ESTIMATING THE PRICE OF HOUSES USING MACHINE LEARNING

Mr. G. Hari Hara Kumar [1], M. Lahari [2], K. Jana Priyanka [3], N. Ajay Kumar [4], M. Samson Raju [5]

[1] Assistant Professor, Department of CSE, Ramachandra College of Engineering, Eluru, A.P
[2,3,4,5] UG Students, Department of CSE, Ramachandra College of Engineering, Eluru, A.P

ABSTRACT

Our Project explores the question of how house prices in five different counties are affected by housing characteristics (both internally, such as number of bathrooms, bedrooms, etc. and externally, such as public schools' scores or the walkability score of the neighborhood). Using data from sold houses listed on Zillow, Trulia and Redfin, three prominent housing websites, this paper utilizes both the hedonic pricing model (Linear Regression) and various machine learning algorithms, such as Random Forest (RF) and Support Vector Regression (SVR), to predict house prices. The models' prediction scores, as well as the ratio of overestimated houses to underestimated houses are compared against Zillow's price estimation scores and ratio. Results show that SVR gives a better price prediction score than the Zillow's baseline on the same dataset of Hunt County (TX) and RF gives close or the same prediction scores to the baseline on three other counties. Moreover, this paper's models reduce the overestimated to underestimated house ratio of 3:2 from Zillow's estimation to a ratio of 1:1. This paper also identifies the four most important attributes in housing price prediction across the counties as assessment, comparable houses' sold price, listed price and number of bathrooms.

1. INTRODUCTION

According to the US Census Bureau, 560,000 houses were sold in the United States in 2016. In addition, 65% of all American families owned houses in 2016. For the Americans who sold and bought these houses, a good housing price prediction would better prepare them for what to expect before they make one of the most important financial decisions in their lives. A recent report from the Zillow Group, a popular housing database website, indicates that house sellers and buyers are increasingly turning to online research in order to estimate house price before contacting real estate agents. Researching how much the house you are interested in is worth on your own can be difficult for multiple reasons. One reason is that there many factors that influence the potential price of a house, making it more complicated for an individual to decide how much a house is worth on their own without external help. This can lead to people making poorly informed decisions about whether to buy or sell their houses and which prices are reasonable. Because houses are long term investments, it is imperative that people make their decisions with the most accurate information possible. Therefore, housing websites such as Zillow, Trulia and Redfin exist to provide estimations of housing valuations based on the houses' characteristics, at no cost.

However, the estimations provided by these housing websites are not always accurate. For example, Zillow states that their housing price prediction algorithm, called “Zestimate”, only estimates 54.4% of houses within the 5% of their actual sale prices. For Trulia, only 48.2% of houses have Trulia-estimated prices to be within the 5% range of their actual sold prices. Therefore, the first question of this project is “Whether we can outperform Zestimate’s prediction score or come close to it”. In this project, we define the prediction score as the percentage of houses whose estimated prices fall within the 5% range of their actual sold prices. Using this project’s datasets and Zestimate as the predictions, we compute Zillow’s prediction scores and use them as the baselines to see how well our own models perform. We choose Zillow’s estimator as a benchmark instead of its competitors’ because Zillow is widely regarded as the most popular housing website due to its large databases of 110 million houses and their 11 years of expertise in pricing estimations. According to Hitwise, a consumer analytics company, Zillow’s market share, based on online visits to the site, is 27.2% in 2016, while the numbers for Trulia and Redfin are 9.4% and 3.7%, respectively. Zillow tends to overestimate their listed properties, meaning the Zestimates are higher than the actual sold prices of the houses. In the dataset
of 1,457 sold houses I collected, the ratio of overestimated houses to underestimated houses is 3 to 2. Hollas, Rutherford and Thomson (2010) studies Zillow’s estimations of single-family houses and finds that 80% of their housing sample gathered from Zillow are overpriced by Zestimate. For a house seller who prices his house based on Zillow’s suggestion, he/she is likely to list his/her house for more than what it is worth. According to Zillow research in 2016, if a house is priced above its true market valuation, it tends to stay on the market five times longer compared to a house that is well-priced, suggesting a string penalty for overpricing houses. Moreover, the same research suggests that houses that have been on the market for two months can lose 5% of its original listed price. Asabere and Huffman (1993) also supports the theory of a reversed correlation between a house’s time on the market and its final sold price. Therefore, the second question of this project is “Whether our models can get rid of this overestimation problem?”. The final question of this project is “What the most important factors affecting housing prices are". In order to answer the three questions listed above, this project proposes using both the hedonic pricing model and various machine learning algorithms.

