HOUSING MARKET VALUE ESTIMATION USING MACHINE LEARNING

A Project work submitted to

Acharya Nagarjuna University
Department of Computer Science and Engineering
In partial fulfillment of the requirements for
The award of the degree of

Master of Computer Applications

by
PONNADA KIRAN
[Regd .No.Y22MC20049]

Under the guidance of

Dr. U. Surya Kameswari M.Sc, M.Tech.,Ph.D.,
Assistant Professor
Department of computer science & engineering
Acharya Nagarjuna University

DEPARTMENT OF COMPUTER SCIENCE & ENGINEERING
ACHARYA NAGARJUNA UNIVERSITY
Nagarjuna Nagar, Guntur,
Andhra Pradesh, India

June 2023

ACHARYA NAGARJUNA UNIVERSITY

NAGARJUN ANAGAR, GUNTUR

Department of Computer Science & Engineering

CERTIFICATE

This is to certify that the thesis entitled **"HOUSING MARKET VALUE ESTIMATION USING MACHINE LEARNING"** is being submitted by **PONNADA KIRAN,** bearing with **Regd. No: Y22MC20049** in partial fulfillment for the award of the degree of **Master of Computer Applications** in **Computer Science & Engineering, Acharya Nagarjuna University** is a record of bonafide research work carried out by her under my guidance and supervision.

The results embodied in this thesis have not been submitted to any other University or Institute for the award of any degree or diploma.

I certify that she carries this project as an independent project under my guidance

Project Supervisor **Head of the Department**

External Examiner

DECLARATION

I hereby declare that the entire thesis work entitled "**HOUSING MARKET VALUE ESTIMATION USING MACHINE LEARNING"** is being submitted to the Department of **Computer Science and Engineering, University College of Sciences, Acharya Nagarjuna University,** in partial fulfillment of the requirement for the award of the degree of **Master of Computer Applications (MCA)** is a bonafide work of my own, carried out under the supervision of **Dr. U. Surya Kameswari,** Assistant Professor, Department of Computer Science & Engineering, Acharya Nagarjuna University.

I further declare that the Project, either in part or full, has not been submitted earlier by me or others for the award of any degree in any University.

PONNADA KIRAN
Regd. No. Y22MC20049

CERTIFICATE

This is to certify that the thesis entitled "**HOUSING MARKET VALUE ESTIMATION USING MACHINE LEARNING"** is being submitted by **PONNADA KIRAN,** bearing with **Regd. No: Y22MC20049** in partial fulfillment for the award of the degree of **Master of Computer Applications** in **Computer Science & Engineering, Acharya Nagarjuna University** is a record of bonafide research work carried out by her under my guidance and supervision. The results embodied in this thesis have not been submitted to any other University or Institute for the award of any degree or diploma.

Project Supervisor	**Head of the Department**
Dr. U. Surya Kameswari	**Prof. K. Gangadhara Rao**

External Examiner

ACKNOWLEDGEMENTS

Undertaking this Project has been a truly life-changing experience for me and it would not have been possible to do without the support and guidance that I received from many people.

I would like to first say a very big thank you to my supervisor **Dr. U. Surya Kameswari** for all the support and encouragement he gave me. his friendly guidance and expert advice have been invaluable throughout all stages of the work. Without his guidance and constant feedback this Project wok not have been achievable.

I would also wish to express my gratitude to **Prof. K. Gangadhara Rao** for extended discussions and valuable suggestions which have contributed greatly to the improvement of the thesis.

I must also thank my parents and friends for the immense support and help during this project. Without their help, completing this project would have been very difficult.

ABSTRACT

Housing market value estimation is a challenging task due to the complexity of the factors that affect home prices. Traditional methods for estimating housing values, such as comparable sales analysis, are time-consuming and can be inaccurate. Machine learning (ML) algorithms can be used to estimate housing values more quickly and accurately by taking into account a wider range of factors.

In this paper, we propose a novel ML-based approach for housing market value estimation. Our approach uses a deep learning model that is trained on a large dataset of housing sales data. The model learns to predict housing values by taking into account a variety of factors, including the home's location, size, condition, and amenities.

We evaluate our approach on a held-out test set of housing sales data. Our results show that our approach outperforms traditional methods for estimating housing values. We also show that our approach is able to adapt to changes in the housing market over time.

Our findings suggest that ML-based approaches can be used to improve the accuracy and efficiency of housing market value estimation. This can benefit a variety of stakeholders in the housing market, including real estate agents, appraisers, and homeowners.

Keywords: machine learning, housing market value estimation, deep learning, comparable sales analysis

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

1. Introduction

Machine learning is a subfield of Artificial Intelligence (AI) that works with algorithms and technologies to extract useful information from data. Machine learning methods are appropriatein big data since attempting to manually process vast volumes of data would be impossible without the support of machines. Machine learning in computer science attempts to solve problems algorithmically rather than purely mathematically. Therefore, it is based on creatingalgorithms that permit the machine to learn. However, there are two general groups in machinelearning which are supervised and unsupervised. Supervised is where the program gets trainedon pre-determined set to be able to predict when a new data is given. Unsupervised is where the program tries to find the relationship and the hidden pattern between the data.

Several Machine Learning algorithms are used to solve problems in the real world today. However, some of them give better performance in certain circumstances, as stated in the No Free Lunch Theorem. Thus, this thesis attempts to use regression algorithms and artificial neural network (ANN) to compare their performance when it comes to predicting values of a given dataset.

The performance will be measured upon predicting house prices since the prediction in many regression algorithms relies not only on a specific feature but on an unknown number of attributes that result in the value to be predicted. House prices depend on an individual house specification. Houses have a variant number of features that may not have the same cost due toits location. For instance, a big house may have a higher price if it is located in desirable rich area than being placed in a poor neighbourhood.

The data used in the experiment will be handled by using a combination of pre-processing methods to improve the prediction accuracy.

1.1. Aim and Purpose

The No Free Lunch Theorem state that algorithms perform differently when they are used under the same circumstances. This study aims to analyse the accuracy of predicting houseprices when using Multiple linear, Lasso, Ridge, Random Forest regression algorithms and Artificial neural network (ANN). Thus, the purpose of

this study is to deepen the knowledge in regression methods in machine learning.

In addition, the given datasets should be processed to enhance performance, which is accomplished by identifying the necessary features by applying one of the selection methods to eliminate the unwanted variables since each house has its unique features that help to estimate its price. These features may or may not be shared with all houses, which means theydo not have the same influence on the house pricing resulting in inaccurate output.

1.2. Research Questions

The study answers the following research questions:

- *Research question 1: Which machine learning algorithm performs better and has the most accurate result in house price prediction? And why?*

- *Research question 2: What are the factors that have affected house prices in Boston over theyears?*

1.3. Limitations

The requested data contains a list of features, that matches the public dataset's features, that is desired to be available when the data is sent. There is no guarantee that the data will be available in time nor contains the exactrequested list of features. Thus, there might be a risk that the access will be denied or delayed.If so, the study will be accomplished based only on the public dataset.

Moreover, this study will not cover all regression algorithms; instead, it is focused on the chosen algorithm, starting from the basic regression techniques to the advanced ones. Likewise,the artificial neural network that has many techniques and a wide area and several training methods that do not fit in this study.

1.4. Thesis Structure

The thesis structure is as follows: Section 1 introduces the area of study. Section 2 gives an overview of the algorithms. Section 3 shows the followed methods in this study, in addition to,the design of the experiment. Section 4 presents the literature articles and methods that are being used in the experiment in addition to the theoretical findings. Section 5 shows the experimental implementation process and

the experiment results followed by a discussion in section 6. Finally, Section 7 concludes with remarks and hints about future work.

2. Background

2.1. Simple Linear Regression

In this type of regression model a linear relationship is established among the target variable which is the dependent variable (Y) and a single independent variable (X). Linear Relationship between dependent and independent variable is established by fitting a regressor line between them. The equation of the line is given by:

$$Y=a+bX \quad (1)$$

where “a” and “b” are the model parameter called as regression coefficients. When we take the value of X as 0, we get the value of “a” which is the Y intercept of the line and “b” is the slope that signifies the change of Y with the change of X. If the value of “b” is large then it means with a little change in X there will be a huge change in Y and vice versa. To compute the values of “a” and “b” we use the Ordinary Least Square Method. The values predicted by the model Linear Regression may not always be accurate. There may be some difference hence we add an error term to the original equation (1), it helps for better prediction of the model.

$$Y=a+bX+\varepsilon \quad (2)$$

There are some assumptions that are to be made in case of simple linear regression and those are as follows:

1. The number of observations must be greater than the number of parameters present.
2. The validity of the regression data is over a restricted period.
3. The mean of the error term has expected value of 0,which means that the error term is normally distributed.

