Housing Market Crash Prediction Using Machine Learning and Historical Data

By

Parnika De

Agenda

- Introduction
- Background
- Rise and Fall of the housing market between 2000 and 2010
- Data Collection
- Data Pre-processing
- Machine Learning Models
  - Linear Regression
  - HMM
  - LSTM
- Results and Discussion
- Conclusion

Introduction

- The objective of this project is to examine the historical data and predict using machine learning techniques whether we are nearing another housing crisis.
- We would investigate few elements of the Housing Crisis of 2008 and then build a dataset.
- Then we would apply ML models (Linear Regression, HMM, and LSTM) on the datasets to achieve the objective.

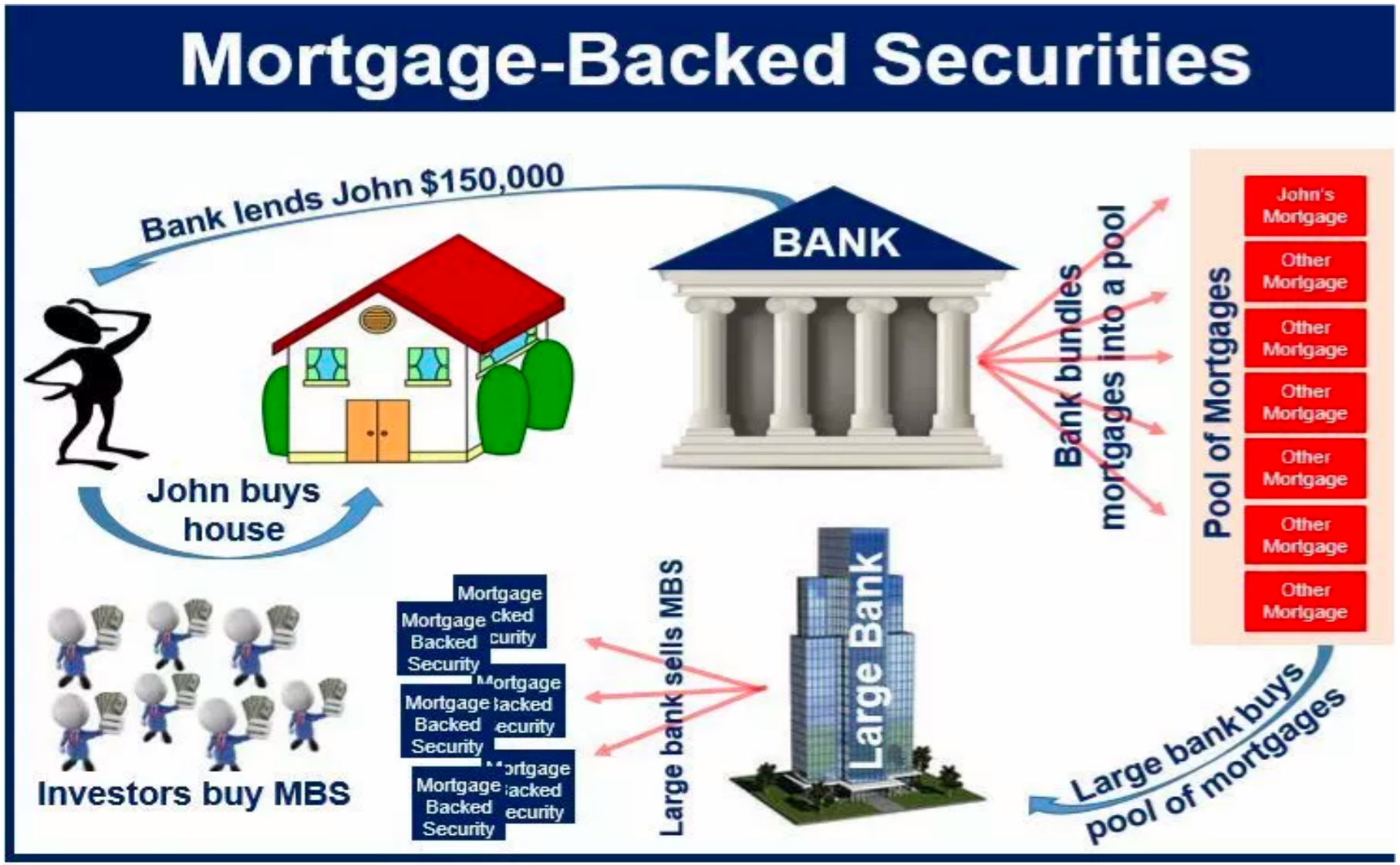
Background

- In the early days buying houses was not as complex as there were not too many layers to buying houses. If people had money, they buy all cash houses otherwise they would take loans from the banks.
- Banks in early days had very strict lending policies and it was impossible for people with low credit history to get loans from banks.
- As the risks were low there the interest that was earned by the banks was also very low.

Background(cont’d)

- During the early 2000s after the dot-com crisis, it was thought that the housing market was the sturdiest market as the housing prices increased throughout the crisis.
- People started investing more money in the housing market.
- Investors who were not buying houses were investing in the housing market through Mortgage-Backed Security(MBS).
- An MBS is a type of asset-based derivative security that derives its value from the underlying asset, the mortgages.
- The investors of MBS receive periodic payments just like other bonds.

Mortgage Backed Securities(MBS)



Rise of the Housing Market

- The mortgages were made very lucrative as the Federal Reserve Bank reduced the interest rates extremely low for short-term loans (ARM).
- People without substantial credit score could now buy houses through subprime loans.
- Mind set of people thinking Housing market is the pillar of investment mainly after the 2000's dot-com crisis.
- More and more people bought houses or invested in the housing market through MBS
- *Result: The Housing market boomed in the early to mid 2000*

Fall of the Housing Market

- In the 2000s the MBS investments started getting very sophisticated.
  - Investment banks started slicing MBS's into tranches.
  - A **tranche** is a slice of a bundle of derivatives. It allows you to invest in the portion with similar risks and rewards.
- Banks were also giving out more sub-prime loans, therefore the MBSs now have a significant amount of subprime loans.
- Subprime lending is the provision of loans to people who may have difficulty maintaining the repayment schedule. Historically, subprime borrowers were defined as having FICO scores below 600.
- *Everything works fine until borrowers of loan starts defaulting.*

Fall of the Housing Market(cont'd)

- Around 2007-2009 when the interest rates were changed for ARM borrowers people started defaulting.
- The mortgage defaulters were huge in numbers therefore it affected the others in the chain of mortgage.
- Investors of MBS started losing money from their investments.
- The banks were also investors in the MBSs; therefore banks also lost a large sum of their investment along with people stopping mortgage payments
- Bank got a taste of all their wrong decision. But it did not stop there because people started losing their jobs.
- In no time the US was in a huge recession along with the countries that invested in US businesses.

Reasons of 2008 Housing Crisis

- The 2008 housing crisis devastated the American economy
- The factors that led us to the 2008 recession
  - Inflated housing prices, that created a housing bubble
  - Relaxed banking policies that led to the high borrowing rate
  - Relaxed overall financial regulation
  - Policies developed by banks to give more subprime mortgages

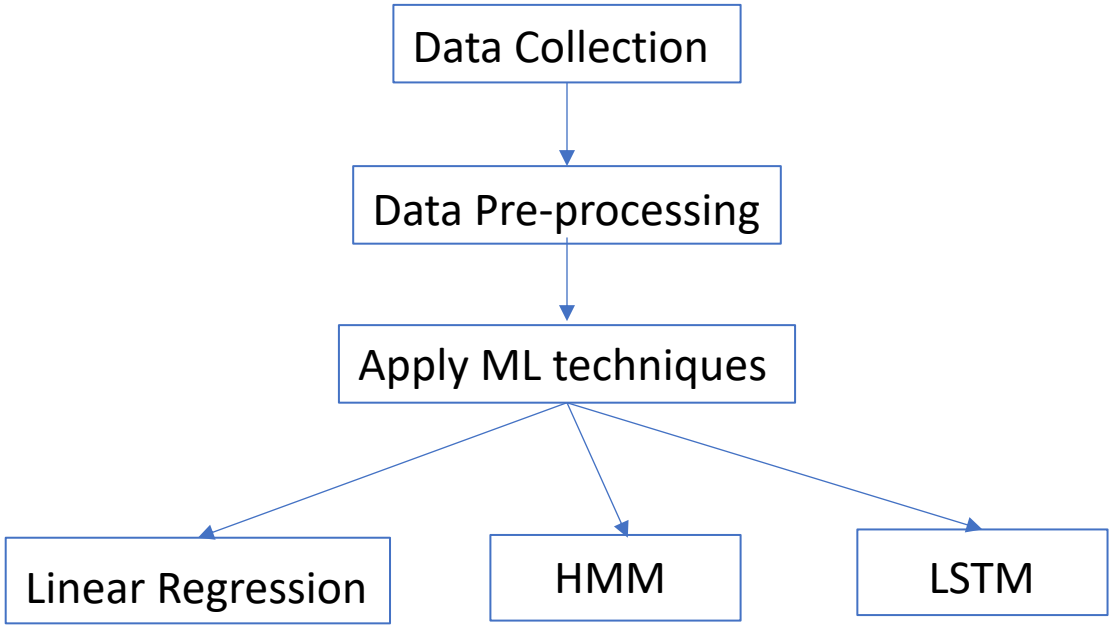
Prediction of Housing Crises

Techniques for predicting can range from simple statistical techniques to more complex deep learning ones. In this project, we make use of the following techniques:

- Linear Regression
- Hidden Markov Model (HMM)
- Long short-term Memory (LSTM)

If crises like these can be predicted before hand then measures can be taken to prevent or lessen the impact of the crisis.

