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 in order to predict housing prices here in Canada.

Keywords—machine learning, the 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 attempts to take a deeper look at the Canadian housing market in order to formulate a strategic approach to buying or selling properties using machine learning and 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 year. Variables such as the CPI (Consumer Price Index), the HPI (House Price Index), and socio-political events such as the 2008 collapse, various recessions, and even Covid-19 are taken into consideration. Though not quite as in-depth as other studies like it across the world, this project attempts to look for patterns and possibly clear up some of the guesswork when it comes to property pricing.

# 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 GDP 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 the purposes of 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 in order to help predict pricing outcomes. In the article titled “House price prediction using machine learning”, the authors take note of the advancing nature of machine learning and its 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 in order 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 article titled “House price prediction using random forest machine learning technique”, the authors use a fairly in-depth random forest regression model in order 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. In the article titled “Combining machine learning models to predict house prices”, the author takes on a rather in-depth study of what can be considered a combination of the two previous articles. He sets out to investigate a multitude of AI models in order to evaluate their application in predicting housing prices. He uses simple linear regression, decision tree regression, random forest regression, and also 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 titled “Housing price prediction via improved machine learning techniques”, the authors once again research a multitude of AI models in order 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 follows the standard seven principles for machine learning.

1. Gathering data

Our data was gathered from two main sources, Stats CREA, and Statistics Canada. We consulted four different data types: the MLS home price index from Stats CREA, the House price index, the Consumer price index, and mortgage interest rates, all from Statistics Canada. These are all reputable sources, with thorough and pertinent data.

1. Preparing the data

To prepare the data, we considered a few complications. Things such as outside factors affecting the price, varying characteristics of the properties, overfitting of the data, and possible unusual formats of the data can skew the results. These were addressed by choosing specific groups of factors to address, using the “One Hot Encoder” function to break out the characteristics into Boolean value tables, and transposing all of the data into one table, respectively.

1. Choosing the model

Based on the problem we are attempting to solve, there are a few options available when it comes to selecting a model: Linear regression, Decision Tree regression, and Random Forest regression to name a few. The model we chose was Random Forest Regression. This is a very versatile tool that is used for predicting continuous outcomes. It works by consulting an ensemble of decision trees, which are each trained on random portions of the data, using random features to create randomness and diversity, leading to more accuracy. At every node in every decision tree, the random forest algorithm [Figure 1] will select the best route based on certain criteria, in this case, the MSE (mean squared error) equation [Figure 2]. Predictions are ultimately made based on the predictions of the individual decision trees by taking an average. This helps to reduce variance and improve accuracy.



Figure 1



Figure 2

1. Training

To train our model, the HPI, CPI, and Interest rates are the selected features. These sets are then divided into training datasets and test datasets. As previously mentioned, each decision tree is independently trained on a random subset of the data and features. This randomness will help to reduce overfitting and produce a more accurate result.

1. Evaluation

To evaluate the model, the mean squared error and the R-squared score are used. The mean squared error calculates the average of the squared differences between the predicted values and the true values. This checks the overall accuracy of the model. The R-squared score is a statistical measure of the relationship between the independent variable and the dependent variable. It checks the fit of the model compared to the sourced data.

1. Hyperparameter tuning

Some possible parameters for tuning the model can include Bayesian optimization or Grid search. Bayesian optimization can be used to try and predict the overall function shape, while grid search cross-references the data to look for optimal combinations. Other potential options for tuning are changing parameters such as the learning rate by adjusting the step size, the rate decay by finding the local minima through a lower learning rate, the momentum by reducing noise using previous iterations, or data size. By modifying these potential parameters, the efficiency and accuracy of the model can be increased.

1. Prediction

At this point, we have a rough idea of where prices are going. Based on the general trend, prices are going up, but at varying rates. This model will seek to predict the variance with a degree of accuracy that will make it easier for homebuyers to make informed decisions.

# Timeline

The timeline that we decided to implement for this project is displayed in the following Gantt chart [Figure 3]. We separated the tasks into smaller, more easily accomplished steps in order to have better control over the process. This timeline does take into account responsibilities outside of this class, as well as the time needed for research and development. The timeline was set up to have little overlap between tasks as we felt this would limit having to juggle too many things at once, also, it works best for the systematic approach we are taking. Ultimately, we felt that this timeline worked best in accordance with both of our schedules and appropriately covered everything we needed.