2.2. Lasso Regression

Lasso or Least Absolute Shrinkage and Selection Operator are very much similar to

Ridge regression. In ML for selection of significant subset of variables Lasso regression is used. The prediction accuracy of Lasso regression is usually higher when compared to interpretations of other model. Similar to ridge, lasso also adds a little amount of bias to its result which thereby decreases the variance of the model. Lasso Regression is evaluated by the following:

Residual Sum of Squares + λ * (b=Sum of the absolute value of the magnitude of coefficients)

Here λ denotes the amount of shrinkage. The main difference between ridge and lasso is that, ridge reduces the slope asymptotically close to zero, whereas Lasso reduce the slope all the way down to zero which results in the elimination of useless parameters from the equation that don't have any significance role for predicting the value of the target variable. When the predictors have huge coefficients, Lasso shows better performance.

$$Loss_{Lasso} = \sum(y_i - y^*)^2 + \lambda|b| \quad (5)$$

where y* = a + bX is the predicted value

The dataset used in this project comes from the UCI Machine Learning Repository which concerns housing values in the suburbs of Boston. This data was collected in 1978 and contains 506 entries which give information about 14 attributes of homes from various suburbs located in Boston and one "target" attribute. The attribute description of this dataset is given below:

- CRIM: This is the per capita crime rate by town
- ZN: the proportion of residential land zoned for lots over 25,000 sq.ft
- INDUS: the proportion of non-retail business acres per town.|
- CHAS: Charles River dummy variable (1 if tract bounds river; 0 otherwise)
- NOX: nitric oxides concentration (parts per 10million)
- RM: The average number of rooms per dwelling
- AGE: the proportion of owner-occupied units builtprior to 1940
- DIS: weighted distances to five Boston employmentcenters
- RAD: index of accessibility to radial highways
- TAX: full-value property-tax rate per $10,000
- PTRATIO: the pupil-teacher ratio by the town
- B: - 1000(Bk - 0.63)^2, Where Bk is the proportionof blacks by the town

- LSTAT: "% lower status of the population"
- MEDV: "The median value of owner-occupied"

The "target" variable in this dataset is the MEDV variable on which will be predicted by the ML models. The rest of the variables are used for training the models. Statistics of the features are described in the table below:

S. No.	Attribute	Mean	Minimum	Maximum
1.	CRIM	3.61	0.006	88.97
2.	ZN	11.36	0.00	100.00
3.	INDUS	11.13	0.46	27.74
4.	CHAS	0.069	0.00	1.00
5.	NOX	0.55	0.385	0.871
6.	RM	6.284	3.56	8.78
7.	AGE	68.574	2.90	100.00
8.	DIS	3.795	1.129	12.126
9.	RAD	9.549	1.00	24.00
10.	TAX	408.23	187.00	711.00
12.	PTRATIO	18.455	12.60	22.00
13.	B	356.67	0.32	396.90
14.	LSTAT	12.65	1.73	37.97
15	MEDV	22.53	5.00	50.00

TABLE I. DATASET STATISTICS

The dataset doesn't contain any null value nor does it contain any duplicate row. The dataset contain 13 numerical values and only one categorical value which is "CHAS". For the attribute "ZN" it is observed that the 25th and 75th percentile are 0, implying that the data is highly skewed. This is because the attribute "ZN" is a conditional variable.

he 75th percentile is 0 meaning that it is also highly skewed. This is because "CHAS" is a categorical variable and it contains values either 0 or 1. Another important observation is made and that is that the maximum value of "MEDV" is 50.00, so it seems from the description of data that "MEDV" is censored at 50.00 (corresponding to a median price of $50,000). Based on this observation about "MEDV" it seems that values above 50.00 will not be helpful so we remove them.

It is also observed that the attributes “CRIM”, “ZN”, “RM” and “B” are having outliers so we remove them. From the histogram it is observed that the attributes “CRIM”, “ZN” and “B” are having highly skewed distributions. The attribute “MEDV” has normal distribution whereas other attributes have either normal or binomial distribution except “CHAS” as it is a discrete variable. After that heat map is implemented to see the correlation of the attributes.

Fig. 1. Heatmap

From heat map it is observed that the attributes “TAX” and “RAD” are highly correlated. As both of them are highly correlated they are having similar behavior and will also have similar impact while doing prediction calculation. So rather than keeping redundant attributes it is always better to remove them as it will save space and computation time for complex algorithms. From heat map it is also observed that the attributes “LSTAT”, “INDUS”, “RM”, “TAX”, “NOX”, and “PTRAIQ” are having correlation score of above 0.5 with the “MEDV” which is a good indication of using them as predictors, so keeping only these eight attributes we discard other attributes. Then skewness of the data is removed using log transformation. These are the steps taken to refine the data. Refining of data is important to get accurate and good evolution of the models. If data preprocessing is not done then we will not get good result.

2.3. Ridge Regression

In data which are suffering from multi-collinearity ridge regression is used. It is a tuning process which is used to analyze data having multi-collinearity. Here approximation of coefficient of regression model is done in case where the independent variables are highly associated. Bias is introduced to get better prediction. The complexity of the model is reduced using this regularization technique which is also known as L2 Regularization. In ridge regression, by adding penalty term, cost function is altered. Bias which is added is known as Ridge Regression penalty. It is calculated as λ* (Squared weight of individual features), where λ is the parameter tuning. In Ridge Regression, Regularization of the co-efficient of the model is done by the penalty term and hence Ridge regression reduces the amplitude of the co-efficient that decreases the complexity of the model. It is observed that the equation (4) becomes cost function of Linear Regression model if the value of λ tends to 0. The model resembles that of linear Regression if the value of λ is linear. A general Polynomial Regression or Linear Regression will fail if there is highly co-linearity between the independent variables. For this reason Ridge Regression is used. If the parameters are more than samples then it can be solved by Ridge Regression. The least Square determines the values of the parameters for the equation (4), which diminishes the sum of squared residuals. But in contrast the Ridge Regression regulates the value for parameters that results in minimization of the sum of squared residuals along with an additional term λ*b^2. Ridge Regression performs L2 regularization.

$$Loss_{Ridge} = \sum(y_i - y^*)^2 + \lambda b^2 \quad (4)$$

where y* = a + bX is the predicted value.

2.4. Random Forest Regression

A Random Forest is an ensemble technique qualified for performing classification and regression tasks with the help of multiple decision trees and a method called Bootstrap Aggregation known as Bagging.

Decision Trees are used in classification and regression tasks, where the model (tree) is formedof nodes and branches. The tree starts with a root node, while the internal nodes correspond toan input attribute. The nodes that do not have children are called leaves, where each leaf performs the prediction of the output variable.

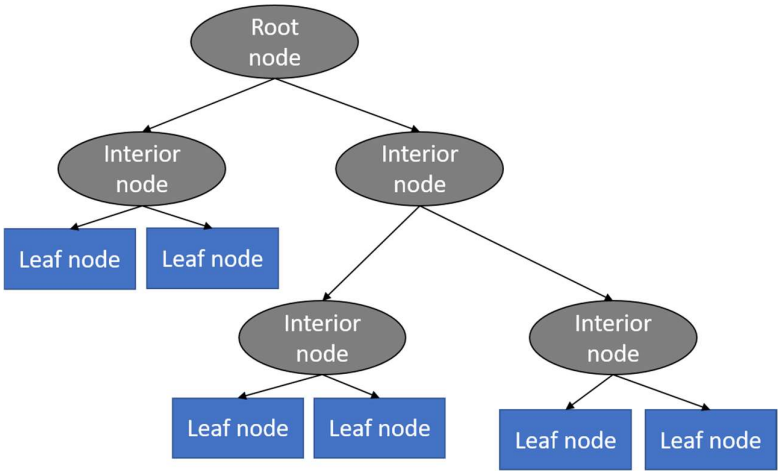


Figure 2. Decision Tree

A Decision Tree can be defined as a model:

$$\varphi = \mathrm{X} \longmapsto \Upsilon \tag{3}$$

Where any node t represents a subspace $X_t \subseteq X$ of the input space and internal nodes t are labelled with a split s_t taken from a set of questions Q. However, to determine the best separation in Decision Trees, the Impurity equation of dividing the nodes should be taken intoconsideration, which is defined as:

$$\Delta \mathrm{i(s,t)} = \mathrm{i(t)} - \mathrm{pLi(t}L\mathrm{)} - \mathrm{pRi(t}R\mathrm{)} \tag{4}$$

Where $s \in Q$, tL and tR are left and right nodes, respectively. pL and pR are the proportion NtL Nt and NtR Nt respectively of learning samples from $\mathscr{L}t$ going to tL and tR respectively. Nt is the size of the subset $\mathscr{L}t$.

