Predicting Real Estate Housing Prices

Aniket Malsane[1], Swapnil Shah[2], Avnish Kanungo[3], Dhrithi Alva[4], Deepika Vasili[5]

Abstract - It could be very important to evaluate the contemporary repute of the marketplace and are expecting its performance over the fast time period a good way to make appropriate monetary choices. This venture affords the occasion of artificial neural community-based models to guide land buyers and residential developers during this important mission. This report describes the choice variables, design method, and therefore the implementation of those fashions. The models utilize ancient marketplace performance information sets to instruct the artificial neural networks which will predict unexpected destiny performances. An application example is analysed to demonstrate the version skills in analysing and predicting the market overall performance.

***Key Words*:**

1.INTRODUCTION

Traditional house fee prediction is predicated on fee and sale charge assessment lacking an regular preferred and a certification process. Therefore, the supply of a residence charge prediction model enables replenish an essential facts gap and enhance the performance of the actual property marketplace. A house is regularly the maximum vital and most costly buy an man or woman makes in his or her lifetime. Ensuring owners have a trusted thanks to display this asset is extraordinarily vital. The Zestimate turned into created to offer purchasers the most quantity information as feasible about homes and therefore the housing market, marking the first-time clients had access to the present sort of home value statistics for free of charge. Making the roles of humans easy is what machines are tuned to do. Therefore the use of Machine Learning Algorithms, we've analyzed the value styles of housing in Amsterdam and primarily based on that the use of various algorithms, we've come up with prediction lines that healthy these patterns.

1.1 Methodology

- Load Pandas DataFrame containing housing data
- Do some simple data exploration /visualisation
- Refine or clean the data of any missing values.
- Split the data in train and test sets .
- Find the optimal model parameters using scikit-learn's GridSearchCV.
- Predict cross validated estimates of y for each data point and plot on scatter diagram vs true y for the following algorithms.

1.2 Algorithms to be used

Random Forest Regression:

A Random Forest is an ensemble method capable of acting each regression and category duties with the use of more than one decision bushes and a method called Bootstrap Aggregation, commonly called bagging. The simple concept in the back of that is to mix multiple decision trees in figuring out the very last output in place of relying on man or woman decision timber. Approach :

• Pick at random K records factors from the education set.

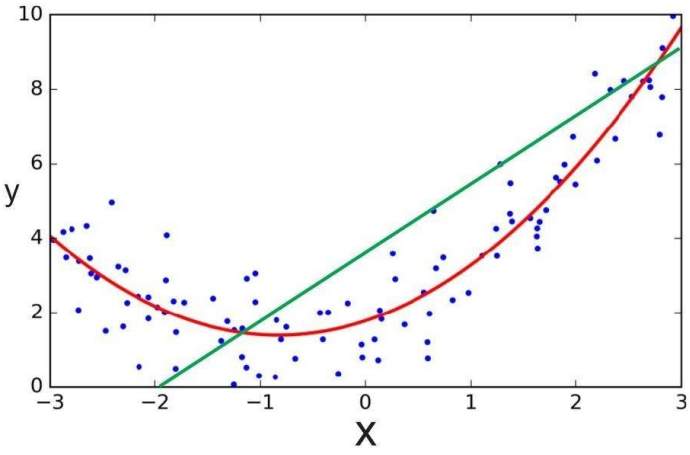
• Build the decision tree associated with those K facts factors.

• Choose the variety Ntree of bushes you want to build and repeat step 1 & 2.

• For a new facts point, make each considered one of your Ntree bushes expect the price of Y for the statistics factor, and assign the brand new information factor the average across all of the predicted Y values.

2. Polynomial Regression

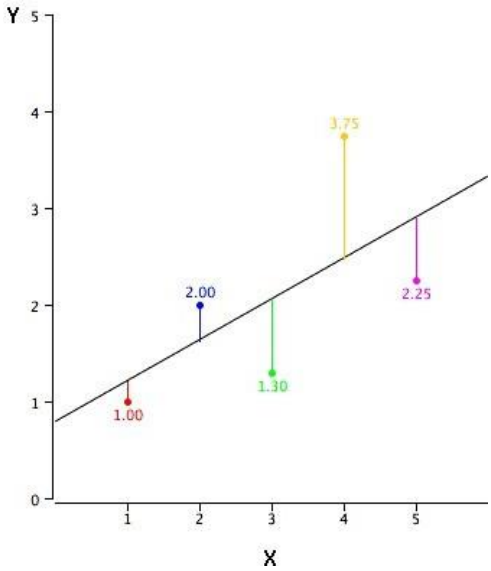
Polynomial regression is a shape of linear regression wherein the relationship among the impartial variable x and based variable y is modeled as an nth diploma polynomial. Polynomial regression fits a nonlinear dating among the value of x and the corresponding conditional mean of y.



