Housing Price Prediction Model Using Machine Learning Algorithm- Lasso Regression

Ms. Sonia Kukreja
SCSE
Galgotias University
Greater Noida, U.P, India
sonia.kukreja@galgotiasuniversity.edu.in

Sajag Chauhan
SCSE
Galgotias University
Greater Noida, U.P, India
sajagchauhan7@gmail.com

Satyam Mandal
SCSE
Galgotias University
Greater Noida, U.P, India
satyam.mandal.77@gmail.com

Satya Prakash
SCSE
Galgotias University
Greater Noida, U.P, India
satya1573@gmail.com

ABSTRACT

Machine Learning have played an important role in some past years in image recognition, speech recognition, medical diagnosis, analyzing big set of data. With the help of machine learning algorithm, we have enhanced the security measures, customer services, automatic automobiles systems. Here we have explored, how predictive models can be very useful for predicting the sales price of the house on the basis of various factors. We have analyzed the housing dataset and some of the learning models. In the previous research based on linear regression. It has been found that the accuracy was not certain. In this model, we have used lasso regression to predict the prices as it has features like Framework able to adapt and stochastic for selection of models. The results were impressive as those were able to make a comparison with other existing house price prediction models. This model proves to be an improvement of the estates policies. The research use machine learning methodologies to explore new scenarios of house price prediction.

In this model, there were few models used like XGBoost, Lasso regression. These were used because of their order precision execution. XGBoost also shows that which variable have important effects on sale price. In that view, we suggest a house price prediction model that a real estate agent and buyer can use to get the best deal on basis of different factors and features of the house. This research exhibits a predicting model using lasso regression because of its accuracy and overcoming issue of correlated inputs.

CCS CONCEPTS

Information systems → Data Management systems →Graph based database models

KEYWORDS

Lasso regression, Gradient Boosting, Prediction Model

1 Introduction

Start with a question to buyer to describe their dream villa, apartment, shop, house. They would not start with the height of ceiling and the nearby transportation facilities. But, the competition's datasets proves that there are many more things that affects price negotiation than the other factors of the house like the balcony, number of bedrooms and many more. However, the features in the dataset may have impact on the price of house.

What is Learning? There is an example of Rat learning to avoid the poisonous food on the basis of smell and look, after a period of time. Normally, Rat eats every food after having a look and smell of it. What if it is poisonous. Rat will eat a small part of the food. If rat feels a illness in served thing. Rat stay away from that kind of poisonous thing even after. In future, that food is negatively impacted for the rat. The rat has labelled it as negative. So, rat will not have it again. This happens because of the learning from the past experiences. Similarly, Machine Learning plays an important part like, the creatures have past events for analyzing, exploring or differentiating of food. From view of previous example, what if the positive event is labelled as negative. The similar future events will also be affected. We analyze machine learning model that learns to filter out the spam emails. The trick used was to use the past experience of naming the emails as spam. The previous spam emails were remembered based on their names for future

references. When a new e-mail arrives, it is checked with the past spam emails. If, it gets matched. It will be trashed. Otherwise, goes to the message directory.

There was an another methodology called "Learn By Intuition". This do not suggest the process of training system

– the capacity of unaccounted emails. The advanced ability of a learner is to explore beyond the limits. It sounds like the inductive thinking. The emotional part of learning in previous example of rat shows some certain results. A new implementation of dealing with the Nutrition requirement and the taste.

Responsibilities beyond human abilities: an extra entire group on station that benefit from computer take-in systems are recognized by the research for significant and complex details : galactic facts, restorative chronicles turning under restorative experience, climate divination, genomic data dismemberment, search engines, even e -data with continuously rising sum of e-saved data accessible, therefore becomes clear that there is a possibility of preserve gathered alongside information chronicles about the severe majority of the data covered. This will approach people to bode overly little and often very perplexing. Also complicated data sets will offer an opportunity to reach new horizons by combining projects taking the borderless memory limits and increasing the level of transformation of PCs. Taking serious examples into account are relevant.