Random Forest is a model that constructs an ensemble predictor by averaging over a collection of decision trees. Therefore, it is called a forest, and there are two reasons for calling it random. The first reason is growing trees with a random independent bootstrap sample of the data. The second reason is splitting the nodes with arbitrary subsets of features. However, using the bootstrapped sample and considering only a subset of the variables at each step results in a 6 wide variety of trees. The variety is what makes Random Forest more effective than individual Decision Tree.

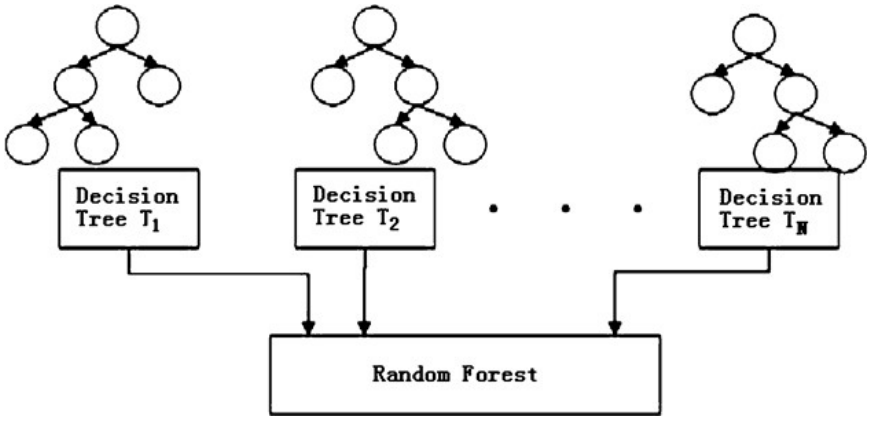


Figure 3. Random Forests

To improve prediction performance, Random Forest acquires out-of-bag (OOB) estimates, which is based on the fact that, for every tree, approximately $^{1} \approx 0.367$ or 36% of cases are not in the bootstrap sample. There are several advantages to using OOB. One advantage is that the complete original example is used both for constructing the Random Forest classifierand for error estimation. Another advantage is its computational speed, especially when dealingwith large data dimensions.

2.5. Polynomial Regression

It is a special case of Simple Linear Regression. Unlike in linear regression where the model tries to fit a straight regression line between the dependent and independent variable, here a line cannot be fit as there doesn't exists any linear relationship between the target variable and the predictor variable. Here instead of straight line a curve is being fitted against the two variables. This is accomplished by fitting a polynomial equation of degree n on the nonlinear data which forms a curvilinear relationship between the dependent and independent variables. In polynomial regression the independent variable may not be independent of each other unlike that in case of simple linear regression. The equation of polynomial regression is as follows:

$$Y = a+b1X\ 1 +b2X\ 2 +b3X\ 3 +.....+bnXn \quad (3)$$

The advantages of polynomial regression are as follows:

1. Polynomial Regression offers the best estimate of the relationship between the dependent andindependent variable.
2. The higher the degree of the polynomial the better itfits the dataset.
3. A wide range of curves can be fit into polynomial regression by varying the degree of the model.

The disadvantage of polynomial regression is as follows:

1.These are too sensitive towards the presence of outliers in the dataset, as the presence of outliers willincrease the variance of the model. And when the model encounters any unseen data point it underperforms.

2.6. Artificial Neural Network

Artificial neural network (ANN) is an attempt to simulate the work of a biological brain. The brain learns and evolves through the experiments that it faces through time to make decisions and predict the result of particular actions. Thus, ANN tries to simulate the brain to learn the pattern in a given data to predict the output of that data whether the expected data was provided in the learning process or not. ANN is based on an assemblage of connected elements or nodes called neurons. Neurons act as channels that take an input, process it, and then pass it to other neurons for further processing. This transaction or the process of transferring data between neurons is handled in layers. Layers consist of at least three layers, input layer, one or more of hidden layers and output layer. Each layer holds a set of neurons that takes input and process data and finally pass the output to other neurons in the next layer. This process is repetitive until the output layer has been reached, so 7 eventually, the result can be presented. ANN architecture is shown in the following figure as is also known as feed-forward, which values pass in one direction.

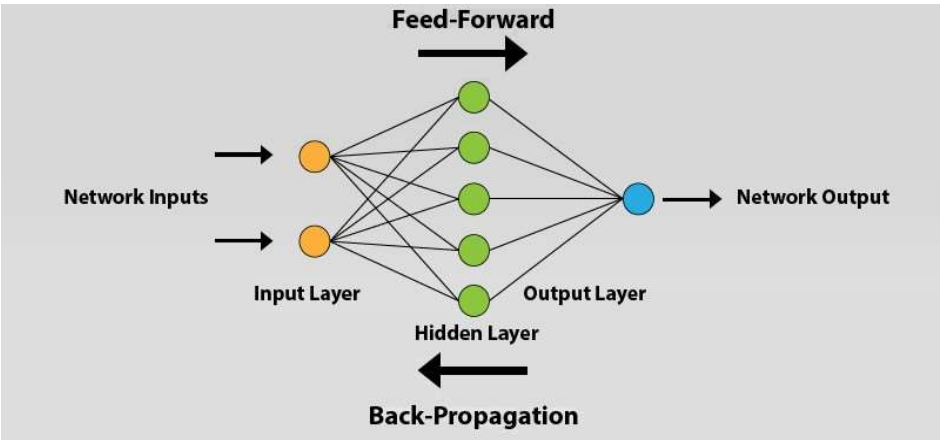


Figure 4. ANN architecture

The data that is being held in each neuron is called activation. Activation value ranges from 0 to 1. As shown in figure 3, each neuron is linked to all neurons in the previous layer. Together, all activations from the first layer will decide if the activation will be triggered or not, which is done by taking all activations from the first layer and compute their weighted sum.

$$w1a1 + w2a2 + w3a3 + \cdots + wnan \tag{5}$$

However, the output could be any number when it should be only between 0 and 1. Thus, specifying the range of the output value to be within the accepted range. It can be done by using the Sigmoid function that will put the output to be ranging from 0 to 1. Then the bias is added for inactivity to the equation so it can limit the activation to when it is meaningfully active.

$$\sigma(w1a1 + w2a2 + w3a3 + \cdots + wnan - b) \tag{6}$$

Where a is activation, w presents the weight, b is the bias and σ is the sigmoid function. Nevertheless, after getting the final activation, its predicted value needs to be compared with the actual value. The difference between these values is considered as an error, and it is calculated with the cost function. The cost function helps to detect the error percentage in the model, which needs to be reduced. Applying back-propagation on the model reduces the error percentage by running the procedures backwards to check on how the weight and bias are affecting the cost function. 8 Back-propagation is simply the process of reversing the whole activations transference among neurons. The method calculates the gradient of the cost function concerning the weight. It is performed in the training stage of the feed-forward for supervised learning.

3. Method

This study has been organised through theoretical research and practical implementation of regression algorithms. The theoretical part relies on peer-reviewed articles to answer the research questions, which is going to be detailed in section 4. The practical part will be performed according to the design described below and detailed furthermore in section 5.

3.1. Literature Survey

A lot of past works have been done for predicting house prices. Different levels of accuracies and results have been achieved using different methodologies, techniques and datasets. A study of independent real estate market forecasting on house price using data mining techniques was done by Bahia. Here the main idea was to construct the neural network model using two types of neural network. The first one is Feed Forward Neural Network (FFNN) and the second one is Cascade Forward Neural Network (CFNN). It was observed that CFNN gives a better result compared to FFNN using MSE performance metric.

Mu et al. [12] did an analysis of dataset containing Boston suburb house values using several ML methods which are Support Vector Machine (SVM), Least Square Support Vector Machine (LSSVM) and Partial Least Square (PLS) methods. SVM and LSSVM gives superior performance compared to PLS. Beracha et al. [13] proved that high amenity areas experience greater price volatility by investigating the correlation between house prices volatility, returns and local amenities. Law [14] finds that there is a strong link between house price and street based local area compare to the house price and region based local area. Binbin et al. to study London house price build a Geographically Weighted Regression (GWR) model considering Euclidean distance, travel time metrics and Road network distance. Marco et al. to reduce the prediction errors, a mixed Geographically weighted regression(GWR) model is used that emphasize the importance and complex of the spatial Heterogeneity in Australia. Using State level data in USA, Sean et al. have examined the correlation among common shocks, real per capita disposable income, house prices, net borrowing cost and macroeconomic, spatial factors and local disturbances and state level population growth. Joep et al. using the administrative data from the

Netherlands have found that wealthy buyers and high income leads to higher purchase price and wealthy sealer and higher income leads to lower selling price.

