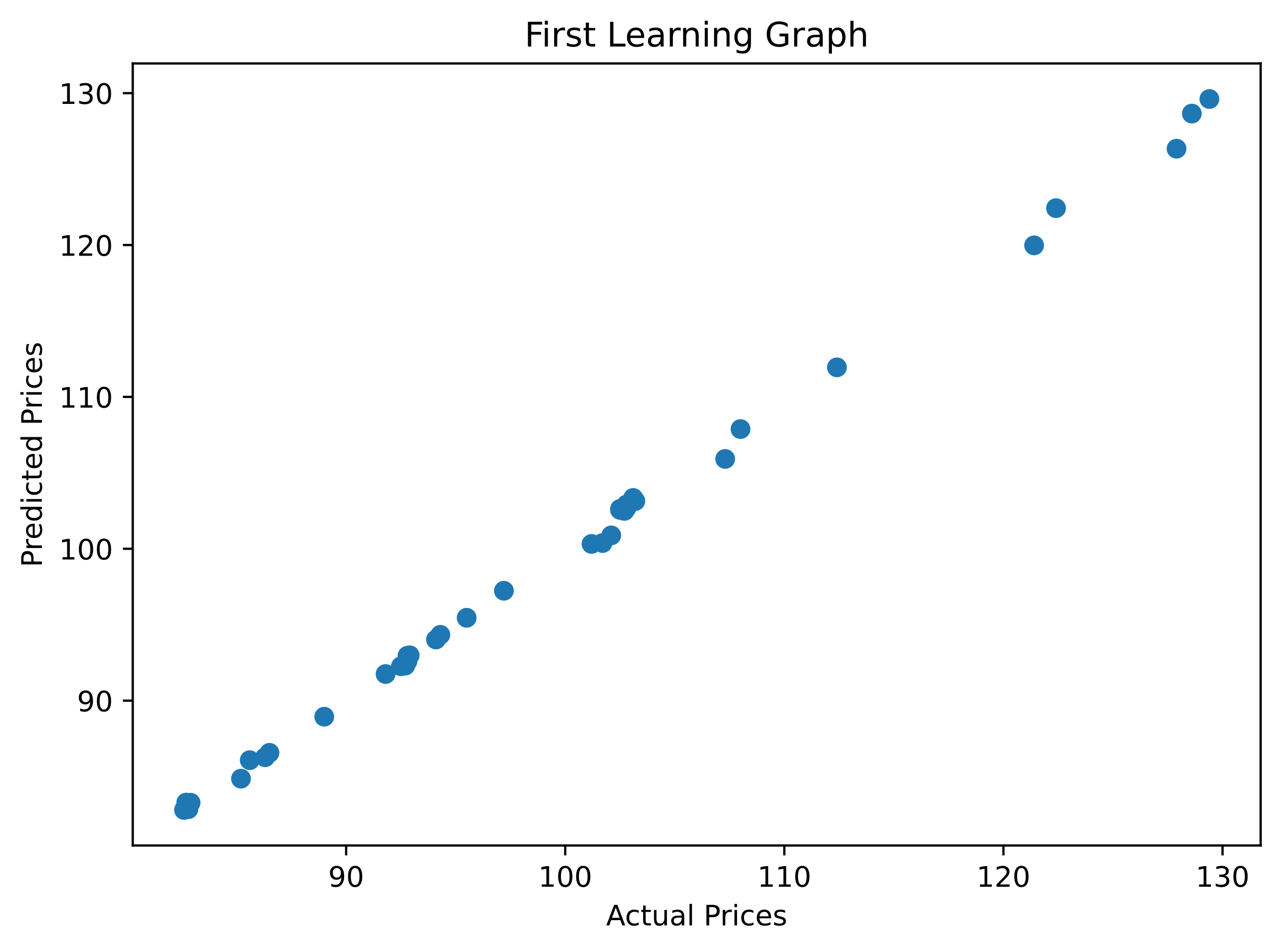


Fig. 3

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]

