House Price Prediction

Approach using Data Mining Techniques.

KANAGADDI SUNIL JOSHI
DEPT. of MATHEMATICS AND STATISTICS
UNIVERSITY at ALBANY,
THE STATE UNIVERSITY OF NEW YORK
sjoshi2@albany.edu

Abstract— **The system which we propose here is a approach for recommendation of housing for a particular student depending on the live data collected from Craigslist in accordance to the Albany area. We used a Regression model and a Prediction scheme (which are related to mining massive data). This system can effectively predict the house price based on input features like number of bedrooms, square feet area and year of construction for university student and analysis which we have implemented in this project. , this method can be applied on a larger scale and also put to test with respect to live data .**

KEYWORDS

SVM , LINEAR REGRESSION, CRAIGLIST.LASO REGRESSION ,MSE

I. Introduction

The gap between when a student is new in a city and find it difficult to look for accommodations within university area takes a lot of time and efforts .The primitive thought of eliminating the dilemma among the students motivated us to use this approach and develop a better solution for existing problem with different housing website like craigslist , zillo ,etc. The problem was to develop a efficient approach .

The problem was to develop a efficient to help student in finding off campus accommodation . Since as student we faced the similar problem in finding the accommodation , so this is the main motivation behind the project we choose to work upon

II. METHODOLOGY

First, we get into description of data collection we used craigslist as main source of data to collect live listing of houses/apartments. We used input filters as site as Albany ,category as apartments ,ma price as $1500 and private bedroom as TRUE , as most of the student are interested in looking for accommodations with similar features .the data we collected is the most recent data as we sorted the result from craigslist housing scrapper by newest and also limited the request to 1000 entries , we then split data to build an model and predict the outcomes . later we applied data mining techniques to predict, test, train and evaluate the model performance to conclude the predictions for a particular set of trained data .

III. IMPLEMENTATION

Once the data Is collected and saved into a pandas dataframe , we built a regression model based on using input feature as number of bedroom , square feet area, year of construction and target value as house price. We then predicted the price outcome by using regression predict method .

A step wise approach could be stated as :

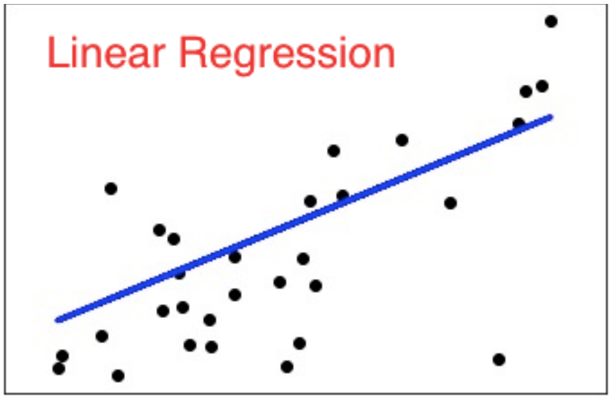
- Data collection
- Data pre-processing –

  The missing values in columns like bedrooms ,area , geotag are filled is replaced mode values of. The respective column .We then removed special character ,emoticons and HTML tag from text data .
- Regression
- Implementation of regression models like
  - Linear regression
  - Lasso
  - SVR
  - Ensemble
  - Decision Tree Regressor
- We use the above stated models as our aim is to predict continuous numerical data we then compare all the models score , MSE(mean squared error) and choose the best fit model .

IV. ALGORITHMS & TECHNIQUES USED.

This project is used a regression model in order to predict the house prices in machine learning the regression models are supervise leaning model that analysis data and then predict the outcome. the basic regression model takes a set of input data and predict for each given input based on best fit line. we also used other regression algorithms like lasso, SVR ,Decision tree regressor. Here we will discuss two major regression algorithms used for this project.

The graphical representation of linear regression model is as described below: -



Linear regression is used to estimate real world values like cost of houses, number of calls, total sales etc. based on continuous variable(s). Here, we establish relationship between dependent and independent variables by fitting a best line. This line of best fit is known as **regression line** and is represented by the linear equation **Y= a *X + b**.