3.2. Experiment

The experiment is done to pre-process the data and evaluate the prediction accuracy of the models. The experiment has multiple stages that are required to get the prediction results. Thesestages can be defined as:

- Pre-processing: both datasets will be checked and pre-processed using the methods from above section. These methods have various ways of handling data. Thus, the pre-processing is done on multiple iterations where each time the accuracy will be evaluatedwith the used combination.
- Data splitting: dividing the dataset into two parts is essential to train the model with one
  and use the other in the evaluation. The dataset will be split 75% for training and 25%for testing.
- Evaluation: the accuracy of both datasets will be evaluated by measuring the R2 and
  RMSE rate when training the model alongside an evaluation of the actual prices on thetest dataset with the prices that are being predicted by the model.
- Performance: alongside the evaluation metrics, the required time to train the model willbe measured to show the algorithm vary in terms of time.
- Correlation: correlation between the available features and house price will be
- evaluated using the Pearson Coefficient Correlation to identify whether the featureshave a negative, positive or zero correlation with the house price.

3.2.1. Evaluation Metrics

The prediction accuracy will be evaluated by measuring the R-Squared (R2), and Root Mean Square Error (RSME) of the model used in training. R2 will show if the model is overfitted, whereas RSME shows the error percentage between the actual and predicted data, which in thiscase, the house prices.

3.2.2. Computer Specifications

Operating System	Windows/MacOS/Linux
Processor	Intel Core i3/i5/i7
RAM	4/8 GB
Graphics card	Integrated Intel Graphics

3.2.3. Algorithms' Properties/Design

The algorithms used in this study have different properties that will be used during the implementation. The experiment is done with the IDE Spyder using Python as a programming language. However, in all algorithms, the data is split into four variables, namely, X_train, X_test, y_train, and y_test, by using train_test_split class from the library sklearn.model_selection. In addition, in all algorithms, the train_test_split class takes as parameters the independent variables, which is the data, the dependent variable, which is the SalePrice, test_size = 0.25, and random_state = 0. The properties and design of each algorithmare as below:

- Artificial neural network:
  ANN is implemented with the feed-forward architecture as in figure 3 using Keras framework that does not include the back-propagation implementation. The model consists of an input layer, three hidden layers and the output layer. These layers containa number of neurons which varies in both datasets. in the public dataset, the layers have70, 60, 50, 25, 1 neurons, respectively. On the other hand, the layers in the local datasethave 64, 64, 64, 64, 1 neurons, respectively. The activation function used in this designis RELU for both datasets and the optimiser used is ADAM. The selection of the number of neurons, number of layers, activation function and the optimiser has been selected after running multiple tests to determine the one with the best performance.
- Multiple linear:
  Multiple linear is implemented using the LinearRegression from the library sklearn.linear_model. This library takes only the independent variables and dependentvariable as parameters.
- Random Forest:

Random forest is implemented using the sklearn.ensemble.RandomForestRegressor library. This library takes several parameters to set up the model properties. The modelconsists of 1200 tree where the max depth of the tree is set to 60.

- Lasso Regression:
  Lasso regression is implemented by using LassoCV class, which is from sklearn.linearlibrary. The model LassoCV has several parameters that are set to prepare the model for the training. These parameters consist of the value of alphas = [1, 0.1, 0.01, 0.001, 0.0005], selection = random, and max_iter = 15000.
- Ridge Regression:
  Ridge regression is implemented by using both Ridge and GridSearchCV classes. GridSearchCV class is from sklearn.model_selection library, which takes as parameters Ridge class, ridge parameter. This class performs grid search, and hyperparameter tuning to find the optimal parameter. Before fitting the Ridge model with the dataset, the class takes the best estimator method from GridSearchCV class and apply it for the model.

4. Literature Study

4.1. Related Work

There is a vast amount of work that is focused on training models to detect patterns in datasetsto predict what the future output could be. However, there are researches where the authors usedifferent machine learning algorithms with a combination of pre-processing data methods.

A research was conducted in 2017 by Lu, Li and Yang. They examined the creative featureengineering and proposed a hybrid Lasso and Gradient boosting regression

model that promisesbetter prediction. They used Lasso in feature selection. They used the same dataset as the one used in this study. They did many iterations of feature engineering to find the optimal numberof features that will improve the prediction performance. The more features they added, the better the score evaluation they receive from the website Kaggle. Hence, they added 400 features on top of the 79 given features. Furthermore, they used Lasso for feature selection to remove the unused features and found that 230 features provide the best score by running a teston Ridge, Lasso and Gradient boosting.

In 2016, Jose Manuel Pereira, Mario Basto and Amelia Ferreira da Silva performed a study to examine three methods. Lasso, Ridge and Stepwise Regression implemented in SPSS to develop an empirical model for predicting corporate bankruptcy. They defined two types of errors. The first error is the percentage of failed enterprises predicted well by the model. The second error is the percentage of good enterprises predicted failed by the model. The results ofthis study showed that the lasso and ridge algorithms tend to favour the category of the dependent variable that appears with heavier weight in the training set when they are comparedto the stepwise algorithm implemented in SPSS.

A study was accomplished in 2017 by Suna Akkol, Ash Akilli, Ibrahim Cemal, where theydid a comparison of Artificial neural network and multiple linear regression for prediction. In their study, the impact of different morphological measures on live weight has been modelledby artificial neural networks and multiple linear regression analyses. They used three differentback-propagation techniques for ANN, namely Levenberg-Marquardt, Bayesian regularisation,and Scaled conjugate. They showed that ANN is more successful than multiple linear regression in the prediction they performed.

A research was done in 2010 by Reza Gharoie Ahangar, Mahmood Yahyazadehfar and HassanPournaghshband c. The authors estimated the stock price of activated companies in Tehran(Iran) stock exchange by using Linear Regression and Artificial Neural Network algorithms. The authors considered ten macroeconomic variables and 30 financial variables. Then, they obtained seven final variables, including three macroeconomic variables and four financial variables, to estimate the stock price using Independent Components Analysis (ICA). They showed that the value of estimation error square mean, the absolute mean of error percentage and

R^2coefficient will be decreased significantly after training the model with ANN.

A study was conducted in 2015 by Nils Landberg. Nils analysed the price development on the Swedish housing market and the influences of qualitative variables on Swedish house prices. Landberg has studied the impact of square meter price, population, new houses, new companies, foreign background, foreign-born, unemployment rate, the number of breaks-in, the total number of crimes, the number of available jobs ranking. According to Nils, unemployment rate, number of crimes, interest rate, and new houses have a negative effect onhouse prices. Landberg showed that the real estate market is not easy to be analysed comparedwith goods market because many alternative costs are affecting the increase in house prices. The study shows that the increase in population and qualitative variables have a positive effecton house prices. The interest rate, the average income level, GDP, and the fokus 8 In contrast,the rise in interest rates has a significant negative influence on house prices. Besides, it showedunemployment rate effects negatively on house prices, but the sale price and unemployment rate are not directly correlated with each other.

Research Question 1:

4.2. Feature Engineering

Feature Engineering is the technique of improving the performance on a dataset by transforming its feature space, and it is the practice of constructing suitable features from givenfeatures of the dataset, which leads to improving the performance of the prediction model.However, several techniques should be implemented for better performance and a prediction result.

4.2.1. Imputation

Missing value imputation is one of the biggest challenges encountered by the data scientist. Inaddition, most machine learning algorithms are not powerful enough to handle missing data. Missing data can lead to ambiguity, misleading conclusions, and results. There are two types of missing values ; the first type is called missing completely at random (MCAR). MCAR can be expressed as:

$$P(R|X, Z, \mu) = P(R|\mu) \quad (7)$$

Where R is the response indicator variables, X are independent of data variables, and Z is latent. The second type is called missing at random (MAR), which can be expressed as:

$$P(R = r|X = x, Z = z, \mu) = P(R = r|X^0 = x^0) \quad (8)$$

$$for\ all\ x^{\mu}, z\ and\ \mu \quad (9)$$

There are two methods of handling missing data, namely ignoring missing data and imputationof missing data. Ignoring missing data is a simple technique which deletes the cases that containmissing data. The disadvantages of this method are that it reduces the size of the dataset, andituses a different sample size for different variables. Imputation of missing data is a techniquethat replaces missing data with some reasonable data values. However, the imputation of missing data method has two types, single imputation, and multiple imputations. Single imputation contains several approaches, such as mean imputation and regression imputation. Mean imputation is the most common approach of missing data replacement. It replaces the missing data with sample mean or median. However, it has a disadvantage which is if missing data are enormous in number, then all those data are replaced with the same imputation mean, which leads to change in the shape of the distribution. Regression imputation is a technique based on the assumption of the linear relationship between the attributes. The advantage of regression imputation over mean imputation is that it was able to preserve the distribution shape.

