MACHINE LEARNING BASED PERSPECTIVES FOR HOUSING MARKET CRASH PREDICTION

SIPRA SAHOO[1], MADHURI RAO[2], SMITA RATH[3,*], DEEPAK KUMAR PATEL[4], MITRABINDA KHUNTIA[5], SHRABANEE SWAGATIKA[6],SUSHREE SANGITA JENA[7], PRABHAT KUMAR SAHU[8]

[1,5,6]Associate Professor, Siksha 'O' Anusandhan Deemed to be University, Department of Computer Science and Engineering , India
[3,4,8]Associate Professor, Siksha 'O' Anusandhan Deemed to be University, Department of Computer Science and Information Technology, India
[2] Senior Assistant Professor, Department of Computer Engineering and Technology, Dr. Vishwanath Karad MIT World Peace University, Pune, Maharashtra, India.
[7]Assistant Professor, Department of Computer Science & Engineering , GITA Autonomous College , Bhubaneswar, Odisha, India
E-mail: [1]siprasahoo@soa.ac.in, [2]madhurirao.iter.soa@gmail.com, [3,*]smitarath@soa.ac.in, [4]deepakpatel@soa.ac.in,[5]mitrabindakhuntia@soa.ac.in, [6]shrabaneeswagatika@soa.ac.in, [7]sushree_cse@gita.edu.in, [8]prabhatsahu@soa.ac.in

ABSTRACT

The recession of 2008, commonly referred to as the "Subprime Mortgage Crisis," was the worst to hit the USA and had an impact on the entire world, costing more than 8.4 million people their employment alone in the USA. From late 2006 to the end of 2009, the crises repeatedly forced innocent people to leave their homes, and the Gross Domestic Product (GDP) dropped as low as 4.3%. In order to assist in estimating future market values and prices so that a real estate market recession can be averted, this working paper analyses old and present prices, levels, and interest rates to determine at what rate the market could have been rescued from falling. A house market crash prediction with machine learning techniques including Linear Regression, Hidden Markov Model, and Long Short Term Memory is presented here.

Keywords: *Hidden Markov Model; Linear Regression, Long Short-Term Memory; Mortgage-Backed Securities (MBS); House Market Crash*

1. INTRODUCTION

Home values declined by 31.8% in 2007, and although interest rates increased in step with market value, home values eventually started to decline, making it impossible for borrowers to repay their subprime mortgages. Lenders suffered losses as a result of the thousands of delinquent borrowers. Mortgage-backed securities (MBSs), which were backed by these hazardous subprime loans, were transferred by lending institutions to other institutions in an effort to profit from rising house prices. When that did not occur, MBSs also lost value, resulting in the failure of numerous banks. Following the fall of the real estate market, banks and companies lost faith in one another, which decreased stock values. The Reserve Primary Fund came under attack because investors were withdrawing money too quickly, therefore the federal government of the USA began to use the term in both of its houses as "Bank bailout." They were concerned that the Fund's stock market investments would cause it to fail, much like Lehman Brothers did. Despite having experienced a great number of recessions in the past, the United States continues to have one of the fastest growing economies in the world. The biggest blow to the USA came in 2008 as a result of bad bank practices, subprime mortgages and their market value, sales of mortgage securities, and market betting even while subprime mortgage rates continued to rise. For people who cannot afford prime mortgages, there exist subprime mortgages. The following are the causes: not being able to pay the down payment on a house because your credit score is less than 650. There are ripple effects, and the crisis caused numerous institutions, including Bear & Stearns and Lehmann Brothers, to fail.

Leaving people without work, in housing, and with a financial loss for five years. Backward banking practices and the exploitation of current mortgage credit directives for market and asset investment were to blame for the housing crisis of 2008. If we are aware of a bubble bust and foresee it far in advance, these bubbles could have been avoided

Highlights:

- In this research, four methodologies to forecast and comprehend the market's dynamic are examined namely - Keynesian, Farmerian, MBS, and traches.
- We construct three machine-learning models: LSTM (Long Short-Term Memory), HMM (Hidden Markov), and linear regression.
- Using historical data and focusing on the financial slump of 2008, the prediction of the housing market meltdown was made with LSTM, HMM, and Linear Regression.
- This research provides a new edge over complicated financial structures that can assist economist in making robust financial predictions and products.