In this equation –

Y – Dependent Variable

a – Slope

X – Independent variable

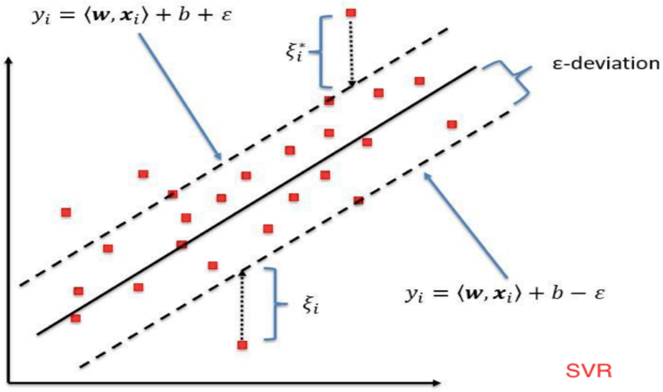
b – Intercept

These coefficients **a** and **b** are derived based on minimizing the sum of squared difference of distance between data points and regression line.

Assuming line of best fit for a set of points is –

y = a + b * x

The next technique used for regression is SVR, which is basically to decide a decision boundary at 'e' distance from the original hyper plane such that data points closest to the hyper plane or support vectors are within that boundary line.


$y_i = \langle \boldsymbol{w}, \boldsymbol{x}_i \rangle + b + \varepsilon$
ξ_i^*
ε-deviation
ξ_i
$y_i = \langle \boldsymbol{w}, \boldsymbol{x}_i \rangle + b - \varepsilon$
SVR

The comparison of SVR and Linear Regression Approaches are compared using the accuracy score and Mean square error.

V. RELATED WORK

Our Problem as stated above is best solved using both the SVR and the Linear regression models for the prediction. Though Different approaches can be used for solving this problem, we chose to use these techniques for a appropriate , fast , efficient and better deserved output.

The Predictions are made based on Sklearn model predict method which is an array of house prices , in addition to the price we can also predict number of bedrooms, square feet area using the same regression models.

. VI. CONTRIBUTION

We as a Team Contributed individually and we were motivated to develop this project with a whole single mindset and this motive made us put in equal efforts towards the course of the project. We individually Split up the Task of developing the system into Individual Modules primitively. Out of which Shubham pal was assigned to do the Model Building ,Pratik was assigned the task of Data pre-processing.

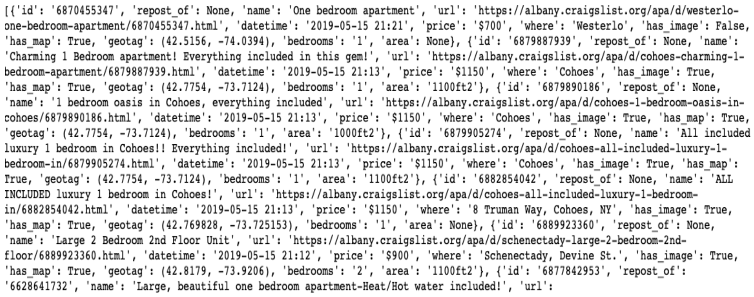
I was involved in Data collection. So my contribution to this system was that I collected data from Craigslist website using python CraigslistHousing scraper , in which i filtered data using parameters like site as Albany , maximum price as $1500 , private bedrooms as True , housing type as apartment. The data collected was sorted by newest which means every time we get most recent house listings.

I then limited the requests to 1000 , as Craigslist doesn't have API so its not fair to pull large amount of data using the scraper and also it's time consuming as well.

Finally this similar approach is applied on other housing websites like Zillow, Trulia and gather as much new listings as we can so that we can make better predictions for price and other target variables like bedrooms, square feet area.

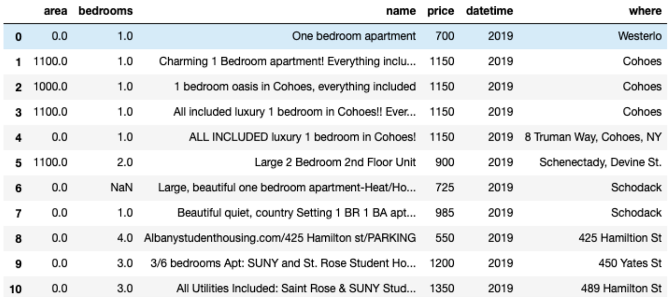
VII. RESULTS & ANALYSIS

1.) Snapshot of the raw data collected from CraigslistHousing.



