International Journal of Scientific Research in Engineering and Management (IJSREM)
Volume: 04 Issue: 08 | August -2020 ISSN: 2582-3930

House Price Prediction using Supervised ML

Narasingh Pratap, Vipin Kumar ,Mr Amit Kumar

Department of Information Technology

Rajkiya Engineering College Ambedkar Nagar Uttar Pradesh

Abstract: House price is chief argument of people in the world. This paper gives brief information about how to predict house prices using machine learning technique. In this paper we use Support vector Machine algorithm and KNN algorithm with python libraries to predict the house price faithful. We involve variegated basis to evaluate the original price of houses like -Sale price of house, Overall material and finish quality of house, Ground living area in feet, size of garage in car capacity, size of garage in square feet, Total square feet of basement area. Paper also gives numerical and graphical data to predict the actual price of house. This paper issue the information about which dataset is used in our proposed model to predict the price of houses using machine learning technique. After using method of this paper people will be able to predict actual price of house.

Keywords:house price, SVM, KNN, Actual price;

1.Introduction:

House/home is one of the most necessary part of people in their life. Everybody want to lives in their own house. People can notimagine their complete life without house. If anyone does not has his own house then, he want to buy house for his family and himself. Price of houses varies from location to location . this paper provide the complete information to buyer and seller for buying and selling price of houses. Price of house affected due various factor these are living area, overall finish and material quality of house to ground living area in feet, size of garage in car parking capacity, size of whole garage where car will be park, distance of hospital from home, distance of railway station and airport, environment of that area etc.In this paper machine learning technique will use to solve the problem house buyer and seller to predict the actual price of house. Support Vector machine are used to solve the house price prediction problem. Datasets contains high training data and remaining less testing data to calculate accuracy of the model.

1.1 Supervised Machine learning:Supervised machine learning is the subcategory of machine learning .In supervised learning a supervisor will available who gives reward after completion of task. In supervised learning input and output both will be given by supervisor. Supervised learning split dataset into two part training and testing dataset.