2. LITERARTURE REVIEW

Learning about the investors, markets, CDOs, Mortgage-Backed Securities, S&P credit scores, the bank bailout, Tranches system, how subprime mortgage rates were high while the market was simple, and the flawed bank policies will help you better comprehend what actually transpired before it happened. [1]. For a deeper understanding of the financial crisis and how banks played a big role in it, we will be looking into the history of such policies. In this section, we'll also examine how using relevant data sets, machine-learning models could be used to anticipate future financial crises similar to the one of 2008. [2]. The issue arose when consumers began obtaining subprime mortgages without having their credit scores verified. The average credit score should be over 700, but since these loans could be obtained with a credit score as low as 400, anyone could suddenly purchase a home. Because the banks were hesitant to lend to one another, the US government and the gift federal bank were forced to purchase these subprime mortgages laden with BBB and BB tranches valued at 95% from the start [3]. This caused tension between the banks for having more business, so started competing with each other producing vast numbers of Brookers and agents to find the clients, the home prices started climbing high the sky made it may fair for the private investors and CDOs [4]. A highly structured and problematic instrument in the banking industry that may be used to wager for or against the market, which is a pool of loans and other assets that are offered to institutional investors as a collateralized debt obligation (CDO). Housing prices' unexpected inflation caused a housing bubble to burst. Fraudulent financial techniques that offered extremely high loan rates. Financial laws and regulations contained flaws. Instead of examining profiles, the FED and large banks have policies in place to disseminate more subprime mortgages. Machine Learning could be a very important tool for predicting house market crash compared to the traditional tools of assessment and prediction. **Table 1** presents a survey of machine learning techniques that have been explored in various financial areas. In the case of the 2008 market crisis, machine learning was a rumored term in the market. It was not a real time use for forecasting the market. The later market used Shapley values to study the market, but isn't multidimensional enough as CDOs still existed. The neglecting attitude and faulty policies of banks, not doing a back search of the client and not maintaining the given credit score of 750+ raised the subprime loans instead of prime loans for all AAA tranches and filled them with BB and BBB then selling the swap bonds. Now the Shapley values is a machine learning model for checking the client and credit score, bank balance, salary, expenditure to verify if they are eligible for a prime loan. Now not every time Shapley values accuracy is the point, it can be faked through. In addition, the crisis may arise again. Adding Linear Regression is to find the market trend analysis of GDP growth. The Hidden Markov Model (HMM) has already been used in weather forecasting before; this is also used to predict market crashes. LSTM model is something, which is still in the works by many economists, does not calculate behavioural economics, and so will not be sufficient, so we are trying to see if it works with 30 years data [3].

Research Question RQ 1:
Why is HMM, LSTM and LR considered specifically in this study for house market crash predictions?
Machine Learning Techniques can be classified as supervised, unsupervised and time series based for this problem. If historical crashes are labelled then supervised learning techniques such as Liner Regression, Random Forest and Neural Network can be explored. For detecting anomalies,

unsupervised learning models such as Hidden Markov Model, K-means and auto-encoders could be useful. LSTM is a tome series model that could be useful for deep learning sequence analysis. In this paper we present an exhaustive study of machine learning approaches for understanding the impact of analysis in house market crash prediction. Hence, HMM, LSTM and LR are considered specifically.

3. PROPOSE MODEL

The goal is simple verifiable lodging costs, the mortgage rates and the houses sold and its dataset to utilize ML procedures to forecast the upcoming recession (if there any) and how much is the recession quality with respect to 2008. **Figure 1** depicts our proposed model for this study.

3.1 Datasets

The datasets utilized in this study are a combination of a few datasets that contained data from various governments, such as mortgage interest rates and the number of US states' information on housing expenses. We slithered through a website to find out the total number of single-family homes sold. We explored the following dataset from Housing cost, Mortgage rate and Total number of house sold. We used Pandas, a Python information-examination module, to encode these datasets. The data sets covered the years 1990 through 2020, and the time span was one month. To combine each dataset on the dates, various information designs were applied; we then combined all the dates into a single arrangement. We have the month-year for knowing and acknowledging the month raise sale of each year from 1990 [10, 11]. The following observations were noted from the dataset considered for this study- One family house sold: The average of all sales percentages for that month's single-family home sales in each state.

Rate: The mortgage rate value at the beginning of each month, which serves as the monthly mortgage rate.
Observation date: The observation date, or first day of the month, is the day on which we compute the mortgage value for that month.
Total houses sold: the total number of homes sold by all the states each month.
Month and year: Calculating the month and the year is step one.
Period: Period uses splitting for later applications in date-time formats.

3.2 Machine Learning Models

Machine Learning models such as LSTM and LR are widely explored in many prediction problems such as stock market prices prediction. ML Models are categorized as supervised, unsupervised and reinforced. In this study three techniques are applied namely- Long Short Term Memory, Hidden Markov Model and Linear Regression, each of which is explained below.

3.2.1 Long Short Term Memory (LSTM)

It is a special form of RNN that functions like a neural network, and it does not only make simple associations. Since there may be gaps in a period series and a time interval between significant events, this model is in charge of characterising, processing, and forecasting forecasts on time-series information. Since there may be gaps in the unclear interval between important events in a period series [12] [13]. LSTM are suitable for characterising, managing, and making forecasts on time-series information. An input gate, an output gate, and a forget door are all components of an LSTM network. The cell, which recalls values across irregular time intervals, is one of the three gates that regulate the flow of data into and out of the Phone. [14,15]. Recurrent neural networks (RNNs), such as the LSTM model, are frequently employed in time-series arrangements for deep learning. Despite the fact that its design is comparable to RNN, LSTMs have critique associations rather than only feed-forward linkages. RNNs bear a significant issue of evaporating inclination, which prompted the prevalence of LSTM. The key factor causing the inclination to evaporate was the requirement for data to sustain a flow consecutively across all cells before reaching the current cells. Going long distances often contaminated data by being copied by junk values (smaller numbers than 0). Similarly, unlike traditional RNNs, LSTMs can benefit from long-distance circumstances. LSTMs aid in preventing errors from being back propagated across layers and time. By maintaining a steadier blunder, they enable intermittent nets to continue learning across numerous time steps. Additionally, LSTM is superior to RNNs and Hidden Markov Models since it is not sensitive to hole length. Since there may be gaps in the unclear interval between key events in a period series, LSTMs are suitable for characterising, managing, and making forecasts on time-series information [13]. An LSTM network typically has a cell, an input gate, an output gate, and a forget door. The three gates regulate the flow of data into and out of the phone, and the cell recalls values over variable time intervals. The cell keeps track of the interactions between each element in the information structure. The input gate then determines how much additional new data stream can enter the cell. Further, the forget gate then regulates the duration of the data stored inside the cell. Then, the output gate checks the extent to which the qualities within the cell are being used to record the outcome toward the cell after it. There are associations all through the LSTM gates. Loads of these associations, which should be picked up during preparation, decide the mechanism of the gates [4].