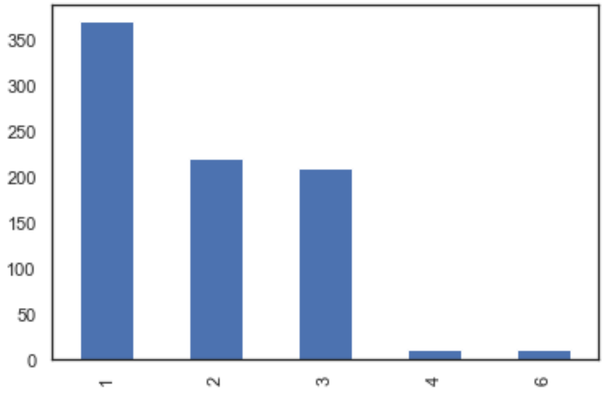
```
[{'id': '6870455347', 'repost_of': None, 'name': 'One bedroom apartment', 'url': 'https://albany.craigslist.org/apa/d/westerlo-
one-bedroom-apartment/6870455347.html', 'datetime': '2019-05-15 21:21', 'price': '$700', 'where': 'Westerlo', 'has_image': False,
'has_map': True, 'geotag': (42.5156, -74.0394), 'bedrooms': '1', 'area': None}, {'id': '6879887939', 'repost_of': None, 'name':
'Charming 1 Bedroom apartment! Everything included in this gem!', 'url': 'https://albany.craigslist.org/apa/d/cohoes-charming-1-
bedroom-apartment/6879887939.html', 'datetime': '2019-05-15 21:13', 'price': '$1150', 'where': 'Cohoes', 'has_image': True,
'has_map': True, 'geotag': (42.7754, -73.7124), 'bedrooms': '1', 'area': '1100ft2'}, {'id': '6879890186', 'repost_of': None,
'name': '1 bedroom oasis in Cohoes, everything included', 'url': 'https://albany.craigslist.org/apa/d/cohoes-1-bedroom-oasis-in-
cohoes/6879890186.html', 'datetime': '2019-05-15 21:13', 'price': '$1150', 'where': 'Cohoes', 'has_image': True, 'has_map': True,
'geotag': (42.7754, -73.7124), 'bedrooms': '1', 'area': '1000ft2'}, {'id': '6879905274', 'repost_of': None, 'name': 'All included
luxury 1 bedroom in Cohoes!! Everything included!', 'url': 'https://albany.craigslist.org/apa/d/cohoes-all-included-luxury-1-
bedroom-in/6879905274.html', 'datetime': '2019-05-15 21:13', 'price': '$1150', 'where': 'Cohoes', 'has_image': True, 'has_map':
True, 'geotag': (42.7754, -73.7124), 'bedrooms': '1', 'area': '1100ft2'}, {'id': '6882854042', 'repost_of': None, 'name': 'ALL
INCLUDED luxury 1 bedroom in Cohoes!', 'url': 'https://albany.craigslist.org/apa/d/cohoes-all-included-luxury-1-bedroom-
in/6882854042.html', 'datetime': '2019-05-15 21:13', 'price': '$1150', 'where': '8 Truman Way, Cohoes, NY', 'has_image': True,
'has_map': True, 'geotag': (42.769828, -73.725153), 'bedrooms': '1', 'area': None}, {'id': '6889923360', 'repost_of': None,
'name': 'Large 2 Bedroom 2nd Floor Unit', 'url': 'https://albany.craigslist.org/apa/d/schenectady-large-2-bedroom-2nd-
floor/6889923360.html', 'datetime': '2019-05-15 21:12', 'price': '$900', 'where': 'Schenectady, Devine St.', 'has_image': True,
'has_map': True, 'geotag': (42.8179, -73.9206), 'bedrooms': '2', 'area': '1100ft2'}, {'id': '6877842953', 'repost_of':
'6628641732', 'name': 'Large, beautiful one bedroom apartment-Heat/Hot water included!', 'url':
```