2. LITERATURE SURVEY

The components that influence the land cost must be considered and their effect on cost has additionally to be demonstrated. An examination of the past information uncovered that the costs demonstrate a non-direct trademark. It is construed that building up a basic direct numerical relationship for these time-arrangement information is found not reasonable for anticipating. Thus, it wounds up basic to build up a non-direct model which fits the information trademark to dissect and estimate future patterns.

1. **AUTHOR:** R. Manjula

CONTENT: R. Manjula have come across with some calculation called as arrange plunge calculation which radically decreases the calculation workload, limiting the number of highlights while choosing the main essential ones. Organizations

like"Zillow.com","magicbricks.com", consists of a vast dataset of houses whose costs they anticipate utilizing machine learning. One of the procedures they utilize is Linear and multivariate regression, profound learning to take in the idea of models from the past outcomes.

2. **AUTHOR:** Nissan Pow

CONTENT: Nissan Pow anticipated both soliciting and sold costs from land properties in view of highlights, for example, topographical area, living territory, and number of rooms, and so forth. Extra geological highlights, for example, the closest police headquarters and re station were removed from the Montr'eal Open Data Portal.

They used techniques, like direct regression, Support Vector Regression (SVR), k-Nearest Neighbors (kNN), and Regression Tree/Random Forest Regression. Their result stated the soliciting cost with a mistake from 0.0985 utilizing an outfit of kNN and Random Forest calculations.

3. **AUTHOR:** Eduard Hromada

CONTENT: Eduard Hromada speaks about the distinctive strides from gathering the information from different promotions and land sites and sending out it into different classifications which is additionally broke down and confirmed. After the confirmation of information, the product device makes measurements plans which will examine relations among checked factors and depict the land showcase as indicated by clients A-Z request. Machine learning algorithms are applied for the analysis of real data on the new housing market of Santiago, Chile. Their goal is to look at the prescient execution of the Neural Network, Random Forest and Support Vector Machine approaches with conventional Ordinary Least Squares Regression.

4. **AUTHOR:** Li Li and Kai-Hsuan Chuet

CONTENT: Li Li and Kai-Hsuan Chuet has studied that real estate price variation has complicated the behaviors non-linearly and some uncertainty. Author has used mathematical model free feature of neural network algorithm. They have used back propagation neural system (BPN) and outspread premise work neural system (RBF) two plans are utilized to set up the nonlinear model for genuine homes value variety expectation of Taipei, Taiwan in view of driving and concurrent monetary lists. The mean supreme value and root mean square blunder two lists of the value variety are chosen as the execution list. Thus, based on the research author has concluded that the variation of house price trend is not that accurate.

Lastly, as different research is done by authors using various Machine Learning Algorithms, it is seen than predicting real estate cost is a complex study. The study shows various results obtained from each of the papers, but the missing factor is that it does not foresee future costs of the houses specified by the client. Because of these results, the hazard in interest is a condo or zone increments extensively. To limit this mistake, clients tend to procure a broker which again expands the cost of the procedure. This prompts the alteration and improvement of the current framework

3. EXISTING SYSTEM

- As different research is done by authors using various Machine Learning Algorithms, it is seen than predicting real estate cost is a complex study. The study shows various results obtained from each of the papers, but the missing factor is that it does not foresee future costs of the houses specified by the client.
- Because of these results, the hazard in interest is a condo or a zone increments extensively. To limit this mistake, clients tend to procure a broker which again expands the cost of the procedure. This prompts the alteration and improvement of the current framework.