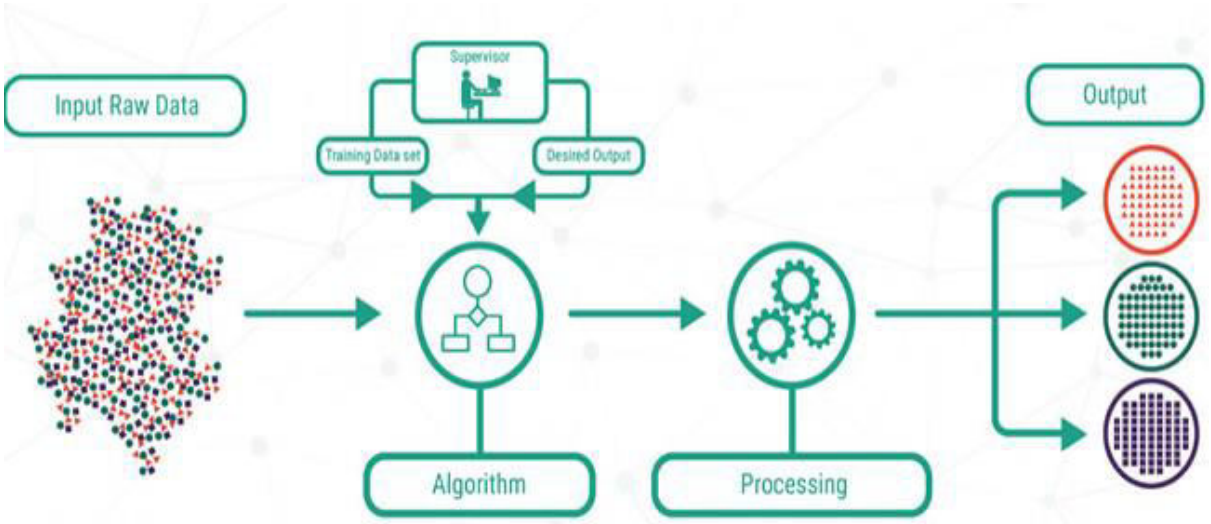


Fig:-1.1 supervised learning

Supervised learning has two subpart:-

1.1.1Classification

1.1.2.Reggression

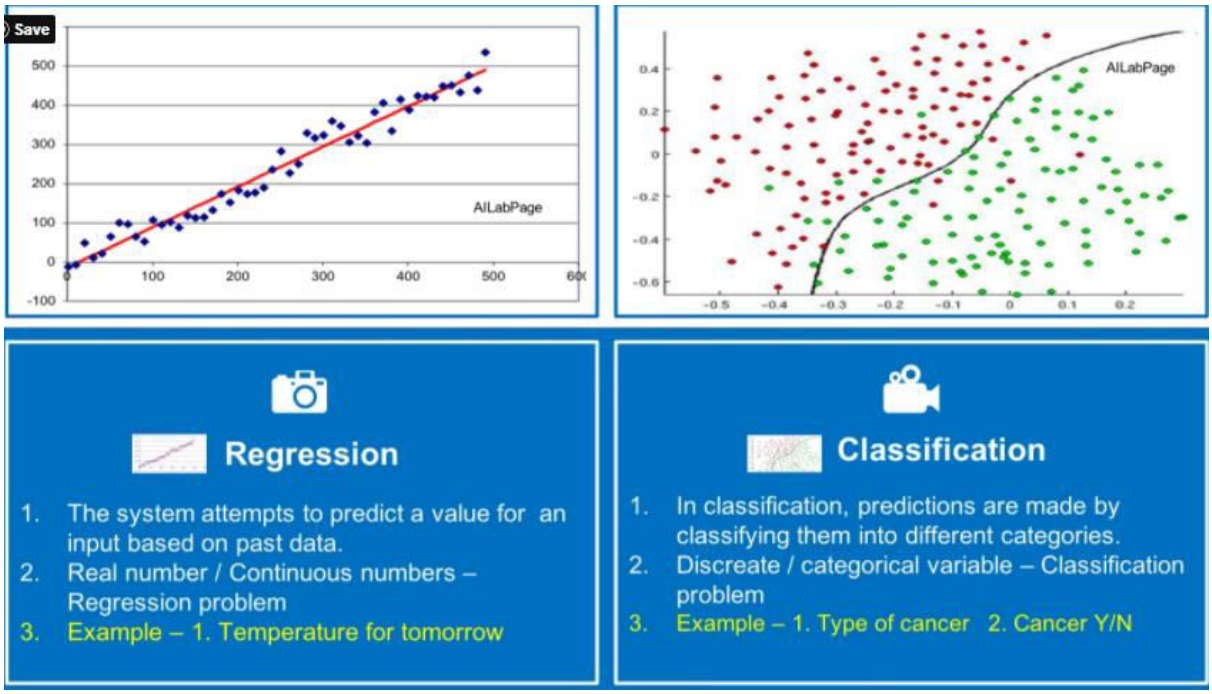


Fig :-1.2 Classification VS Regression

1.1.3 Use of classification algorithm in House Price Prediction:

All classification algorithm can be use in house price prediction task using machine learning technology some well known classification algorithm are given below in table.

Table 1.1

Serial No.	Classification Algorithm
1	Naïve Bayes Classifier
2	Nearest Neighbour
3	Support Vector Machine
4	Decision Tree
5	Boosted Tree
6	Random Forest
7	Neural Network

Some of these algorithm given in above table provides better accuracy in house price prediction solution but some of these algorithm provides very less accuracy, authors should be careful while utilising of these classification algorithm in house price prediction.

1.1.4Use of Regression algorithm in house price prediction:

Regression algorithm are useful for house price prediction some popular regression algorithm are given below in table.

Table 1.2

Serial No.	**Regression algorithm**
1	Linear Regression
2	Logistic Regression
3	K-Means

These Regression algorithm are effective for result in house price prediction process. Regression algorithm produces good accuracy.

Authors can use both classification and regression algorithm in same problem to find the actual house prices. After using these method house buyers and seller will able to make decision for actual price of house. And both buyers and sellers will be satisfies after using these method for result of house prices.

2.Related Work: There are lot of paper related to house price prediction title on the various journals. but we use most related paper with our paper. Various Papers which have published, most of them used machine learning technique to solve the problem of house sellers and buyers.

Durganjali et. Al[1],(2019)proposed house price prediction method ,this paper focused to predict the house prices in metro cities, it also consider the real state market price for prediction of house price ,in model population affect the deals of houses.in this paper authors include decision tree,naïvebayes classification,ada boost method,logistic regression to predict the house prices.Adaboost and decision tree provides more accuracy than other methods and these will very useful for price prediction.