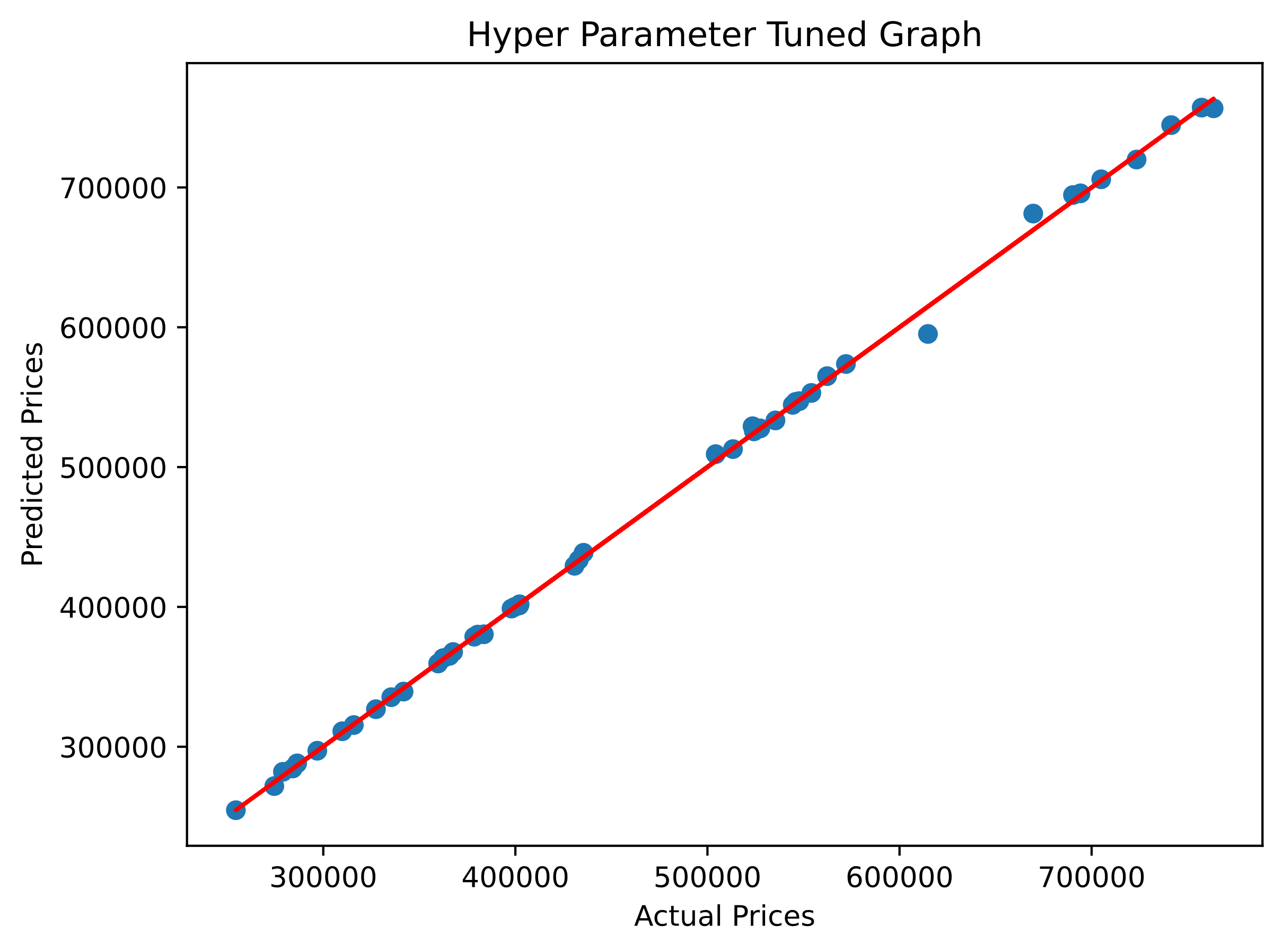


Fig. 4