Ordinary Least-Squares Regression:

In facts, regular least squares (OLS) is a kind of linear least squares method for estimating the unknown parameters in a

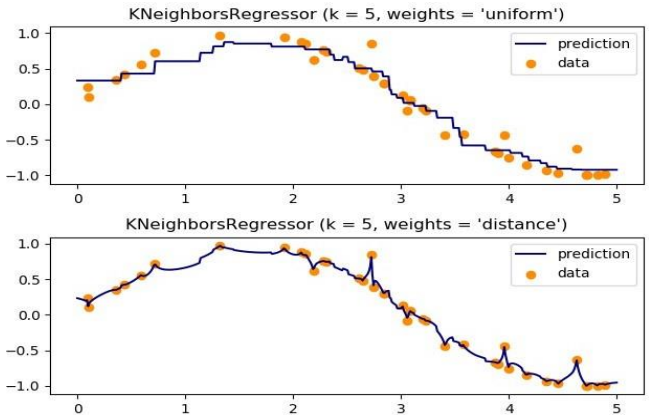
linear regression model. OLS chooses the parameters of a linear function of a set of explanatory variables by the precept of least squares: minimizing the sum of the squares of the differences between the observed structured variable (values of the variable being observed) in the given dataset and those expected via the linear function.



2.1 KNN Regression:

The K-Nearest Neighbors (KNN) algorithm is a easy, clean-toimplement supervised machine getting to know algorithm that can be used to solve both type and regression problems.

The KNN algorithm assumes that comparable things exist in close proximity. In other words, similar matters are close to to every other. KNN captures the idea of similarity (now and again referred to as distance, proximity, or closeness) with some arithmetic we'd have learned in our childhood—calculating the distance between factors on a graph. There are different approaches of calculating distance, and one manner is probably optimum relying at the problem we are solving. However, the directly-line distance (additionally known as the Euclidean distance) is a famous and familiar desire.



2.2 TEST CASE

We are taking first 100 entries in the dataset for the test case to train and test the model. We can get specific outputs by roaming taking the cursor to a specific point in the output graph.

input

surface;rooms_new;zipcode_new;price_new ;latitude;longitude 0

138;4;1060;420000;40.8046725;-73.9634204

1 130;5;1087;550000;52.3555895; 5.0005613

2 116;5;1061;425000;52.3730438; 4.837568

3 92;5;1035;349511;52.4168952;4. 90676664

127;4;1013;1050000;52.3967 89;4.8766067

.. ...

95 84;3;1096;539011;52.3363412;4 .9190299

96 127;5;1081;587500;52.3337697; 4.8522405

97 171;5;1060;515000;40.8046725; -73.9634204

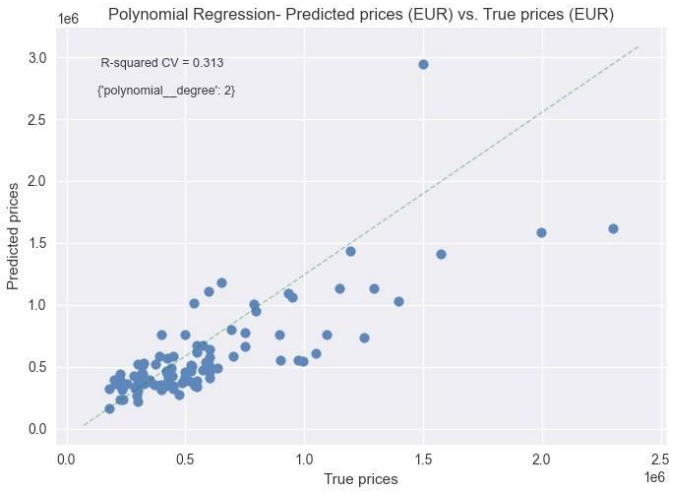
98 169;6;1060;495000;40.8046725; -73.9634204