Parasichet.al[2],(2018) perform kaggles competition ,it uses variable like house area,building manufacturing year etc, method of this paper is neural network, lasso training, residual regressor. Due to less dataset standard deviation result is very high. The Dahn Phan et.al[3],(2018) proposed model for prediction of house cost for Melbourne city in Australia. It takes historical data for house buyers and sellers then they predict actual price of house.in this paper author uses principle component analysis, decision tree, neural network and polynomial regression it gives better accuracy .these model are able to solve to the price related problem of house sellers and buyers in Melbourne city Australia. Suraya et.al[4],(2019) this paper uses Automated machine learning technique with real state data to know the price of data with high accuracy, Genetic Programming with Automated machine learning algorithm gives correct output for price prediction. This model is proposed for city Petalingjaya,Selangor in Malesiya. After using these model anybody can know real state house price of Petalingjayamalesiya.These model suggest high approach for research path on AML with another data. Mansi et.al[5],(2020) provide paper for different regression technology and know that how these are necessary for prediction of house price,this paper usese graphical and neumerical value for solution.this paper suggest suitable data set for machine learning algorithm which produces best result.this paper utilise cross validation and k-fold cross validation methodology for better result.this paper find ,what factors can affect the cost of houses and how it will minimise . this paper provide better result for house sellers and buyers to predict the house prices into particular region.

Ayushet.al[6],(2018) prepared paper for increasing and decreasing the house prices in day by day ,it predict the cost of house using basic requirement of any house buyers, it uses linear regression ,forest regression, neural networks, boosted regression to predict the result. This paper will helpful for buyers that gives cost to sellers for good condition house not for bad condition house. And sellers will also able know actual cost of their house. Feng et.al[7],(2019) proposed ARIMA model with deep learning for large amount of daily data of real state, this paper uses relu function with tensoflow frame work these method analyse that relationship between real state data and house cost are non- linear and it can be perform using tensorflow framework. Marlon et.al[8].(2019) proposed twitter trend data to find the cost of houses, it takes choice and bags of words from twitter for expected result, this paper will able to know the cost of house varies due tweet on twitter related to this real state and house . people choice affected due to twitter trend related to house topic. Nehal et.al[9],(2018) describes method for house buyers and sellers of Mumbai India, authors compare past data of house cost and it relate with people daily life and income then anybody know cost of house also he can know that he will able in future for his own house with these income or not, it utilise linear regression algorithm and it find that people can buy house on their interest. Timothy et.al[10],(2017) provides in their paper market value of houses for sellers and buyers it uses Random Forest ,Lasso, K-Means and clustering algorithm in machine learning for solutions .Random forest model gives better accuracy in this paper and anybody can predict the house cost with these models. Zhen et.al[11],(2019) analyse second hand house cost in their paper, It uses more characteristic for the prediction of house prices floors, decoration, housing type, face, traffic ,elevator etc affect the value of houses it uses multilevel linear regression ,decision tree, optimal choice for house cost ,this paper has prepared Chengdu Jinang China it will very helpful for those buyers who want to buy second hand house.

After studied all these papers we know that support vector machine and Naïve Bayes algorithm are good for house price prediction problem, we uses support vector machine and naïve bayes algorithm with supervised machine learning.

For better prediction of house prices we collect last three years research papers to get help in our paper.

3.Methodology: In this project we use two supervised learning algorithms.

i. Support Vector Machine
ii. K – Nearest Neighbors

3.1 Support Vector Machine

It is a supervised machine learning algorithms. The SVM is used for classification and regression challenges. The SVM is mostly use classification problem. We plot the data items as a dot in n dimensional space where n is the number of features. After that we perform the classification . In classification we find the hyper-plane which differentiate the two classes.

