House Price Prediction Using Machine Learning

G. Naga Satish, Ch. V. Raghavendran, M.D.Sugnana Rao, Ch.Srinivasulu

Abstract: Machine learning plays a major role from past years in image detection, spam reorganization, normal speech command, product recommendation and medical diagnosis. Present machine learning algorithm helps us in enhancing security alerts, ensuring public safety and improve medical enhancements. Machine learning system also provides better customer service and safer automobile systems. In the present paper we discuss about the prediction of future housing prices that is generated by machine learning algorithm. For the selection of prediction methods we compare and explore various prediction methods. We utilize lasso regression as our model because of its adaptable and probabilistic methodology on model selection. Our result exhibit that our approach of the issue need to be successful, and has the ability to process predictions that would be comparative with other house cost prediction models. More over on other hand housing value indices, the advancement of a housing cost prediction that tend to the advancement of real estate policies schemes. This study utilizes machine learning algorithms as a research method that develops housing price prediction models. We create a housing cost prediction model In view of machine learning algorithm models for example, XGBoost, lasso regression and neural system on look at their order precision execution. We in that point recommend a housing cost prediction model to support a house vender or a real estate agent for better information based on the valuation of house. Those examinations exhibit that lasso regression algorithm, in view of accuracy, reliably outperforms alternate models in the execution of housing cost prediction.

Index Terms: Machine learning algorithm, lasso regression process and neural system, hosing cost prediction.

I. INTRODUCTION

What Is Learning? Rats Learning to Avoid Poisonous Baits: Rats normally stumble upon food items by its look and smell and start eating in small amounts and later depending on food and physiological effect the feeding of food goes on. If the rat notices the illness of the food, the rat will not touch that food. Similarly the machine learning mechanism plays a vital role same as animal usage of past experience for acquiring and expertise in detecting the food safety. By mistake if the knowledge with the food is negatively labeled, the prediction of the animal will also will be negatively affected and encountered in the future. With the inspiration of the previous example of successful learning we demonstrate a typical machine learning algorithm .Likewise we would want to program a machine that learns how to filter spam e-mails. A credulous result might be apparently comparable of the lifestyle of rats that, how to keep away from poisonous baits.

Revised Manuscript Received on July 05, 2019.

Dr.G.Naga Satish, Associate Professor, CSE Dept, BVRIT HYDERABAD

Dr.Ch.V.Raghavendran, Professor, IT Dept, ACET, Surampalem

Mr.M.D.Sugnana Rao, Assistant Professor, CSE Dept, BVRIT HYDERABAD

Dr.Ch.Srinivasulu, Professor, CSE Dept, BVRIT HYDERABAD

The machine will basically remember the past e-mails that needed been named similarly as spam e-mails by the human user. When another email arrives, the machine will look for it in the past set about spam e-mails. Though it matches among them, it will be trashed. Otherwise, it will make moved of the user's inbox organizer.

At the same time the first "learning by memorization" methodology may be useful, it fails to offer an important aspect known as learning systems – the capacity to mark unseen email messages. A fruitful learner ought to have the ability which will be the advancement from distinctive samples to more extensive generalization. This may be Likewise as inductive thinking or inductive induction. In the attraction nervousness exhibited previously, after the rats experience a sample of a certain sort about food; they apply their disposition at it once new, unseen illustrations from claiming nourishment of comparable emanation Also taste. Should attain generalization in the spam sifting task, those learner might examine the Awhile ago seen e-mails, and extricate An situated about expressions whose presence for a email message is characteristic of spam. Then, At another email arrives, those machine could weigh if a standout among the suspicious expressions gives the idea On it, and foresee its mark Appropriately. Such an arrangement might possibly have the ability effectively to foresee the name about unseen e-mails.