Flowchart


Data Collection
Data Pre-processing
Apply ML techniques
Linear Regression
HMM
LSTM

Datasets

The dataset that we will be using are:

- Mortgage interest rate [12]
- Housing price [11]
- Total number of houses sold [13]

Data pre-processing

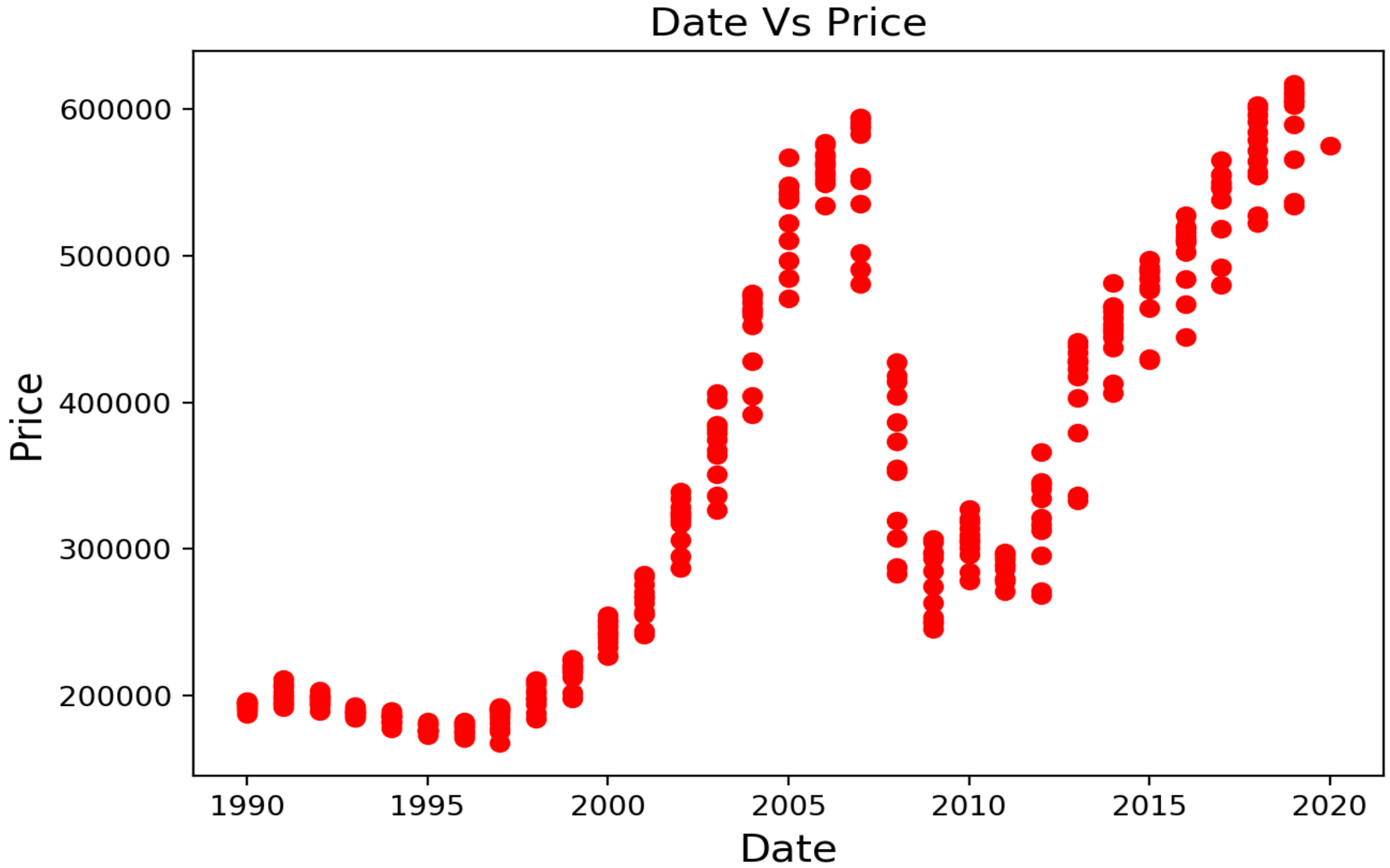
The merging of these data sets and data preprocessing is done through a python data manipulation library, Pandas.

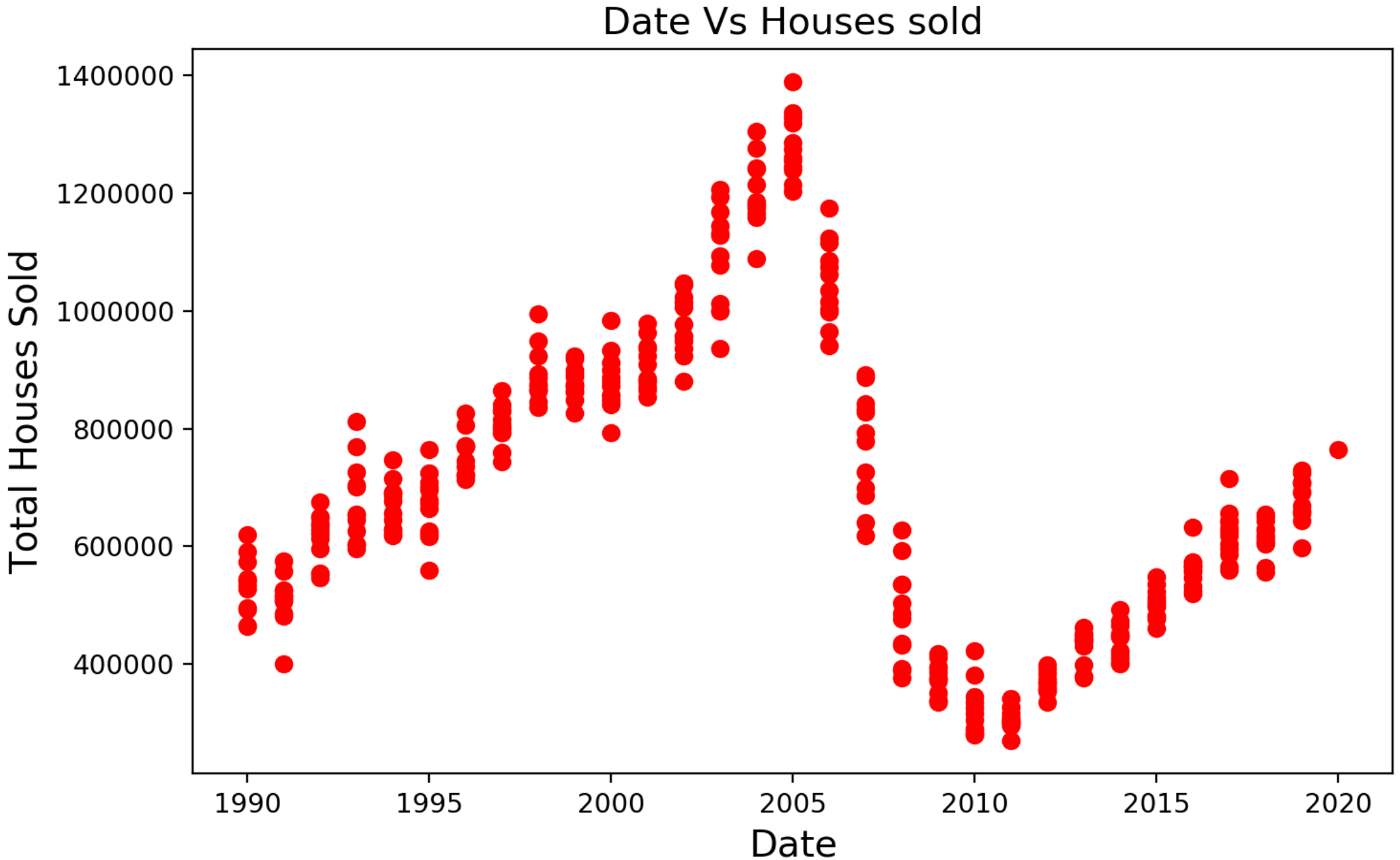
	Date	CA	Alameda
2	Jan-90	$194,952	$226,149
3	Feb-90	$196,273	$219,306
4	Mar-90	$194,856	$225,162
5	Apr-90	$196,111	$229,333
6	May-90	$195,281	$232,291
7	Jun-90	$194,410	$231,250
8	Jul-90	$193,088	$232,916

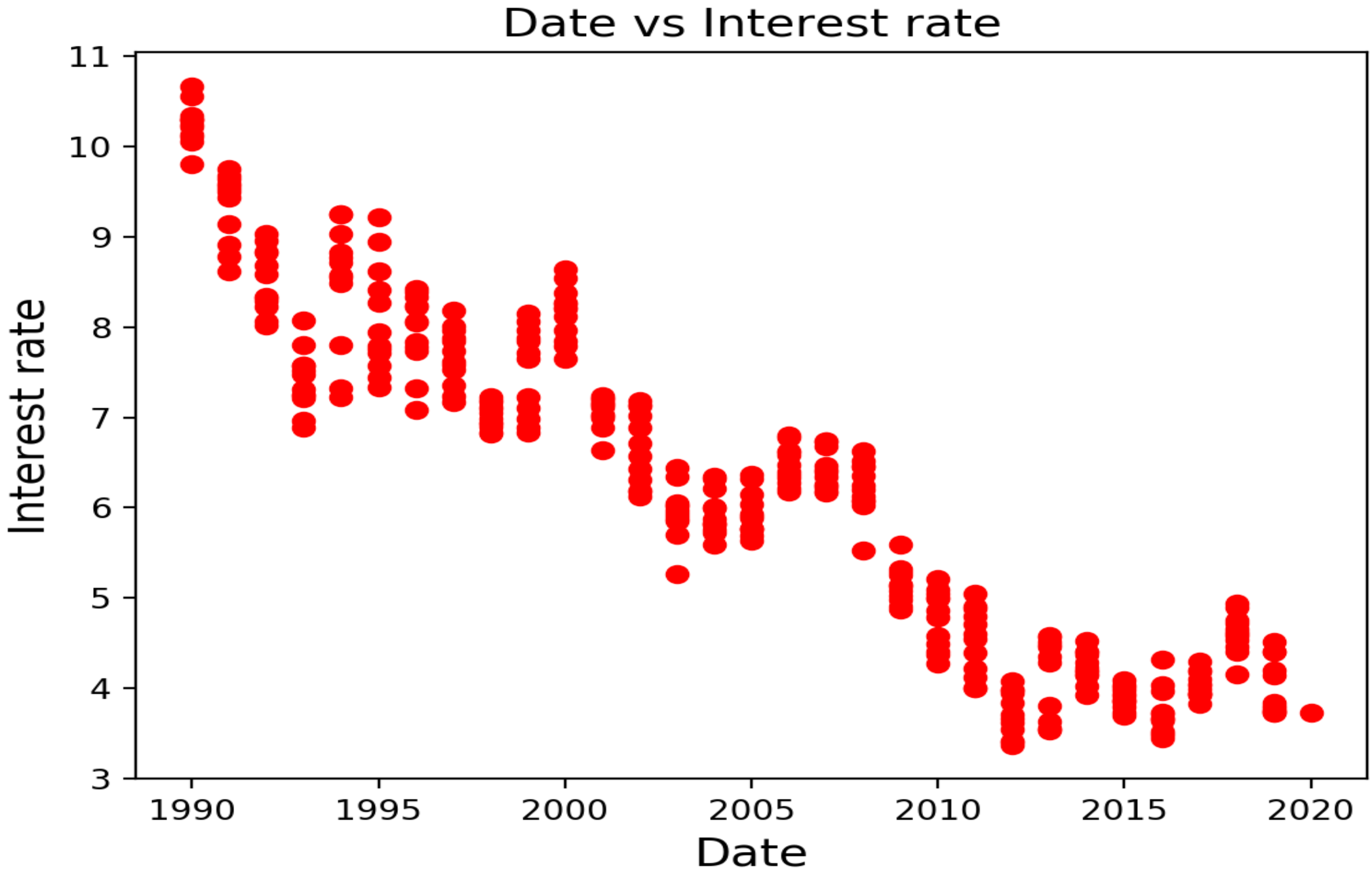
	Month	Year	house_sold
2	Jan.	2020	764000
3	Dec.	2019	708000
4	Nov.	2019	692000
5	Oct.	2019	707000
6	Sept.	2019	725000
7	Aug.	2019	708000
8	July	2019	660000
9	June	2019	729000

	observation_date	MORTGAGE30US
2	1971-04-02	7.33
3	1971-04-09	7.31
4	1971-04-16	7.31
5	1971-04-23	7.31
6	1971-04-30	7.29
7	1971-05-07	7.38
8	1971-05-14	7.42
9	1971-05-21	7.44
10	1971-05-28	7.46
11	1971-06-04	7.52
12	1971-06-11	7.52
13	1971-06-18	7.54
14	1971-06-25	7.54