supervised vs unsupervised, as in consideration implies an interaction for learners or setting, one stop giving that it could differentiate the tasks for that relation as specified by that nature. Think of the email services as it would consider spam message vs aberrance detection considering as an illustrative example as well. For the task of spam detection, we think of an environment to the electronic-messages in which spam / not-spam label may be issued are prepared by the learner. The learner should further strengthen the support for claiming such training in order to assess a tenet for labelling a recently arrived email message. On the other hand, the tasks of fraud capturing, each remaining student receive a detailed procedure of electronic messages (non labeled) about example (planning) and reader's about detecting "unusual" transcripts.

2 Literature Survey

In past decade, the world economical crises have shaped the system with more focus on literacy and strategies circle. These were going to have a positive impact on the assets cost and lodging costs. As these were the one of the reason for clinching alongside the monetary movements. According to Lamer in 2007, some example established of 8 of 10 post during World war 2.

Varga's, Silva's argue that shifts in the cost of cottage assume a key role in majoring the phases of the business. If overabundance demand is to respond, rapidly driving ostensible house costs upwards as the wealth booms, growth and task in the harbour division grows rapidly. The decline in private money reduces exacerbated interest in those withdrawal periods. Ostensible house costs are also evident. As householders will not be able to reduce their expenses, ostensible house costs typically drop sluggishly. The bulk of abidance will be accomplished by lessens clinched alongside the amount of bargains, leading to a decrease in the construction segment and the vocation of lodging built. Moreover, true house costs decrease rapidly during withdrawal and subsidence Similarly, general inflationary trends decrease true home price with muggy observe costs.

Some editors have lately made claimed findings that property costs may allow slight moulding to assess yield. (Forni in 2003; stock and Watson in 2003; Das in 2010;). The division of lodging production refers to an costly and only nearby monetary activity recorded within the GDP. As a consequence, with regard to example, it represents an extensive portion of the economy's general wealth, house cost variances will make a point of GDP growth (Case etc, 2005). With regard to example, these body of evidence with distinct assets may also provide an sign of the future plan from claiming extension (Gupta Also Kabundi, 2010) for the growth of house costs. In general, the exact assessment of the way house costs are produced could make both home and fiscal strategy members a suitable apparatus.

In relation to U.S. house prices, there is massive literature publishing. In addition, strauss case (2007) uses an automatic regressive dispersed slack (ARDL) system, carrying 25 determinants for specific condition of the chose Reserve 's eighth part with conjecture of genuine lodging cost growth. They learn that a benchmark AR model can be beaten by ARDL models. On the 20 largest u, Rapach and strauss (2009) extend the same analysis. Faced along urban decay because of the deindustrialization, produced creativity, agent of government. States rely on ARDL models that look at variables at the local, territorial or national level. Once again, the creators draw comparative solutions regarding fact that joint forecasts on algo's for various slack systems are combined.

3 Design Approach

3.1. Linear Regression :

Straight relapse endeavors to demonstrate the connection between two factors by fitting a direct condition to watched information. One variable is view as illustrative variable, and the other is view as reliant variable. For instance, a modeler should relate the loads of people to their statures utilizing a direct relapse model.

- One variable, indicated x, viewed for indicator, logical, or free factor.
- another variable, signified y, viewed for reaction, result, or ward factor

As the other segment λ maintains the screening for punishment. When λ = 0 ,coefficient becomes Similarly as fundamental straight backslide. When λ = ∞: continually as coefficient becomes zero. At moment that λ >0 .

3.2. Multiple Regression :

Multiple regressions states that it utilized to audit if exists a factually essential affiliation the center of bundle of factors. It gets utilized for finding designs out of people sets of data.

3.3. The Cost Function :

Consequently suppose, you extended the size of a particular shop, the spot you anticipated that those arrangements may an opportunity to be higher. Be that despite extending the size, the deals in the shop didn't grow that much. Something like that those cost associated Previously, growing those range of shop, accommodated you negative results. Thus, we need on limit these expenses. Here, we introduce a cost work, which is essentially used in portraying- Figures and Tables.