Procedure 1: What information will remain in the cell state and what information needs to be discharged must be decided by the LSTM. Either the forget gate or the sigmoid layer decides this. [14]. It takes into consideration about h_{t-1} and x_t of the LSTM which gives outputs as a number between 0 and 1 for each number in the cell state after examining C_{t-1}. 1 denotes keeping all information, while a 0 denotes not keeping any information [8].

Procedure 2: The LSTM network's next layer determines what data should be saved in each cell. It's completed in two steps. The input gate layer determines what data or values need to be updated first. Further, layer generates $C_t\sim$ and the layer then generates a candidate vector and adds it to the state.[13].

Procedure 3: Then, the previous state cell C_{t-1} is updated to the most recent state of the cell.C_t. All the necessary elements for this came from the previous steps. The value obtained by multiplying the input layer value by the candidate vector plus the previous state's value times the forgotten layer's output value. Beneath the figure depicts this [13].

Procedure 4: The output layer then sends the value of the current cell state value to the following cell. Additionally, this is carried out in two processes. First, the sigmoid layer selects the components of the cell state that will be sent to the output layer. The cell state is then multiplied with the sigmoid layer's output after being delivered through the layer tan h (to have values between 1 and 1).

$$i_t = \sigma\ (w_i\ [\ h_{t-1},\ x_t] + b_i\ ) \qquad \text{Eq.(1)}$$

$$f_t = \sigma\ (w_f\ [\ h_{t-1},\ x_t] + b_f) \qquad \text{Eq. (2)}$$

$$o_t = \sigma\ (w_o\ [\ h_{t-1},\ x_t] + b_o) \qquad \text{Eq. (3)}$$

Here, i_t , signifies the input gate. f_t , o_t expresses the forget gate and output gate respectively whereas σ here signifies the sigmoid function. w_x indicates the weight for the respective gate(s), h_{t-1} indicates the output of the previous LSTM block at t-1. x_t indicates the input at the current timestamp and b_x denoteste bias for the respective gate Eq.(1) represents the input gate, while Eq.(2), represents the forget gate and Eq.(3) represents the output gate. Eq(1) indeed suggests that new information will be stored in the cell state, whereas Eq.(2) recommends what needs to be deleted from the cell state. Eq.(3) serves as an activation to the final output of the LSTM block and is expressed in Eq.(6).

$$\check{C}_t = \tanh\ (w_c\ [h_{t-1},\ x_t] + b_c\ ) \qquad \text{Eq. (4)}$$

$$c_t = f_c * C_{t-1} + i_t * \check{C}_t \qquad \text{Eq.(5)}$$

$$h_t = O_t * \tanh\ (c^t) \qquad \text{Eq.(6)}$$

Here c_t as derived in Eq. (5) represents the cell state or memory at time stamp t. In Eq. (6) $\check{C}_t$ denotes the candidate for cell state at time stamp (t).Here Keras was explored to construct and model the LSTM network, therefore the specifics of the various LSTM variants are not necessary. The LSTM model for prediction using the housing dataset is presented next[8]. The neural network library Keras, which has a TensorFlow backend, was used to construct the LSTM model for this project. Given that Python is the primary programming language used for this project, this deep learning package is helpful for it [14]. Keras' latest Consequently, before moving on to the project data, we initially utilise *t* on a tiny dataset to make predictions in order to understand about the dynamics and operation of Keras. We decided to apply Keras on the project dataset after thoroughly understanding the ideas required for this project. We split the dataset into training and testing data for the project, with testing data making up 15% of the total dataset. After that, we processed the training and test data by supplying it to the Keras sequence generator's time series generator [15]. A Sequential model from Keras was selected for the LSTM network construction, and to that model an LSTM network was built with an input shape of 51 and 50 hidden nodes within the LSTM cell. The input shape describes the input to the network's first layer. The size of the epoch for the neural network is then determined. The number of epochs is 1000 for this. One loop across the entire training data makes up an epoch. A single epoch would result in underfitting because t would have little opportunity to learn from the training set. Additionally, using too many epochs might result in overfitting since there would be too much learning from the training data that might not generalise to the testing data. The model is then adjusted to the number of epochs and the cleaned training data. After creating the model and training it, we test it using data to observe how it behaves before making predictions. Then, extend the forecast to include a one-year housing market forecast.