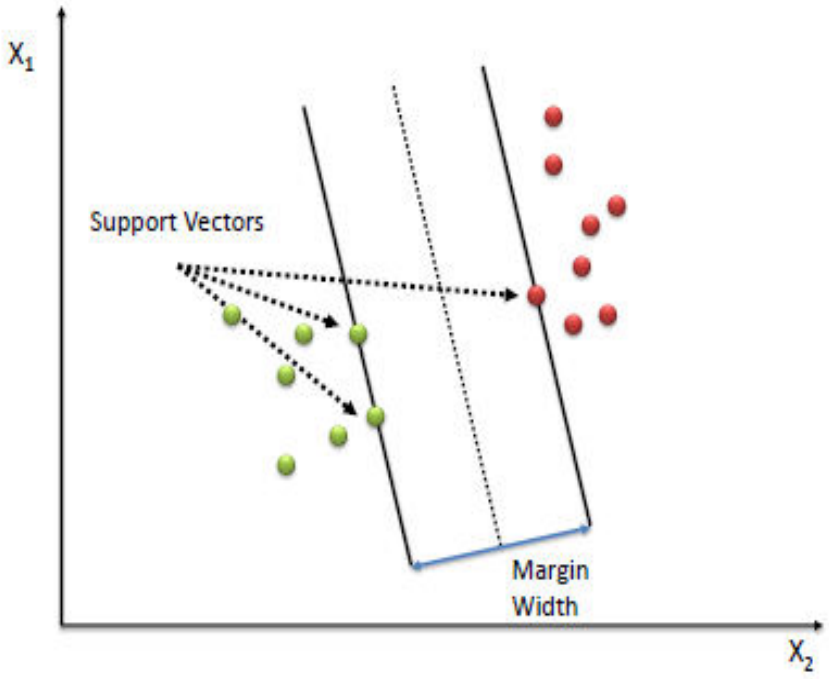


Fig 3.1 SVM

3.1.1 Maximal Margin Classifier

The maximum margin classifier is a hypothetical classifier that describes how support vector machine works.

The X is numerical variable and n is number of features . If you have two input variables ,then this is the form of two dimensional space. A hyper-plane is a line that differentiate the input variable space. You can visualize a line in two dimension space. For example:

A0+ (A1*X1) + (A2*X2) = 0 Where A1 , A2 is the coefficient that determine the slope of the line , A0 is the intercept and X1, X2 are the two input variable.If the equation return the value greater than 0 then it belongs to the first class.

If the equation return the value less than 0 then it belong two the second class.

3.2 K- Nearest neighbors

The KNN is a supervised machine learning algorithm. The KNN is used for solving the classification and regression problem. KNN is non-parametric algorithm that means it does not make any assumption on underlying data.

3.2.1 Working of KNN

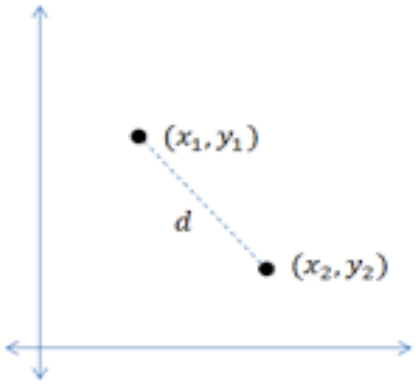
The KNN follow the following procedure

Step-1 : Select the number (K) of neighbors.

Step-2: Find the Euclidian distance of K number of neighbors.

(Suppose we have two data points A and B. The coordinate of point A is (X1,Y1) and the coordinate of point B is (X2,Y2) then the Euclidian distance between A and B is d.

where



$$d = \sqrt{(x_1 - x_2)^2 + (y_1 - y_2)^2}$$

Fig 3.2 Euclidian distance between two point

Step-3: Take the K nearest neighbors after calculating Euclidian distance.

Step-4: In each category count number of data points.

Step-5: Choose the data point which have number of neighbor Is maximum.

Step-6: The model is ready

4 Experiment and Results

The goal of this project is predicting price of the houses on the basis of feature. In this data set the training data have 1460 rows and 31 columns.

The IDE which I used in this project is Pycharm .
The following operation is performed in this project.

4.1 Data Analysis

First we check which type data are present in our training data set and find that data have two type .

4.1.1Numerical Data

The following features have numerical type data
OverallQual, GrLivArea, GarageCars, GarageArea etc.

4.1.2Categorical Data

The following features have Categorical type data
MSZoning, Street, Electrical, HouseStyle etc.

4.2 Data Visualisation

4.2.1 Correlation Matrix Heatmap

The Correlation Matrix Heatmap represent the relationship among the features of training data.