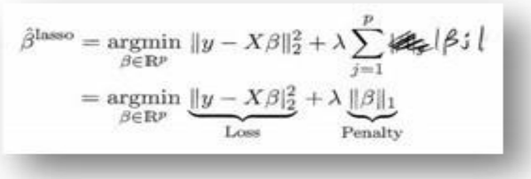
$$J(\theta_0, \theta_1) = \frac{1}{2m}\sum_{i=1}^{m}\left(h_\theta\left(x^{(i)}\right) - y^{(i)}\right)^2$$

3.4. Lasso Regression :

Lasso regression states that details would be analyzed between certain recurrence models that may be available. For a sample, the other regression model is shown and the formula is added as a comparison.

1. LASSO states for (Least-Absolute Shrinkage and Selection Operator).
2. This regression has major standout between regular schedules that makes miserly models in the region for immense number for highlights, the plane seems expansive whichever of the accompanying among the things.
3. Improving the model's overfit tendency. Sever 10 variables can be calculated by abundance. This sufficiently contributes to machine research.
4. The role of guaranteeing an enormous amount or billions of characteristics could increase.

Minimization goal is an addition of LS Obj + λ (sum of outright esteem of coefficients). In comparison, λ seems to be the switching factor as determines the target for regularization, their spot LS Obj remains for the necessary squares criterion that may be nothing yet straight relapse target without normalization. With such increasing content of λ, the inclination will develop and the gap will decrease with respect to the example of the calculation of shrinkage increments (λ)The lasso regression estimate is defined as ∞: What's more, we'd like to change the ones below two plans while λ is in the centre of the two borders for fundamental straight reverse reversals.



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

Applying the model for y coefficient and then for X. Then, joining and considering those coefficients.

Gradient boost is a computer taking in relapse technique. Even positioning concerns, which creates a prediction model from claiming helpless prediction models in a group's framework.

A prescient model's consistency can be seen in two separate ways: Potentially, by persistently having characteristic construction. To add the boosting statistic straight well out.

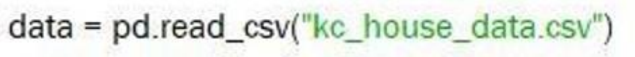
There is a crucial amount in which projections are raised.

- Gradient Boosting
- XG Boosting
- Ada Boosting

Faster computation will be a champion among those. The larger part able Taking in considerations familiar in the last one twenty quite a while. It may have been planned to arrange issues, yet all the it very well may be created should backslide as well. The motivation to slope boosting may have been A strategy. That joins those yields about huge parts "frail" classifiers to handle A fit "board of trustees. " a weak approach (Decision tree) lead to individual their slip cost is primary better than unpredictable speculating.

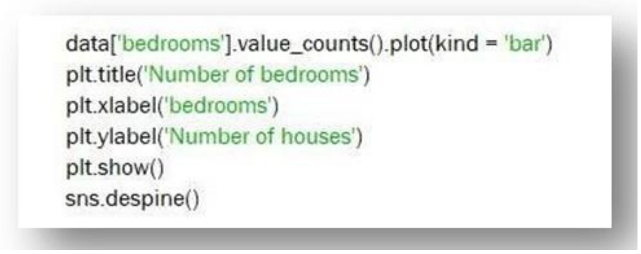
4 Implementation

Read the data for plotting graphs:



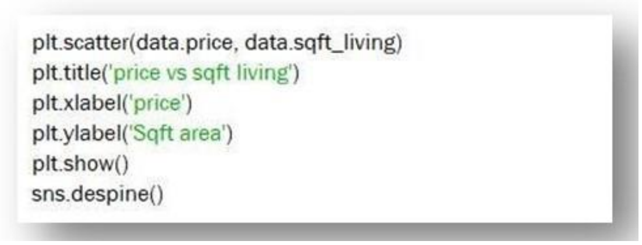
```
data = pd.read_csv("kc_house_data.csv")
```

Using Fetched data, graph between number of houses vs number of bedrooms:



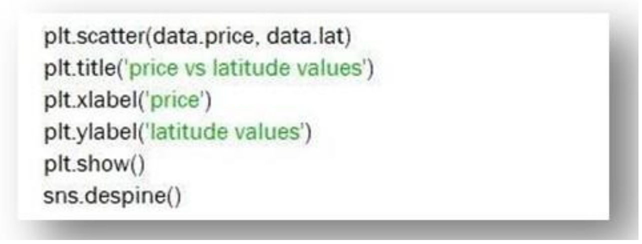
```
data['bedrooms'].value_counts().plot(kind = 'bar')
plt.title('Number of bedrooms')
plt.xlabel('bedrooms')
plt.ylabel('Number of houses')
plt.show()
sns.despine()
```

Graph between Price and Living area:



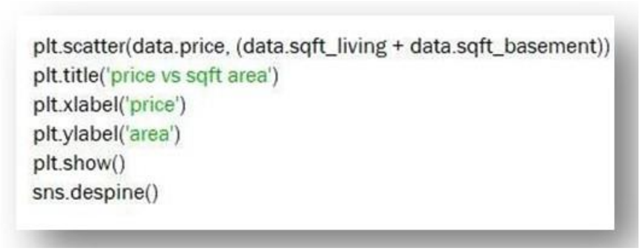
```
plt.scatter(data.price, data.sqft_living)
plt.title('price vs sqft living')
plt.xlabel('price')
plt.ylabel('Sqft area')
plt.show()
sns.despine()
```

Graph between Price and Latitudes:



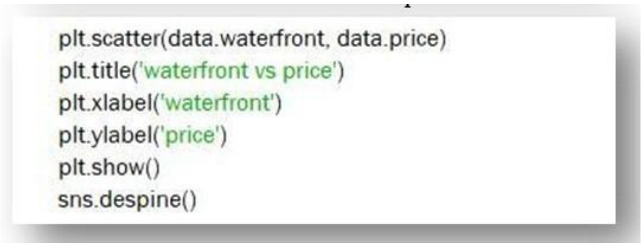
```
plt.scatter(data.price, data.lat)
plt.title('price vs latitude values')
plt.xlabel('price')
plt.ylabel('latitude values')
plt.show()
sns.despine()
```

Graph between Price and Area:



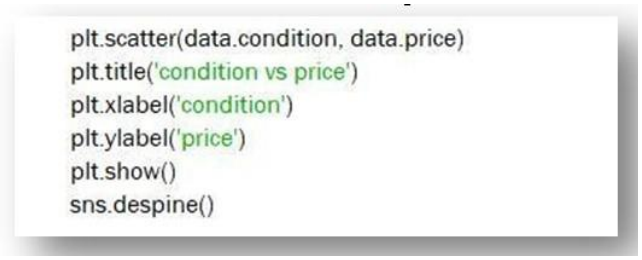
```
plt.scatter(data.price, (data.sqft_living + data.sqft_basement))
plt.title('price vs sqft area')
plt.xlabel('price')
plt.ylabel('area')
plt.show()
sns.despine()
```

Graph between Price and Waterfront:



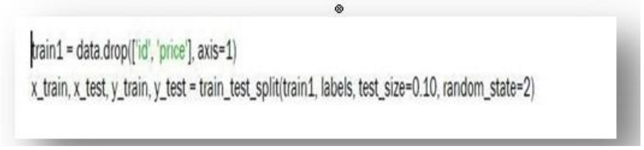
```
plt.scatter(data.waterfront, data.price)
plt.title('waterfront vs price')
plt.xlabel('waterfront')
plt.ylabel('price')
plt.show()
sns.despine()
```

Graph between Condition and Price:



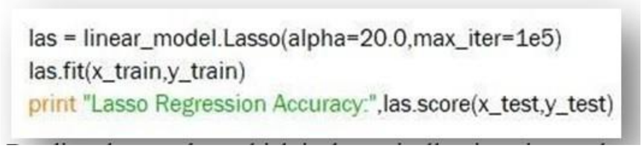
```
plt.scatter(data.condition, data.price)
plt.title('condition vs price')
plt.xlabel('condition')
plt.ylabel('price')
plt.show()
sns.despine()
```