3.2.2 Hidden Markov Model (HMM)

In this paradigm, the present state that is provided for usage has no relationship to historical data or data for the future. Given that there are observed states, The likelihood of the hidden state and the likelihood of a change from one state to the next are understood using HMM. HMM computation will provide the annual temperature (Hot/Cold), providing insight into the size of a tree's ring as it develops (Small, Medium, and Large) [16]. Most significantly, we will receive the state change grid, which will enable us to determine whether it is feasible to anticipate whether a given year's hot weather will be followed by a subsequent year of cold weather. In essence, if it is chilly this year, it is possible that it will be hot or cold the following year [18]. The state transition matrix is what is meant by this. Additionally, we will have a perception framework that will provide us with the likelihood that it will be hot or cold based on whether the tree ring is small, medium, or large [17]. For the Housing market linear Regression used to understand the elements.

Consider a Set of States: $\{ S_1, S_2, ...S_N\}$ Eq(7)

A set of sequences of state is generated as in Eq(8) when processes move from one state to another.

Sequence of States : $\{ S_{i1}, S_{i2}, ..., S_{ik}, ...\}$ Eq(8)

In HMM, the probability of each of the subsequent states depends only on what was in the previous state as depicted in Eq. (9).

$$P(S_{ik}|S_{i1}, S_{i2}, ...,S_{ik-1})=P(S_{ik}|S_{ik-1}) \quad \text{Eq(9)}$$

HMM is a factual demonstrating method, which gets its name from Andrey Markov, who thought of Markov chains. Gee is an assortment of Markov chains that delivers the likelihood of the following grouping relying upon the current states. Given that the current state is known, according to the Markov model, the past and future have no link [16]. Given that there are observed states and probabilities of the change starting with one state and moving on to the next, HMMs are used to predict the likelihood of a secret state. We experimented using these two stages in CS 297. To understand how the HMM works, a model from Prof. Stamps' research was used to code the HMM [17]. Sci-kit HMM was utilized to learn how to code and show the lodging dataset's information. HMM computation will provide the annual temperature (Hot/Cold), providing insight into the size of a tree's ring as it develops (Small, Medium, and Large). Most importantly, we will receive the state change grid, which will enable us to determine whether it is possible to predict whether it will be hot or cool the following year when it is hot one year. In essence, if it's chilly this year, it's possible that it will be hot or cold the following year [17]. The state transition matrix refers to this. Additionally, we will have a framework for perception that will provide us with the likelihood of the temperature being hot or cold based on the size of the tree's ring, whether it is little, medium, or large. Following a Markov transition matrix, a Markov chain transitions from one step to the next that is a n*n matrix. The Markov transition matrix is a matrix with probability of moving from one state to another. The states in HMM that are invisible and can generate one of the M observations which are visible states $\{V_1, V_2 , ..., V_m \}$. Hence in HMM following probabilities are often to be specified.

Matrix of transition Probabilities:

$$A= (a_{ij}) , a_{ij} = P(S_i | S_j) \quad \text{Eq. (10)}$$

Matrix of observational probabilities:

$$B = (b_i (V_m)) , b_i (V_m) = P(V_m | S_i) \quad \text{Eq.(11)}$$

Vector of initial probabilities :

$$\pi = (\pi_i), \pi_i = P (S_i) \quad \text{Eq.(12)}$$

Hence HMM is represented as in Eq.(13).

$$M =(A,B,\pi) \quad \text{Eq.(13)}$$

An observation matrix is a n m matrix that provides the likelihood that a certain state will occur given the observation. In this instance, we would also be citing this passage from Prof. Stamp's study [16]. The initial state distribution matrix (π), which will allow us to begin the calculation for the following steps, is also required in addition to the state transition matrix (A) and observation emission matrix (B). These matrices are all row stochastic, which means that in each n-fold addition of a row, it will always equal 1. After then, a series of observations will be made. Three problems that are covered in the Proofs Stamp paper must be solved in order to quantify the hidden states. Python was used for the coding of the aforementioned issues. We first prepared the data in order to solve the aforementioned issues discussed in the paper. After gathering the data, we multiplied the A matrix to determine the alpha pass or forward algorithm [17]. in relation to the likelihood of occurrence. The beta-pass, also known as the backward method, was computed after the alpha-pass. By starting the matrix backwards, the backward algorithm is calculated. The sole difference between the calculation and the alpha pass is the matrix's starting point. We would use these two matrices to calculate the gamma and di-gamma after computing them. The best-fit values for the model are discovered using di-gammas. We use the alpha matrix for all the values prior to gamma and the beta matrix for all the values subsequent to gamma in order to construct di-gammas. To obtain the di-gamma, first subtract the beta matrix from the alpha matrix. Once the di-gammas have been identified, the HMM model is scaled and the original matrix is updated with the estimated values. As a result, the A, B, and Pi values are updated after each calculation. These modified numbers would be used in our calculations for the following steps [16]. The second phase of the challenge started when we had coded the HMM algorithm and understood how it operated. We applied it on the housing dataset using the HMM from the hmmlearn.hmm module of Sci-kit Learn [17]. To develop the model for this, we included a percentage difference in price to the housing dataset. A column stack of diff percentages, prices, number of houses sold, and rate data were utilised to develop this model. After that, we built the model with Gaussian HMM.