ID	Date	month	year	price	rate	total_house_sold	period
1	1990-01-01 00:00:00.000000	1	1990	194952	10.05	620000	1/1990
2	1990-02-01 00:00:00.000000	2	1990	196273	10.31	591000	2/1990
3	1990-03-01 00:00:00.000000	3	1990	194856	10.34	574000	3/1990
4	1990-04-01 00:00:00.000000	4	1990	196111	10.56	542000	4/1990
5	1990-05-01 00:00:00.000000	5	1990	195281	10.67	534000	5/1990
6	1990-06-01 00:00:00.000000	6	1990	194410	10.29	545000	6/1990
7	1990-07-01 00:00:00.000000	7	1990	193088	10.11	542000	7/1990
8	1990-08-01 00:00:00.000000	8	1990	192180	10.29	528000	8/1990
9	1990-09-01 00:00:00.000000	9	1990	189979	10.22	496000	9/1990
10	1990-10-01 00:00:00.000000	10	1990	187630	10.24	465000	10/1990
11	1990-11-01 00:00:00.000000	11	1990	192020	10.13	493000	11/1990
12	1990-12-01 00:00:00.000000	12	1990	190375	9.81	464000	12/1990
13	1991-01-01 00:00:00.000000	1	1991	192054	9.75	401000	1/1991
14	1991-02-01 00:00:00.000000	2	1991	194806	9.56	482000	2/1991
15	1991-03-01 00:00:00.000000	3	1991	202666	9.59	507000	3/1991


Date Vs Price
Price
600000
500000
400000
300000
200000
1990
1995
2000
2005
2010
2015
2020
Date


Date Vs Houses sold
Total Houses Sold
1400000
1200000
1000000
800000
600000
400000
1990
1995
2000
2005
2010
2015
2020
Date


Date vs Interest rate
Interest rate
Date
11
10
9
8
7
6
5
4
3
1990
1995
2000
2005
2010
2015
2020

Linear Regression

- Linear regression is a supervised learning technique that models linear relationship between the dependent or scalar and the independent or explanatory variables.
- When there is one independent variable, then the modelling technique is called simple linear regression. It is of the form

$$\mathbf{y} = \boldsymbol{b_0} + \boldsymbol{b_1}.\boldsymbol{x_1}$$

- When there is more than one explanatory variable for a scalar then it is called multiple or multivariate linear regression. It is of the form

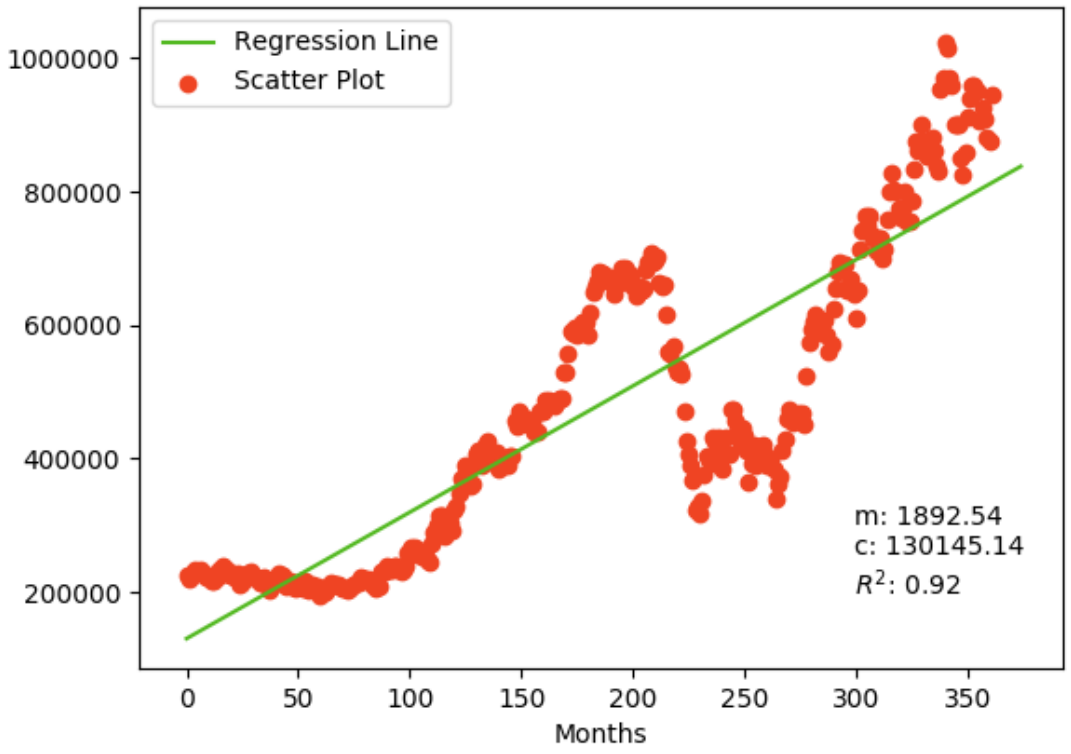
$$\mathbf{y} = \boldsymbol{b_0} + \boldsymbol{b_1}.\boldsymbol{x_1} + \boldsymbol{b_2}.\boldsymbol{x_2} + \boldsymbol{b_3}.\boldsymbol{x_3} + \ldots + \boldsymbol{b_n}\boldsymbol{x_n}$$

Linear Regression(cont’d)

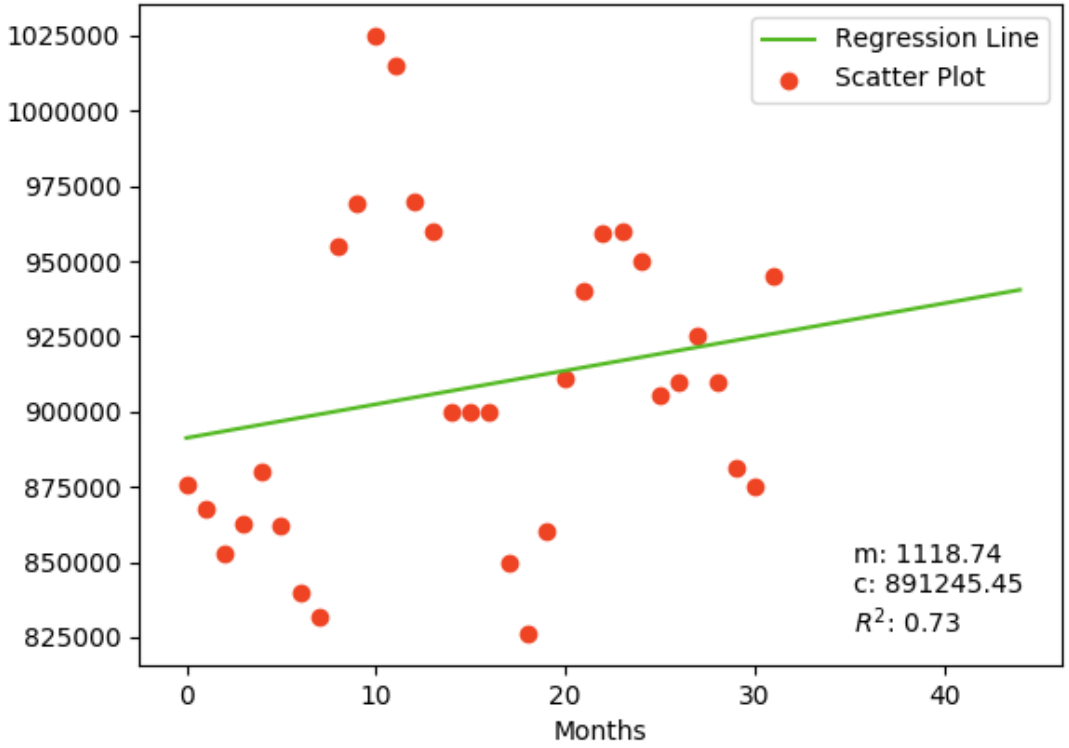
- In this project we have used both simple (CS 297) and multiple linear regression.
- For both the model the dependent variable is the house price and the independent variable is date for the simple linear regression model.
- We started with simple linear regression to understand the dynamics of the house price related to time. In this part I coded the algorithm instead of using sci-kit learn.