4.2.2. Outliers

Outliers are noisy data that they do have abnormal behaviour comparing with the rest of the data in the same dataset. Outliers can influence the prediction model and performance due to its oddity. There are three types of outliers, which are point, contextual, and collective outliers. Point outlier is an individual data instance that can be considered as odd with respect to the rest of the data. The contextual outlier is an instance of data that can be regarded as odd ina specific context but not otherwise. An example of contextual is the longitude of a location. Acollective

outlier is a collection of related data instances that can be considered as abnormal with respect to the entire dataset. In supervised, the detection of outliers can be accomplished visually, where a predictive model is built for normal against outliers' classes. Dean De hasinvestigated the public dataset and he suggests to remove certain outliers from the public datawhen he said "I would recommend removing any houses with more than 4000 square feet fromthe data set" [30]. Another example of detecting outliers is by using Isolation forest, which hastwo stages, training, and testing. The training is to create the isolation trees and then to recordthe anomaly score of each entry in the testing stage. This method has shown a promising result.

4.2.3. Binning

Binning is a technique purposed to reduce the impact of statistical noise, to prevent overfitting,reduce overall complexity and make the model more robust. An interval with all observed values is split into smaller sub-intervals, bins, or groups. Also, binning can be consideredas a form discretisation, which is a technique to cut a continuous value range into a finite number of sub-ranges, where a categorical value is associated with each of them. Although there are several binning methods, this report is limited to equal-width, equal-size, and multi- interval discretisation binning. Equal-width binning is an approach where the whole range of predictor values is divided into a pre-specified number of equal-width intervals. Equal-size binning is an approach where the variety of predictor values is split into intervals in a way thatbins contain an equal number of observations. However, the width of bins depends on the density of observations. Multi-interval discretisation binning is based on entropy minimisationheuristic search for recursively splitting of continuous range into sub-intervals. The entropy function is defined as:

$$Ent(S) = -\sum P(Ci\ ,\ S)\ \log(P(Ci\ ,\ S))\ k\ i=1 \tag{10}$$

Where Ci are predictor classes in input dataset S, where maximising information entropy of the partition induced by T: $S1 \in S$, $S2 = S - S1$ is given by:

$$E(T,\ S) = |S1\ |\ |S|\ Ent(S1\ ) + |S2\ |\ |S|\ Ent(S2\ ) \tag{11}$$

Once cut point T is found for compete interval of S, the process is repeated for sub-intervals recursively until there is no substantial improvement in entropy.

4.2.4. Log Transformation

A log transformation is a method that is used to handle skewed data. It is used to make data conform to normality, it reduces the impact of the outliers, due to the normalisation of magnitude and to reduce the variability of data.

4.2.5. One-hot Encoding

One-hot encoding is a technique that is used to convert categorical features to a suitable formatto be used as an input in Machine Learning algorithms. It transforms a single variable with n observations and d distinct values to d binary variables, where each observation indicating the presence as 1 or absence as 0. In one-hot encoding, the categories are represented as independent concepts.

4.2.6. Feature Selection

Feature Selection is an important technique that is used to handle high-dimensional input data and overfitting caused by a curse of dimensionality by selecting a relevant feature subset based on mutual information criterion. Moreover, feature selection has many advantages, such as improve the prediction performance by reducing dimensionality in the dataset. It speeds up the learning process and leads to a better understanding of the considered problem. However, there are many useful methods for feature selection, such as Mutual Information (MI) and Conditional Mutual Information (CMI) . Mutual information is used for quantifying the mutual dependence of random variables, and it can be considered as the amount of information shared by two variables. MI is given as:

$$I(X;Y) = \sum \sum p(xy) log\ p(xy)\ y\ p(x)p(y)\ x \in X\ \in Y \qquad (12)$$

Where $x \in X$ and $y \in Y$ are the possible value assignments of X and Y, and log is used base 2. Conditional Mutual Information measures the limited dependence between two random variables given the third [37].

$$I(X; Y|Z) = \sum p(z)\ z \in Z \sum \sum p(xy|z) log\ p(xy|z)\ y\ p(x|z)p(y|z)\ x \in X\ \in Y \quad (13)$$

If $I(X; Y|Z) = 0$, X and Y are conditionally independent given Z.

4.3. Evaluation Metrics

Several evaluation metrics measure the performance of machine learning algorithms such as Mean Squared Error (MSE), Root Mean Squared Error (RMSE), R-Squared, and Mean Absolute Error (MAE). However, in this study, the performance of the algorithms is measured by using RMSE and R-Squared. Root Mean Square Error (RMSE) is used as an evaluation metric in machine learning to measure the performance of the model. However, RMSE is similar to the Mean Square Error (MAE). Where all errors in MAE have the same weight, but RMSE penalises the variance, which means it gives more weight to the errors that have large absolute values than that have small absolute values. Therefore, when RMSE and MAE are calculated, RMSE is always bigger than MAE. RMSE is more sensitive to the errors than MAE; therefore, using RMSE for measuring the performance is better than MAE [38]. RMSE can be calculated as the square root of the sum of squared errors$\sum$ ($yi - \hat{yi}$) n 2 i over the sample size n. RMSE can be presented as:

$$RMSE = \sqrt{\sum (yi - \hat{yi}\ )\ n\ 2\ i=1\ n} \quad (14)$$

It can be observed from the equation when the sum of squared errors is closer to zero, RMSE is closer to zero. Therefore, when RMSE is zero, it means there are no errors between the actualvalue y and the predicted value $\hat{y}$.

R-Squared or as known as coefficient determination is a statistical measurement that is used inmachine learning to measure how close the data are to the fitted regression line. R-Squared takes the value between 0 and 1. Where 1 indicates the prefect score and 0 is imperfect. In addition, R-Squared is calculated by measuring the deviations of the observations from their predicted values over the measurement of the deviations of the observations from their mean. R-Squared can be presented as:

$$R\ 2 = 1 - \sum (yi - \hat{yi}\ )\ 2\ i \sum (yi - \bar{y})\ 2\ i \quad (15)$$

Where y is the model, $\bar{y}$is the mean of the model, and $\hat{y}$ is the prediction model of

y. It can be observed when the sum of squared errors $\sum_i (y_i - \hat{y})^2$ Is closer to zero and R-Squared is closer to one [39].

4.4. Research Question 1 Results

Several studies have been performed on or between multiple machine learning algorithms in order to predict and compare the prediction accuracy of the models. These studies indicate thatan algorithm has performed better in their prediction results such as Artificial Neural network(ANN), which gives better accuracy results when it is compared with multiple linear regression. However, the percentage of root square (R2) decreases after training the ANN model which means that ANN is prone to overfitting that it might have an impact on the prediction accuracy in the final evaluation.

Data pre-processing is an essential part of preparing the data to be used by the training model.It does improve the prediction accuracy according to . However, several problems might arise while dealing with the data. For instance, managing the missing values is a difficult task that can be solved in anyways. Hence, it requires a set of iterations before settling on a final solution. Moreover, outliers are noisy values within the dataset. Its existence affects the trainingmodel and preferably to be removed. It can be removed either by following suggesting of which is to be tested alongside Isolation Forest. However, these studies evaluate the accuracy of the prediction with evaluation metrics and not the time need to train a model with the algorithm. In this study, the performance in terms of time is taken into consideration.

Research Question 2

4.5. Factors

There are many factors that influence the house prices. Some of these factors influence positively, whereas others have a negative impact on house prices. The factors addressedin this section are Crime, interest, unemployment, inflation rate.

4.5.1. Crime Rate

Crime rate is the number of crimes that are perpetrated in a period of time. There are differenttypes of crimes, such as burglary, vandalism, theft, assault, and robbery. However, the high level of crime rate leads people to move out to another place where the level of crime rate is low. The number of people who are willing to buy

houses in neighbourhoods that have a high level of crime rate will decrease, and that leads to a decrease in house prices. Therefore, the house prices will decrease in neighbourhoods that have a high level of the crime rate .

