Forecasting Financial and Market Bubble

A Project Report
Presented to
The Faculty of the College of
Engineering

San Jose State University
In Partial Fulfillment
Of the Requirements for the Degree
Master of Science in Software Engineering

By
Khurd Aditi, Krishnamoorthy Akhil, Paruchuru Arihant Sai, Patil Sayali
May 2020

APPROVED

Professor Nima Karimian, Project Advisor

ABSTRACT

Forecasting Financial and Market Bubble
By Khurd Aditi, Krishnamoorthy Akhil, Paruchuru Arihant Sai, Patil Sayali

There was a catastrophic market and business crash in the year 2008 where millions of investors and people were driven to the state of penury within a fortnight. The reason behind the whole crash was when the financial giants involved in hedge fund trading and more mortgage-based loans, and when housing prices started falling rapidly, the homeowners neither had money to pay nor they were able to sell their house. The idea is to assess the picture behind every large acquisition, huge market deals and unwarranted soaring of mortgage or related prices, and how it might contribute to/affect the market bubble. The bigger picture also involves guiding the end user whether to buy/sell a stock or a property at a particular time.

Acknowledgments

The authors are deeply indebted to Professor Nima Karimian for his invaluable comments and assistance in the preparation of this study.

Forecasting Financial And Market Bubble

Aditi Khurd, Akhil Krishnamoorthy, Arihant Sai Paruchuru, Sayali Patil
Computer Engineering Department
San José State University (SJSU)
San José, CA, USA
Email: {aditi.khurd, akhil.krishnamoorthy, arihantsai.paruchuru, sayali.patil}@sjsu.edu

***Abstract*—The catastrophic market and business crash shortly known as 2008 Financial Crisis where millions of investors and people were driven to the state of penury within a fortnight. The reason behind the whole crash was when the financial giants involved in hedge fund trading and more mortgage based loans, and when housing prices started falling rapidly, the homeowners neither had money to pay nor they were able to sell their house. The idea is to assess the picture behind every large acquisition, huge market deals and unwarranted soaring of mortgage or related prices, and how it might have contributed to the market bubble. The bigger picture also involves guiding the end user whether to buy/sell a stock or a property at a particular time. We input every dealing which over the period of time might contribute to bubble burst from the data sets involving shares, housing and mortgage loans as the convergence of these three domains will help us detect the market bubble. This would help forecast slumps in financial sector, and would help in taking necessary steps to determine uncertain amount of products, services and takeovers.**

***Index Terms*—Forecast, Home Value, Linear Regression, LSTM, Stock Price**

I. Introduction

Nobody predicted the stock market would crash in the last quarter of 2007 solely because the banks were participating in hedge funds and lower interest rates. The idea is to evaluate the larger picture behind the crash, thus helping the end user to anticipate any crashes and also to guide the user to buy / sell stocks / hypothecary. The project also includes the study of the market from 1984 to the present, where market anomalies, rising and falling share prices are found over the decades of fortune 500 companies. With the varying market sentiment in the past decade, this application would not only help the investors to take informed decisions but would also aid to avoid market crash, by evaluating anomalies in the market using machine learning and the economists and capitalists can take actions in favor of a healthy market.

II. Problem Statement / Project Architecture

With the application having to deal with multiple components, the client side is now built with React.Js [14], a front-end library developed by Facebook Inc., which our customer/client would be interacting with, and which our product would be displaying results. The backend is currently built upon GoLang [13], which is ahead of back-end technologies with added memory safety and garbage collection. The calls to and forth would be asynchronous, since React.JS follows an asynchronous pattern of calls. The added requirement would be importing node modules and GoLang modules. Once the API call is issued, machine learning algorithm is implemented on data to perform data- cleaning, data-preprocessing and forecasting. Data-cleaning on our data-set is performed in order to handle missing values. The whole front-end and back-end is individually dockerized further on. The database would be mongoDB, since it serves as optimal database since our data-set has the probability to surpass million columns. Having narrowed down to Partitioning and Availability of CAP theorem (Consistency, Availability and Partitioning), the front-end would be deployed to AWS separately, and the back-end along with middle-ware (machine learning model) together would be deployed to the AWS instance, and to support partitioning factor of CAP theorem, a load balancer would be implemented to avoid single point of failure [Fig 1].

