House Price Prediction System

Prof. Pallavi Sambhare[1] Sejal Tidke[2] Siddhi Hattimare[3] Tanushree Bajpai[4] Vaidehi Parekar[5]
[1]Assistant Professor
[1,2,3,4,5]Department of Information Technology
[1,2,3,4,5]G.H.Raisoni College of Engineering, Nagpur, India

Abstract — In this study, it is suggested to evaluate the success rate of integrating machine learning methods such as leveraging artificial neural pathways to regression. Regression methods used in this paper include multifactorial linear, Ridge, Random Forest, and Least Minimum Selection Operator. This study also attempts to figure out the main factors that impact housing prices in Malmö, Sweden by examining the associations between several variables. Two datasets— public and local—were used in this analysis. Both of these include housing costs for Malmö, Sweden, and Ames, Iowa in the United States. For evaluating the precision of the prediction, the root square and mean square root error values of the training model are examined. The appropriate data has been separated into two parts, pre-processing methods are implemented and the evaluation is executed. One component shall be used for the training period, and the other after the test phase. Also claimed was a binning methodology that raises the accuracy of the models. This thesis strives to establish that Lasso outperforms different techniques while training on the public dataset. The correlation diagrams show how the variables are connected. Also according to the research, lending, repo rates, to criminal behaviour, and deposits all have a detrimental effect on housing costs. where the year, inflation, and unemployment rate have favourable impacts on the value of real estate.

Keywords: Artificial Neural Network, Machine Learning, Lasso Regression, Ridge Estimation, Random forest Regression, House Price Prediction

I. INTRODUCTION

A component of artificial intelligence called machine learning combines algorithms and technological advances to gather insights from data. Techniques related to machine learning are applicable in big data since it would be impossible within a guide analyze such a huge amount represents data. Machine learning proposes algorithmic remedies to issues regarding computer science. As alternative to ones that use only mathematical. It focuses around developing algorithms which have made machine learning conceivable. However, machine learning can be broken down both main categories comprise guided and irrational learning... An In order to be capable to arrive at predictions when fresh data is provided, and the programme is trained on the particular set under close monitoring. The programme searches seeking linkages and unsupervised connects between the data which remain hidden.

Presently For addressing issues in the real world, numerous approaches to machine learning can be utilised. But several of these yield greater effectiveness with specific conditions defined by the theory of the lack of free lunch. As consequently, such concepts evaluates the performance of the use of several requirements described by. Since The forecast in many techniques of regression rely not only on a specific attribute nevertheless, on an unknown number of factors which By forecasting housing costs, the performance will be assessed compared to the value what can be displayed. According to particulars of a property, house prices vary. The total number of features in houses fluctuates nevertheless it might not all cost the same. Depending to its location. In this case, the price of an enormous if an asset is positioned in a desirable, price community in contrast to a decaying one, its worth may be higher. The experiment's data will be handled applying a number of methods of pre-processing with the aim to boost the forecasting's accuracy. In order to figure out a link be these elements and the sale price additional details will be additionally consisted of to the national dataset.

II. LITERATURE SURVEY

There is an outstanding the quantity of research devoted to instructing systems to figure trends in datasets to foresee potential future consequences. However, there are examples where the authors combine several algorithms for machine learning with data pre-processing to reach closer their intended goals. Lu, Li, and Yang conducted a study in 2017. After looking towards new feature engineering, for the purpose to validate the equal prediction, a hybrid Bloom and gradient-boost regression equation was designed. They relied on Lasso to figure out attributes. They availed the most of the same dataset the fact that this study conducted. They answered numerous Feature development processes to identify the primary features which will increase the rightness of the prognosis the projections. The score rating they receive from the site called Kaggle grows when more features get added. So they expanded the 79 available features through 400. Likewise, people hired Feature selection lasso to disable the redundant and concluded that 230 aspects gave the highest point result by doing an experiment on enhancing employing Ridge, Lasso, and Gradient. Jose Manuel Pereira, Mario Basto, and Maria Faria da Silva conducted a study in 2014. Published their articles during the year 2016 that assessed the 3 methods. An approach to forecasting based on Lasso, Ridge, and Stepwise Regressions in SPSS the Business default became a reality. 2 unique mistake variations were specified. The first is a percentage of ineffective firms that the model appropriately predicted oversight. The percent of companies that are profitable in the instance anticipated would not succeed the following inaccuracy. The outcomes of this study proved that, when compared alongside the SPSS linear strategy, the type of the variable under consideration that is more prominently exhibited in the training set proved more likely for it to be selected by the lasso and ridge algorithms. In a 2017 study, Suna Akkol, Ash Akilli, and Abdullah Cemal compared multiple linear regression methods with artificial neural networks and their applications for prediction. In this investigation, several linear regression analysis and neural network simulations were used to estimate the impact of a number of morphological components on live weight. They used three different back-