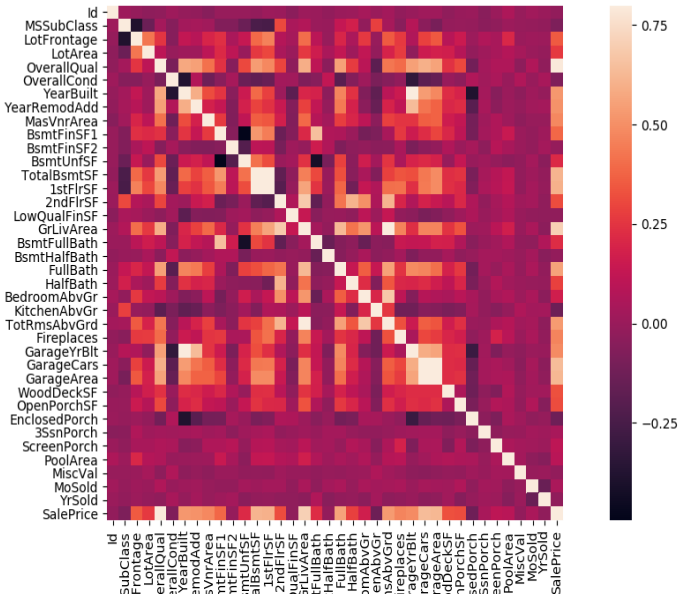


Fig-4.1 Correlation Matrix Heatmap

In this heatmap we can see that influencing percentage of feature to SalePrice. We can see that in this heatmap some feature is most influence to the SalePrice and some feature is less influence to the SalePrice. So I select the top most influencing feature for creating machine learning model.

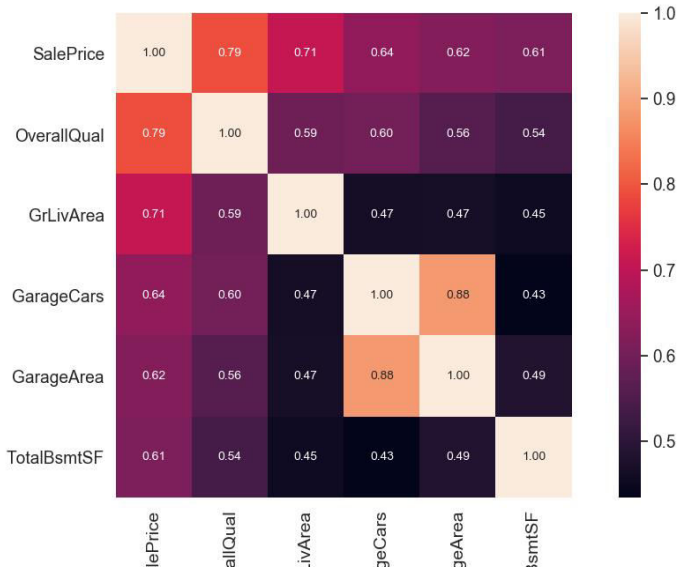


Fig-4.2 relationship between saleprice and feature

This heatmap represent the relationship among the SalePrice and most influencing features.

Table 4.1

Features	**Influencing %**
OverallQual	79
GrLivArea	71
GarageCars	64
GarageArea	62
TotalBsmtSF	61

4.2.2 OverallQual v/s SalePrice

In This section we will see the relationship between OverallQual and SalePrice.

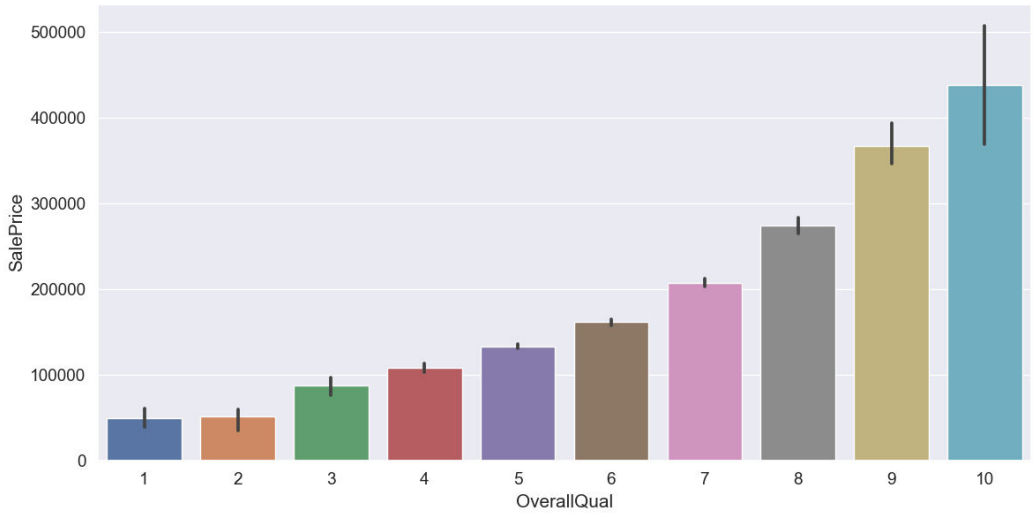


Fig-4.3 Overall quality vs saleprice

In the graph the numbering on x-axis show the quality ranking of the home . The number 1 is represent the house quality is very bad and the number 10 is represent the quality of the house is best. The y-axis of this graph represent the price (in dollar) of the home . In this graph we can see that if the quality is increases then the price is also increases .