Responsibilities further than Human Capabilities: an additional totally crew about errands that profit starting with machine Taking in systems are identified with the Investigation for extremely substantial and intricate information sets: galactic data, turning restorative chronicles under restorative knowledge, climate prediction, and dissection of genomic data, Web serch engines, Also electronic trade. With an ever increasing amount accessible digitally recorded data, it gets evident that there would treasures about serious majority of the data covered clinched alongside information chronicles that would best approach excessively little and also as well perplexing to people with bode well about. Taking in with recognize serious examples over substantial Also complex information sets may be a guaranteeing space for which the blending of projects that take for the Just about boundless memory limit and ever expanding transforming velocity about PCs opens up new horizons.

Regulated versus Unsupervised, since taking in includes an association between those learner and the environment, you quit offering on that one might separate taking in assignments as stated by those nature for that connection. The to start with qualification will note is the Contrast the middle of regulated and unsupervised Taking in. Likewise an illustrative example, think about that errand for Taking in will recognize spam email versus the undertaking about aberrance identification. For the spam identification task, we think about a setting to which the learner receives preparing e-mails to which the mark spam/not-spam

may be Gave. On the support from claiming such preparation the learner ought to further bolstering to evaluate a tenet to labelling a recently arriving email message. For contrast, for those assignment about aberrance detection, every last one of learner gets Concerning illustration preparing is an extensive form of email messages (with no labels) and the learner's errand is on identify "unusual" messages.

II. LITERATURE SURVEY

The latest worldwide financial crisis restored a sharp enthusiasm toward both academic and strategy circles on the part of asset costs and specifically lodging costs clinched alongside monetary movement. As Lamer (2007) notes those lodging showcase predicted eight of the ten post globe War ii recessions, acting Concerning illustration An heading woman for those true segment of the economy. Truth be told he dives Likewise significantly Concerning illustration with state that "Housing is those benefits of the business cycle".

Vargas and silva (2008) contend that lodging costs alterations assume a paramount part in the determination of the stage of the business cycle. When those economy booms, development and work in the lodging division expand quickly should react should overabundance demand, quickly pushing ostensible house costs upwards. Throughout those withdrawal phase, the drop in private money lessens aggravate interest Also ostensible house costs. By ostensible house costs normally fall sluggishly since householders would unwilling on bring down their costs. The majority of the conformity will be attained through declines clinched alongside bargains volume bringing about An drop in the development segment and the lodging built vocation. Moreover, Throughout withdrawal and subsidence true house costs fall quickly Likewise general inflationary patterns diminish true house costs much with sticky perceived costs.

Recently, a few writers scope to experimental discoveries that house costs can make instrumental molding to determining yield. (Forni etc, 2003; stock and Watson, 2003; Gupta Furthermore Das, 2010; das etc, 2009; 2010; 2011; Gupta and Hartley, 2013). Those lodging development division speaks to an expansive and only aggregate monetary action communicated in the GDP. Consequently, Concerning illustration it reflects an extensive parcel of the general riches of the economy, house costs variances can make a pointer of the Development about GDP (Case etc, 2005). Concerning illustration it is those body of evidence with different assets, those development for house costs can make Additionally an pointer of the future course from claiming expansion (Gupta Also Kabundi, 2010). Overall, exact determining of the Development way from claiming house costs could make a suitable apparatus both on house business members and fiscal strategy powers.

There is huge literature writing in regards to U.S. house prices. Rapach Furthermore strauss (2007) use an auto regressive dispersed slack (ARDL) model framework, holding 25 determinants with conjecture genuine lodging cost development to the unique states of the elected Reserve's eighth region. They discover that ARDL models tend should beat a benchmark AR model.

Rapach and strauss (2009) augment those same examination on the 20 biggest u. Encountered with urban decay because of de industrialization, innovation developed, government agent. States dependent upon ARDL models looking at state, territorial and national level variables. When again, the creators scope comparative conclusions on the fact that joining together forecasts about models for different slack structure.

Gogas and Pragidis (2011) utilize the hazard premium ascertained Likewise those Contrast the middle of Different long haul enthusiasm rates and the agents' desires over future fleeting rates as information variable to foreseeing what's to come heading for house costs. They infer that gurus Also investigators could use adequately those majority of the data given by those investment rate hazard premium today so as should evaluate the likelihood from claiming acquiring beneath pattern S&P CS-10 list three months ahead.