4.5.2. Interest Rate

Interest rate is the ratio of a loan that is charged as an interest to the borrower. However, the Swedish Central Bank has divided the interest rates into three categories, which they are Repo, Lending, and deposit rates. Repo rate is the rate that is determined by the central bank of a country when it comes to lending money to other banks in case of any shortfall of funds inother banks . The lending rate is the amount of money that banks charge others for lendingits money. The deposit rate is the amount that banks pay for deposit holders in order to lure customers into putting their money in the bank. However, the interest rate plays a major role in clarifying the fluctuations in house prices in Sweden. In addition, the low-interest rate has participated in increasing house prices.

4.5.3. Unemployment Rate

The unemployment rate is the ratio of people that do not have a job or looking for a job. However, the high level of unemployment is harmful to society because it reduces the GDP, itcauses a loss of people's wealth, and it leads to lower taxes and higher expenses, which affect the government. In addition, the unemployment rate affects as well as the house prices. High level of unemployment indicates that there are many people who do not have jobs and cannot afford to buy houses, which leads to decrease the house prices in order to lure people into buying houses.

4.5.4. Inflation Rate

The inflation rate is the percentage of increasing or decreasing in prices of goods and services.However, the high level of inflation reduces the purchasing power of the currency. Therefore,increasing of inflation rate leads to increase house prices .

4.6. Correlation

Correlation analysis defines the strength of a relationship between two variables, which can be between two independent variables or one independent and one dependent variable. The strength of the relationship can be distinguished based on direction and dispersion strength as in figure 4.

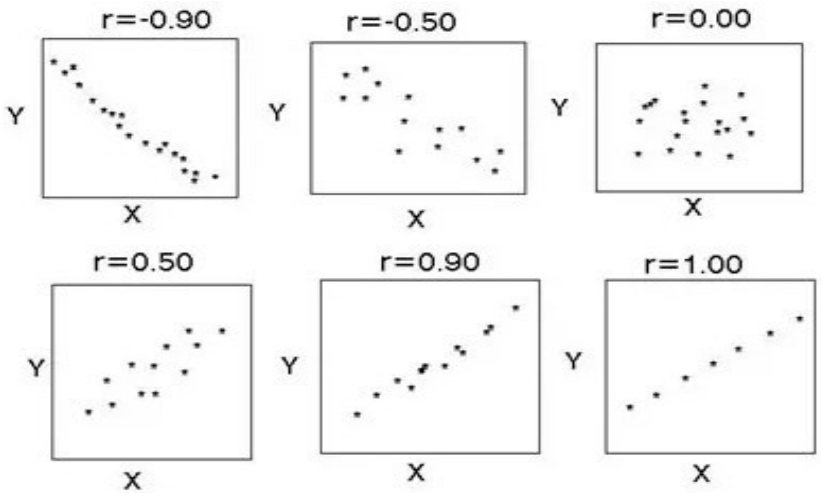


Figure 5. Correlation strength of the value of R

However, the correlation can be presented as a numerical value, which is called the correlationcoefficient. The value of the correlation coefficient ranges between 1 and -1. However, the correlation coefficient with a positive sign indicates that the two variables are positively correlated, which means when a variable increases the other increases as well. The correlation coefficient with a negative sign indicates that the two variables are negatively correlated, whichmeans when a variable increases the other decreases. In addition, when the value of the correlation coefficient is closer to either positive one or negative one means the strength of therelationship between the two variables is strong. However, when the value of the correlation coefficient is closer to zero means the strength of the relationship between the two variable is weak, where the zero value indicates to non-relationship between variables.

4.7. Research Question 2 Results

The theoretical results show that many factors have an impact on house prices, such as the unemployment rate, the total number of crimes, interest rate, GDP, and population. However, it is difficult to analyse the real estate market compared with goods market, because the real estate market has more variables to examine. However, the theoretical study shows that the increase in population, inflation and qualitative variables have a positive effect on house prices. As well as the theoretical study shows that there are some factors that affect houseprices negatively. The crime rate has a significant negative impact on house prices,

according to . Interest rate plays a major role in the fluctuation of house prices , and it has a negative correlation with house prices, which means low-interest-rate leads to increase the house price , Unemployment rate affects the GDP of the country, and it affects the government of the country leading to low taxes and high expenses . Unemployment rate affects negatively house prices, which means high level of unemployment rate leads to decreasehouse prices as reported by.

5. Experiment

5.1. Developing a Model

In this section of the project, we will develop the tools and techniques necessary for a model to make a prediction. Being able to make accurate evaluations of each model's performance through the use of these tools and techniques helps to reinforce greatly the confidence in the predictions.

5.1.1. Defining a Performance Metric

It is difficult to measure the quality of a given model without quantifying its performance on the training and testing. This is typically done using some type of performance metric, whether it is through calculating some type of error, the goodness of fit, or some other useful measurement.

For this project, we will calculate the *coefficient of determination*, R^2, to quantify the model's performance. The coefficient of determination for a model is a useful statistic

in regression analysis, as it often describes how “good” that model is at making predictions.

The values for R^2 range from 0 to 1, which captures the percentage of squared correlation between the predicted and actual values of the target variable.

- A model with an R^2 of 0 is no better than a model that always predicts the *mean* of the target variable.
- Whereas a model with an R^2 of 1 perfectly predicts the target variable.
- Any value between 0 and 1 indicates what percentage of the target variable, using this model, can be explained by the features.

A model can be given a negative R2 as well, which indicates that the model is arbitrarily worse than one that always predicts the mean of the target variable.

```
# Import 'r2_score'

from sklearn.metrics import r2_score

def performance_metric(y_true, y_predict):
    """ Calculates and returns the performance score
between
        true (y_true) and predicted (y_predict) values
based on the metric chosen. """

    score = r2_score(y_true, y_predict)

    # Return the score
    return score
```

5.1.2. Shuffle and Split Data

For this section we will take the Boston housing dataset and split the data into training and testing subsets. Typically, the data is also shuffled into a random order when

creating the training and testing subsets to remove any bias in the ordering of the dataset.

```
# Import 'train_test_split'
from sklearn.model_selection import train_test_split

# Shuffle and split the data into training and testing subsets
X_train, X_test, y_train, y_test =
train_test_split(features, prices, test_size=0.2,
random_state = 42)

# Success
print("Training and testing split was successful.")
```

```
Training and testing split was successful.
```

5.1.3. Training and Testing

You may ask now:

What is the benefit to splitting a dataset into some ratio of training and testing subsets for a learning algorithm?

It is useful to evaluate our model once it is trained. We want to know if it has learned properly from a training split of the data. There can be 3 different situations:

1) The model didn´t learn well on the data, and can't predict even the outcomes of the training set, this is called underfitting and it is caused because a high bias.

2) The model learn too well the training data, up to the point that it memorized it and is not able to generalize on new data, this is called overfitting, it is caused because high variance.

3) The model just had the right balance between bias and variance, it learned well and is able predict correctly the outcomes on new data.

5.2. Analyzing Model's Performance

In this third section of the project, we'll take a look at several models' learning and testing performances on various subsets of training data.

Additionally, we'll investigate one particular algorithm with an increasing 'max_depth' parameter on the full training set to observe how model complexity affects performance.