Results of Simple Linear Regression

Alameda: full

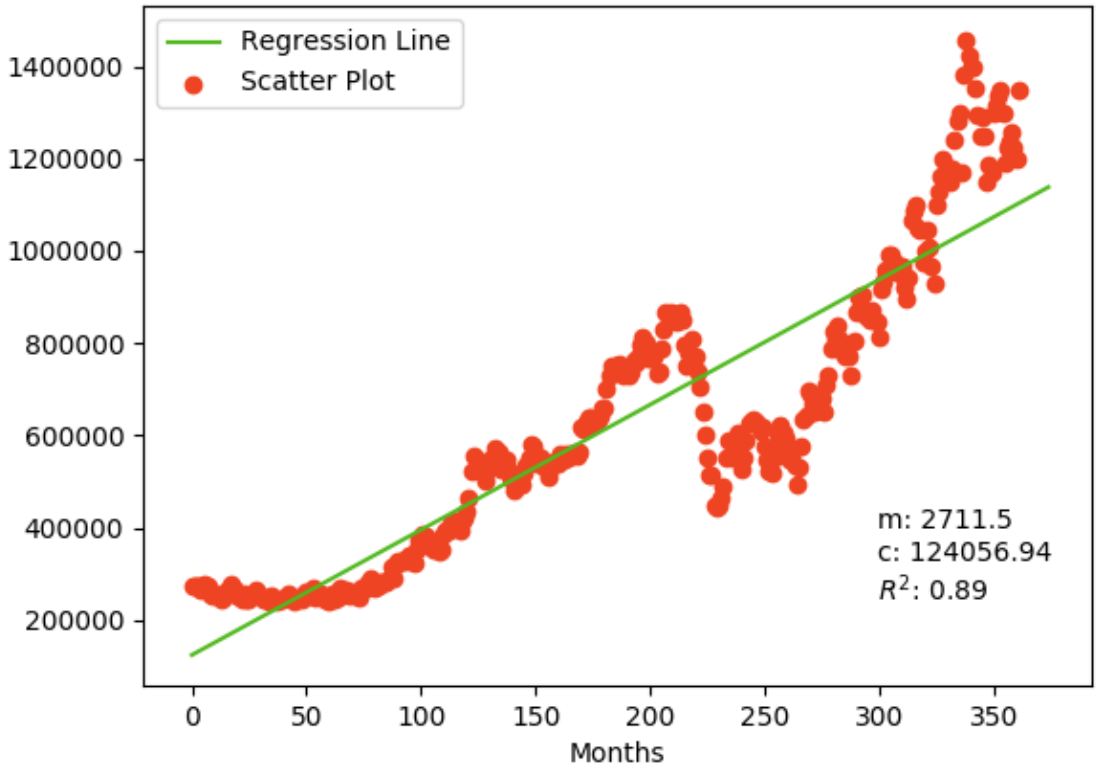


Alameda: recent

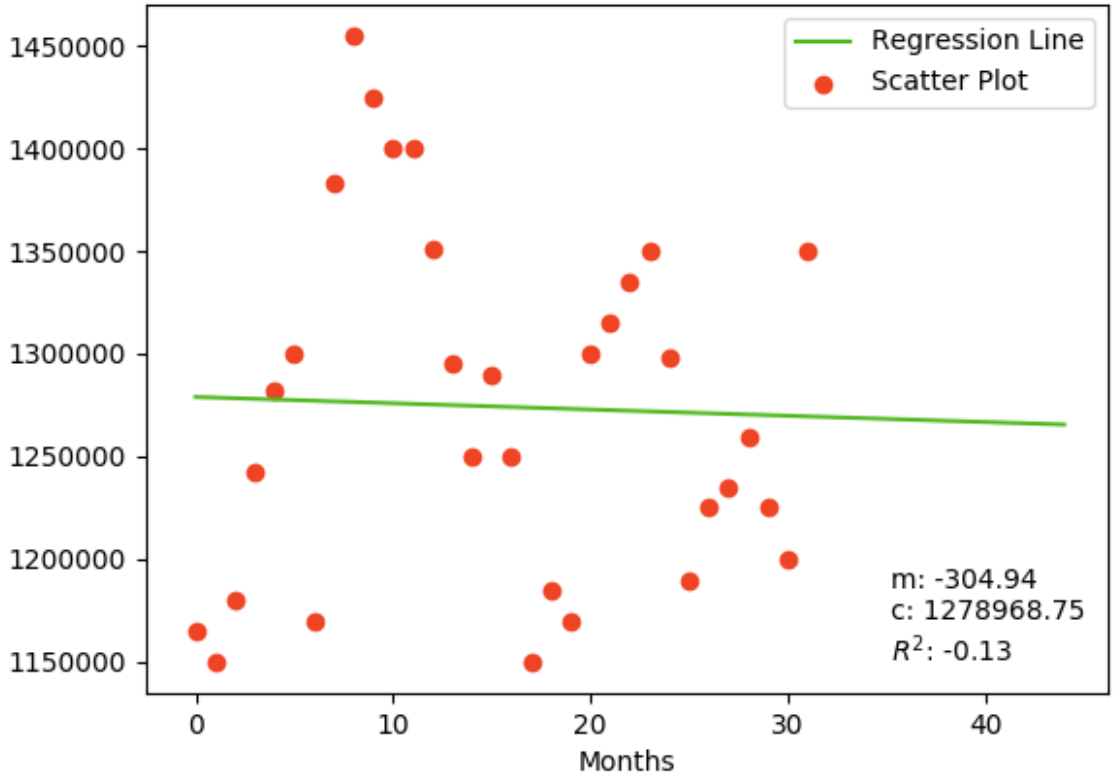


Results of Simple Linear Regression

Santa Clara: full

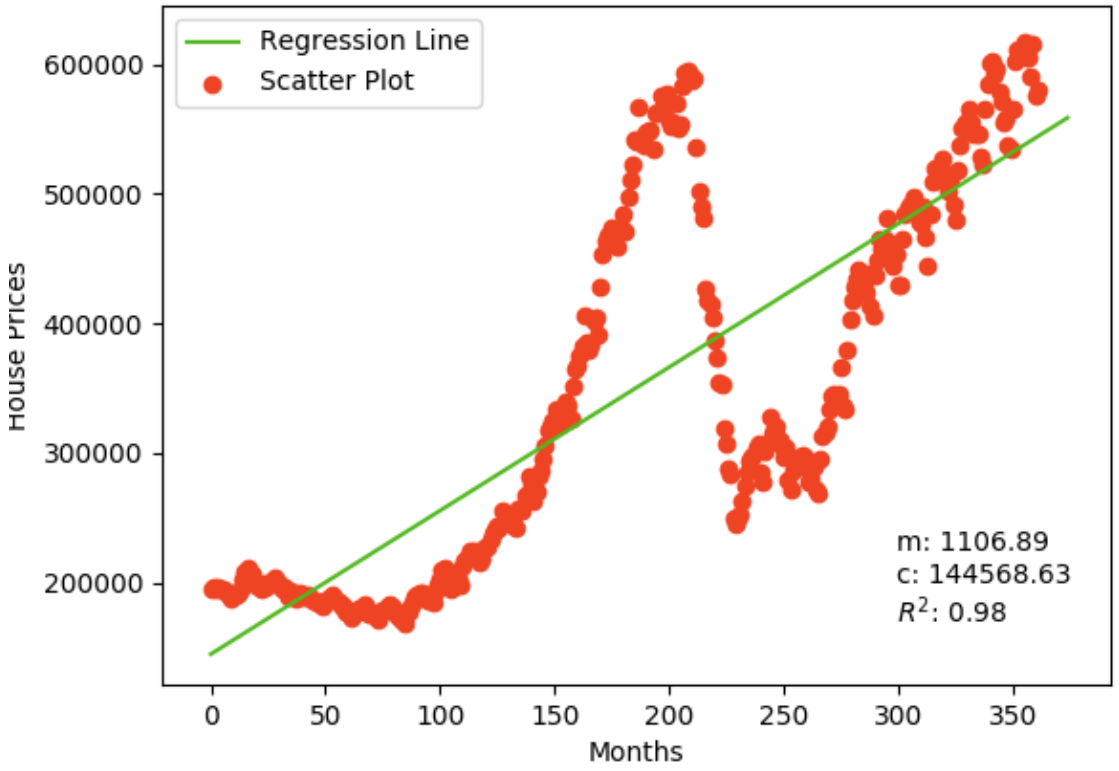


Santa Clara: recent

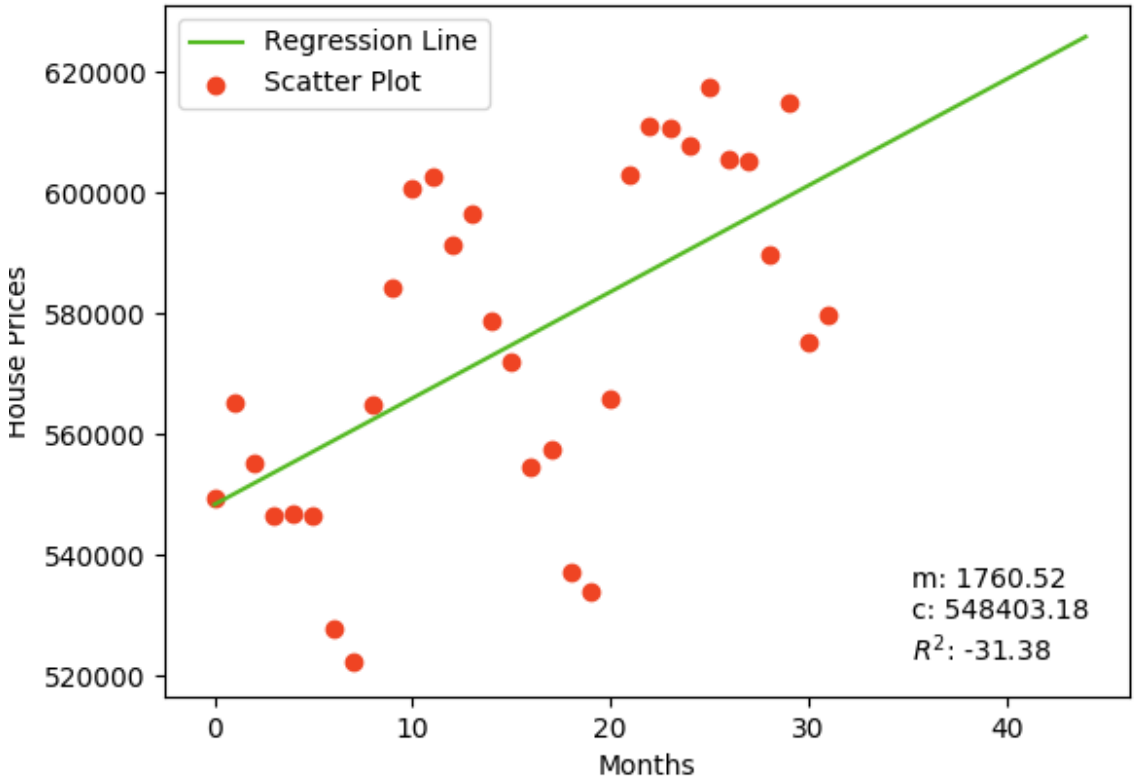


Results of Simple Linear Regression

CA: full



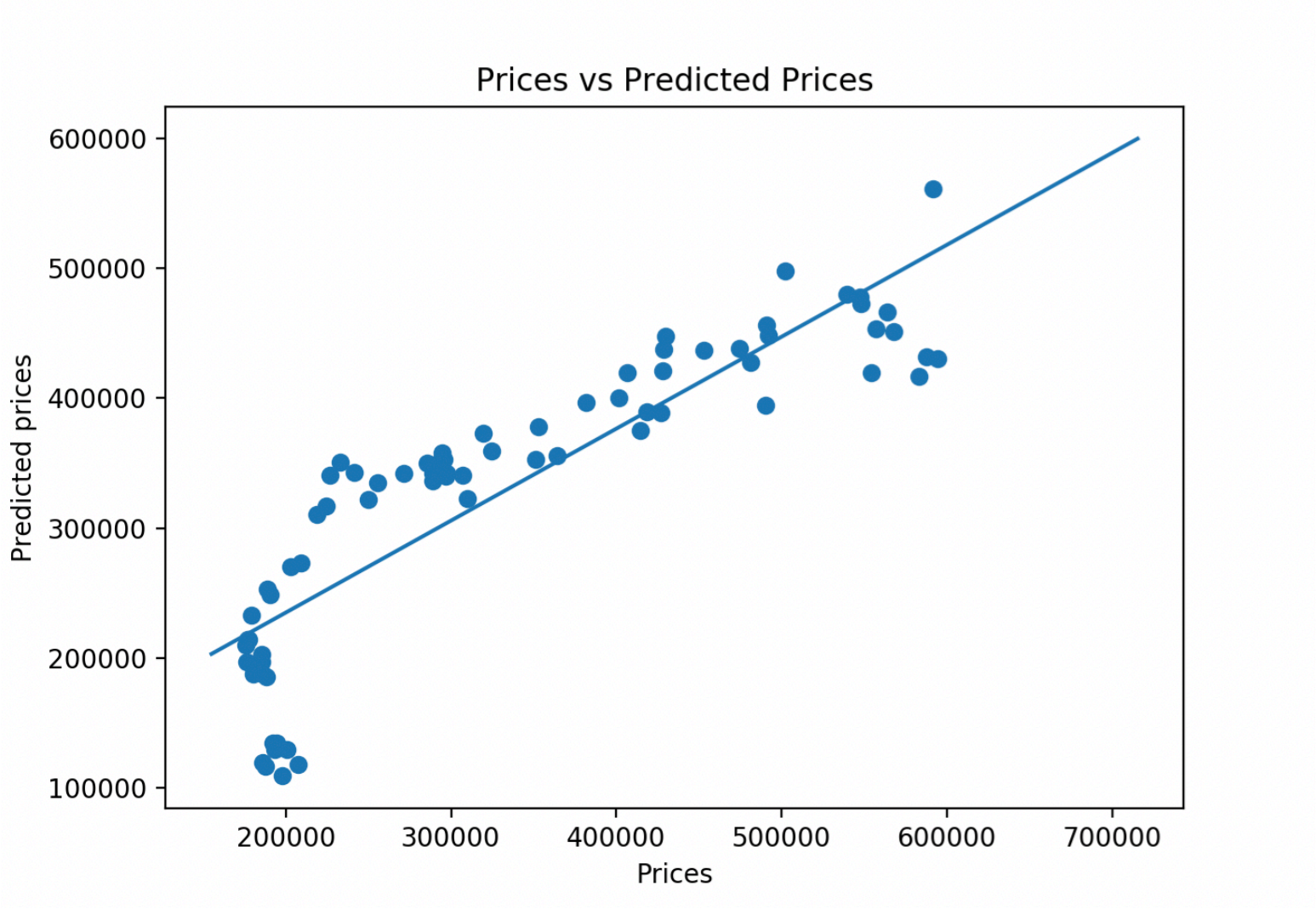
CA: recent



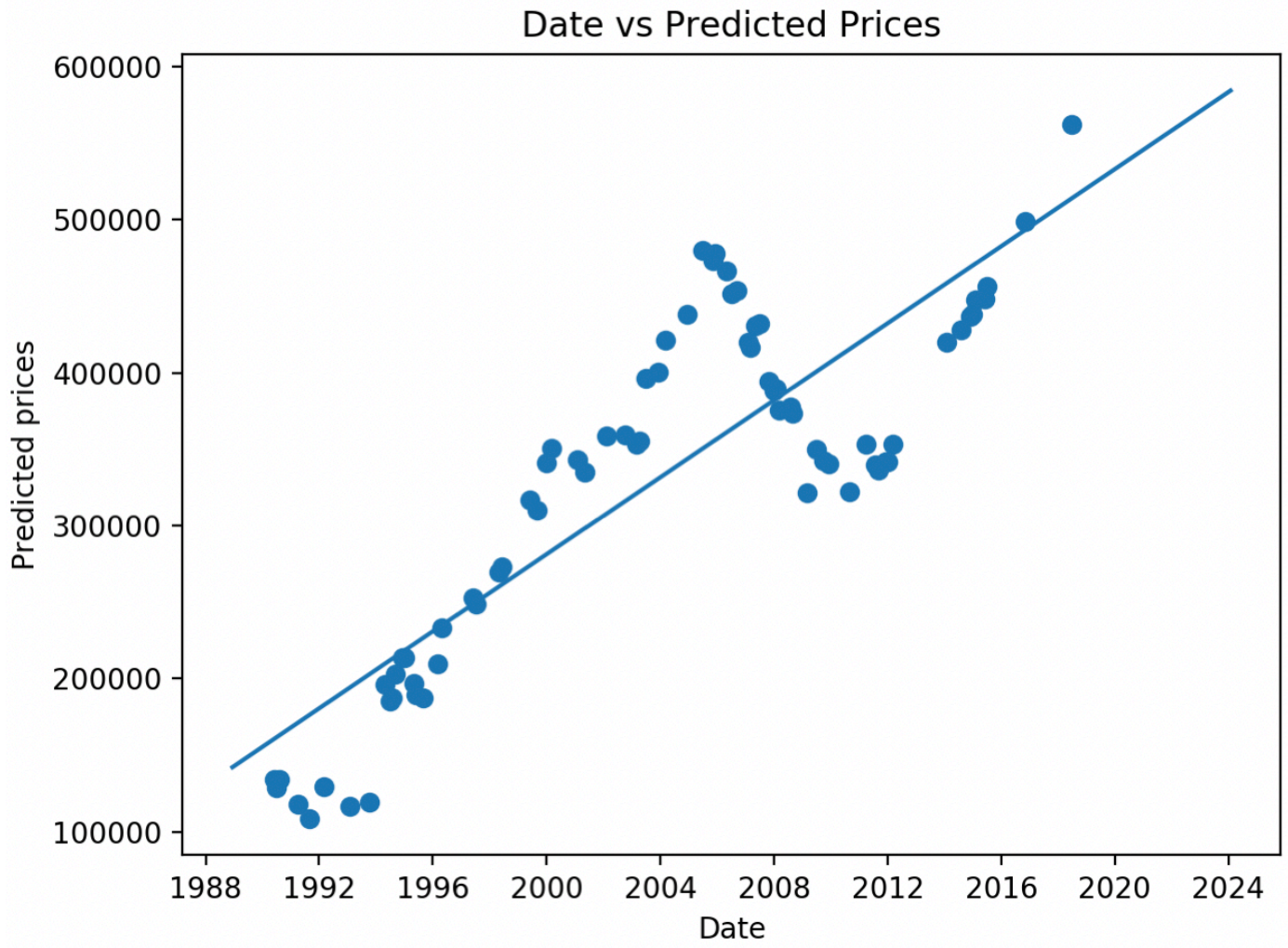
Multiple Linear Regression

- For this model, the dependent values are still the housing prices, but the independent values are date, mortgage rates and the total number of houses that were sold during that period.
- Multiple linear regression was coded using the Python Sci-kit Learn library, the dataset was divided into training and testing set, with 20% of the data being in the testing set.
- Then we fit the data into the model to see the relationship between the actual observed data and the predicted data.
- After the model was created, we calculated the RMSE score to see the error value in the model and the R2 goodness of fit to see how well the model fits the data.