propagation techniques to train ANNs: Levenberg-Marquardt, Bayesian consistency, and Scaled conjugate. They found that, for their prediction the position, an ANN outperforms several types of linear models. In 2010, a study was done out by Shah Gharoie Ahangar, Muhammad Yahyazadehfar, and Hassan Pournaghshband. The authors applied linear regression techniques and neural network simulation to estimate the share value of companies that had been Shiraz is active in international affairs Iranian exchange of 13 shares. 30 financial and 10 macroeconomic influences are taken for consideration by the authors. Then, using Integrated Components Analysis (ICA), they 7 essential values were attained like 3 macroeconomic factors and 4 financial variables, in order to figure out the price of a shares. They pointed out that after building the system with ANN, the estimation of the squared-error mean value, the absolute mean squared error %, and R 2 coefficient will all drastically decrease. Nils Landberg conducted a study in 2015. Nils analyzed the price development on the Swedish property market as well as the effects of qualitative influences on value of homes. Landberg has investigated effects of cost, population, and square footage, new homes, new companies, foreign the past, migrant's population, rate of joblessness, and burglary the total number of crimes, and the ranking of the number of Vacancies. According to Nils, many different kinds of factors affect housing prices badly including the severity of infractions, interest rates, the rate of joblessness, and a wider volume of construction projects. Landberg shown how difficult the real estate market is. In contrast to the services industry, Landberg showed, the market for housing is more difficult to evaluate due to an array of costs that influence the upsurge in the cost of property. Contrary to the investigation, increasing growth rates and advantageous impacts of qualitative characteristics on housing prices. Accentuation, GDP, average income, and inflation 8 on the other side, the rise in the rate of interest has caused a profoundly negative impact overall housing values. Further studies have revealed that while the unemployment rate has a detrimental effect on valuations of homes, the sale price and unemployment rate don't extremely connect.

III. PROPOSED METHODOLOGY

The theoretical inquiry and practical use of methods of regression assisted to organise this work. Peer-reviewed documents are used in the theoretical portion that addresses the research questions, which are clarified in further detail in section 4. The practical portion will be carried out in adherence to the layout to be outlined below and in greater depth in section 5. Nearly all of the literature review focuses on articles with full texts made obtainable online, free-of charge articles, and peer-reviewed journals from the Kristianstad College the database search engine Summon, as well as the search websites; the printed version of Research Gate, and the To Data Science journal collection.

The goal of the literature study is to lay an adequate basis for machine learning's use of regularisation, regression, and neural network training approaches, as well as how each of these techniques can be precisely employed in predicting home values. The literature study includes a review of associated research plus the feature techniques of engineering used in the current study. Further study standards that analyze how effective of the algorithms. Furthermore, the factors that were utilized were adapted to the area dataset. Performing the experiment will prepare the data and test the model prediction reliability. For achieving the expected outcomes, the experiment must go over several phases. The steps taken are listed as follows: - Preprocessing: Using the techniques from section, the two datasets will be validated and preprocessed.

These methods can treat data in numerous ways. For such, the previous processing takes place in a couple of iterations, with the accuracy evaluated using the chosen combination each time. - Data splitting: It is of the utmost importance to split a data set into two equal parts so that the simulator can be validated across the two and updated on one. 75 percent of the data is employed for training, and a quarter will be used for testing. - Examination: The RMSE and coefficients of R2 utilized in training the model, and the contrasts across the real prices on the test data and the prices predicted by the model, will be evaluated to assess the accuracy of both data. 10 - Performance: Added to the metrics for evaluation, the total amount of time required to train the model will be evaluated so we can see how the technique varies in terms of time. - Correlation: The correlation measured by the Pearson Coefficient will be utilized to figure out the whether the features' correlation is -ve, +ve, or 0 given and the value the residence in question Survey Metrics The accuracy of predictions will be verified by looking at the (RSME) and the coefficient of r (R2) of the training model. R2 denotes if RSME indicates the % fitting error between the genuine and shown data, in this scenario for the price of a home, the model is over fitted.