*A screenshot of a project

Description automatically generated*

Figure 3

# Conclusion

For many people, setting out to purchase a house can be a daunting undertaking. Among all the factors that one would consider, the cost of the house is possibly the largest concern. It has already been established that the pricing can vary wildly depending on the time of the decade, time of the year, or even global socio-political events. Events such as recessions, pandemics, or market collapses can greatly affect house pricing and there are not many reliable ways to help draw any patterns or conclusions. Predicting housing prices using a random forest model is a good way to try and make sense of the pattern. Not only is this a powerful method of drawing conclusions based on random data and features that correlate, but it is also very good at predicting continuous outcomes. Learning this method as a key tool in machine learning allows us to better appreciate the complexities and subtleties of AI development. It is a great foundation to build on while learning some of the basic first steps. Its potential is limitless as computers get more powerful and data availability becomes more widespread.

# 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 Graphs

Appendix B: First Output of Demo Model

Appendix A: Initial Data Graphs

The initial data shows upward trends after a basic normalization of our House Price Index data [Figure 4], Interest Rate data [Figure 5], and Cost Price Index specifically for shelters [Figure 6].

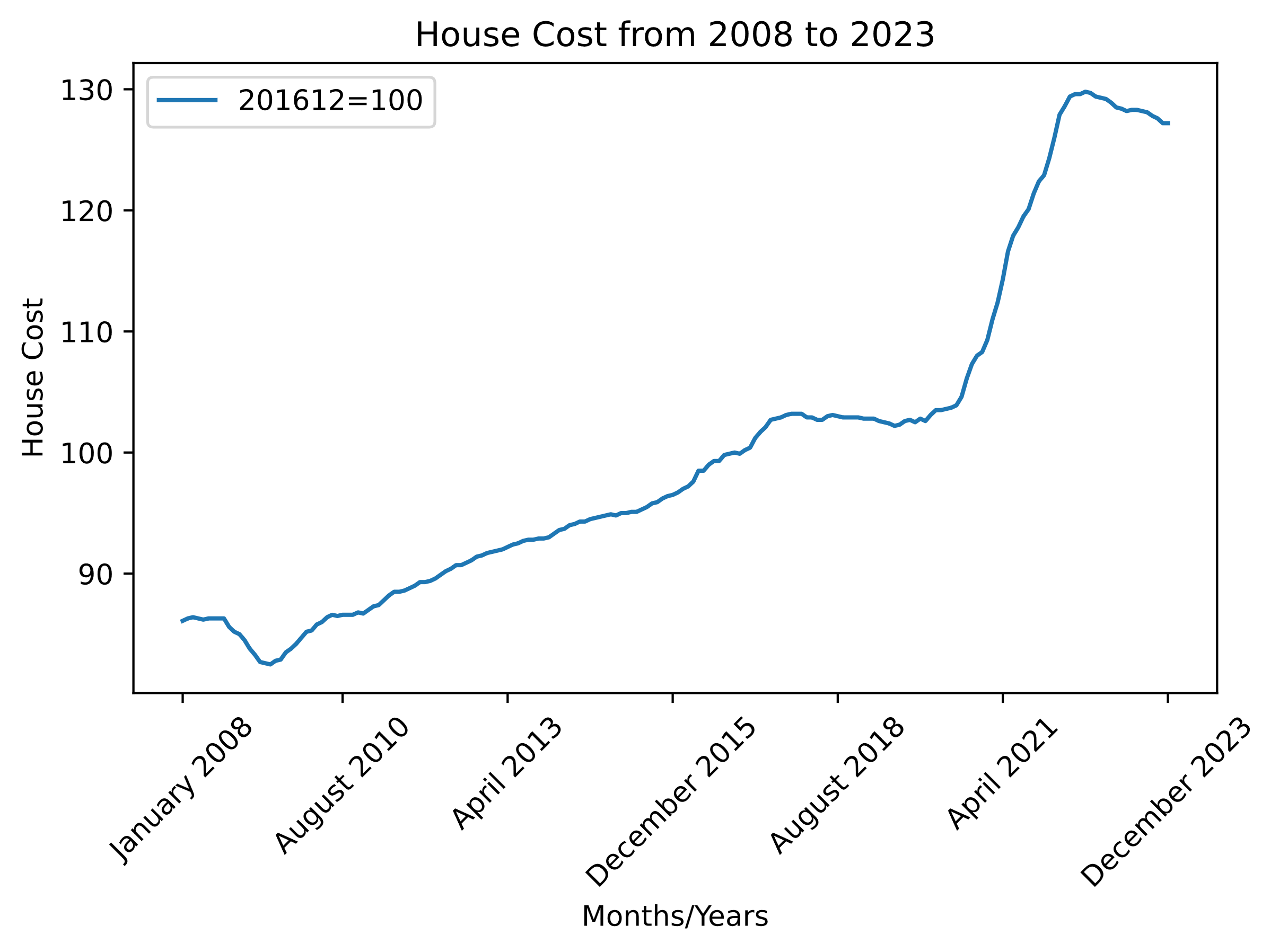


Figure 4

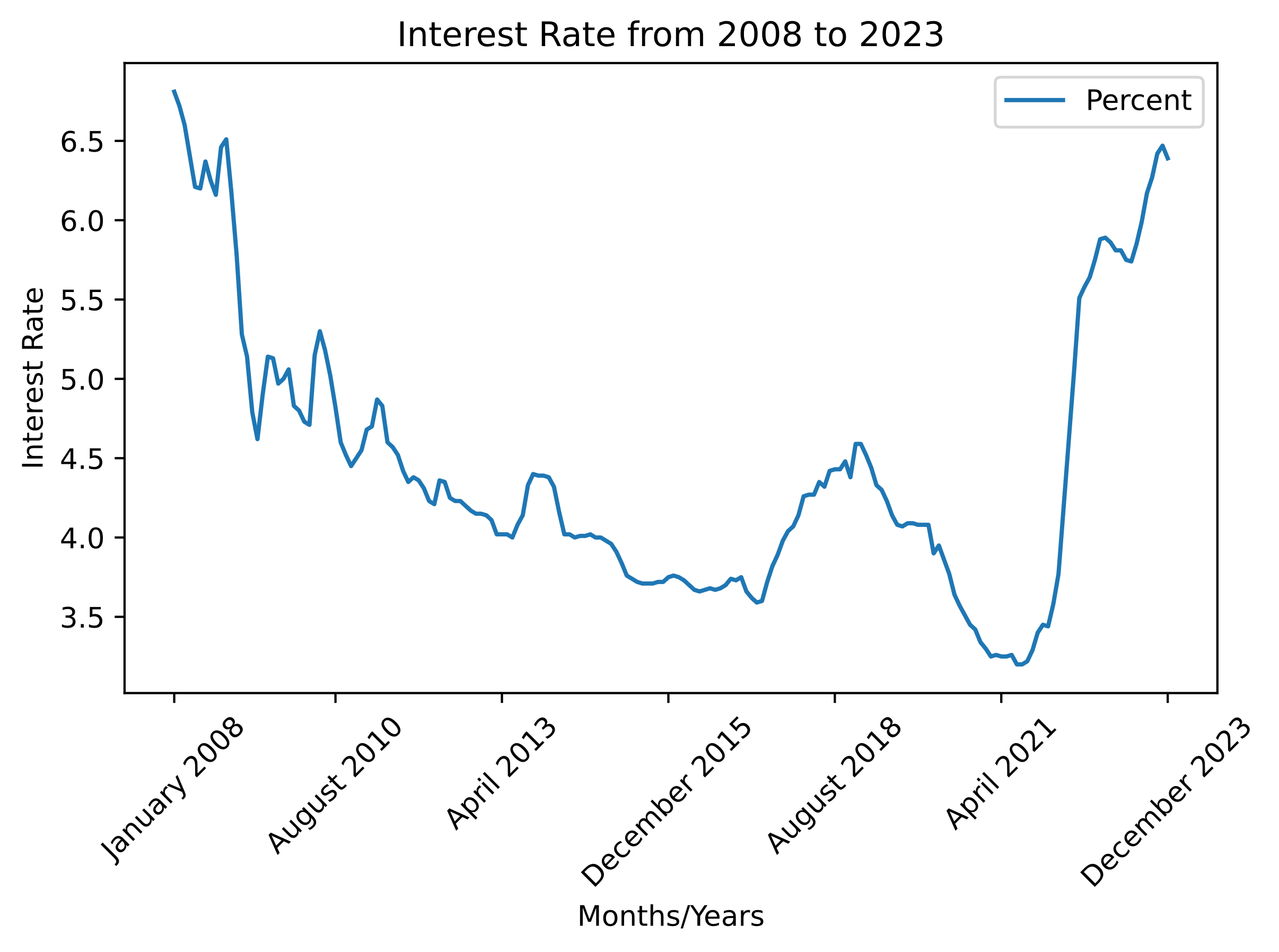


Figure 5

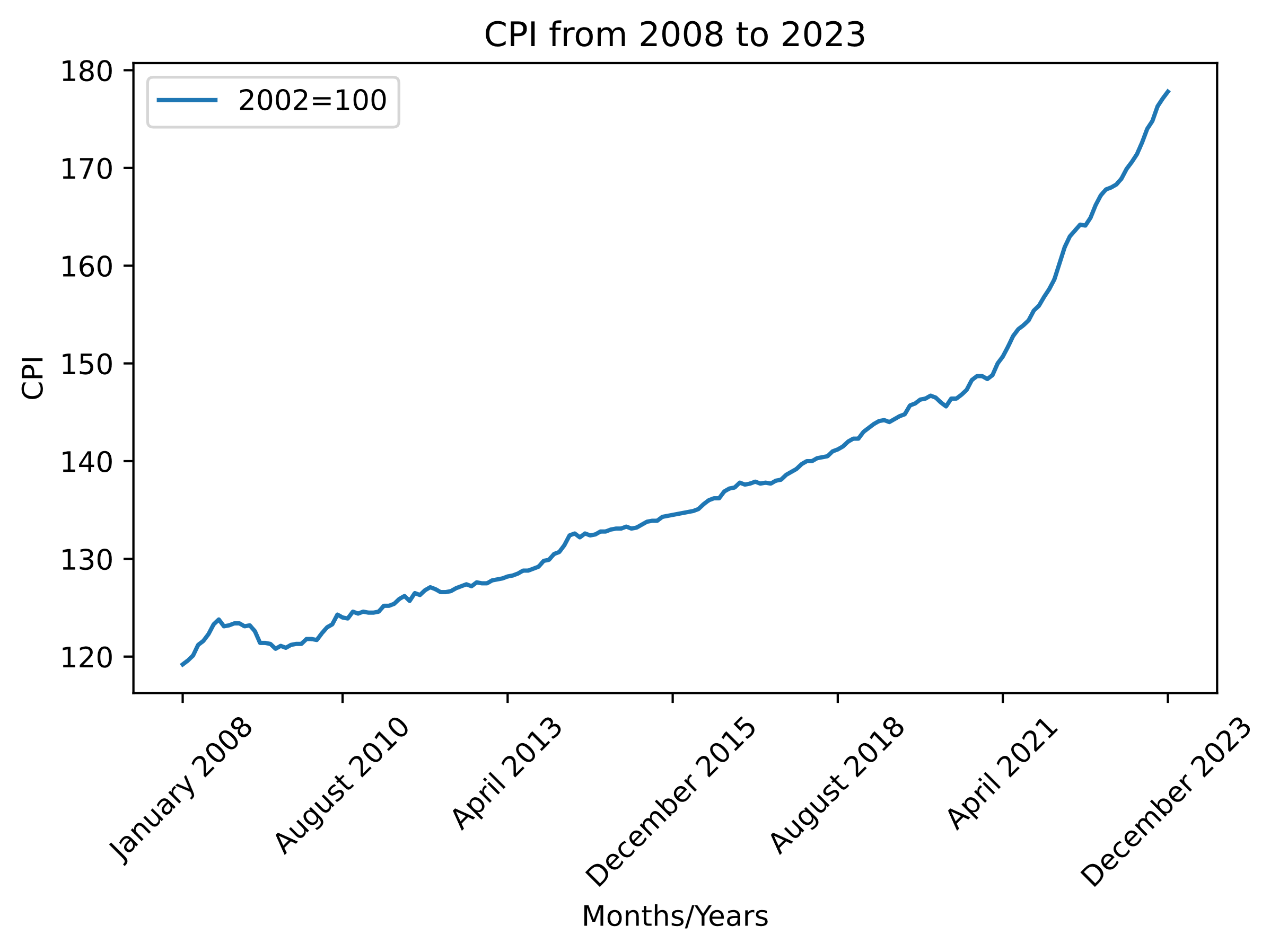


Figure 6

Appendix B: First Output of Demo Model

The code used for our initial output graph needs a lot of work. There have been lessons learned. We need to look at breaking out the features and hyperparameter tuning. The first graph produced does not indicate any of what we are trying to find yet. The first graph [Figure 7] produced has a 30% accuracy, this is due to the features not being broken out properly and needing to fine-tune the data being fed in.