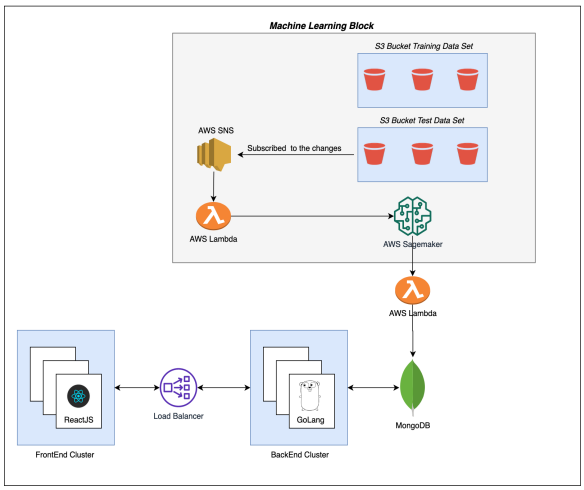

Fig. 1. Architecture

III. Method(s) / System Design

The project will be accommodated with an ability to replicate the whole system in a local environment by dockerizing the system. Each system will dockerized individually, so to replicate the system at individual level and the whole system will be provided with docker-compose file to launch the entire stack [Fig 2].

To Launch the entire stack

- docker compose up -d

The data-set has 6217 rows and 8 columns and gives an overview of the land and home values of some major cities

of the United States. It includes data for cities from the year 1984 to 2020. The data was pre-processed to take care of the null values, conversion of strings to floats, label encoding and feature extraction is performed. A graphical display of various features of data using Tableau is created to study the dependencies and characteristics of features. We could get a good overview of home values to later compute upon them to predict the home values for following year.

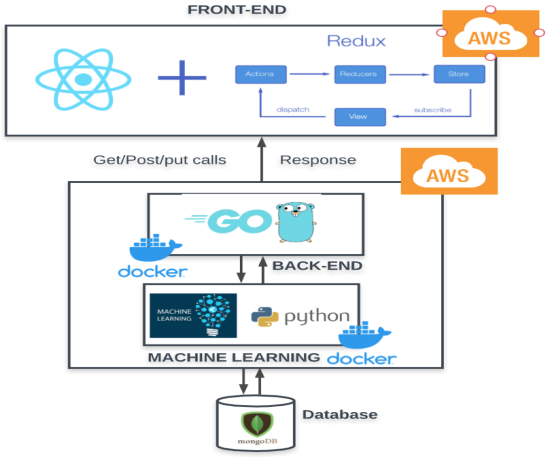

Fig. 2. System Design

Component	Description
Frontend cluster	– These are the interactive parts of the project where user will be able to interact with the with product. – They are only meant to display data in such a manner which can help a user to visualize the results generated by the ML model.
Backend cluster	– This system drives the user personalization and medium for the frontend to able to access data which is stored in our DB. – This system needs to scale if we are concurrently serving large number of requests as they are important for user experience.
Live data feed	– Live data feed is a real time system which process latest tweets regarding fortune 500 companies or about major cities we target. – This system will help us deciding price appreciation and depreciation based how they appear in twitter. – For ex: One of our top 500 company is Google and if google makes some good acquisition or invests in housing, it affects its share and also the result of our analysis.
Machine Learning	– This part can be termed as critical part of our project, because in this part our dataset and ML model both sit together and work in harmony. – The preprocessed dataset sits in AWS s3 buckets and this is provided in huge chunks to AWS sagemaker though AWS Lambda where our ML model is running. Whenever a live data makes a change it is processed and stored in buckets which in turn runs against the model to evaluate new results. – The accuracy of the model will decide on the accuracy of our results. The model plays an extensive role in attracting users to use the system
Document DB	– After the data has been processed by and run against the model in AWS sagemaker the generated results are stored in document DB. – This part of the system will interact with backend and is also a part of user experience.

Fig. 3. System Design Component Description

IV. Evaluation Methodology / Materials

An evaluation methodology sets the tone for the quality of the project. It is a tool which help us evaluate the quality of the project that we have achieved and whether it satisfies the required criteria. Forecasting Market Bubble is a typical data science project whose quality of the project will depend on how accurate we were able to predict the market bubble and how accurate we were able to predict land and house prices for a quality investment from a user's perspective. The main parts of the project are data sets and algorithms which are the key parts of this project to drive the quality of the project and meet the evaluation criteria successfully.