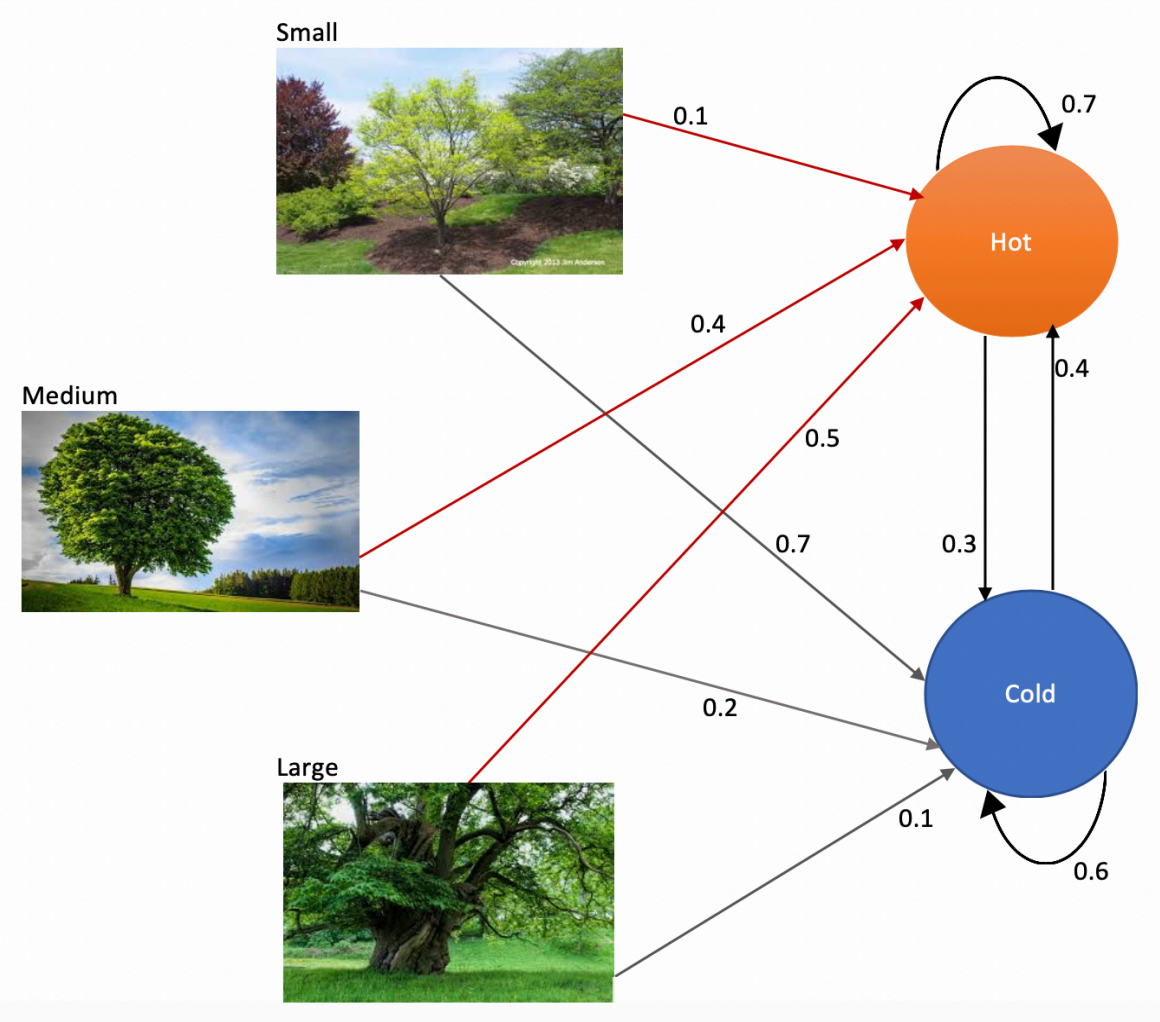
Results of Multiple Linear Regression



Results of Multiple Linear Regression



Hidden Markov Model(HMM)



What is the temperature of a year(Hot/Cold)?

Given:
A: State transition matrix

$$\begin{array}{c} \\ H \\ C \end{array} \begin{array}{c} \begin{array}{cc} H & C \end{array} \\ \begin{bmatrix} 0.7 & 0.3 \\ 0.4 & 0.6 \end{bmatrix} \end{array}$$

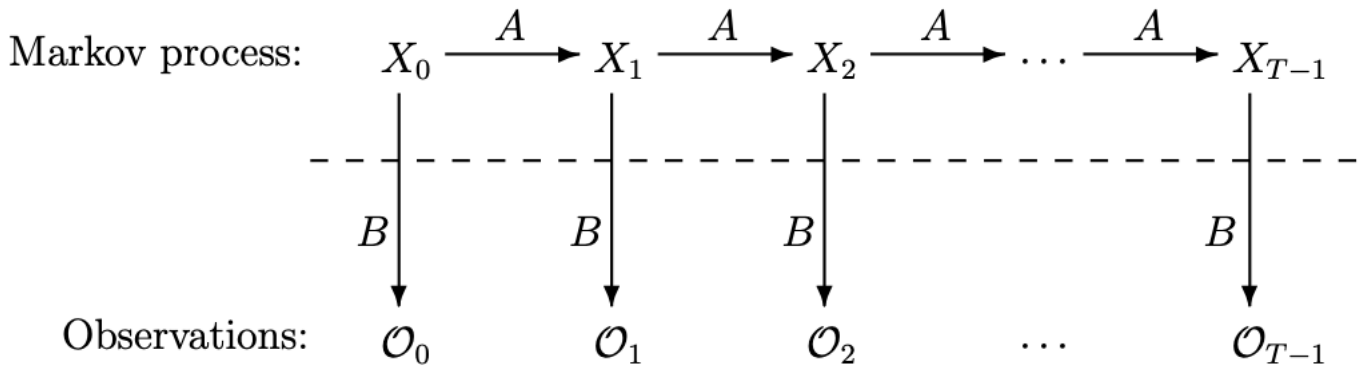
B: Observation emission matrix

$$\begin{array}{c} \\ H \\ C \end{array} \begin{array}{c} \begin{array}{ccc} S & M & L \end{array} \\ \begin{bmatrix} 0.1 & 0.4 & 0.5 \\ 0.7 & 0.2 & 0.1 \end{bmatrix} \end{array}$$

π: Initial state distribution matrix

$$\begin{array}{cc} H & C \\ [0.6 & 0.4] \end{array}$$

Hidden Markov Model(HMM) cont'd



X_i represent the hidden state sequence. The Markov process—which is hidden behind the dashed line—is determined by the current state and the A matrix. We are only able to observe the O_i , which are related to the (hidden) states of the Markov process by the matrix B.

Hidden Markov Model(HMM) cont’d

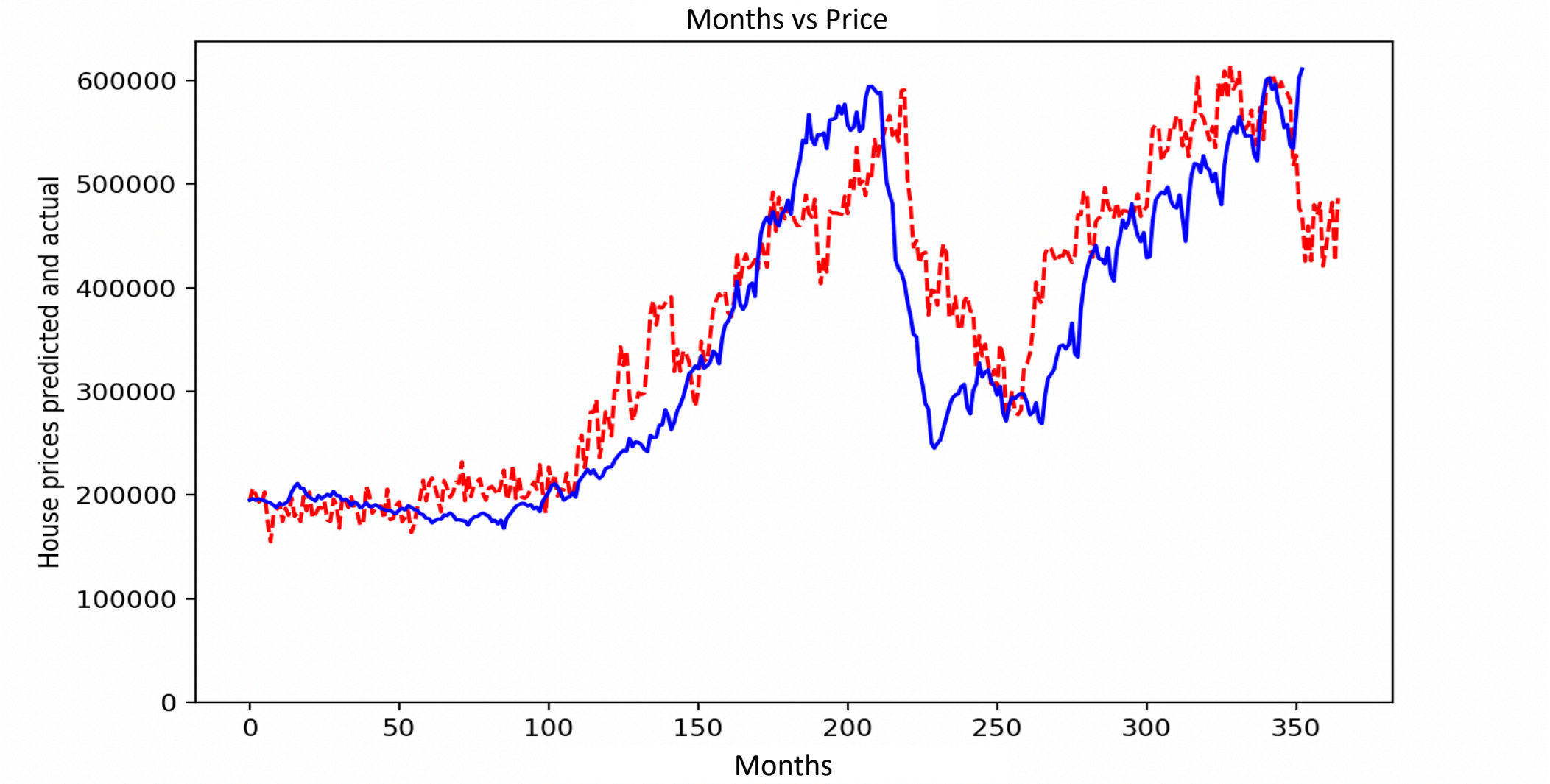
There are three fundamental problems for HMMs:

- Given the model parameters and observed data, estimate the optimal sequence of hidden states.
- Given the model parameters and observed data, calculate the model likelihood.
- Given just the observed data, estimate the model parameters.