3.2.3 Linear Regression

This is a managed learning procedure that builds a direct connection between the reliant or scalar and the free or logical factors. In the event that there is one autonomous variable, the displaying strategy *s* is *called* straightforward direct relapse [18]. For this situation, the scalar variable *s* is reliant upon only one illustrative variable. Simple linear regression model is expressed by Eq. (14).

$$y = b_o + b_1.x_1 \qquad \text{Eq.(14)}$$

The point of having more than one illustrative variable for a scalar is called different or multivariate Linear Regression. in this model, the scalar variable has more than one illustrative variable. The Multiple Linear Regression model is given as Eq.(15) [18].

$$y = b_o + b_1.x_1 + b_2x_2 + b_3x_3 \ldots\ldots b_nx_n \qquad \text{Eq. (15)}$$

Above condition displays, 'y' as the reliant variable x_1 , $x_2, x_n$are the free factors. The mistake term s the steadyb_0, t s likewise called the y-capture and b_1, b_2, b_n are the loads of the autonomous variable. In this task, we have utilized both simple and multiple linear regression. For both the model, the reliant variable ‘s’ the house cost and the autonomous variable ‘s’ date for the basic direct relapse model. We began with straightforward direct relapse to comprehend the elements of the house value connected with time and the time has impacted the real estate market and coded the calculation as opposed to utilizing sci-unit learn (utilized later for numerous relapse) [18]. The dataset utilized for this ‘s’ the lodging cost dataset where ‘t’ contains house costs for all provinces of California and normal of all California house costs.Limiting the number of squares, blunders or residuals is the main target of this technique. For every x and y value, which represent the information sources and outcomes, the squared aggregates are determined. To scale the variables and reduce error values, a preparation rate is used [18]. This is repeated until the fewest total squares of errors are achieved, and no more progress is imaginable. The coding reveals the methods used to locate the hotel cost least squares regression model. After identifying the best-fit line, we can extrapolate it as far as is necessary to create an expectation. In order to understand how the line looks for the thirty years of verified data, we used the computation on the general lodging cost dataset for basic straight relapse. From that point forward, we extrapolated data for the previous two years in order to examine the hotel expansion beginning in 2018 [19]. In addition to determining the best fit line, we also calculated the Root Means Square Error (RMSE) score, which typically falls between 0 and 1. The difference between noticed and anticipated quality is estimated using RMSE. A

RMSE score of 0 suggests that there is no relationship between the predicted and noticed features, while a score of 1 suggests that the anticipated and noticed qualities are nearly identical. A RMSE score of 1 indicates improbability since it suggests that both pieces of information are identical, which is never going to happen. A model with an RMSE of 0 means that the expected value is completely different from the observed value, which should never be the case as this would indicate that the model was not properly created [18]. Next, we forecasted the cost of lodging using several straight regressions. However, the free qualities are current, contract rates, while the dependent qualities in this part are still the lodging charges. Additionally, the total number of homes sold during that time. The Python Sci-pack Learn package was used to code various direct relapses. In order to do this, the dataset was divided into two sets: a preparing set and a testing set, with 20% of the data in the testing set [19]. The information was then incorporated into two independent models, which were created to examine the relationship between the actual noticed information and the predicted information. After creating the various models, we calculated the RMSE score to assess the model's flaws and the R2 decency of fit to assess how well the model fits the data. The code piece below demonstrates how the aforementioned advancements were encoded [18].

4. RESULTS AND DISCUSSIONS

So after cramming through all the three. models we can see LSTM being the most efficient and giving most accurate results for prediction without being over fit , whereas HMM and linear regression gives the dynamism of the market and the market trends but doesn't give anything about whether there's going to be a recession or not. Although the time taken by the LSTM is the highest as it understands the trend and time series and takes time; whereas HMM and Linear Regression gives the results and prediction output in seconds. Let us see and analyse prices and household numbers before prediction.