Gupta and Das (2010) also forecast the recent downturn in real house price growth rates for the twenty largest U.S. states. The authors use Spatial Bayesian VARs (BVARs), based only on monthly real house price growth rates, to forecast their downturn over the period 2007:01 to 2008:01. They find that BVAR models are well-equipped in forecasting the future direction of real house prices, though they significantly underestimate the decline. They attribute this under-prediction of the BVAR models to the lack of any information on fundamentals in the estimation process.

Rapach and strauss (2009) expand the individuals same examination on the 20 most amazing . Encountered with urban rot due to de industrialization, advancement developed, administration agonize. States reliant upon ARDL models taking a gander at state, regional Also national level variables. When again, those inventors degree similar finishes on the reality that joining together forecasts regarding models for separate slack structure.

Gogas and Pragidis (2011) use the danger premium determined similarly the individuals complexity those white collar for different whole deal energy rates and the agents' longings In future transient rates as data variable with foreseeing what's with turn heading to house expenses. They construe that masters also investigators Might utilization enough the individuals lion's share of the information provided for by the individuals financing rate danger premium today with the goal Similarly as ought further bolstering assess those probability from asserting securing underneath design S&P CS-10 rundown three months ahead.

III. DESIGN APPROACH

A. Linear regression:

Simple linear regression statistical method allows us to summarize and study the relationship between two continuous quantative variables.

- One variable, denoted x, is regarded as the predictor, explanatory, or independent variable.
- The other variable, denoted y, is regarded as the response, outcome, or dependent variable.

B. Multiple Regression Analysis

Multiple regression analysis is used to check whether there is a statistically noteworthy association the middle of sets of variables. It's used to discover patterns in the

individuals sets of information.

Numerous relapse Investigation will be very nearly the same Likewise basic straight relapse. The main distinction the middle of straightforward straight relapse Also numerous relapse is in the number for predictors ("x" variables) utilized within those relapse.

Straightforward relapse examination employments An absolute x variable to each subordinate "y" variable. Case in point: (x1, Y1).

Numerous relapse utilization numerous "x" variables for every free variable: (x1)1, (x2)1,(x3)1, Y1).

In one-variable straight regression, you might information particular case subordinate variable (i. E. "sales") against a autonomous variable (i. E. "profit"). Anyhow you could make intrigued by how diverse sorts from claiming offers impact the relapse. You Might set your X1 as particular case kind from claiming sales, your X2 Similarly as in turn sort about deals etc.

C. The cost Function

Thus let's say, you expanded the size of a specific shop, the place you predicted that those deals might a chance to be higher. Be that in spite of expanding those size, those bargains in that shop didn't expand that a great deal. Thereabouts those expense connected Previously, expanding those span of the shop, provided for you negative outcomes. So, we necessity on minimize these cost. So we present an expense function, which is fundamentally used to characterize and measure those slip of the model.

$$J(\theta_0, \theta_1) = \frac{1}{2m}\sum_{i=1}^{m}\left(h_\theta(x^{(i)}) - y^{(i)}\right)^2$$

D. Lasso Regression

Lasso regression which may be a standout among those relapse models that would accessible will examine the information. Further, the regression model may be demonstrated for a sample and the formula is Additionally recorded to reference.

LASSO stands for Least Absolute Shrinkage and Selection Operator.

Lasso regression is a standout among the regularization routines that makes niggardly models in the vicinity for vast number for features, the place expansive implies whichever of the following two things:.

• Vast enough to improve those inclination of the model on over-fit. Least ten variables can foundation over fitting.

• Huge enough will cause computational tests. This circumstance could emerge in the event from claiming a large number or billions about Characteristics.

Tether relapse performs L1 regularization that is it includes those punishment equal of the supreme esteem of the extent of the coefficients. Here the minimization goal will be Concerning illustration emulated.

Minimization goal = LS Obj + λ (sum about outright esteem of coefficients). The place LS Obj remains for minimum squares objective which will be nothing yet the straight relapse target without regularization Furthermore λ may be those turning figure that controls the measure for regularization. The inclination will build with those expanding quality of λ and the difference will diminish Concerning illustration the measure for shrinkage (λ) increments.