Disadvantages:

- It cannot capture hidden relations between the price and features of houses and does not give overall better estimations.
- It doesn't perform well when we include features such as zip code, longitude, which are not linearly related to house price.
- So, we found few models that performs much better and captures the hidden information in those features and project the accuracy rates that each of these models gives.

4. PROPOSED SYSTEM

In the present real estate world, it has turned out to be difficult to store huge amount of information and concentrate them for one's own prerequisite. Likewise, the separated information ought to be helpful. The framework makes ideal utilization of the Linear Regression Algorithm. It makes use of such information in the most effective way.

The direct relapse calculation satisfies customer by expanding the exactness of their decision and diminishing the danger of putting resources into a home. A tons of highlights that could be added to make the framework all the more generally satisfactory.

One of the real future extensions is including home database of more urban areas which will give the client to investigate more domains and achieve an exact choice. More factors like subsidence that influence the house costs should be included. Top to bottom subtle elements of each property will be added to give plentiful points of interest of a coveted domain. This will help the framework to keep running on a bigger level.

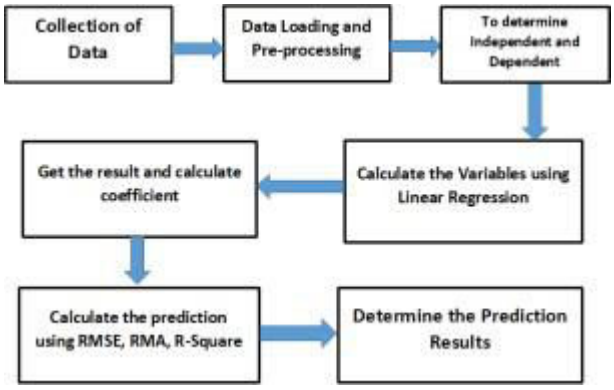

Fig.1 Architecture

5. RESULTS

Fig:2 Home Page

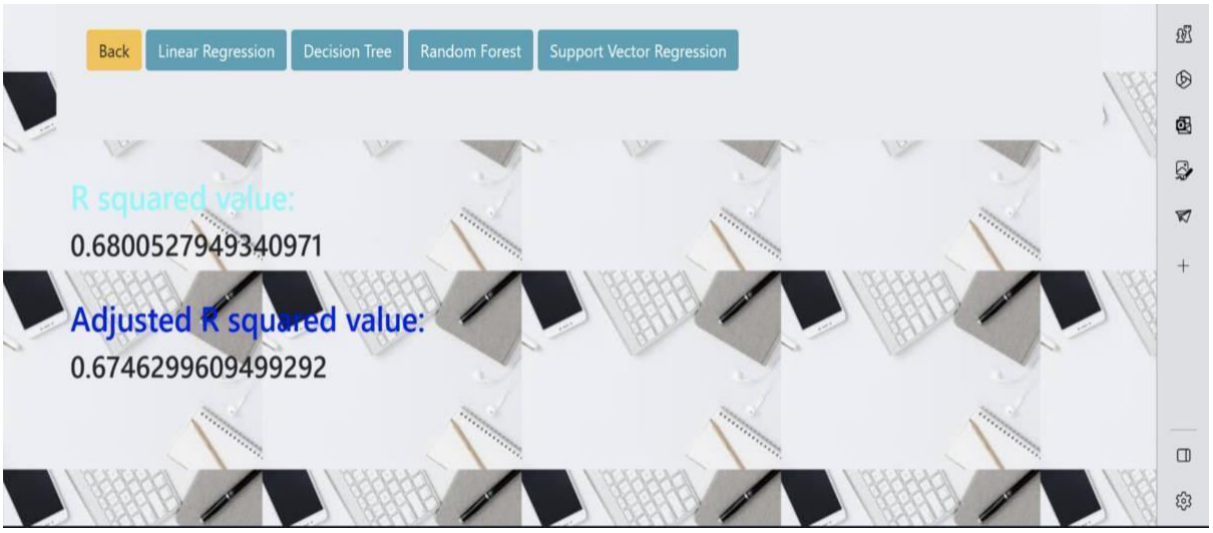

Fig:3 Linear Regression Page

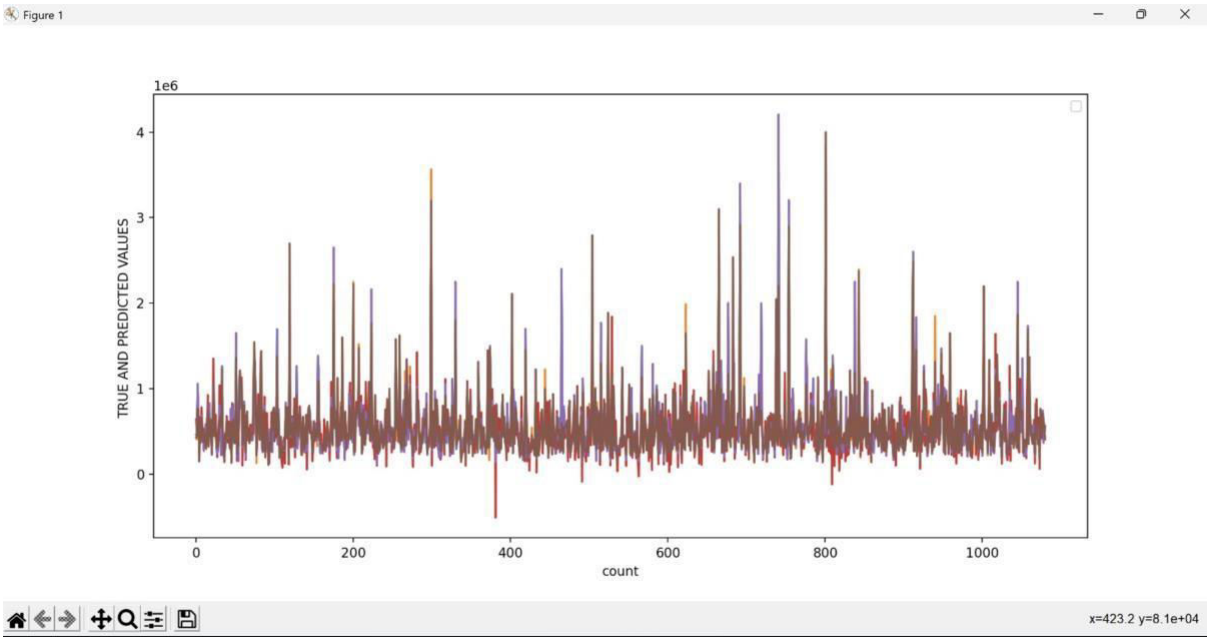

Fig:4 Output Screen for Linear Regression

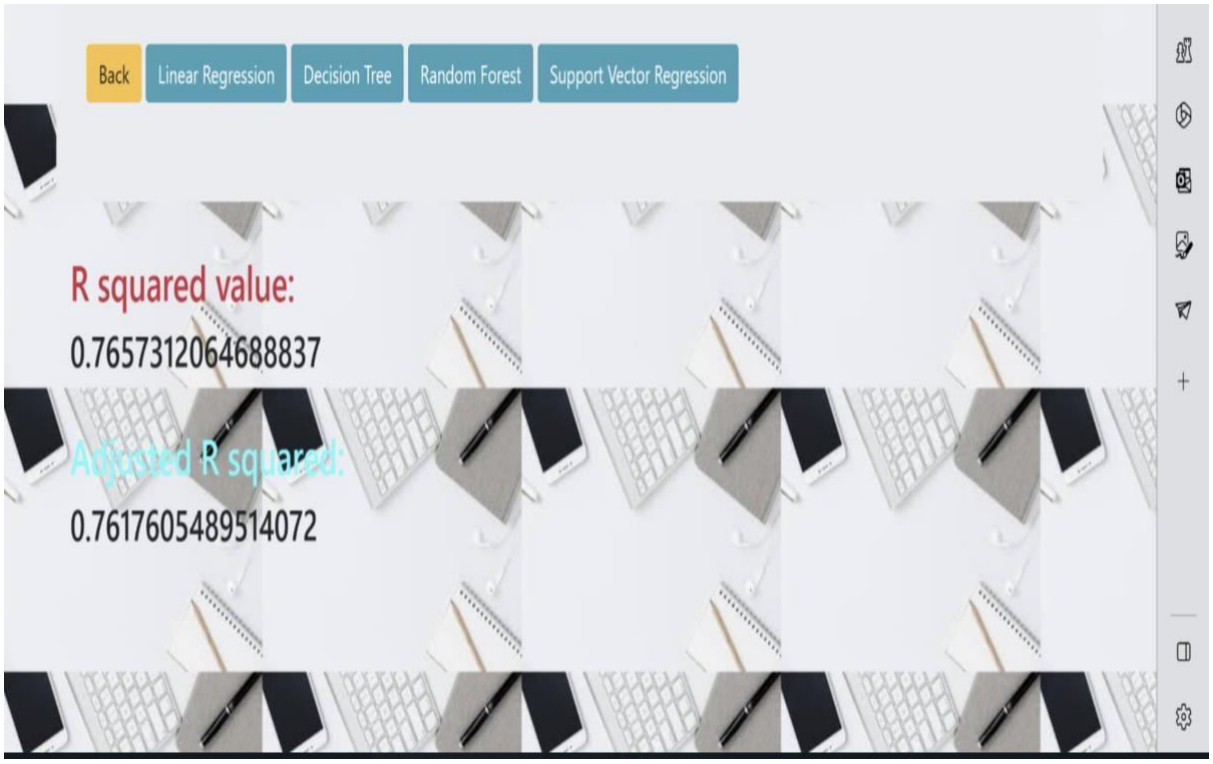

Fig:5 Decision Tree

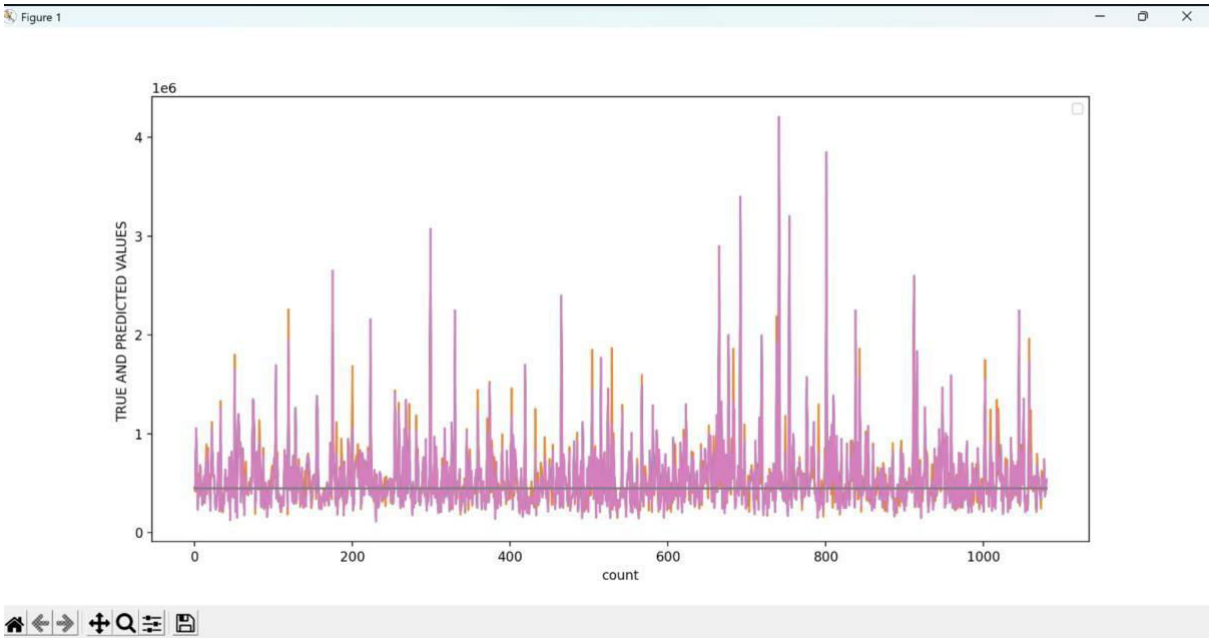

Fig:6 Output Screen for Decision Tree

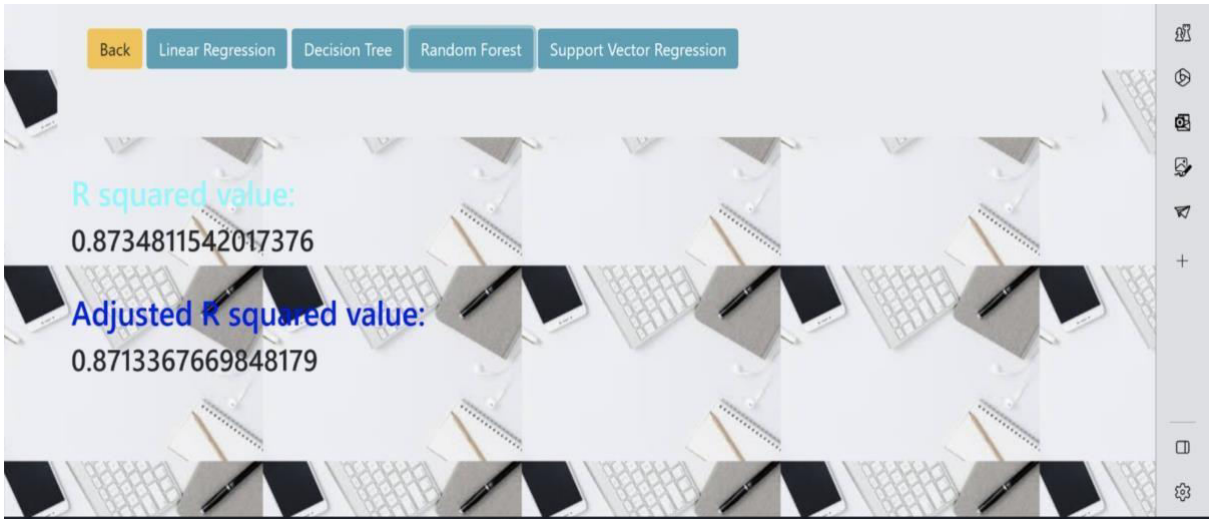

Fig:7 Random Forest