Feeding with Training data:



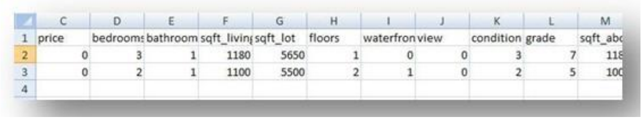
```
train1 = data.drop(['id', 'price'], axis=1)
x_train, x_test, y_train, y_test = train_test_split(train1, labels, test_size=0.10, random_state=2)
```

Model Training with Lasso Regression:

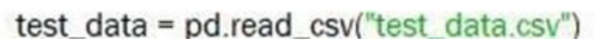


```
las = linear_model.Lasso(alpha=20.0,max_iter=1e5)
las.fit(x_train,y_train)
print "Lasso Regression Accuracy:",las.score(x_test,y_test)
```

Reading data dynamically through excel sheet:

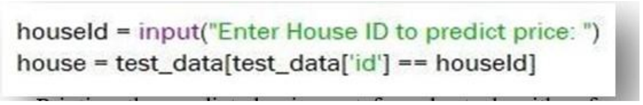


	C	D	E	F	G	H	I	J	K	L	M
1	price	bedrooms	bathroom	sqft_living	sqft_lot	floors	waterfron	view	condition	grade	sqft_abo
2	0	3	1	1180	5650	1	0	0	3	7	118
3	0	2	1	1100	5500	2	1	0	2	5	100
4											



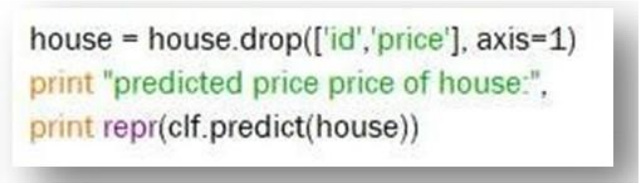
```
test_data = pd.read_csv("test_data.csv")
```

We take input (id) to predict price:



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

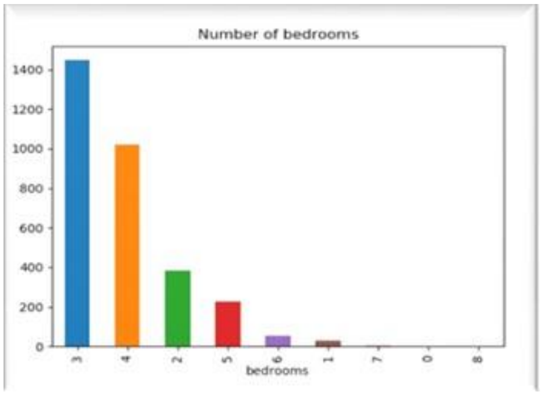
Printing predicted cost through highly efficient algorithm:



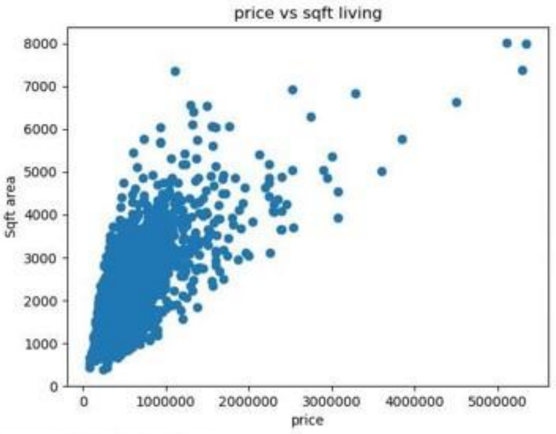
```
house = house.drop(['id','price'], axis=1)
print "predicted price price of house:",
print repr(clf.predict(house))
```