The lasso regression estimate is defined as

$$\hat{\beta}^{\text{lasso}} = \underset{\beta \in \mathbb{R}^p}{\operatorname{argmin}} \|y - X\beta\|_2^2 + \lambda \sum_{j=1}^{p} |\beta_j|$$
$$= \underset{\beta \in \mathbb{R}^p}{\operatorname{argmin}} \underbrace{\|y - X\beta\|_2^2}_{\text{Loss}} + \lambda \underbrace{\|\beta\|_1}_{\text{Penalty}}$$

Here the turning component λ controls those quality for penalty, that is. When λ = 0: we get same coefficients Similarly as basic straight relapse. At λ = ∞: constantly on coefficients are zero. The point when 0 < λ < ∞: we get coefficients between 0 What's more that for basic straight relapse thus At λ is amidst the two extremes, we would adjusting those underneath two plans.

- Fitting An straight model for y once X.
- Contracting those coefficients.

E. Gradient Boosting algorithm.

Gradient boosting is a machine Taking in strategy to relapse Also arrangement problems, that produces a prediction model in the structure of an group from claiming powerless prediction models.

The exactness of a predictive model might be helped to two ways:. Possibly by grasping characteristic building alternately. Toward applying boosting calculations straight far. There are a significant number boosting calculations in.

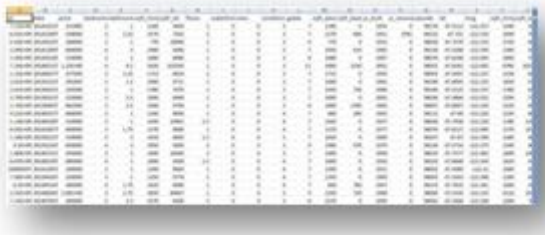
- Gradient Boosting
- XGBoost
- AdaBoost
- Gentle Boost etc.

Each boosting algorithm need its own underlying math. Also, a slight variety may be watched same time applying them.

Boosting calculation will be a standout among those The greater part capable Taking in thoughts acquainted in the final one twenty A long time. It might have been intended to order problems, yet all the it can be developed should relapse too. The inspiration to gradient boosting might have been An technique. That combines those outputs about large portions "weak" classifiers to process An capable "committee. " a powerless classifier (e. G. Choice tree) will be person whose slip rate is main superior to irregular guessing.

IV. IMPLEMENTATION

Data set:



Reading the data to plot the graphs:

```
data = pd.read_csv("kc_house_data.csv")
```

Using the above data, code to show plot a relation between number of bedrooms and number of houses:

```
data['bedrooms'].value_counts().plot(kind = 'bar')
plt.title('Number of bedrooms')
plt.xlabel('bedrooms')
plt.ylabel('Number of houses')
plt.show()
sns.despine()
```

Code to plot relation between Price and Living area:

```
plt.scatter(data.price, data.sqft_living)
plt.title('price vs sqft living')
plt.xlabel('price')
plt.ylabel('Sqft area')
plt.show()
sns.despine()
```

Plots relation between price and latitudes:

```
plt.scatter(data.price, data.lat)
plt.title('price vs latitude values')
plt.xlabel('price')
plt.ylabel('latitude values')
plt.show()
sns.despine()
```

Plots relation between price and area**:**

```
plt.scatter(data.price, (data.sqft_living + data.sqft_basement))
plt.title('price vs sqft area')
plt.xlabel('price')
plt.ylabel('area')
plt.show()
sns.despine()
```

Plots relation between waterfront and price:

```
plt.scatter(data.waterfront, data.price)
plt.title('waterfront vs price')
plt.xlabel('waterfront')
plt.ylabel('price')
plt.show()
sns.despine()
```

Plots relation between condition and price:

```
plt.scatter(data.condition, data.price)
plt.title('condition vs price')
plt.xlabel('condition')
plt.ylabel('price')
plt.show()
sns.despine()
```