Looking at the graph in **Figure 2** we can see how the number of households were at peak in the year 2005 with all the AAA tranches filled with subprime loans. That is where the subprime rate also started increasing and then drastically fell in 2007-2008[5]. The above graph as in **Figure 3**depicts how the R started to decline gradually over time but it can be observed carefully that there's a rise in R 2006-2008 especially when the ABS (Asset-backed securities) e the subprime rates were around 6% and rising. However, after 2008 there is quite a fall of 'n', which is the Mortgage rate as the economy was tried to restore by cutting down the rate [5]. Now this **Figure 4** shows the houses sold in the consecutive years, the selling of houses skyrocketed as people could easily get loans for housing with the policies developed by the bank. We can also observe that there was a drought in the housing market around 2009-2010. From all the three graphs, we can have the remark that the houses' pricing was very high as the demand and selling of the houses were high. Later the supply and demand triangle twisted during and after 2008 as there were too many houses available but very little number of buyers as the consequence the prices drastically dropped being the very reason leading to recession [5]. The above graph as in **Figure 5** shows the relation between original prices and predicted prices by the linear regression model. If we know the ideal situation then all the dots would be on the line and slope being 1; but the slope, we have is 0.6 and These values were rather well predicted, but there is still room for improvement because they are so near to the line. Nevertheless, as you can see in **Figure 6** the regression model is just to show the dynamism of the market, but misses to show fall in the prices for the future. It is increasingly showing the positive slope, so this model can. **Figure 7** depicts the sub plots of LR on the original predicted price. **Figure 8** depicts the scatter plot of prices compared to as predicted by Linear Regression methodology. Here we see that come values are not effectively predicted and therefore accuracy seems to have declined. The next model prediction we are looking at is by Hidden Markov Model (HMM). The graph above as depicted by **Figure 9** is showing a drop in the house prices for the next coming year, which is obtained with the help of HMM. The predictions though have a better accuracy but deviations are still visible clearly. **Figure 10** and **Figure 11** depict the results of LSTM Methodology applied for this prediction problem. From Figure 15 it is observed that the prediction made is quite close to the testing data. Additionally, by a wide margin, LSTM has produced the best r2 scores and can be concluded that it is the best model for this task. Graph represented in **Figure 11** shows the prediction of house prices of the Testing set of the dataset by LSTM model. The green line indicates the prediction for the next year. As the green line is down it can be inferred that there is going to be a fall in the prices. However, if we talk about the seasonal variation; the summer usually has a higher price so here we see that the prices are going to fall according to the model. **Table II** presents a summary of the results and observations gathered from the help of this study. In this paper three methodologies namely Linear Regression, Hidden Markov Model and Long Short Term Memory were studied for predicting a possible house market crash. The summary of our results as provided in Table 1 help in endorsing the fact that Long Short Term Memory as a Machine Learning methodology is efficient with a better R^2 , though training cost may be high. Linear Regression as a concept fails to identify the patterns efficiently and though efficient in terms of training time cannot always find when the prices will fall.

Table 1: Center Survey of Machine Learning Techniques in Financial Areas

Authors	Machine Learning	Financial Areas
Nayak et al (2015) [5]	Support Vector Machine	Stock Market price Prediction
Rundo,et al(2019)[6]	Advanced Markov Model	Adaptive Trading System
Lei et al (2020) [7]	Reinforcement Learning	Financial signal representation and algorithmic trading
Hasan et al (2006) [8]	Hidden Markov Model	Stock Market Forecasting
Ho et al (2020) [9]	Support Vector Machine, random Forest, Gradient Boosting Algorithms	Predicting property prices

Table 2: Comparative Results of LR, LSTM and HMM

Name of the Model	How efficient is it?	Training Time	R^2	Prediction
LR	Medium	Low	0.6	House prices will rise by the time
HMM	Medium	Low	0.70	House prices will fall
LSTM	High	High	0.96	House prices will fall slightly