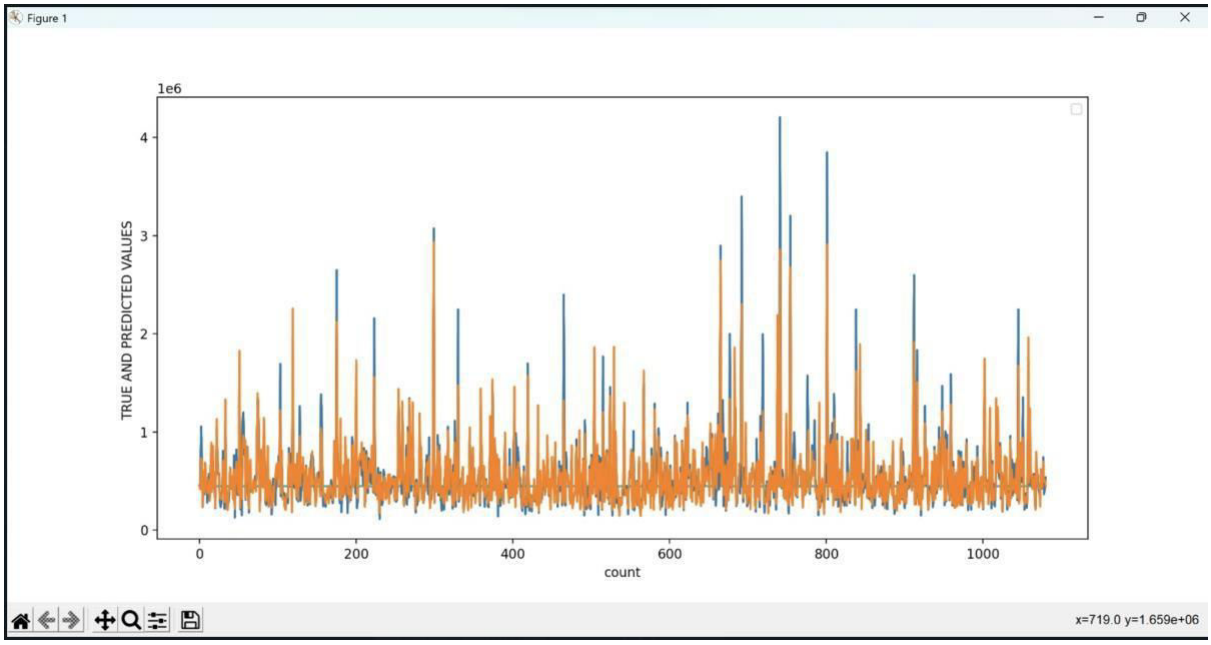

Fig:8 Output Screen for Random Forest

Fig:9 Support Vector Machine

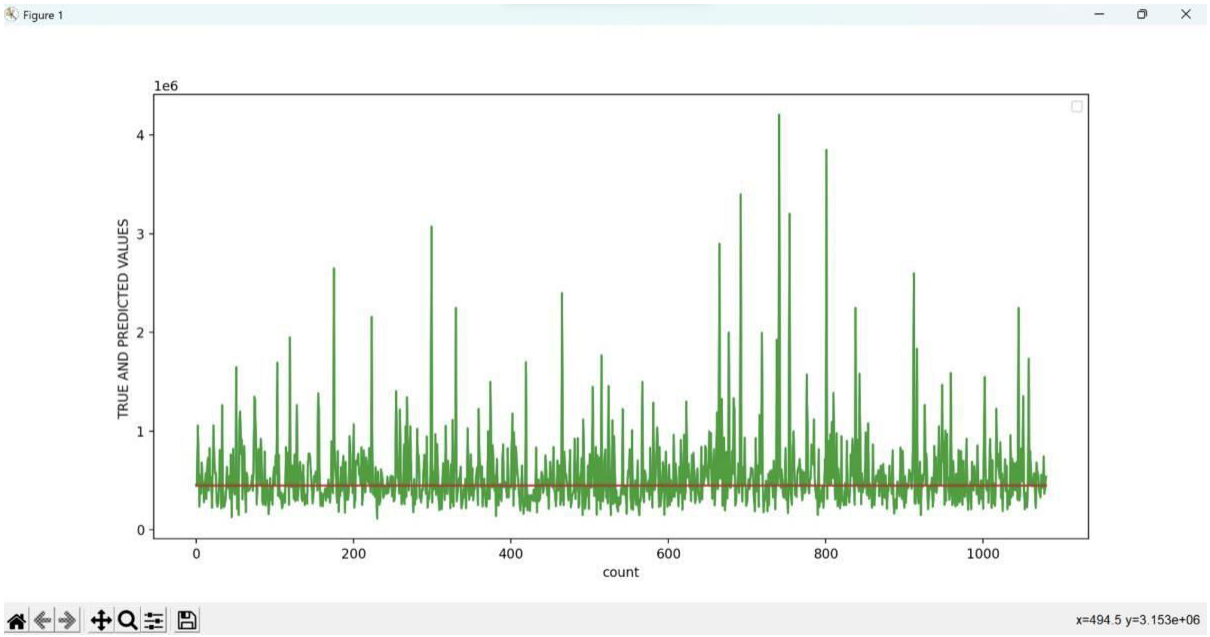

Fig:10 Output Screen for Support Vector Machine

6. CONCLUSION

In the present real estate world, it has turned out to be difficult to store huge amount of information and concentrate them for one's own prerequisite. Likewise, the separated information ought to be helpful. The framework makes ideal utilization of the Linear Regression Algorithm. It makes use of such information in the most effective way.

The direct relapse calculation satisfies customer by expanding the exactness of their decision and diminishing the danger of putting resources into a home. A tons of highlights that could be added to make the framework all the more generally satisfactory. One of the real future extensions is including home database of more urban areas which will give the client to investigate more domains and achieve an exact choice.