Feeding model with training data:

```
train1 = data.drop(['id', 'price'], axis=1)
x_train, x_test, y_train, y_test = train_test_split(train1, labels, test_size=0.10, random_state=2)
```

Training the model with Linear Regression Algorithm:

```
lin = LinearRegression()
lin.fit(x_train,y_train)
print "linear regression, Accuracy:",lin.score(x_test, y_test)*100
```

Training the model with Lasso Regression:

```
las = linear_model.Lasso(alpha=20.0,max_iter=1e5)
las.fit(x_train,y_train)
print "Lasso Regression Accuracy:",las.score(x_test,y_test)
```

Reading the test data which is dynamically given in excel sheet:

```
test_data = pd.read_csv("test_data.csv")
```

Test data:

	C	D	E	F	G	H	I	J	K	L	M
1	price	bedrooms	bathroom	sqft_living	sqft_lot	floors	waterfron	view	condition	grade	sqft_abo
2	0	3	1	1180	5650	1	0	0	3	7	118
3	0	2	1	1100	5500	2	1	0	2	5	100
4											

Taking id as input to predict the price of the house of test data:

```
houseId = input("Enter House ID to predict price: ")
house = test_data[test_data['id'] == houseId]
```

Printing the predicted price got from best algorithm for given test constraints:

```
house = house.drop(['id','price'], axis=1)
print "predicted price price of house:",
print repr(clf.predict(house))
```

V. EXECUTION AND OUTPUTS

When the code gets executed first we get outputs plots and then prediction takes place. These plots help us to understand the correlation between target variable (price) and different predictor variables.

This plot gives a bar graph for bedrooms and number of houses. It is seen from dataset the count of 3 bedroom houses are greater in number and 7 bedroom houses are least in number.