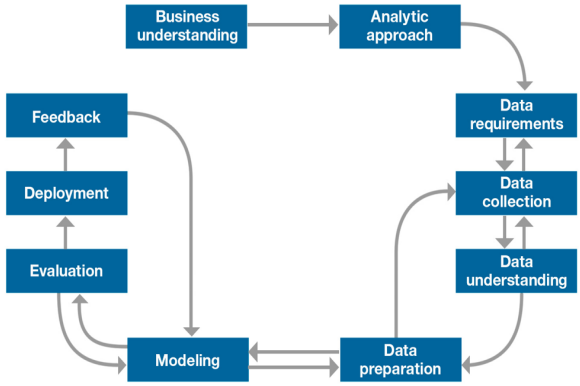

Fig. 4. Evaluation Methodology

The evaluation methodology process in Fig 4 was applied by providing the collected data sets to the different machine learning models to test out the accuracy of each ML model which will inform through which model our project should proceed. The financial forecasting is achieved through data sets which have quarterly points and therefore are always advised to use supervised ML model and exclude non-linear models.

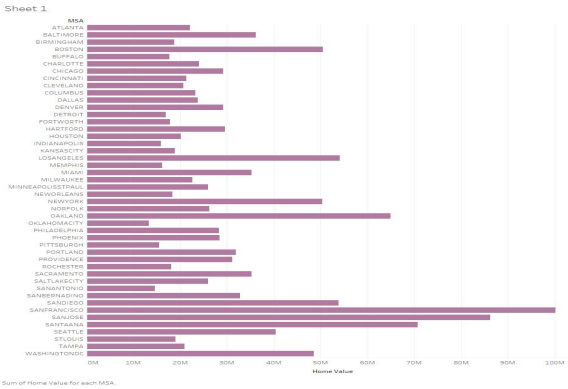

Fig. 5. Distribution of Hand Values according to the Cities

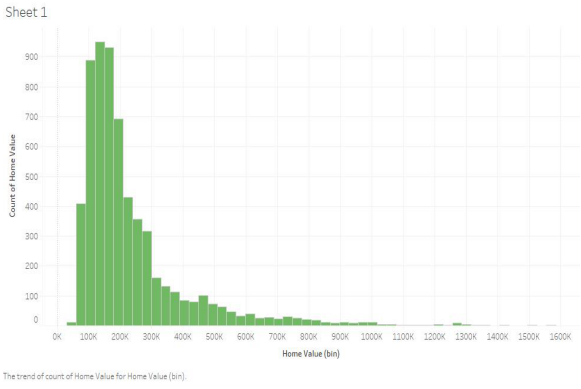

Fig. 6. Histogram of Home Values

Initially, we forecasted the home values by looking at history and taking the average of last n values. This formed the baseline system. This baseline method proved to be a good foundation to implement other machine learning algorithms.

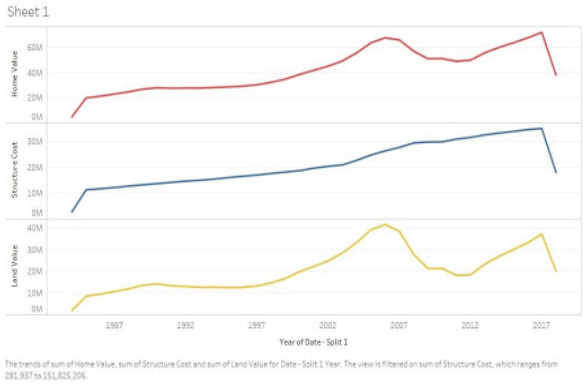

Fig. 7. Year vs Land Value, Structure Cost and Home Value

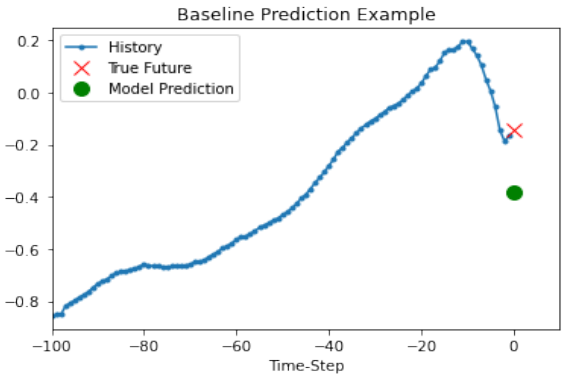

Fig. 8. Baseline Prediction

1. Linear Regression: The first algorithm that is applied is the Linear Regression which comes under the Predictive modelling. Using this algorithm, the possible output (Y) for a given input (X) is forecasted by taking into consideration the previous values. Linear Regression looks for statistical relationship between the continuous variables [Fig 9]. After fitting the Linear Regression algorithm to our model, the results gave a very high RMSE. So we decide to take more independent variables into consideration.