4.2.3 LivingArea V/s SalePrice

In this section we will see the relationship between GrLivArea and SalePrice.

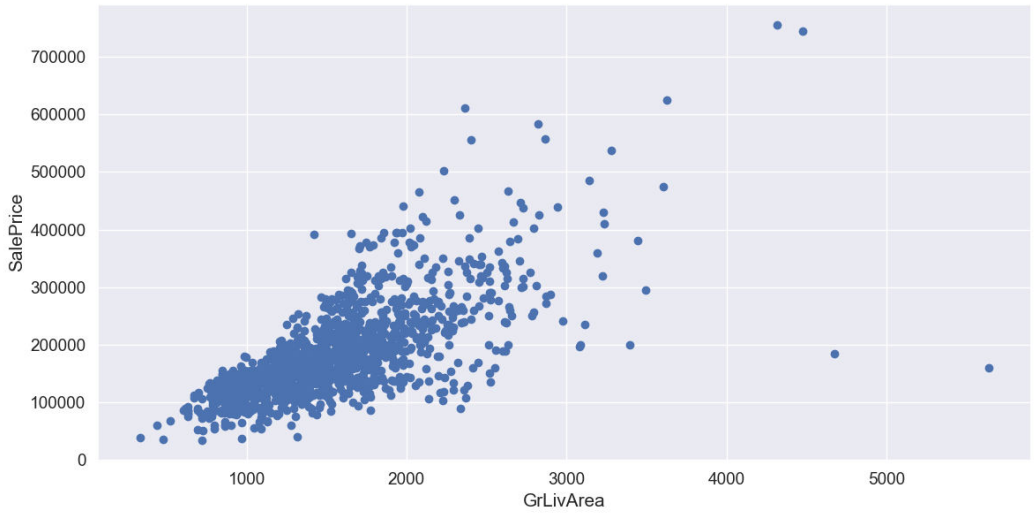


Fig 4.4 Living area vs saleprice

In the graph the x-axis show the GrLivArea(in square feet) and the y-axis show the SalePrice Of the house. The maximum house is available 764 square feet to 2160 square feet.

4.2.4 GarageCars v/s Saleprice

In this section we will see the relationship between GarageCars and saleprise.

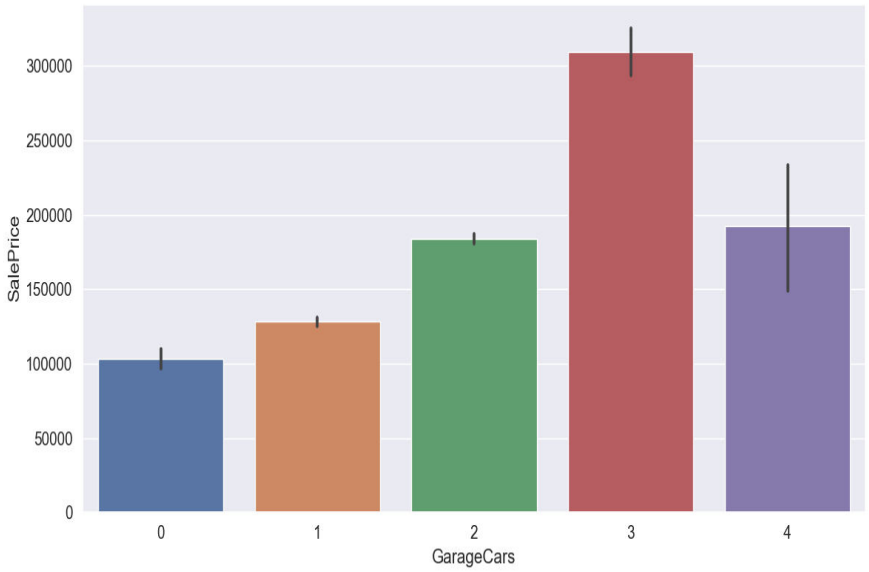


Fig 4.5 garage cars vs sale price

The x-axis represent the number of cars and the Y-axis represent the SalePrice of the house . In this graph we can see that if the number of GarageCars is increases from 0 to 3 then the price is also increases and when the GarageCars capacity is 4 then the price of the home is decreases . So we can say that the most demandable home is that which have GarageCars capacity is 3.