99 265;5;1081;2300000;52.333769 7;4.8522405

Output

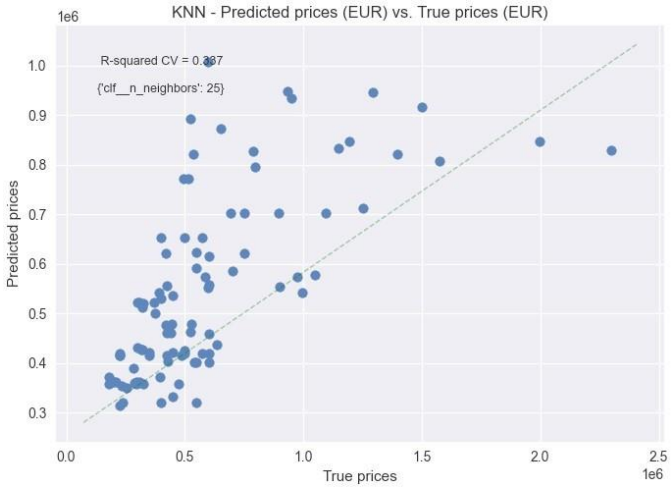
The Output contains score for all the algorithms used which acts as a metrics for performance of the algorithms

1) Linear Regression



Mean Score: 0.429

2) KNN



3) Polynomial Regression

Mean Score : 0.313

4. Random Forest Regression

Mean Score = 0.571

3. RESULT AND DISCUSSION

1. Random Forest Regression

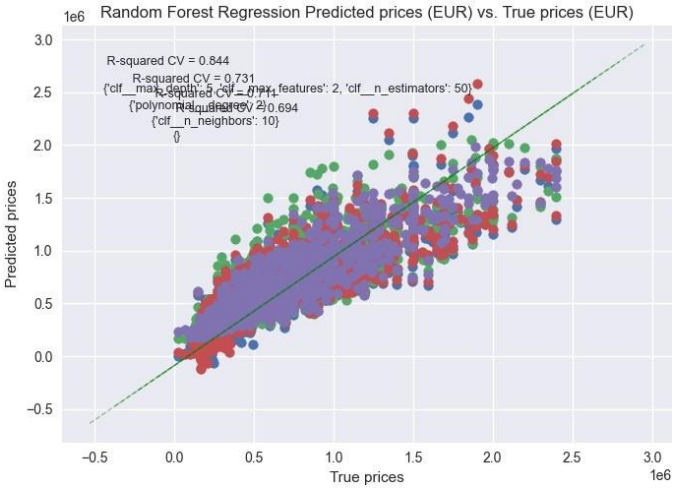
Optimal Parameters:max_depth: 5,min_child_weight: 6,gamma: 0.01, colsample_bytree: 1, subsample: 0.7