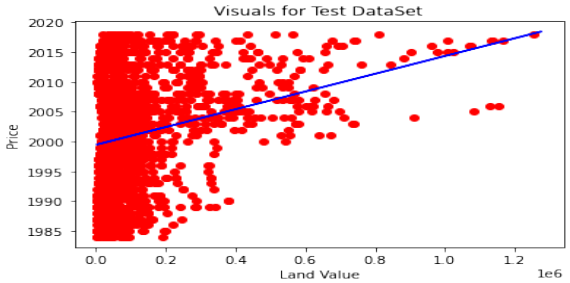

Fig. 9. Linear Regression

2. Multiple Regression The goal of using Multiple Regression is to minimize the error. The dependency of the prediction of home values on more than the history of home values is taken into consideration. It checks the contribution of various independent variables to the way regression describes the data. But Multiple Regression gave us negative values for about 13 cities and hence no accurate results [Fig 10]. We researched and concluded that the presence of timestamp in our data resulted in inaccurate results and proceeded to a time series algorithm.

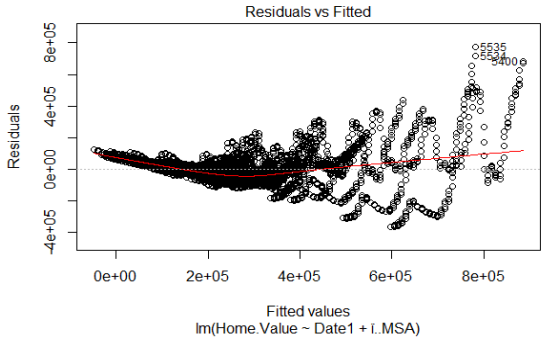

Fig. 10. Residual vs Fitted

3. Time Series Forecasting Time series forecasting is a great tool in finance and stock related predictions due to the temporal nature of data. Time series adds a time dimension to the data. It involves fitting the models to the historical data and predicting the future where the future is completely unavailable [Fig 11].

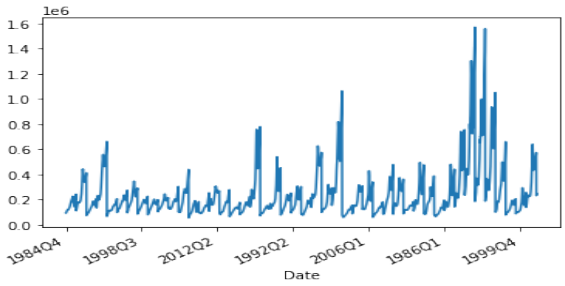

Fig. 11. Time Series

A Recurrent Neural Network usually works best with time series as it works on it step-by-step maintaining a record of what it has learned so far. Sequential memory is one of the feature of RNN that makes it suitable for time-stamped data. Long Short-Term Memory(LSTM) is a variation of RNN used for complex deep learning algorithms. They are capable of remembering recurrent patterns selectively for a long time.

a. Univariate LSTM First, the model learns only using one feature, Home Value history, indexed with date. The model takes last 100 observations from the training data in order to learn the patterns and predict the future values. After a batch size of 256 observations, the internal parameters to the model are changed for training purposes. The total of 40 epoch runs for 200 steps each [Fig 12]. This model gave predictions better than the baseline approach with a data loss of 0.09 after learning.

b. Multivariate LSTM Additional features are considered to improve the learning process of model. We considered the home values as well as home price index for training the model which resulted in better results [Fig 13].