Requirements concerning computers the ability of the system employed the duration of time needed the model is being measured to train it. During the experiment. By employing GPU resources rather than CPU resources, certain libraries speed up model learning. Table 1. Specifications for computers Windows 10 operating system, Core i7 7700k processor, 16 GB of RAM, and a 1080 TI OC graphics card Different properties of the algorithms used for this study will be employed during implementation. The experiment was carried out using Python as the programming language plus the IDE Spyder. Each algorithm's attributes and architecture are listed below: Artificial neural network (11), the feed-forward architecture shown in figure 3 is utilised for constructing ANN utilising the Keras framework, which does away with the use of backpropagation. Three layers of concealment, an output layer, and a data input layer contribute to the model. In these, there are different numbers of neurons in the two datasets. The dataset that is provided has seventy, sixty, fifty, twenty-five, and one neurons in each layer. He layers of data in the local dataset, on the other hand, each have sixty-four, sixty-four, sixty-four, and 0 units. The activation function used in this design for each data set is RELU, but the optimiser is Adams. The number of nerve cells, number of layers, activation function, and algorithm have been found after sprinting several studies to determine the one that fared the best. -The longitudinal regression tool found in the sklearn.linear_model library serves to implement multiple linear. With this library, the only variables that may be used as parameters are independent and dependent variables.- Implementing an any Forest: The any forest is

perform using the sklearn. ensemble. Random Forest Regress or library. This library requires a specific number of parameters to control the model attributes. The 1200 shrub in the model with a maximum dimension of 60. To prepare the some model for training, a number of settings have been changed.

The options for the Grid SearchCV class located in the sklearn.model_selection put are Ridge parameter, ridge class. For determining the right parameter, this class employs hyper parameter improving and grid search. The class uses the GridSearchCV. Class's best estimator process before training its Ridge structure to the dataset.

IV. System framework

The artificial neural network (ANN) algorithm is slower compared to various algorithms, for example those in [appendix F], after training, especially when analyzing huge dimensions of data. ANN frameworks use the CPU to process any information that has been provided to them. The framework may utilise GPU resources opposed to processing power because to the additional GPU users. When utilising a significant amount of axons in an ANN, it has helped to quickly speed up the conditioning process. The real-life findings of the regional and public datasets demonstrated that applying the same method involving a single characteristic to two different datasets delivers results that vary. In the local dataset and the public dataset, Lasso regression gets the best overall score. Random Forest regression gets the best overall score. Nevertheless we made use of both individual and public datasets share the same algorithms and qualities. Via various files reveals the difficulty it is when applying the same method of preprocessing to several datasets. The results indicated that the two datasets scored differentially when it comes to of accuracy. Though fundamentally similar the local and public datasets includes a distinctive yet unequal features. Also, the desired accuracy could vary if the implementation used a different design. The correlation strength between the local and publicly accessible data also varies. Owing to the accuracy of prediction in both datasets, the correlation indicates the larger the link, the better the accuracy. Thereby, the local dataset needs more features to increase the relationship strength and increase its chances of coming up with a proper prediction model.

V. Conclusion

The investigation contrasts the use of algorithms for regression and artificial neural networks to forecast the worth of homes in Ames, Iowa, USA, and Malmö, Sweden. The public data accomplished promising findings given that it was full of aspects and had strong correlation, however the local data produced inferior results while the same prior process technique was put to use due to the fact it had less information as well as a lower correlation than the public data. Therefore, it needs to add additional factors to the local data, likely ones that have a strong correlation with the cost of homes. Although Ant delivered the best RMSE values scores, Lasso had the highest total score. The ultimate outcomes of the investigation demonstrated that offers predictions that are more accurate than the other applied algorithms. Even with less pleasantly to prices and year, crime, money, loans, and repo rates only significantly negatively effect home prices.

Acknowledgement

We are grateful to the Docent and Associate Professor Qinghua Wang, individual without their wisdom and knowledge in this subject we would have been unable to finish our research, ultimately guided us in this. Thanks to Svensk Maklarstatistik for providing details on Malmö home prices. Thank you to Niklas Gador, our examiner, for providing input and comments on the midway seminar presentation.

References

[1] brå. [Online].; 2020. Available from: www.bra.se.
[2] scb. [Online].; 2020. Available from: www.scb.se
[3] David HW, William GM. No Free Lunch Theorems for Optimisation. IEEE TRANSACTIONS ON EVOLUTIONARY COMPUTATION. 1997 April; I(1): 67-82.
[4] Svensk Mäklarstatistik. [Online].; 2020. Available from: www.maklarstatistik.se.