More factors like subsidence that influence the house costs should be included. Top to bottom subtle elements of each property will be added to give plentiful points of interest of a coveted domain. This will help the framework to keep running on a bigger level.

REFERENCES

1. R Manjula, Shubham Jain, Sharad Srivastava and Pranav Rajiv Kher, "Real estate value prediction using multivariate regression models," IOP Conference Series: Materials Science and Engineering, 2017.
2. V.Sampathkumara, M.Helen Santhib and J.Vanjinathan, "Forecasting the land price using statistical and neural network software," 3rd International Conference on Recent Trends in Computing, 2015.
3. Nihar Bhagat, Ankit Mohorkar and Shreyas Mane, "House Price Forecasting using Data Mining," International Journal of Computer Applications, 2016.
4. Eduard Hromada, "Mapping of real estate prices using data mining techniques," Czech Technical University, Czech Republic, 2015.
5. Pallav Ranka and Prof. Kripa Shanker, "Stock Market Prediction using Artificial Neural Networks," Indian Institute of Technology, Kanpur (208016), India.
6. Adyan Nur Alfiyatin and Ruth Ema Febrita, "Modeling House Price Prediction using Regression Analysis and Particle Swarm Optimization," International Journal of Advanced Computer Science and Applications, 2017
7. Li Li and Kai-Hsuan Chu, "Prediction of Real Estate Price Variation Based on Economic Parameters," Department of Financial Management, Business School, Nankai University, 2017.
8. Nissan Pow, Emil Janulewicz and Liu Dave, "Applied Machine Learning Project 4 Prediction of real estate property prices in Montreal," 2016.
9. Dr. Swapna Borde, Aniket Rane, Gautam Shende and Sampath Shetty, "Real Estate Investment Advising Using Machine Learning," IRJET, 2017.
10. "Property Valuation using Machine Learning Algorithms: A Study in a Metropolitan-Area of Chile," AMSE Conference Santiago, Chile, 2016.
11. Mansurul Bhuiyan and Mohammad Al Hasan, "Waiting to be sold: Prediction of Time- Dependent house selling probability," IEEE International Conference on Data Science and Advanced Analytics, 2016.
12. Wan Teng Lim, Lipo Wang,Yaoli Wang and Quing Chang, "Housing Price Prediction Using Neural Networks," IEEE 12th International Conference on Natural Computations, Fuzzy Systems and Knowledge Discovery, 2016.
13. Muhammad A. Razi and Kuriakose Athappilly, "A comparative predictive analysis of neural networks (NNs), nonlinear regression and classication and regression

tree(CART) models," Western Michigan University, 2005.
14. Youness El Hamzaoui and Jose Alfredo Hernandez Perez, "Application of articial neural networks to predict the selling price in the real estate valuation process," Morelos, Mexico, 10th Mexican International Conference on Articial Intelligence, 2011.
15. Ruben D. Jaen, "Data Mining: An empirical Application in Real Estate Valuation," Florida International University, 2002.