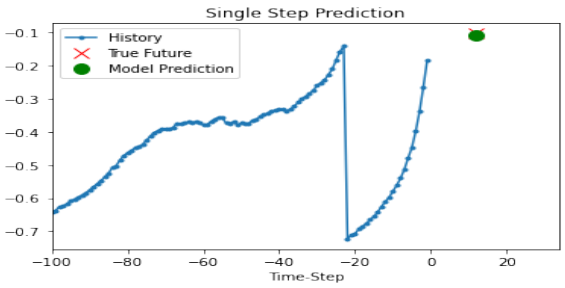

Fig. 12. Univariate LSTM

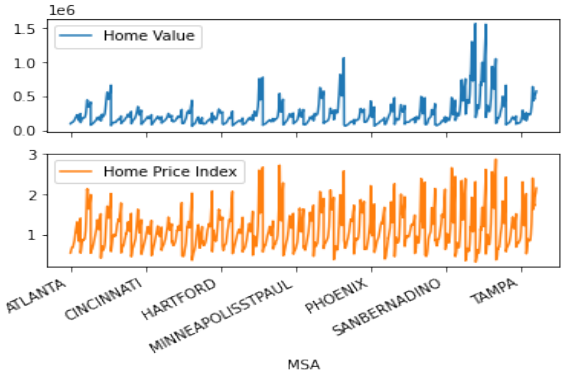

Fig. 13. Multivariate LSTM

V. Results

1. Preprocessing or data cleaning: Preprocessing of dataset is a stage where we have cleaned and formatted the data which can be used to process through an algorithm. We cleaned redundant values, and filled in missing values during the pre-processing. After the dataset is prepared, we have divided the dataset into two sets which are mainly called training and testing sets which we have used in training and model evaluation. The results we have achieved with the processed data set is a singular value for every metropolitan city, which is the result of the forecast, that gives the value for the next one year.

2. Feature extraction or generation: In this process, we had extracted the important features and also developed new features based on the base features already in the dataset which meets the project requirements more closely. We were able to carve out quarter rates for structures in major cities from daily rates datasets which meet the project requirements more closely.

3. Comparison of various available supervised models: The various algorithms we used, considered given the complexity of the project, are NN (Nearest neighbor), K-Neighbors, Naive Bayes, Random forest, time-series, Multiple regression, linear regression and Gradient boosting processor algorithms. The outputs through these algorithms are evaluated by calculating the RMSE (root mean square) values. RMSE is a mean of all the errors, a higher RMSE denotes higher deviation of result produced by the model from the actual path. By evaluating multiple algorithms, the time-series algorithm gave us less Root mean square (RMSE) value, hence the output of time

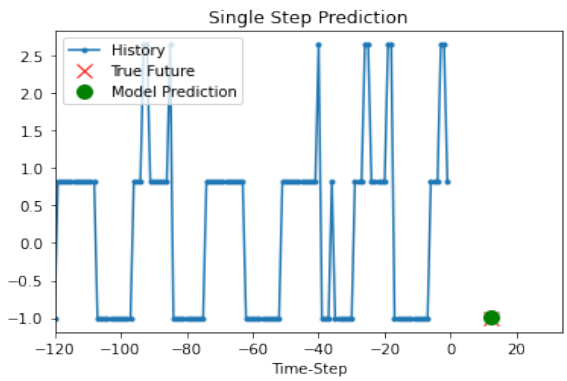

Fig. 14. Multiple Predictions

series algorithm is considered.

4. We have imported the processed data in one of MongoDB collections in mLab. When the users enter any city name on UI, then they will see three Bar Graphs. First Bar Graph represents the variation of Land Value [Fig 17], second represents the variation of Home Value [Fig 18] and third represents the variation of Structure Cost [Fig 19]. The data used is for the period from 1985 to 2018. As Land Value and Structure Cost sum up to total Home Value, these graphs will help users understand the variation of individual values which will help them make the decision whether to buy a land or not in a particular city.

5. There are two dashboards - one is to search the price of lands and properties of a city for the last 25 years, and then a forecast is made predicting the home value for the next 1 year. The user enters a metropolitan city name (out of 46 cities in our database) [Fig 15]. An auto-suggestion feature is provided as the user types the name of a city. For instance, if the user types “BU”, cities including Buffalo, Columbus and Pittsburgh are suggested as all of these cities have BU in common [Fig 16].

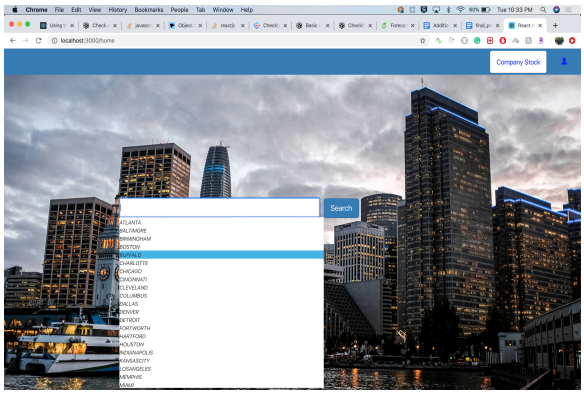

Fig. 15. Searching an MSA

The dashboard page is where in, the user’s search results navigate to. The dashboard page has 3 main entities. It has land value, home value and structure cost [Fig 17, 18, 19]. These prices are provided for the last 35 years, ranging from 1984 Q1 to 2018 Q2 in a form of a bar graph, and the prices are in dollars.