4.2.5 GarageArea v/s SalePrice

In this section we will see that the relationship between GarageArea v/s SalePrice.

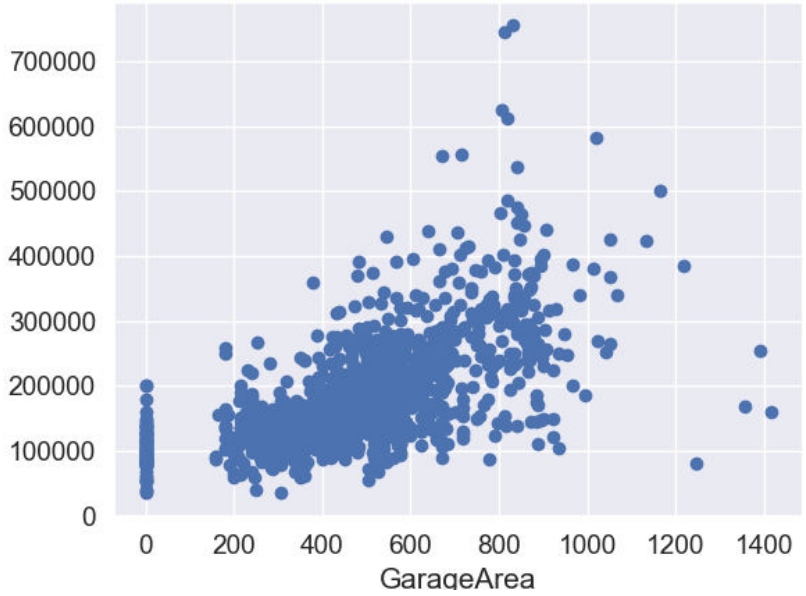


Fig -4.6 garage area vs sale price

The x-axis represents the GarageArea and the Y-axis represents the SalePrice of the house. All the dot show a particular house GarageArea and its selling price. The maximum home belongs to 200 square feet to 800 square feet.

4.2.6 TotalBsmtSF v/s SalePrice

In this section we will see that the relationship between TotalBsmtSF v/s SalePrice

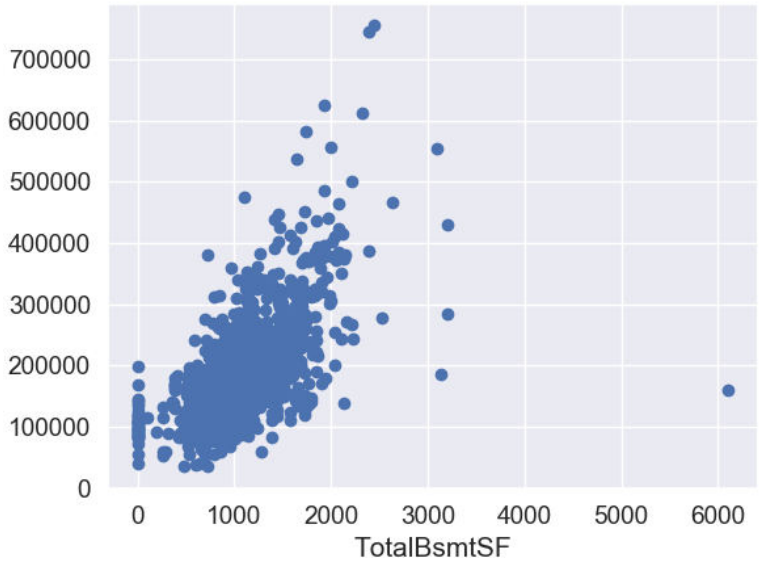


Fig 4.7 total basement vs saleprice

The x-axis represent the TotalBsmtSf area and the y-axis represent the SalePrice of the house. Every dot represents a particular home and the maximum dot is available approximately 445 square feet to 1876 square feet.

4.3 Data cleaning

Data cleaning is a process of removing unnecessary data from the data set . here first we select the data which is most related to the sale price using heatmap matrix and then drop the remaining data from the data set.