2.) After Processing of Data



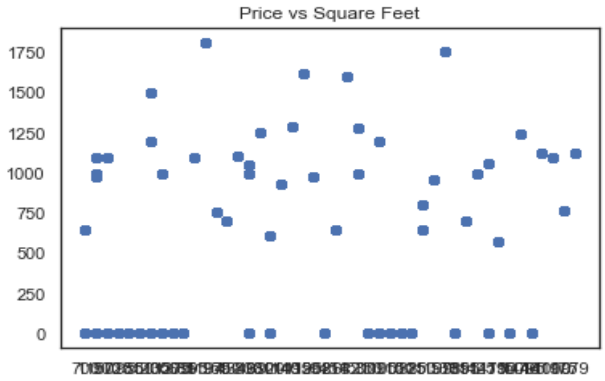
	area	bedrooms	name	price	datetime	where
0	0.0	1.0	One bedroom apartment	700	2019	Westerlo
1	1100.0	1.0	Charming 1 Bedroom apartment! Everything inclu...	1150	2019	Cohoes
2	1000.0	1.0	1 bedroom oasis in Cohoes, everything included	1150	2019	Cohoes
3	1100.0	1.0	All included luxury 1 bedroom in Cohoes!! Ever...	1150	2019	Cohoes
4	0.0	1.0	ALL INCLUDED luxury 1 bedroom in Cohoes!	1150	2019	8 Truman Way, Cohoes, NY
5	1100.0	2.0	Large 2 Bedroom 2nd Floor Unit	900	2019	Schenectady, Devine St.
6	0.0	NaN	Large, beautiful one bedroom apartment-Heat/Ho...	725	2019	Schodack
7	0.0	1.0	Beautiful quiet, country Setting 1 BR 1 BA apt...	985	2019	Schodack
8	0.0	4.0	Albanystudenthousing.com/425 Hamilton st/PARKING	550	2019	425 Hamiltion St
9	0.0	3.0	3/6 bedrooms Apt: SUNY and St. Rose Student Ho...	1200	2019	450 Yates St
10	0.0	3.0	All Utilities Included: Saint Rose & SUNY Stud...	1350	2019	489 Hamilton St

3.) Formed a histogram to see the count of bedrooms .



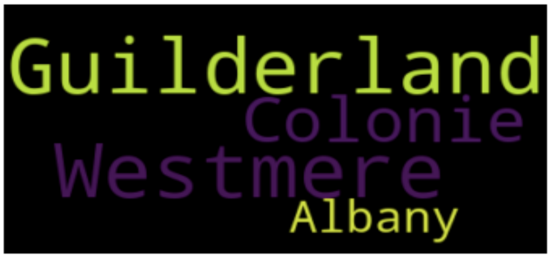
As we can see count of 4-6are very few so that doesn't make in feature selection for training the model , so we removed them.

Also plotted a scatter plot of price vs square feet area to see the correlation between them, it turns out that they have postive coreelation.

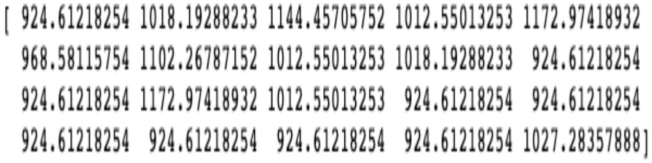


4) Word Cloud of the listings .

Basically the word cloud is to see where most of the house listings are located.



6.) Snapshot of the Predicted house prices.



```
[ 924.61218254 1018.19288233 1144.45705752 1012.55013253 1172.97418932
  968.58115754 1102.26787152 1012.55013253 1018.19288233  924.61218254
  924.61218254 1172.97418932 1012.55013253  924.61218254  924.61218254
  924.61218254  924.61218254  924.61218254  924.61218254 1027.28357888]
```

When We compare the Predicted Outcomes,the outputs are almost the same,but SVR gaves us the highest model score comapred to other models.

VIII . FUTURE SCOPE

The method we propose is still naive and would pop up a lot of errors while coding the actual logic and ample time is needed to implement this and analyze and apply the systematic approach, as every time we scrape we get new data , so which is going to impact the ,models score and predictions. Better techniques would be to see whether other features like geocodes, location will also impact on predicting house price.

IX. Advantages & Disadvantages

We considered only 3 input features to train the model, that makes sense as house price depends majorly on number of bedrooms, square feet area and year of construction, which are strongly correlated with the target variable price.

Since it is a low-scale implementation, we used a fixed set of data by limiting to 1000 data points, but in case we use a whole new large scale of data, we might build an accurate model and predict the house price based on several other features.

X. CONCLUSION

Hence,after analysing the data and making predictions I can say that this approach is really useful in different applications like real estate industries, insurance companies and banking where house price plays a key role so this method can be implemented in large scale to deploy.

References

1) *https://medium.com/coinmonks/support-vector-regression-or-svr-8eb3acf6d0ff*

2) https://towardsdatascience.com/create-a-model-to-predict-house-prices-using-python-d34fe8fad88f

3) *https://www.tutorialspoint.com/machine_learning_with_python/machine_learning_with_python_algorithms.htm*