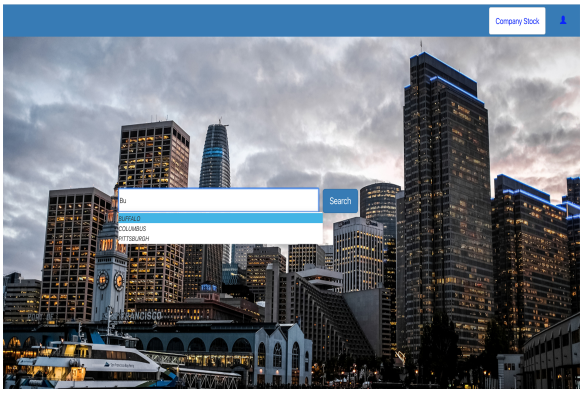

Fig. 16. Autosuggestion of MSA

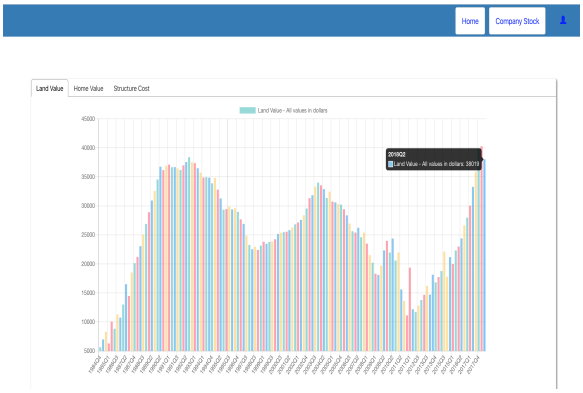

Fig. 17. Land Value

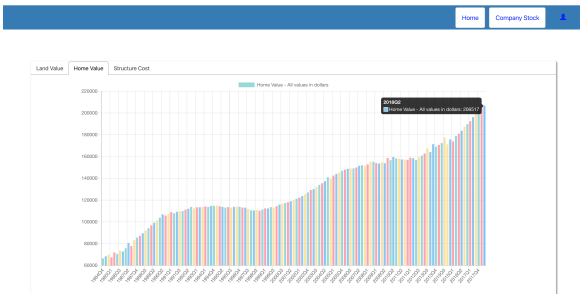

Fig. 18. Home Value

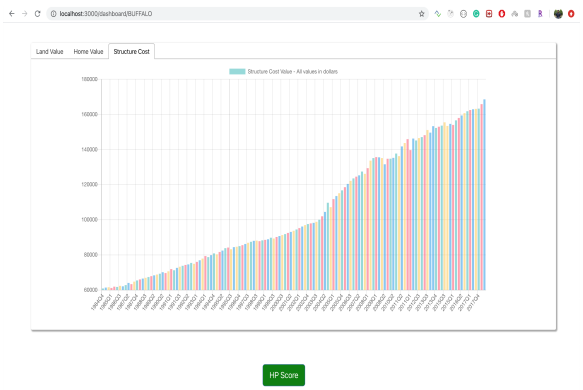

Fig. 19. Structure Cost

The pivotal feature of this application, the prediction of home value is provided in the next page, and the user can navigate to the results page by clicking on the button “HP Score” [Fig 19]. The results page has the following: home values ranging from time period 2018 Q3 to 2020 Q1 in a form of a bar graph [Fig 20]. The prediction home value is provided below the graph [Fig 21].

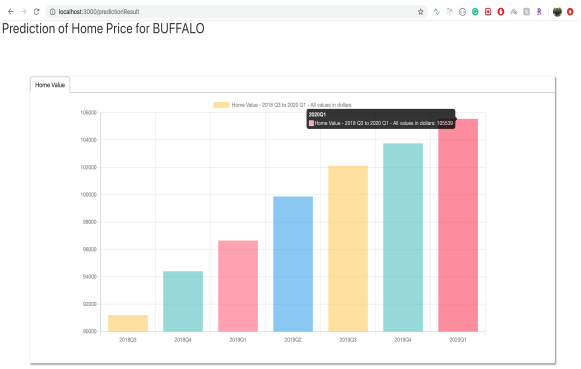

Fig. 20. Home Value from 2018Q3-2020Q1

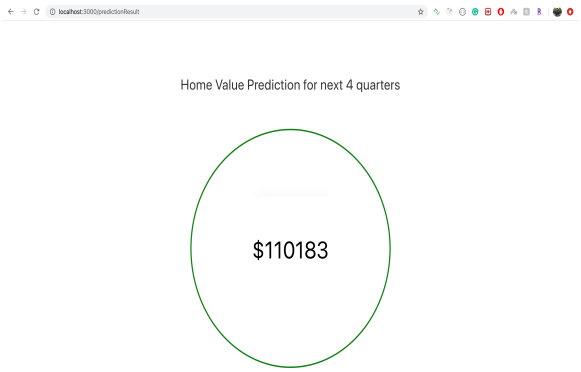

Fig. 21. Home Value prediction for next 1 year

6. The stock dashboard is where users can search for any company name and get the stock results. Users are also provided with a feature of entering a company name. When users enter any company name on UI, they will see the information related to the stock price of that company, Market Cap value and average volume traded per day over a range of 60 days [Fig 23].