4.4 Filling Missing Value

Some time in dataset have some value is missing and it create a problem .For solving this problem we fill these vacant blocks . we fill mean or median value from that feature.

4.5 Result

I used two machine learning models in this project . The first model is Support Vector Machine (SVM) and the accuracy of SVM is 68%. The second machine learning model K- nearest neighbors algorithms and the accuracy of KNN is 34%. The SVM gives better result so I select the SVM for prediction

Table 4.2

Serial No.	Model	Accuracy
1.	Support Vector Machine	68%
2.	KNN	34%

After using both model we can say that support vector machine algorithm with given data sets is more accurate than Naïve Bayes algorithm .

Conclusion and future work

We have mentioned the step by step procedure to analyze the dataset and correlation matrix between the parameters. Thus we select the most influencing features from the dataset and drop the dataset which are less influence to the house price. .Model which are used in this project Support Vector Machine and KNN both can predict the house prices ,but both does not produces same accuracy.in this project we provide basic need of customer during house bargaining time, amount of houses varies on the basis of their need. First model Support Vector Machine provides high accuracy and it will be more useful for house buyers and sellers in future. Second model KNN gives less accuracy for given datasets ,it will not be useful for house buyers and sellers for long time dues it's less accuracy. This project will be helpful for the

people in future because it contains basic need of customer and data set in dollar form.

6.Reference:

[1]"HouseResalePricePredictionUsingClassificationAlgorithms"
https://ieeexplore.ieee.org/document/8882842[last accessed 13 2020]

[2]" Predicting Sales Prices of the Houses Using Regression Methods of Machine Learning"https://ieeexplore.ieee.org/document/8482191[last accessed 13 2020]

[3] "Housing Price Prediction using Machine Learning Algorithms: The Case of Melbourne City, Australia"https://ieeexplore.ieee.org/document/8614000[last accessed 13 2020]

[4] "Automated Machine Learning based on Genetic Programming: a case study on a real house pricing dataset"https://ieeexplore.ieee.org/document/8970916[last accessed 13 2020]

[5]" Prediction of House Pricing Using Machine Learning with Python"https://ieeexplore.ieee.org/document/9155839[last accessed 13 2020]

[6] "House Price Prediction Using Machine Learning And Neural Networks"https://ieeexplore.ieee.org/document/8473231[last accessed 13 2020]

[7]" House Price Prediction Approach based on Deep Learning and ARIMA Model" https://ieeexplore.ieee.org/document/8962443[last accessed 13 2020]

[8] "Predicting Housing Market Trends Using Twitter Data"https://ieeexplore.ieee.org/document/8789862[last accessed 13 2020]

[9]"Real Estate Value Prediction Using Linear Regression"

https://ieeexplore.ieee.org/document/8697639[last accessed 13 2020]

[10]" A Spatio – Temporal Hedonic House Regression Model" https://ieeexplore.ieee.org/document/8260698[last accessed 13 2020]

[11] "Model Research on Forecast of Second-Hand House Price in Chengdu Based on XGboostAlgorithm"https://ieeexplore.ieee.org/document/8935894[last accessed 13 2020]

[12]"to know about superviesd Machine learning"https://www.google.com/search?q=supervised+machine+learning&sxsrf=ALeKk01-VgD1NrxTbLOAq45fcVt0mB2Gqw:1597303009894&source=lnms&tbm=isch&sa=X&ved=2ahUKEwixioO00ZfrAhWowjgGHakUDR8Q_AUoAXoECBUQAw&biw=1366&bih=625#imgrc=HlqSRaBMBBIpQM[last accessed 13 2020]

[13]" to know about svm"https://www.analyticsvidhya.com/blog/2017/09/understaing-support-vector-machine-example-code/ [last accessed 13 2020]

[14] "svm image"https://www.google.com/search?q=support+vector+machine+image+classification&rlz=1C1GGRV_enIN751IN751&sxsrf=ALeKk01AYBWUhwAfmgt4WjS5e7vAD-w0qA:1597401538039&source=lnms&tbm=isch&sa=X&ved=2ahUKEwie7vO5wJrrAhXEwTgGHYaqBLIQ_AUoAXoECA4QAw&biw=1366&bih=576#imgrc=o_DkwiMWNXZQUM

[last accessed 13 2020]

[15] "supervised ml image"https://www.pinterest.com/pin/281334307958456610/ [last accessed 13 2020]