5 Results

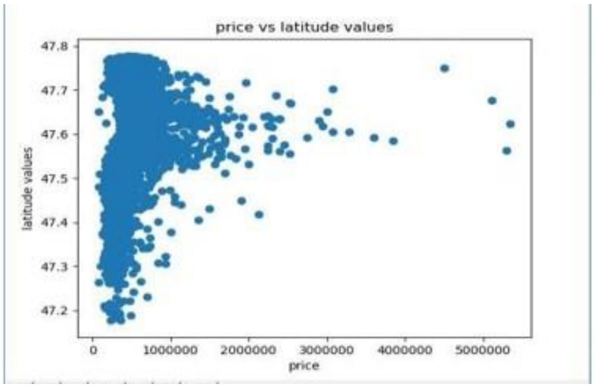
Through the above code, we get the predicted prices of house and some plots. Those graphs help to correlate between price and various predictor variables.



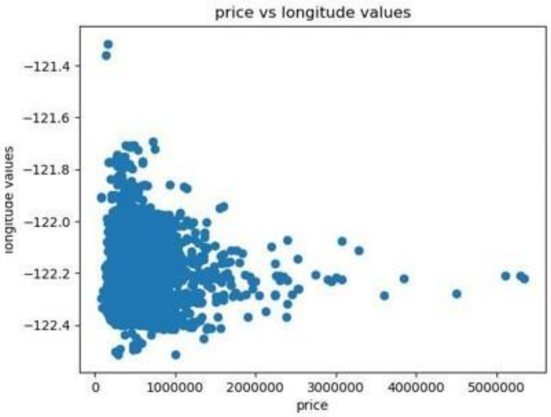
No. of houses vs No. of bedrooms



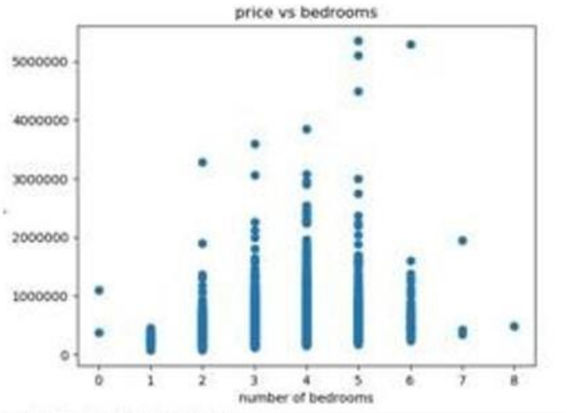
Price vs SQFT living



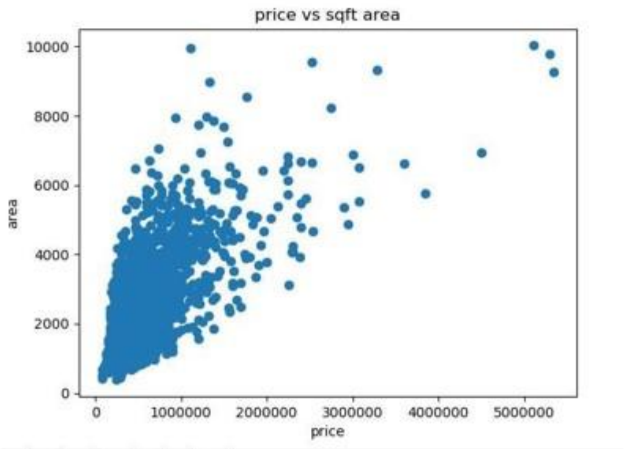
Price vs Latitude Values



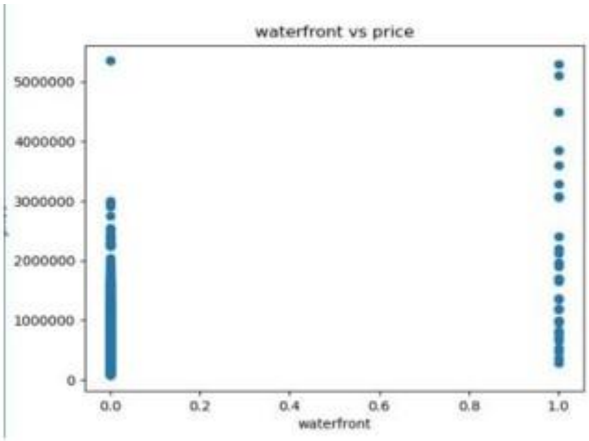
Price vs Longitude Values



Price vs Bedrooms

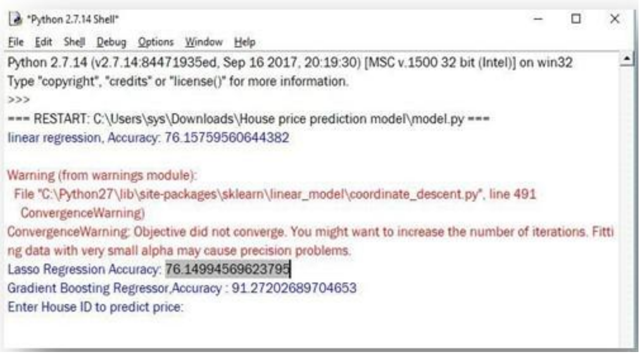


Price vs SQFT Area



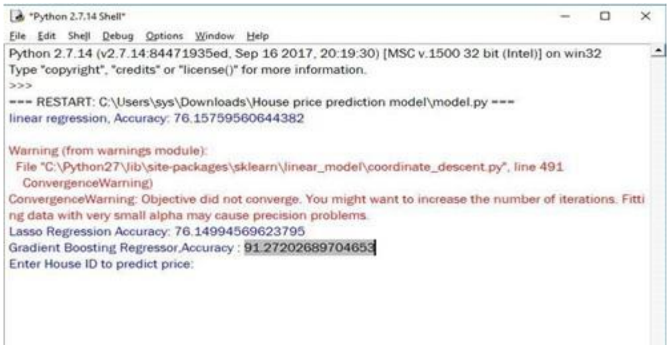
Waterfront vs price

Lasso Regression Accuracy will be:



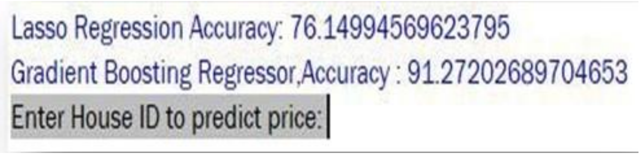
```
Python 2.7.14 (v2.7.14:84471935ed, Sep 16 2017, 20:19:30) [MSC v.1500 32 bit (Intel)] on win32
Type "copyright", "credits" or "license()" for more information.
>>>
=== RESTART: C:\Users\sys\Downloads\House price prediction model\model.py ===
linear regression, Accuracy: 76.15759560644382

Warning (from warnings module):
  File "C:\Python27\lib\site-packages\sklearn\linear_model\coordinate_descent.py", line 491
    ConvergenceWarning)
ConvergenceWarning: Objective did not converge. You might want to increase the number of iterations. Fitti
ng data with very small alpha may cause precision problems.
Lasso Regression Accuracy: 76.14994569623795
Gradient Boosting Regressor,Accuracy : 91.27202689704653
Enter House ID to predict price:
```

Gradient Boosting Regression Accuracy will be:



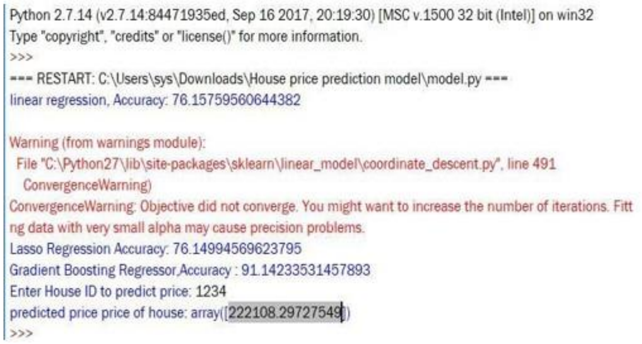
```
Python 2.7.14 (v2.7.14:84471935ed, Sep 16 2017, 20:19:30) [MSC v.1500 32 bit (Intel)] on win32
Type "copyright", "credits" or "license()" for more information.
>>>
=== RESTART: C:\Users\sys\Downloads\House price prediction model\model.py ===
linear regression, Accuracy: 76.15759560644382

Warning (from warnings module):
  File "C:\Python27\lib\site-packages\sklearn\linear_model\coordinate_descent.py", line 491
    ConvergenceWarning)
ConvergenceWarning: Objective did not converge. You might want to increase the number of iterations. Fitti
ng data with very small alpha may cause precision problems.
Lasso Regression Accuracy: 76.14994569623795
Gradient Boosting Regressor,Accuracy : 91.27202689704653
Enter House ID to predict price:
```

For other inputs:



```
Lasso Regression Accuracy: 76.14994569623795
Gradient Boosting Regressor,Accuracy : 91.27202689704653
Enter House ID to predict price:
```

Output (Predicted):



```
Python 2.7.14 (v2.7.14:84471935ed, Sep 16 2017, 20:19:30) [MSC v.1500 32 bit (Intel)] on win32
Type "copyright", "credits" or "license()" for more information.
>>>
=== RESTART: C:\Users\sys\Downloads\House price prediction model\model.py ===
linear regression, Accuracy: 76.15759560644382

Warning (from warnings module):
  File "C:\Python27\lib\site-packages\sklearn\linear_model\coordinate_descent.py", line 491
    ConvergenceWarning)
ConvergenceWarning: Objective did not converge. You might want to increase the number of iterations. Fitt
ng data with very small alpha may cause precision problems.
Lasso Regression Accuracy: 76.14994569623795
Gradient Boosting Regressor,Accuracy : 91.14233531457893
Enter House ID to predict price: 1234
predicted price price of house: array([222108.29727549])
>>>
```

6 Conclusion

Here we are proposing a model which provides customer and buyer with some gander at potential lodging value forecasts for a new, best method. In addition, some other strategies are analyzed in the form of XG help while reaching to the proposed prediction. In our model, straight former works mean that works have been used, anything like that future value forecasts will tend toward all the more sensible values. By embracing those ideas for the prediction system. In addition, it will take near by to two days or a week to prepare our data collection. We could use several processors along with computations attached, as opposed to conducting the computations sequentially, which could also minimize the readiness time. Including all functionalities in the model, as confront to be in the list, we can offer choices for customers with selecting a preferred location. On second thought, locale have to generate those high demanding requirements.

ACKNOWLEDGMENTS

We have done our hardest to convey the details in different words and explanations as simply as possible. So that, it will be comprehended by the big spectrum of researchers, analyst and students for further studies.

We complete research study under guidance of **Ms. Sonia Kukreja**. We recognize the wise guidance, support and knowledge provided by our guide.

Simple acknowledgment must not pay the obligation We are obliged to our parents, during the entirety of this endeavor, to provide direct/ indirect assistance.

REFERENCES

[1] https://www.kaggle.com/c/house-prices-advanced-regression- techniques

[2] https://medium.com/@ageitgey/machine-learning-is-fun-80ea3ec3c471

[3] http://www.wired.co.uk/article/machine-learning-ai-explained

[4] https://towardsdatascience.com/create-a-model-to-predict-house-prices-using-python-d34fe8fad88f

[5] https://www.coursera.org/lecture/ml-foundations/predicting-house-prices-a-case-study-in-regression-aI5W6

[6] https://medium.com/analytics-vidhya/predicting-house-prices-using-classical-machine-learning-and-deep-learning-techniques-ad4e55945e2d

[7] https://escholarship.org/uc/item/3ft2m7z5

[8] https://www.researchgate.net/publication/340006049_MACHINE_LEARNINGHOUSE_SALE_PRICES_PREDICTION_USING_LINEAR_REGRESSION

[9] https://www.ijitee.org/