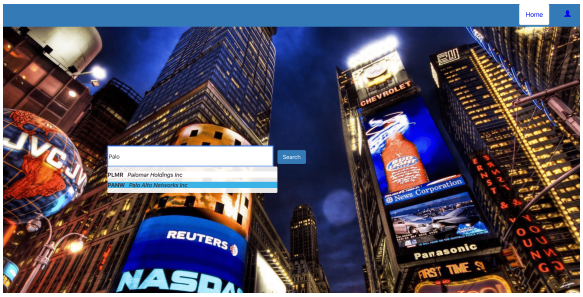

Fig. 22. Stock Search

An auto-suggestion feature is provided in the search bar. For instance, if the user types “AMAZ”, companies including Amazon, Amazing Oil company, Amazonas Florestal Limited, etc are suggested [Fig 22]. All the companies stock prices are real time. A total of 69477 data points, which includes US stocks, Chinese stocks, Mutual Funds data and ETF (Exchange-traded fund) are provided.

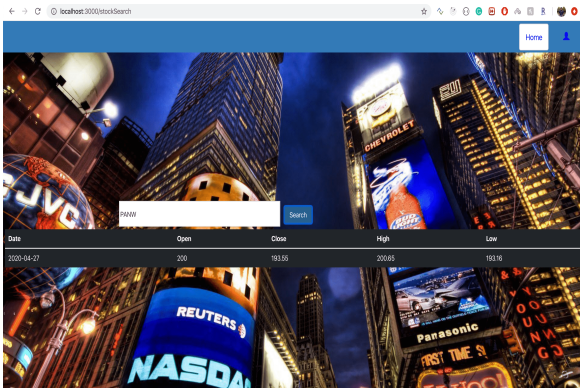

Fig. 23. Stock Information

VI. Discussion

1. User can login and logout of his private session.

2. User can search a stock price and look at its history on its rise and falls, and can see statistical analysis from the stock’s initial public offering (IPO) till date.

3. User can search for land price and look at its history of the rise and falls and can see take informed decisions whether to buy the land/stock with the statistics (considering booming/slowing economy, market sentiment and market bubble) the application provides.

4. User can search for a city to get the details about the variation in land prices over the past few years. The server module of this application will communicate to database, fetch required data and send it to the client.

5. The search made by the user will be saved and it will be displayed on user’s dashboard in terms of graphs expressing his search history and recent trend of land price variation.

6. Client-side module of this application will be user friendly giving end users a rich UI experience. Various graphs on the user’s dashboard will enable users to efficiently visualize the land and stock price variations.

VII. Related Works

The catastrophic market and business crash, shortly known as the recession, in the year 2008 where millions of investors and people were driven to the state of penury within a fortnight. The reason behind the whole crash was when the financial giants involved in hedge fund trading and more mortgage-based loans, and when housing prices started falling rapidly, the homeowners neither had money to pay nor they were able to sell their house. The records from the years 1800 to 1940 show that there were drops in the financial market ranging from 16.4 percentage to 66.5 percentage [6]. On studying such historical records mentioning the occurrence and bursts of the bubble, it is observed that these bubbles are an implication of irrational human behavior.

This irrational human behavior can be described as a situation where everyone who is part of the market starts bidding for an asset and due to the competition, that follows the bidding keeps on soaring and there comes a point where the underlying true value of the asset is lost. Such a situation has been termed as irrational exuberance. This exuberance is caused mainly due to an asset bubble. The asset bubble refers to the soaring of asset prices. The global finances have noted many such instances of bubbles and booms. Some appearances are the South Seas Bubble, John Law’s Mississippi Company and Florida Land Boom of the 1920s.

Such an unstable financial system is a great risk, especially for developing countries. The financial phenomenon like recession, bubbles, booms should be analyzed as they affect the majority of the world’s livelihood. As a result, many attempts have been made to capture and study the behavior of these bubbles so that they can be identified and predicted. The goal of the financial bubble forecasting is to assess every situation indicating large acquisitions, huge market deals and an unexpected rise in market prices.