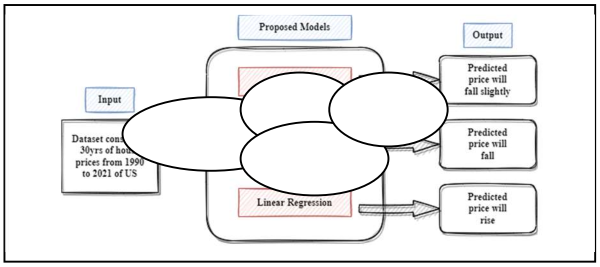

Figure 1: Proposed Model System

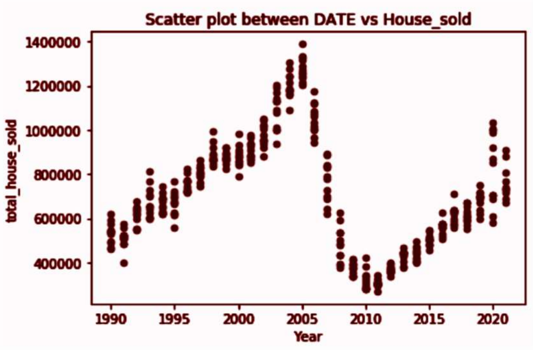

Figure 2: Number of House sold

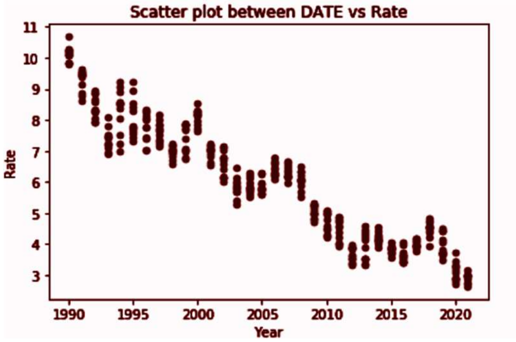

Figure 3: Scatter Plot of Rate

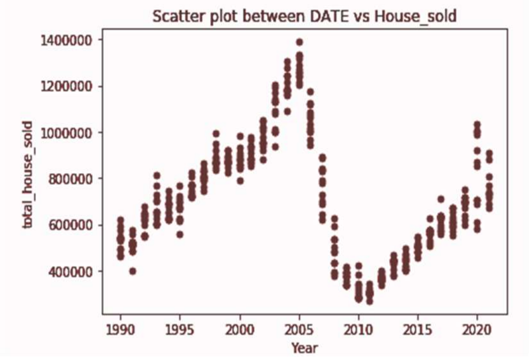

Figure 4: House sold in consecutive years

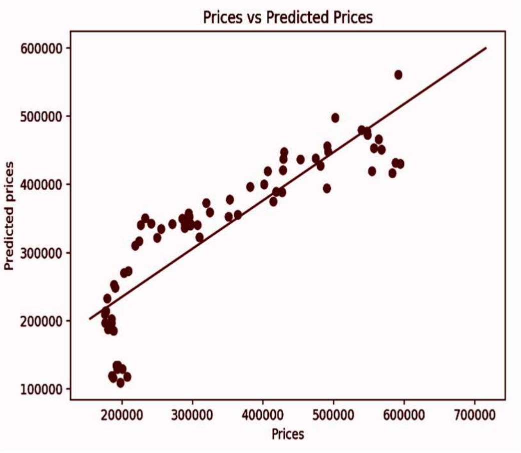

Figure 5: Linear Regression Model of Prediction

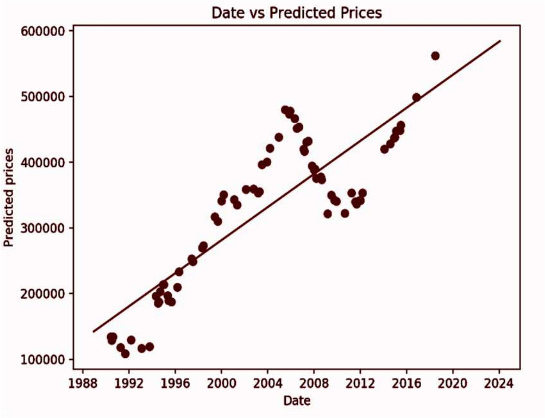

Figure 6: Depiction of Dynamism of the Market by Linear Regression Model

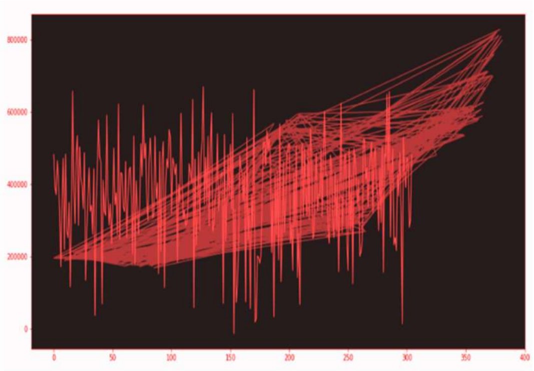

Figure 7: Sub Plot of linear Regression on original and predicted price

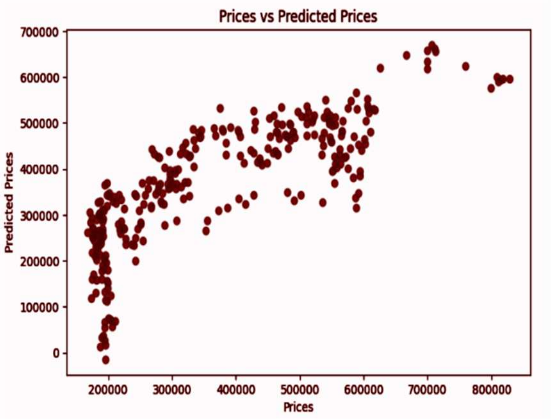

Figure 8: Sub Plot of linear Regression on original and predicted price.