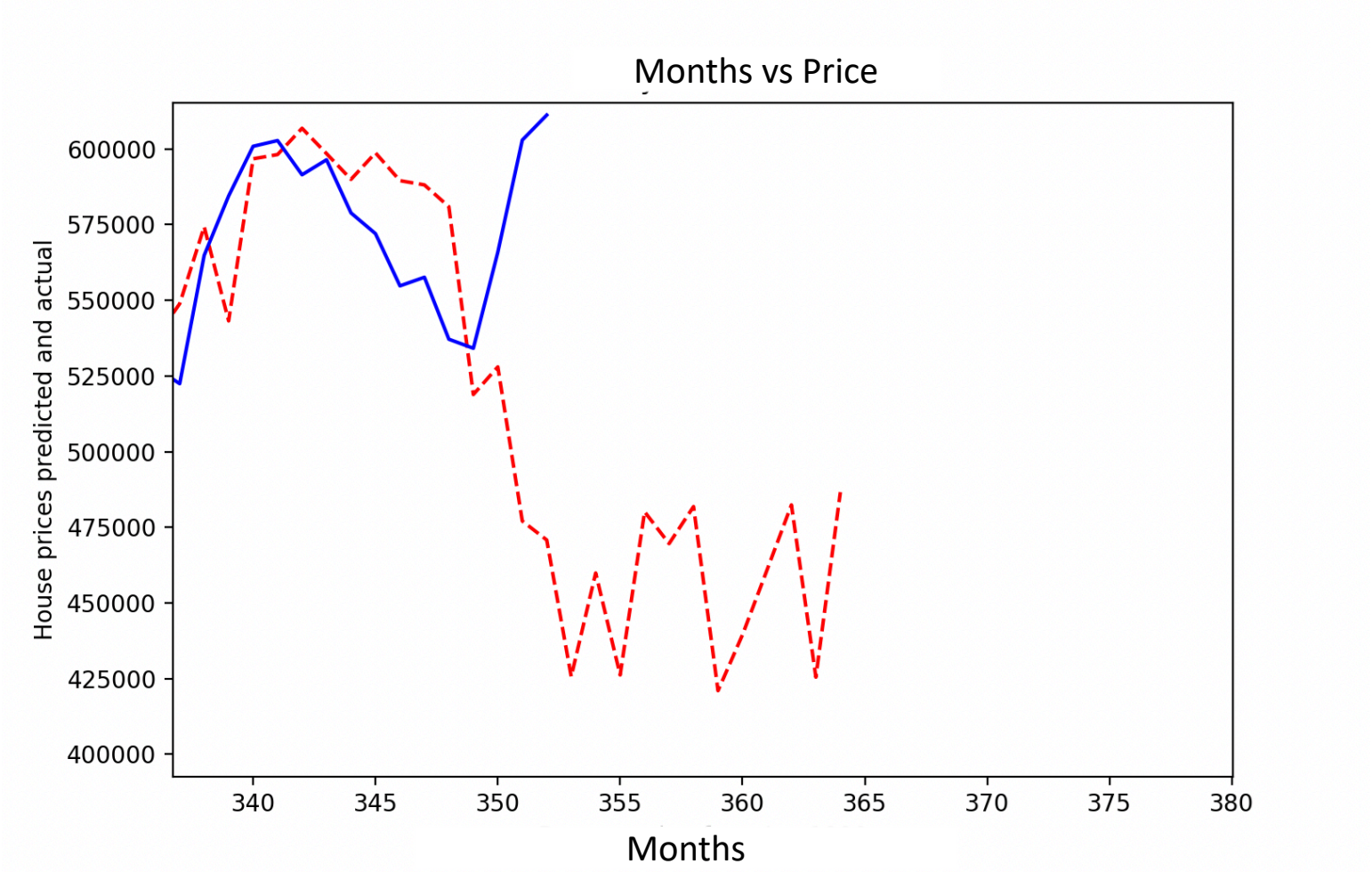
HMM coding

- We used the HMM from hmmlearn.hmm module of Sci-kit learn to apply it to the housing dataset.
- We have added a percentage difference in price to the housing dataset to build the model.
- Therefore the data used to build this model is a column stack of diff_percentages, prices, num_of_houses_sold, rate.

Result of Hidden Markov Model(HMM)

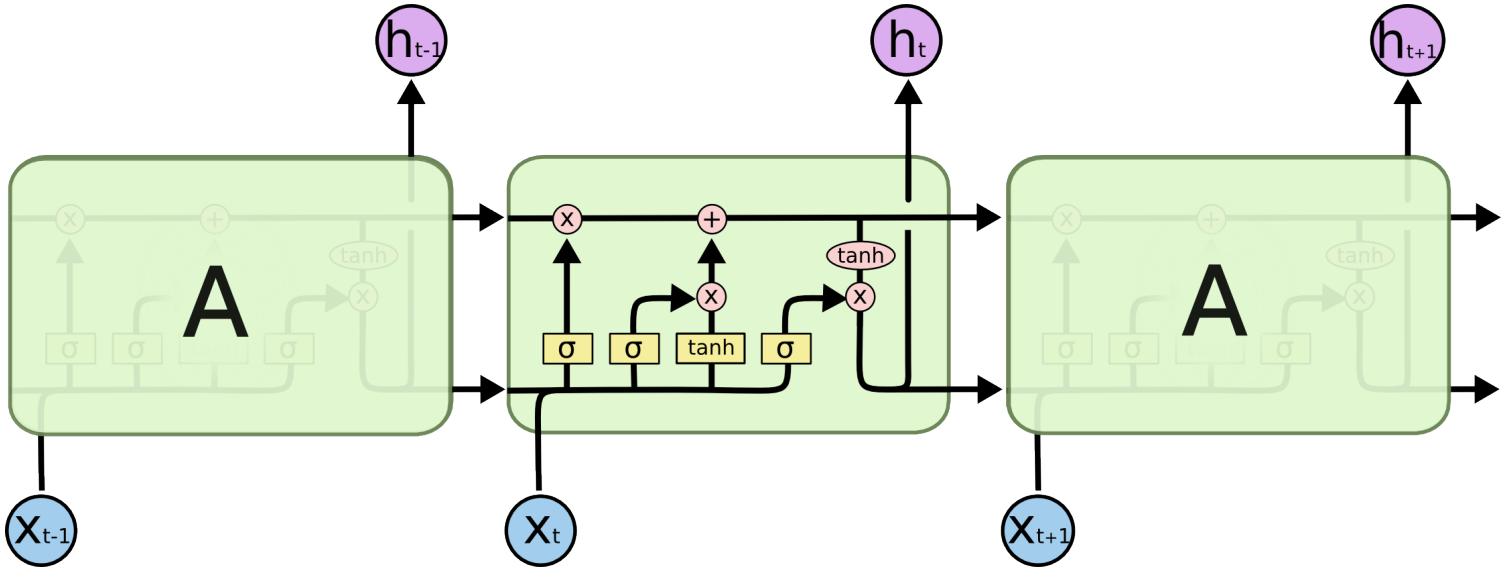


Result of Hidden Markov Model(HMM)



Long-short Term Memory(LSTM)

- Long short-term memory (LSTM) is a type of recurrent neural network (RNN) that is mostly used in the field of deep learning.
- LSTMs help preserve the error that can be backpropagated through time and layers.
- Also, not being sensitive to gap-length makes LSTM superior than RNNs and Hidden Markov Models.
- LSTMs are well-suited for classifying, processing and making predictions on time series data, since there can be gaps of unknown duration between important events in a time series.

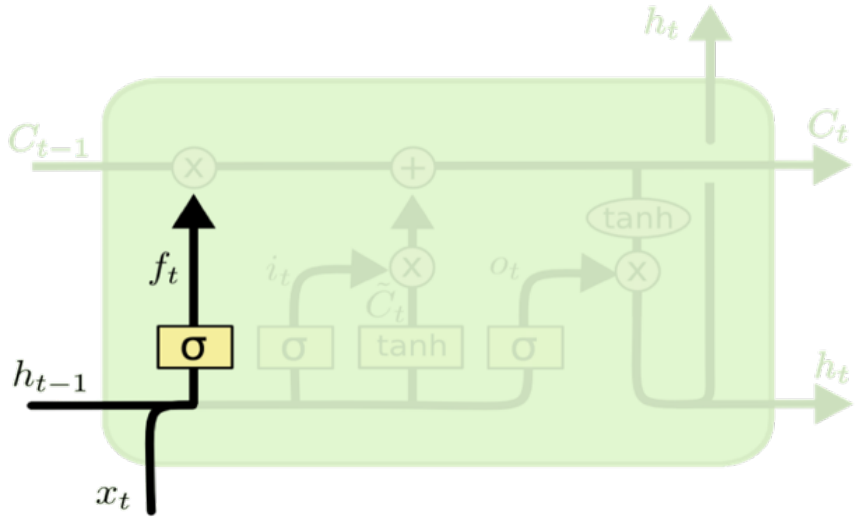


Long-short Term Memory(LSTM)

- An LSTM network typically has a cell, an input gate, an output gate and a forget gate.
- The cell remembers values over arbitrary time intervals and the three gates regulate the flow of information into and out of the cell.
- The cell keeps track of the dependencies between all the elements of the input sequence.
- Next the input gate checks the amount of new information flow into the cell.
- Then the forget gate controls how long the information can stay in the cell.
- Finally, the output gate checks the amount to which the values in the cell are used to compute the final output to the next cell.

LSTM Walkthrough

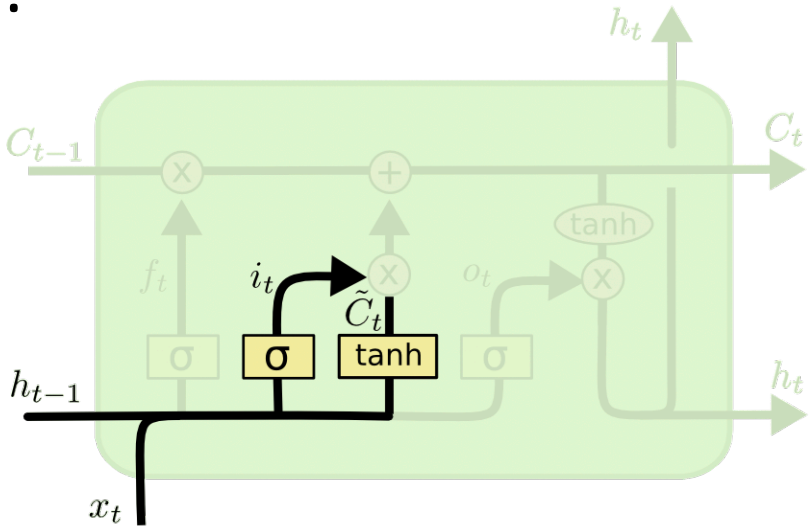
- LSTM must decide on what information is going to stay in the cell state and what information needs to be dumped.
- This decision is made by the forget gate or the sigmoid layer.



$$f_t = \sigma\left(W_f \cdot [h_{t-1}, x_t] \; + \; b_f\right)$$

LSTM Walkthrough(cont'd)

- The next layer of LSTM decides what information is to be stored in the cell state of LSTM network. This is done in two parts. First the input gate layer decides what information/values needs to be updated.
- Then the tanh layer creates $\tilde{C}_t$ a candidate vector, that is added to the state.