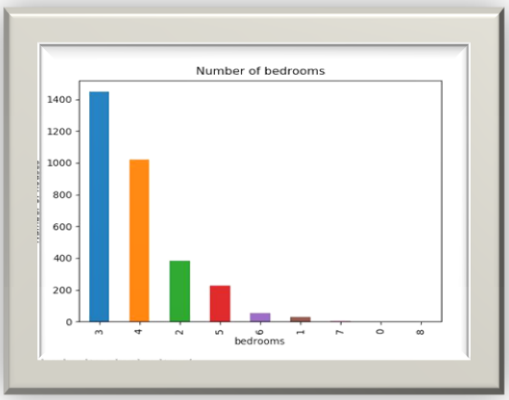


Fig – 1 : No. of houses vs no. of bedrooms

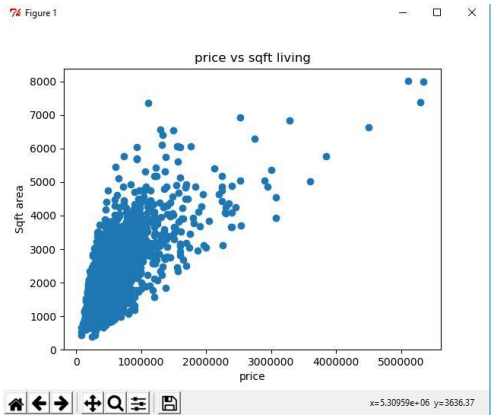


Fig-2: Price vs Sqft living

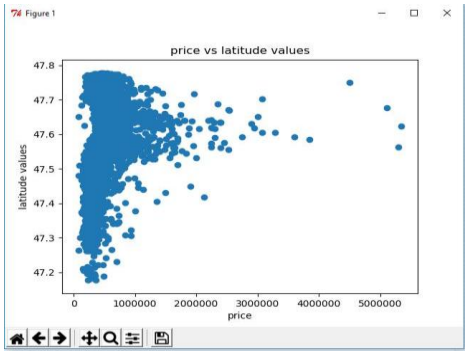


Fig3: Price vs Latitude values

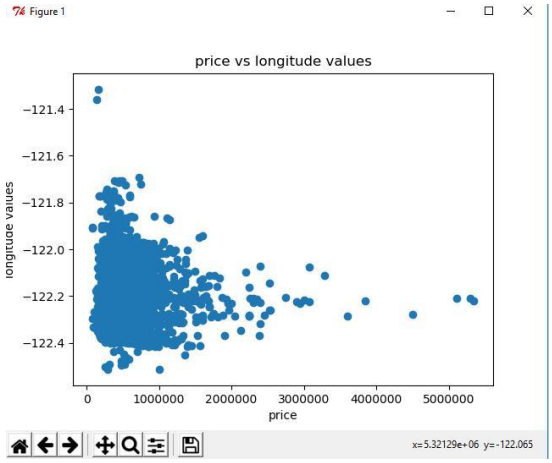


Fig – 4: Price vs Longitude values

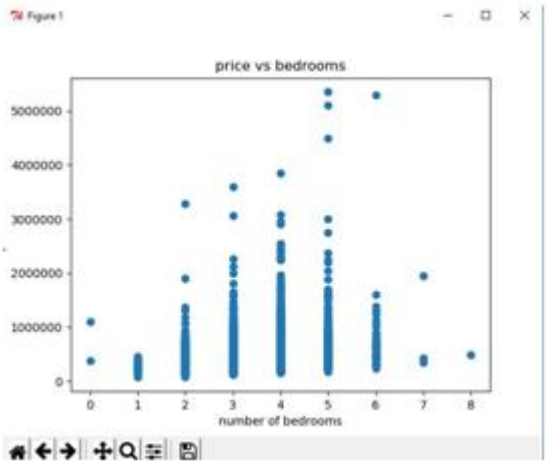


Fig –5 Price vs Bedrooms

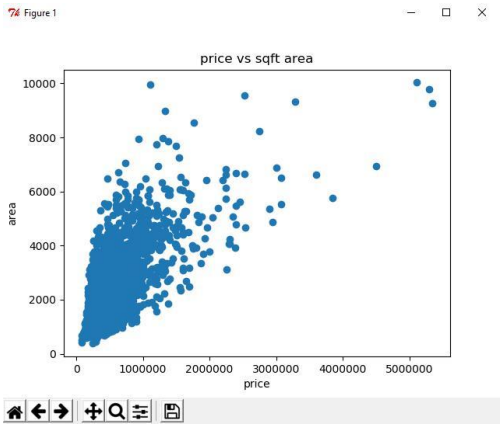


Fig –5 Price vs sqft area

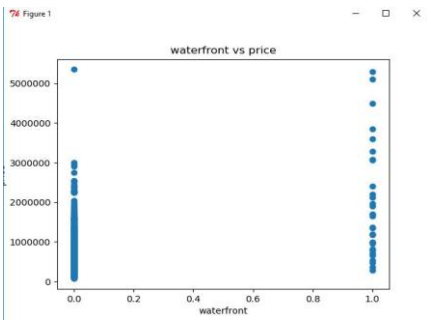
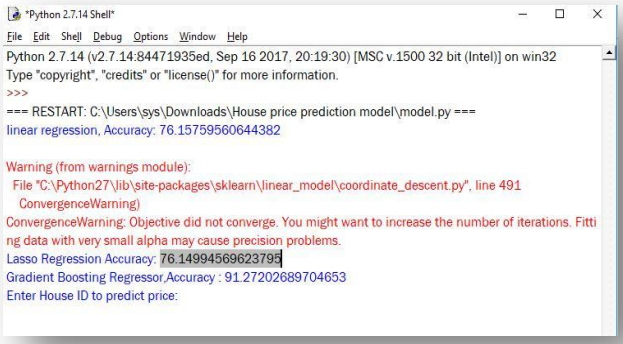


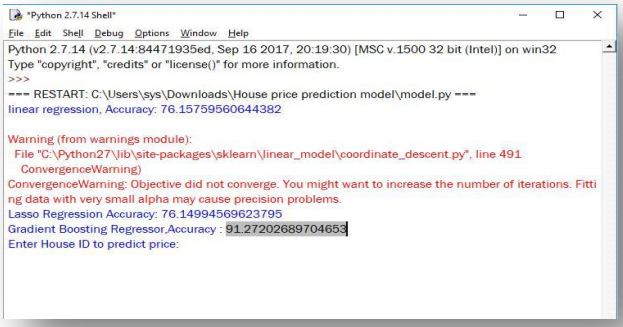
Fig -7: Price vs Total area

Linear Regression accuracy:

Lasso Regression accuracy:

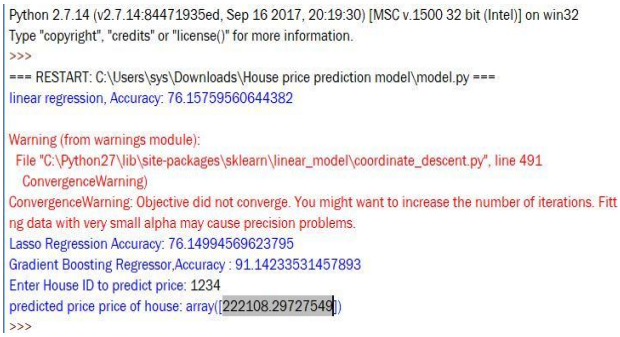


Gradient Boosting Regression accuracy:



Taking input ID for prediction:

Getting predicted price as output:



VI. CONCLUSION

We have managed out how to prepare a model that gives users for a novel best approach with take a gander at future lodging value predictions. A few relapse strategies have been investigated Furthermore compared, when arriving during a prediction strategy In light of XG support. Straight former imply works bring been utilized within our model, something like that that future value predictions will have a tendency towards All the more sensible values. We concocted an approach with use similarly as considerably information as time permits for our prediction system, by adopting those ideas from claiming gradient boosting.

Inspite of Hosting generated all the attempting provision that met our introductory requirements, there are Different upgrades that could be produced later on. These incorporate upgrades we didn't settle on because of constrained duration of the time. A real worry for the prediction framework may be the stacking period. Moreover, our data set takes more than one day should prepare. As opposed performing the computations sequentially, we might utilize various processors and parallel the computations involved, which might possibly decrease the preparation time Furthermore prediction period. Include All the more functionalities under the model, we can give choices for client with select a district alternately locale should produce those high temperature maps, as opposed to entering in the list.

REFERENCES

1. https://medium.com/@ageitgey/machine-learning-is-fun-80ea3ec3c471
2. https://www.sas.com/en_us/insights/analytics/machine-learning.html#machine-learning-importance
3. http://www.wired.co.uk/article/machine-learning-ai-explained
4. https://deeplearning4j.org/ai-machinelearning-deeplearning
5. David E. Rapach , Jack K. Strauss " Forecasting real housing price growth in the Eighth District states"
6. Vasilios Plakandaras+ and Theophilos♦, Rangan Gupta*, Periklis Gogas "Forecasting the U.S. Real House Price Index"
7. Gupta and Das (2010) Forecasting the US Real House Price Index: Structural and Non-Structural Models with and without Fundamentals
8. Rangan Gupta "Forecasting US real house price returns over 1831–2013: evidence from copula models"

AUTHORS PROFILE

Dr.G.Naga Satish is working as Associate Professor in the Department of CSE at BVRIT HYDERABAD. He has More than 17 Years of teaching experience. He has published more than 34 Publications in reputed International Journals.

Dr.Ch.V.Raghavendran is working as Professor in the Department of IT at Aditya College of Engineering & Technology, Surampalem. He has More than 22 Years of teaching experience. He has published more than 30 Publications in reputed International Journals. He authored for four books in Computer Sciences.

Mr.M.D.Suganan Rao is working as Assistant Professor in the Department of CSE at BVRIT HYDERABAD. He has More than 13 Years of teaching experience. His Area of Interest is Machine Learning, Analytics etc.,

Dr.Ch. Srinivasulu is working as Professor and HoD in the Department of CSE at BVRIT HYDERABAD. He has 22 Years of experience in Teaching and Industry. His Area of Interest is Soft Computing, Parallel Algorithms etc.,