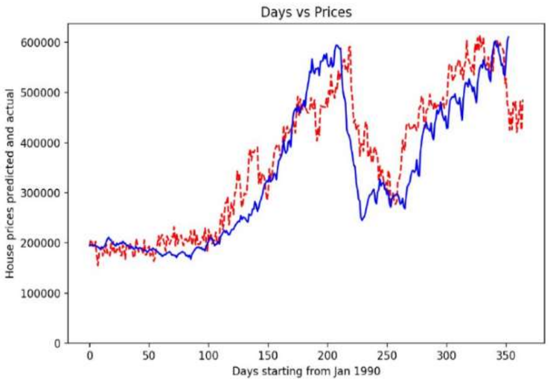

Figure 9: House prices actual vs predicted by HMM

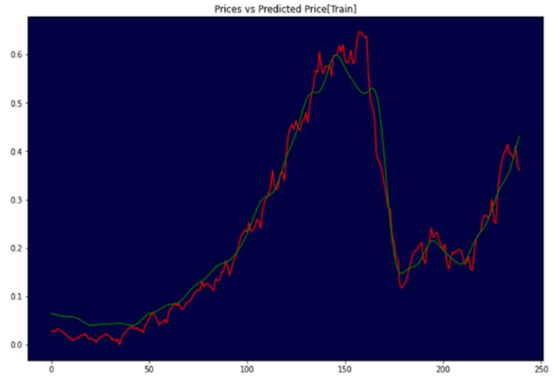

Figure 10: Graphical representation of training and prediction using LSTM

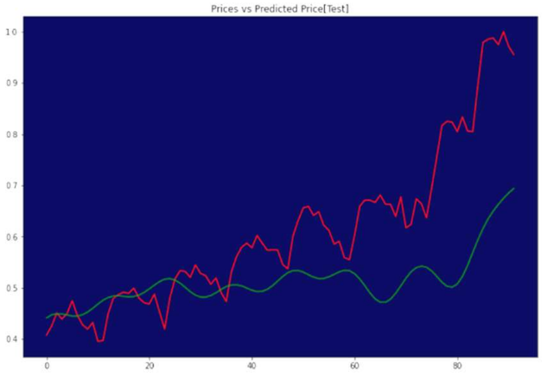

Figure 11: Graphical representation of testing and prediction using LSTM

5. CONCLUSION

Three machine learning models were examined in this paper namely Linear Regression, Hidden Markov Model and Long Short Term Memory for predicting house market crash based on the dataset of 2007-2008 crash. Simulation results clearly demonstrate that the LSTM Model is significantly better than Hidden Markov Model and Linear Regression Mode, LSTM though is complicated methodology. Currently the study was conducted for a dataset of the USA only. Therefore, the

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

applicability of these models for predicting market crashes for other economies need to be studied as well. The housing market and the current economic downturn are strongly intertwined. As compared to other approaches employed in this work or other papers, the LSTM method for prediction is by far the best method for prediction and is quick and effective, though it is complex. Given that we can forecast using both theoretical models and machine learning, it will be possible to combine these two fields in the future for improved outcomes.

REFERENCES:

[1] Fowler, L., "Forecasting New Housing Starts Using Real GDP and Average 30-Year Fixed Mortgage Rates", *Linfield University Student Symposium: A Celebration of Scholarship and Creative Achievement.* Event. Submission 40. https://digitalcommons.linfield.edu/symposium/2019/all/40, 2019.

[2] Crouhy, M. G., Jarrow, R. A., & Turnbull, S. M. , "The subprime credit crisis of 2007", *The Journal of Derivatives*, *16*(1), 2008, 81-110.

[3] Farmer, R. E., "The stock market crash of 2008 caused the Great Recession: Theory and evidence", *Journal of Economic Dynamics and Control*, 2012,*36*(5), 693-707.

[4] Demyanyk, Y., & Hasan, I., "Financial crises and bank failures: A review of prediction methods", *Omega*, *38*(5), 2010, 315-324.

[5] Nayak, R.K.; Mishra, D., Rath, A.K. A Naïve, "SVM-KNN based stock market trend reversal analysis for Indian benchmark indices", Appl. Soft Comput.,2015, 35, 670–680.

[6] Rundo, F., Trenta, F., Di Stallo, A., Battiato, S., "Advanced Markov-Based Machine Learning Framework for Making Adaptive Trading System", Computation, 2019, 7, 4.

[7] Lei, K., Zhang, B., Li, Y., Yang, M., Shen, Y., "Time-driven feature-aware jointly deep reinforcement learning for financial signal representation and algorithmic trading", Expert Syst. Appl., 2020, 140, 112872.

[8] Hasan, M.R. and Nath, B. , "Stock market forecasting using Hidden Markov Model: A New Approach", 5th Intl. Conf. on Intel. Sys. Design and Appl., IEEE, 2016.

[9] Ho, W. K. O., Tang, B. S., & Wong, S. W., "Predicting property prices with machine learning algorithms",Journal of Property Research, 38(1), 2020, 48–70.

[10] De,P. "Housing Market Prediction Using Machine Learning and Historical Data", 2020.

[11] Nyman, R., & Ormerod, P., "Predicting economic recessions using machine-learning algorithms", arXiv preprint arXiv:1701.01428, 2017.

[12] Kirkpatrick II, C. D., & Dahlquist, J. A. , "*Technical analysis: the complete resource for financial market technicians",* FT press, 2010.

[13] Olah, C., "Understanding LSTM networks", 2015, https://colah.github.io/posts/2015-08-Understanding-LSTMs/

[14] Ding, G., & Qin, L., "Study on the prediction of stock price based on the associated network model of LSTM", *International Journal of Machine Learning and Cybernetics*, *11*(6), 2020, 1307-1317.

[15] Dhanasekar, D., Di Troia, F., Potika, K., & Stamp, M., "Detecting encrypted and polymorphic malware using hidden Markov models", In *Guide to Vulnerability Analysis for Computer Networks and Systems*, pp. 281-299), 2018,Springer, Cham.

[16] Zhang, M., Jiang, X., Fang, Z., Zeng, Y., & Xu, K., "High-order Hidden Markov Model for trend prediction in financial time series", *Physica A: Statistical Mechanics and its Applications*, *517*, 2019, 1-12.

[17] Li, X., Parizeau, M., & Plamondon, R. "Training hidden markov models with multiple observations-a combinatorial method", *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *22*(4), 2020, 371-377.

[18] Tofallis, C., "Least squares percentage regression", *Journal of Modern Applied Statistical Methods*, 2019.

[19] Lo, A. W., & MacKinlay, A. C., "A non-random walk down Wall Street", In *A Non-Random Walk Down Wall Street*. Princeton University Press., 2009.