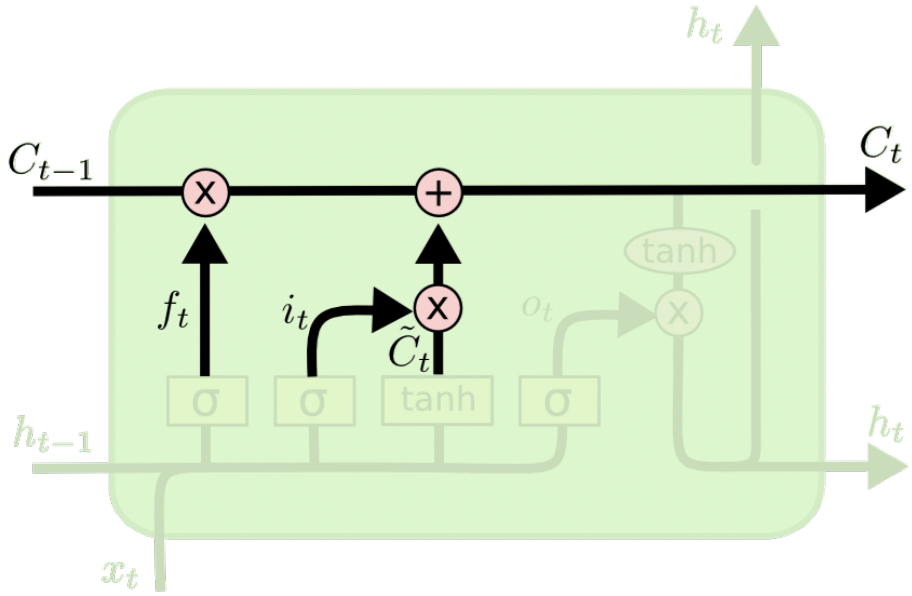


$$i_t = \sigma\left(W_i \cdot [h_{t-1}, x_t] \; + \; b_i\right)$$

$$\tilde{C}_t = \tanh(W_C \cdot [h_{t-1}, x_t] \; + \; b_C)$$

LSTM Walkthrough(cont'd)

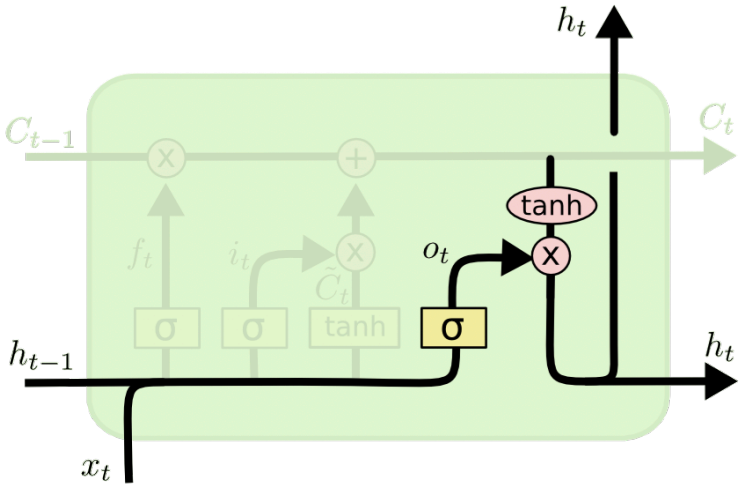
- Next the old state of the cell C_{t-1} is updated to the new cell state C_t. The previous steps gave us all the essential parameters to this.
- The old state is multiplied by the output of the forget layer and then it is added to the value we get from multiplying the input layer value to the candidate vector.



$$C_t = f_t * C_{t-1} + i_t * \tilde{C}_t$$

LSTM Walkthrough(cont'd)

- Finally, the output layer outputs the value of the current cell state to the next cell. This is also done in two steps firstly; the sigmoid layer decides what parts of the cell state is going to the output layer.
- Then, the cell state is sent through the tanh layer (to push the values to be between −1 and 1) and is multiplied by the output of the sigmoid layer.



$$o_t = \sigma\left(W_o\ [h_{t-1}, x_t]\ +\ b_o\right)$$

$$h_t = o_t * \tanh\left(C_t\right)$$

LSTM coding

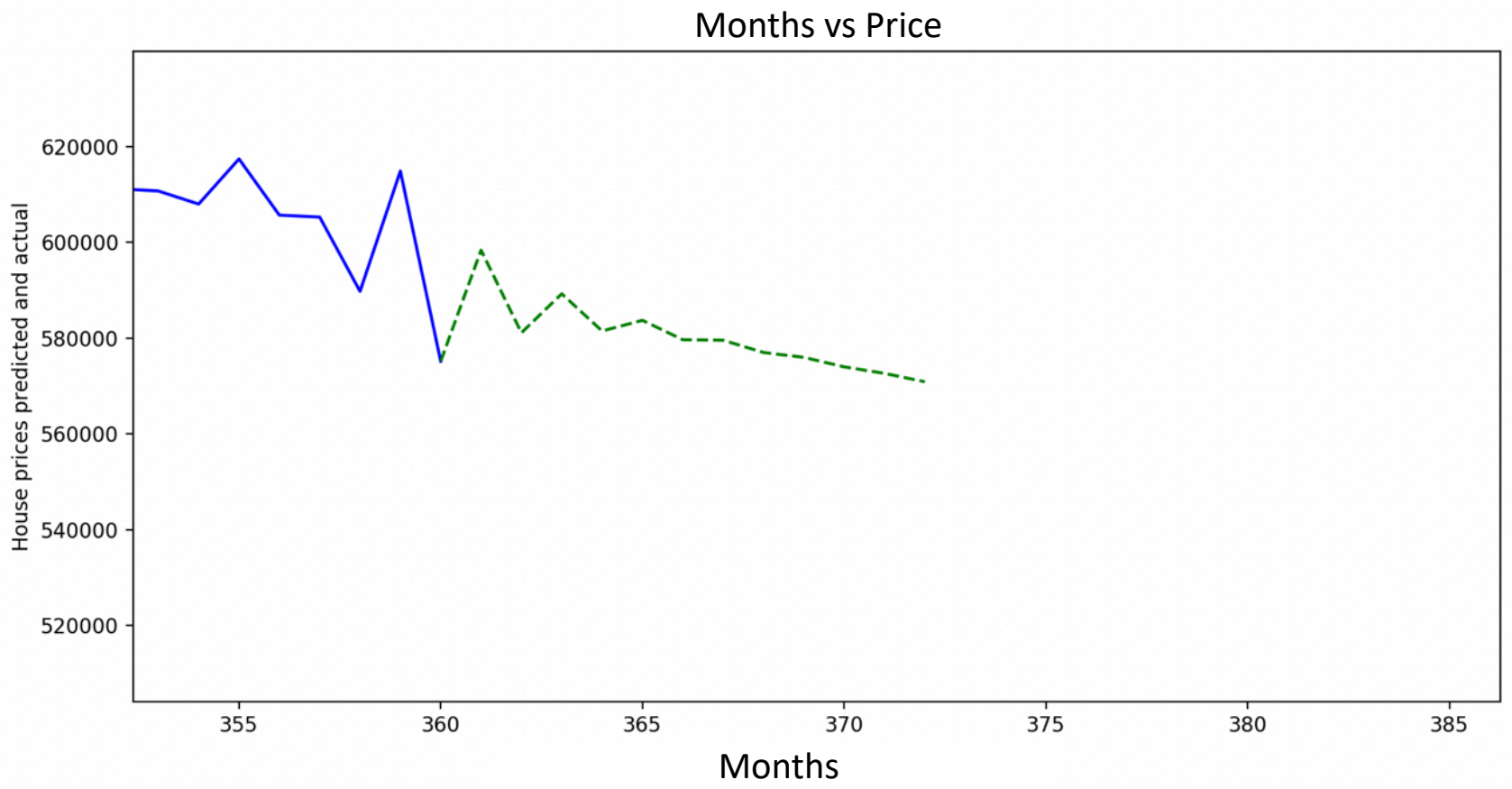
- We divided the dataset into training data and testing data with testing data being 15% of the whole dataset.
- Then we processed the training and testing data by feeding it to the time series generator of Keras sequence generator.
- To build the LSTM network a Sequential model from keras was chosen and to that model LSTM network was added.
- The weights that are given to initial Keras network is uniformly divided within each layer which is given by *init='uniform'*.

```
model = Sequential()
model.add(LSTM(50, activation='relu', input_shape=(look_back, 1),
               init='uniform'))
model.add(Dense(1))
model.compile(loss='mean_squared_error', optimizer='adam')
```

Result of Long-short Term Memory(LSTM)

Result of Long-short Term Memory(LSTM)

Result of Long-short Term Memory(LSTM)



Results and Discussion

Model Name	Prediction	Time to train	Efficiency	R-squared score
Linear Regression	House prices will Eventually rise	Low	Medium	0.76
HMM	House prices will fall	Low	Medium-Low	0.706
LSTM	House Prices will fall Slightly	High	High	0.92

Conclusion

- Financial crisis and housing market crisis are closely tied together and have a huge impact on economy.
- The techniques discussed here can help us to forecast the housing prices for the future. From all the graphs and prediction models, we can foresee that there will be a fall in the house prices for the next year.
- But it won’t be as bad as that of 2008 because the banks this time around are taking every precaution to prevent a crisis like that of 2008.

References

[1] Y. Demyanyk and I. Hasan, “Financial crises and bank failures: A review of prediction methods”, *Omega*, vol. 38, issue 5, pp.315-324, 2010.

[2] E.J. Schoen, "The 2007–2009 Financial Crisis: An Erosion of Ethics: A Case Study", *J. Bus. Ethics*, vol. 147, pp. 805-830, Dec 2017.

[3] M. Zhang and K. Xu, “High order Hidden Markov Model for trend prediction in financial time series”, *Physica A: Stat. Mech. and its Appl.*, vol. 517, pp.1-12, 2019.

[4] M.R. Hasan and B. Nath, “Stock market forecasting using Hidden Markov Model: A New Approach”, *5th Intl. Conf. on Intel. Sys. Design and Appl., IEEE*, 2006.

[5] F.A. Gers, D. Eck, J. Schmidhuber, "Applying LSTM to time series predictable through Time- Window approaches", *Perspectives in Neural Comput., Springer*, vol. 1, pp. 193- 200, 2002.

[6] Y. Hu, X.Sun, X. Nie, Y. Lweand L. Liu, “An Enhanced LSTM for Trend Following of Time Series”, *IEEEAccess,* IEEE, 2019.

[7] Y. Demyanyk, “Quick exits of subprime mortgages” *Fed. Res. Bank of St. Louis Rev.*, vol. 92, 2008.

[8] M.G. Crouhy, R.A. Jarrow and S.M. Turnbull, “The Subprime Credit Crisis of 2007”, *J. of Deriv*, pp. 81-110, 2008.

[9] E.P. Davis, D. Karim, “Could early warning systems have helped to predict the sub- prime crisis?”, *Ntl. Inst. Econ. Rev.*, vol. 206, pp. 35–47, 2008.

[10] R.Nyman and P.Ormerod, "Predicting economic recessions using machine learning algorithms", Dec 2016.

[11] Housing price dataset: https://www.car.org/marketdata/data/housingdata

[12] Mortgage interest rate dataset: https://fred.stlouisfed.org/series/MORTGAGE30US

[13] Total houses sold dataset: https://ycharts.com/indicators/new_homes_sold_in_the_us

[14] M.Stamp, “A Revealing Introduction to Hidden Markov Models”, Oct 2018.

[15] C.Olah, “Understanding LSTM Networks”, Aug 2015.

Thank you