Many tools have been developed for the identification and detection of bubbles. The log-periodic power-law model is one such flexible tool. This model considers faster than the exponential increase in asset prices and the associated accelerated oscillations as a determining factor for the bubbles [2]. This model has proved to be successful in finding out the time windows for two crashes in the Chinese market. The model was able to find out the market peak dates in the ranges of predicted crash dates but only using the data of the market prior to the crash.

VIII. Conclusions

With varying land prices and share prices around United States, the application ‘Forecasting Financial and Market Bubble’ would help the end users take informed decisions by comparing the prices of land prices in the metropolitan cities for the last 30 years. Also, with the prediction functionality of our application, the users can foresee the future land prices, and take a decision before they finalize a purchase. Moreover, with the stock’s API integration, we give share prices of almost every other company across United States. The users can also see how the stock prices fluctuates whenever there is a change in land prices and see how the stocks are performing currently.

Acknowledgement

The authors are deeply indebted to Professor Nima Karimian for his invaluable comments and assistance in the preparation of this study.

References

[1] Yan, G. C. C. (2011). A Financial Engineering Approach to Identify Stock Market Bubble. Systems Engineering Procedia, 2, 153-162.

[2] Jiang, Z. Q., Zhou, W. X., Sornette, D., Woodard, R., Bastiaensen, K., & Cauwels, P. (2010). Bubble diagnosis and prediction of the 2005–2007 and 2008–2009 Chinese stock market bubbles. Journal of economic behavior & organization, 74(3), 149-162.

[3] Johnston, D. E., & Djurić, P. M. (2012, June). Estimating hedge fund risk factor exposures. In 2012 IEEE 13th International Workshop on Signal Processing Advances in Wireless Communications (SPAWC) (pp. 510-514). IEEE.

[4] Virtanen, T., Tölö, E., Virén, M., & Taipalus, K. (2018). Can bubble theory foresee banking crises? Journal of Financial Stability, 36, 66-81.

[5] Ardila, D., Sanadgol, D., & Sornette, D. (2018). Out-of-sample forecasting of housing bubble tipping points. Quantitative Finance and Economics, 2(4), 904-930.

[6] Rapp, D. (2015). A Short History Of Booms, Bubbles, And Busts. In Bubbles, Booms, and Busts (pp. 159-345). Copernicus, New York, NY.

[7] Fedorova, E. A., & Lukasevich, I. Y. (2012). Forecasting financial crises with the help of economic indicators in the CIS countries. Studies on Russian Economic Development, 23(2), 188-194.

[8] Malinkina, A. V. (2017, October). Identification and dating of "bubbles" on financial markets: Comparison of posterior algorithms and monitoring algorithms. In 2017 Tenth International Conference Management of Large-Scale System Development (MLSD) (pp. 1-5). IEEE.

[9] Grebenyuk, E. A. (2017, October). Algorithms for detecting and dating the financial bubbles in real time. In 2017 Tenth International Conference Management of Large-Scale System Development (MLSD) (pp. 1-5). IEEE.

[10] Yan, W., Woodard, R., & Sornette, D. (2010). Diagnosis and prediction of tipping points in financial markets: Crashes and rebounds. Physics Procedia, 3(5), 1641-1657.

[11] Yanjia, C. (2010, August). Analysis of the Causes and Effects of the Global Financial Crisis and Lessons Learned. In 2010 International Conference on Management and Service Science (pp. 1-4). IEEE.

[12] A. Mehra, "Understanding the CAP Theorem," 30 April 2019. [Online]. Available: https://dzone.com/articles/understanding-the-cap-theorem.

[13]"Go, an open source programming language," [Online]. Available: https://golang.org/.

[14]"React, a JavaScript library," [Online]. Available: https://reactjs.org/.

[15] V. Sampathkumara, M. H. Santhib and J. Vanjinathan, "Forecasting the land price using statistical and neural network software," in 3rd International Conference on Recent Trends in Computing 2015 (ICRTC-2015), Delhi, India, 2015.

[16] X.Chen , L.Wei and J.Xu, House Price Prediction using LSTM.

[17] "Quick-Start Guide to mLab," [Online]. Available: https://docs.mlab.com/.

[18] "Docker CLI reference," [Online].Available: https://docs.docker.com/engine/reference/commandline/docker/.

[19] D.Varghese, "Comparative Study on Classic Machine learning Algorithms," 2018. [Online].Available:https://towardsdatascience.com/comparative-study-on-classic-machine-learning-algorithms-24f9ff6ab222