Code Example of Mean Squared Error Output

*# pass in your y test data and check it against your y predicted data*

mse = mean\_squared\_error(y\_test, y\_predict)

*# share the accuracy of the model*

print(f"mean Squared Error: {mse}")

mean Squared Error: 0.30800769230769187

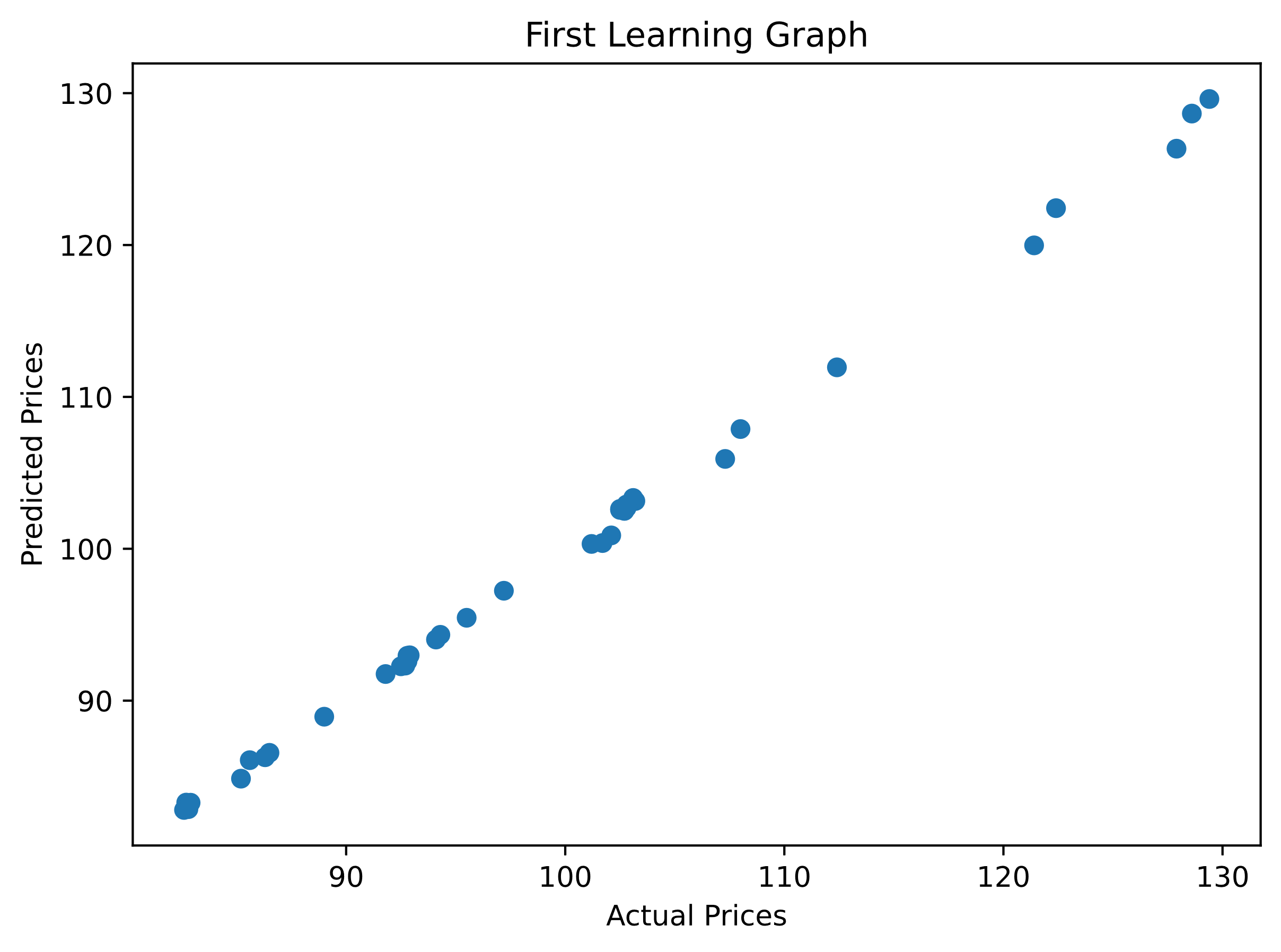


Figure 7