R^2 Score: 0.839

Polynomial Regression

Optimal Parameters: Degrees: 2

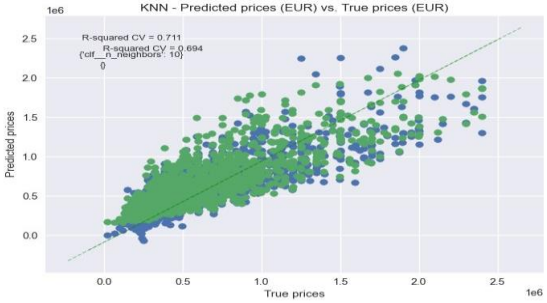
R^2 Score 0.731



KNN Regression

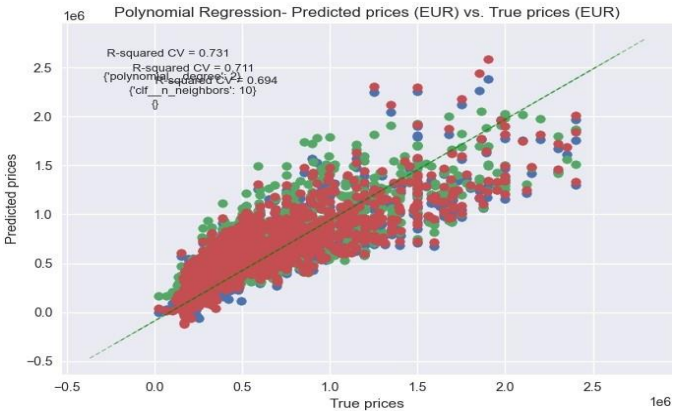
Optimal Parameters: n_neighbors: 10

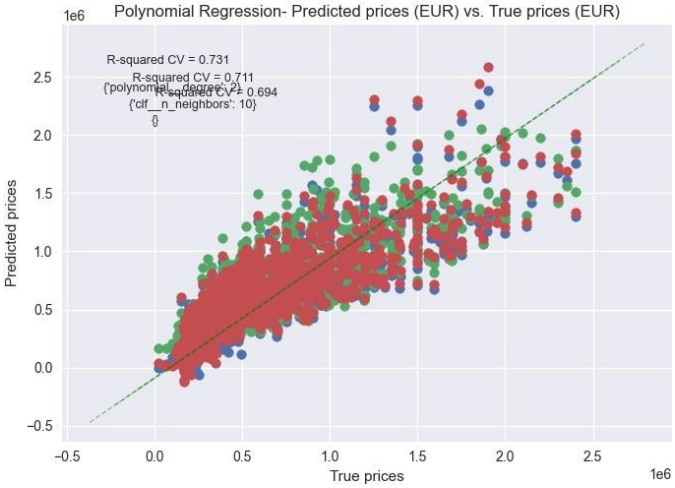
R^2 Score: 0.711

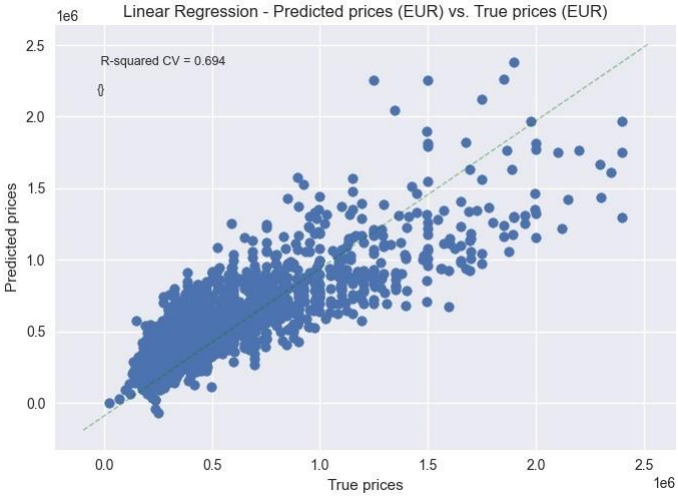


4 Ordinary Least-Squares Regression

R^2 Score 0.694







We got random forest regression having the best score so we can say that this algorithm is best in predicting house/real estate prices. It will make the work easy for many buyers and sellers in the world. They just have to put in their dataset and they can easily find the apt algorithm for their dataset and start predicting.

3. 1 Conclusion

The analyzed fashions protected Linear Regression, Polynomial Regression, Random Forest Regression and KNN algorithms, out of which we located out that the satisfactory in shape changed into the version generated with Random Forest Regression with a R2 rating of 0.839. So we can say that any dataset that is much like the dataset of Amsterdam housing fees used can use Random Forest Regressor with the best parameters based on the dataset to predict the prices efficiently.

This mission can effortlessly be implemented by means of all and sundry and this is the advantage that it'll offer to the humans of the society as nobody gets cheated and absolutely everyone can get just costs for his or her dream lands

3.2 Future Scope

The Project may be accelerated by means of introducing more algorithms like Neural Network MLP Regression, XGBoost regression, Ridge Regression, Lasso Regression, and so forth. Also we will include an increasing number of functions in order that prediction technique becomes more efficient. The fashions proposed inside the task may be used to predict now not simplest housing charges however additionally various different matters whose dataset suits the description of the dataset used.

REFERENCES

- Prediction of Residential Property Prices – A State of the Art https://www.researchgate.net/publication/325473040_Prediction_of_Residential_Property_Prices_-A_State_of_the_Art
- Sampathkumar et al. / Procedia Computer Science 57 (2015)112 – 121.
- Mansural Bhuiyan and Mohammad Al Hasan (2016) "Waiting to be Sold: Prediction of TimeDependent House Selling Probability" IEEE International Conference on Data Science and Advanced Analytics pp468-477
- E. Fix and J. L. Hodges Jr, "Discriminatory analysisnonparametric discrimination: consistency properties," DTIC Document, Tech. Rep., 1951.
- J. Smola and B. Sch¨olkopf, "A tutorial on support vector regression," Statistics and Computing, vol. 14, no. 3, pp. 199–222, Aug. 2004, ISSN: 0960-3174.

  DOI: 10.1023/B:STCO.0000035301.49549.88. [Online]. Available:

  http://dx.doi.org/10.1023/B:STCO.0000035301.49549.88.