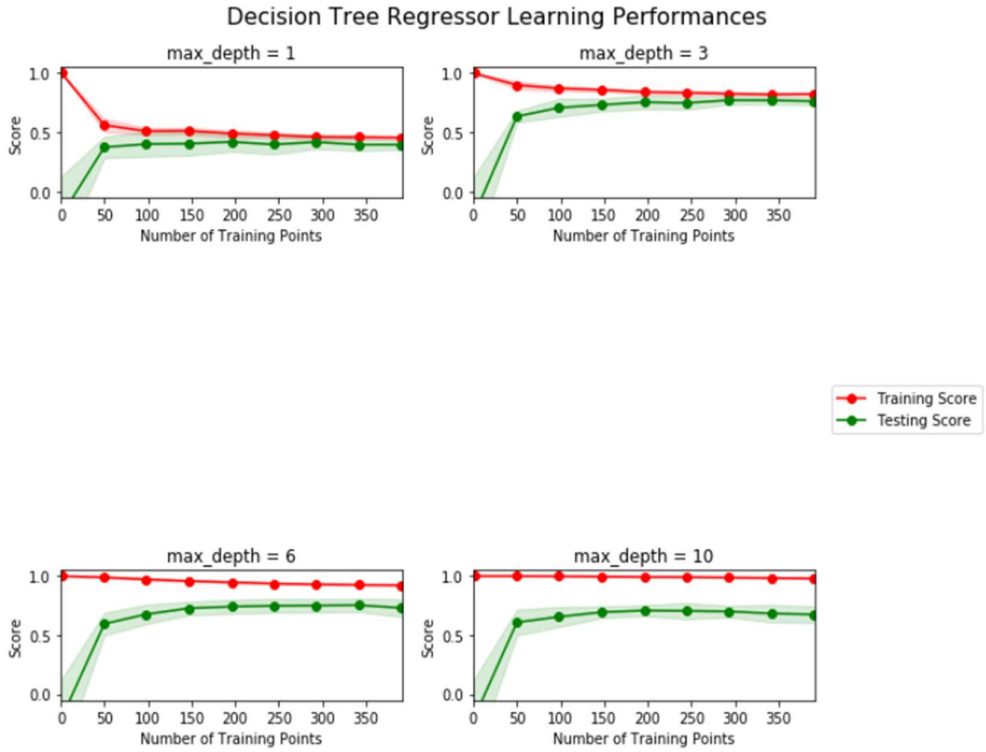
Graphing the model's performance based on varying criteria can be beneficial in the analysis process, such as visualizing behavior that may not have been apparent from the results alone.

5.2.1. Learning Curves

The following code cell produces four graphs for a decision tree model with different maximum depths. Each graph visualizes the learning curves of the model for both training and testing as the size of the training set is increased.

Note that the shaded region of a learning curve denotes the uncertainty of that curve (measured as the standard deviation). The model is scored on both the training and testing sets using R2, the coefficient of determination.

```
# Produce learning curves for varying training set sizes and maximum depths
vs.ModelLearning(features, prices)
```



5.2.2. Learning the Data

If we take a close look at the graph with the max depth of 3:

- As the number of training points increases, the training score decreases. In contrast, the test score increases.
- As both scores (training and testing) tend to converge, from the 300 points treshold, having more training points will not benefit the model.
- In general, with more columns for each observation, we'll get more information and the model will be able to learn better from the dataset and therefore, make better predictions.

5.2.3. Complexity Curves

The following code cell produces a graph for a decision tree model that has been

trained and validated on the training data using different maximum depths. The graph produces two complexity curves — one for training and one for validation.

Similar to the **learning curves**, the shaded regions of both the complexity curves denote the uncertainty in those curves, and the model is scored on both the training and validation sets using the `performance_metric` function.

```
# Produce complexity curve for varying training set sizes and maximum depths
vs.ModelComplexity(X_train, y_train)
```

5.2.4. Bias-Variance Tradeoff

If we analize how the bias-variance vary with the maximun depth, we can infer that:

- With the maximun depth of one, the graphic shows that the model does not return good score in neither training nor testing data, which is a symptom of underfitting and so, high bias. To improve performance, we should increase model's complexity, in this case increasing the max_depth hyperparameter to get better results.

- With the maximun depth of ten, the graphic shows that the model learn perfectly well from training data (with a score close to one) and also returns poor results on test data, which is an indicator of overfitting, not being able to generalize well on new data. This is a problem of High Variance. To improve performance, we should decrease the model's complexity, in this case decreasing the max_depth hyperparameter to get better results.

5.2.5. Best-Guess Optimal Model

From the complexity curve, we can infer that the best maximum depth for the model is 4, as it is the one that yields the best validation score.

In addition, for more depth although the training score increases, validation score tends to decrease which is a sign of overfitting.

5.3. Correlation

The correlation gives an overview of the association strength between available features and the house price. The correlation is being calculated using the Pearson Coefficient Correlation method. In the public dataset, features have an impact on the sale price. Figure 10 shows the correlation with a positive sign in the public data.

Figure 8. Positive correlation

On the other hand, figure 11 shows the features that have a correlation with a negative signwith house price.

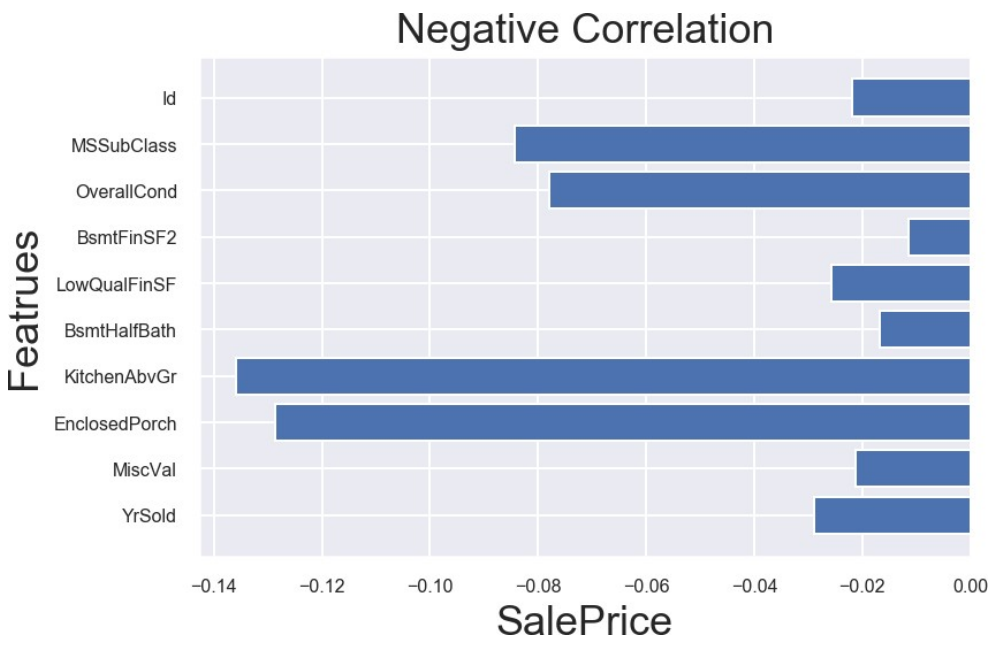


Figure 9. Negative correlation

Since there are not many features in the local data, figure 12 shows the correlation for all features in the local data.

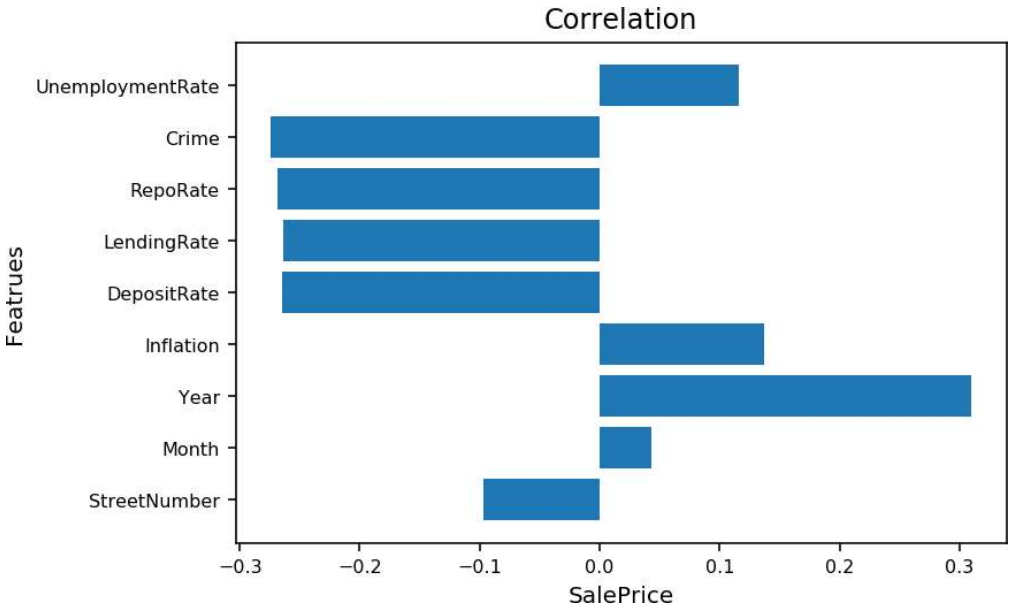


Figure 10. Correlation in data

5.4. Experiment Results

After necessary data pre-processing is done various models are implemented and evaluated. For evaluation of models Train-Test split method is used in this paper. Data is split into 80% and 20%, 80% of the data is used as training data and the rest of the data is used as test data. The first partition which is the training data is used for fitting the model and the second partition which is the test data is used for testing the accuracy of the model. There are various ways of splitting the data like 70-30, 70% for training and 30% for testing or 75-25, 75% for training and 25% for testing etc there is no hard and fast rule, but in this paper the dataset is split into the ratio mentioned above which is 80% for training and 20% for testing. After data set is split the models are implemented which are Simple Linear Regression, Polynomial Regression, Ridge Regression and Lasso Regression and the accuracy of the models are measured. To measure the accuracy of the models the metrics used are RMSE, R-Squared and cross validation. These are described as follows:

5.4.1. Root Mean Square Error (RMSE)

For evaluation of the quality of predictions Root Mean Square Error (RMSE) is used. It measures the standard deviation of the prediction errors (Residuals). Residuals are

basically the measure of the Euclidean distance of the data points from the regression line. To compute RMSE the following formula is used: The absolute fit of the model is shown by RMSE. The lower the value of RMSE is the better the fit of the model is and better the accuracy is achieved.

$$RMSE = \sqrt{\sum n\ (y\text{i} - y^*)2/n}$$

5.4.2. R-Squared

A good measure for evaluation of the fitness of the model is R-Squared. The value of R-Squared ranges from 0 to 1, 0 being 0% accurate while 1 being 100% accurate. The larger the value of R-Squared is the better the fit is achieved. For evaluation of R-Squared the following formula is used: Here SSE is the squared sum of error terms i.e. the sum of the squared residuals. Residuals are the difference between the observed and predicted values. SST is the sum of the squared total and it is the difference of each observation from the overall mean.

$$R2 = 1 - SSE/SST \quad (7)$$

where SSE (Squared sum of error): sum of the squared residuals, which is squared differences of each observation from the predicted value.

$\sum(y_\text{i} - y^*)^2$and,

SST (Sum of Squared Total): squared differences of each observation from the overall mean. $\sum(y_\text{i} - \bar{y})^2$

where y_i is the observed value, y^* is the predicted value and $\bar{y}$ is the mean of the observed values.

5.4.3. Cross-Validation

It is also known as rotation estimation or out of sample testing. It is a resampling technique in which different portions of the data is used for training and testing on different iterations. It is mainly used for predictions and cases where evaluation of the accuracy of the accuracy of a predictive model is needed. After implementation of themodels it is observed that the best accuracy in R-Squared metric is achieved by Lasso Regression which is 88.72% andthe second best accuracy is achieved by Ridge Regression which is 88.28% in the same metric. Then the third best accuracy was

achieved by Polynomial Regression which is 74.27% when the degree of the polynomial is set at 2. The least accuracy was shown by Simple Linear Regression which achieved an accuracy of 73.66%. In case of cross validation the best accuracy is given by the Lasso Regressionwhich is 85.57% and the least accuracy is observed in both Simple Linear Regression and Polynomial Regression both achieving an accuracy of 73.17%. Ridge Regression also performed well and got an accuracy of 85.53% in cross validation metric. In RMSE metric Ridge Regression and Lasso Regression got 2.88 and 2.83 score respectively whereas Linear Regression got score of 4.32 and PolynomialRegression got 4.27 score.

Model Name	R-Squared	RMSE	Cross-Validation
Simple Linear Regression	73.66%	4.329	73.17
Polynomial Regression (degree=2)	74.27%	4.279	73.17
Ridge Regression	88.28%	2.888	85.83
Lasso Regression	88.79%	2.833	85.57

TABLE II. RESULT TABLE

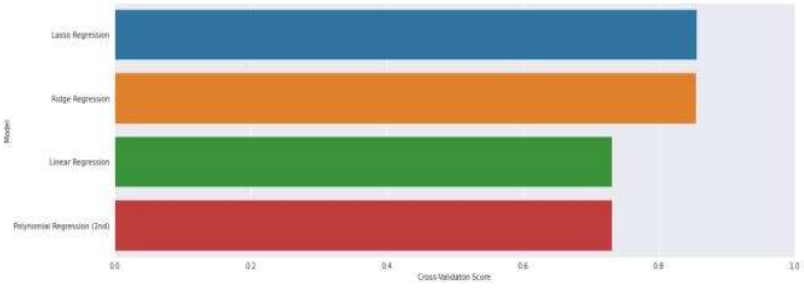


Fig 11: Cross Validation Score

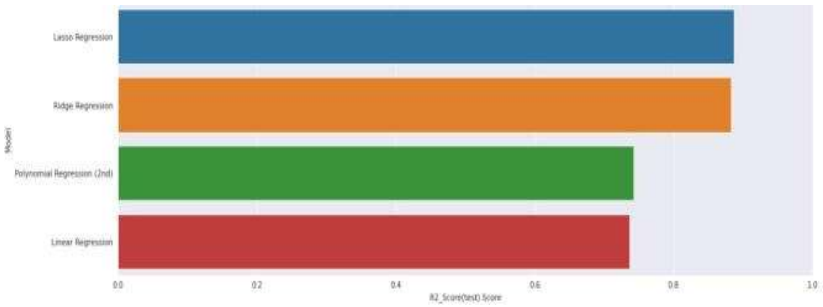


Fig 12: R-Squared

6. Discussion

This study was conducted on two different datasets, public and local. The public data set has 80 features and 1460 rows, and the local dataset has 9136 rows and a total of 13 features and 19 after adding the features presented in this study. However, the final trained model gave promising results regards the house prices.

Training of the ANN has shown that it is slower than other algorithms as in [appendix F], especially when it comes to large-size data. ANN frameworks utilise the CPU to process the data. However, there are additional GPU drives that allow the framework to make use of the GPU resources instead of the CPU. It has helped to speed up the training process rapidly whenusing a significant number of neurons in ANN.

ANN got affected negatively after eliminating the outliers when training the model with the public data. The R2 score got worse, which could fall into several reasons, such as the implemented design of ANN and removing important nodes from the dataset, which were considered as an outlier by the IsolationForest library. However, this gap has been resolved after applying the binning method presented in this study.

Applying the same algorithm that has the same property on two different datasets gives different results as the empirical results of the local and public datasets have shown. Lasso regression has the best score overall in the public dataset, and Random

Forest regression has the best score overall in the local dataset. Although, we applied the same properties for the algorithms in both public and local datasets.

Working with two different datasets prove that it is difficult to use the same pre-processing methods with multiple datasets. The results have shown differences in accuracy and performance between the two datasets. Although the local and public datasets are similar in concept, they have different and unequal features. In addition, if another design used in the implementation, it would result in different prediction accuracy.

Furthermore, the correlation strength varies between the public and local dataset. The correlation shows that the stronger the relationship, the better the accuracy, as shown from theprediction accuracy in both datasets. Thus, the local dataset requires more features that wouldsupport it to raise the correlation strength and have a chance to achieve an accurate prediction model.

Data processing and feature engineering are crucial in machine learning to build a prediction model. Furthermore, a model cannot be made without some data processing. For instance, as shown in the experiment, the model could not be trained before handling the missing values and converting the text in the dataset into numerical values. Another example is that the resultshave shown that the algorithms did not predict accurately before log transformation on the saleprice. Hence, from the experiment, we saw that pre-processing the data does improve the prediction accuracy and matches the result of.

Outliers have been resolved over two ways. They have been handled as suggested by andby using Isolation Forest. Isolation forest made a better outlier's detection which led to a noticeable improvement in the RMSE and R2 scores when looking at the ANN, for example.

Although this study has shown that Lasso made the best prediction, one cannot guarantee thatit will perform the same when used for other purposes than the ones that have been presented in this study. An example of this can be seen from comparing Lasso's output when used in bothpublic and local datasets. The accuracy differs due to the datasets not having the same characteristics.

Getting an overview of the correlation helped to understand the difference between the two datasets. It showed the value of some variables and their effect on the

prediction. A weak dataset may result in an inaccurate prediction or correlation between the features. Results showthat the correlation in the local data is weak and almost considered as zero correlation since thevalues are close to zero.

Question 1 – Which machine learning algorithm performs better and has the most accurateresult in house price prediction? And why?

The practical results show that Lasso is the most accurate algorithm among other algorithms,and it has the best performance after evaluating both RMSE and R2. Lasso achieved 0.052 and
0.13 RMSE scores on public and local data, respectively. Lasso scored 0.8912 and 0.1836 R2 scores on public and local data, respectively. Besides, Ridge and Lasso have achieved 0.0533 and 0.0529. However, Random Forest has performed better than Multiple linear, Ridge and Lasso in the local dataset due to the weakness of the local dataset and the overfitting.

Question 2 – What are the factors that have affected house prices in Boston over the years?

The theoretical result shows that crime rate, interest rate and unemployment ratehave a negative influence on house prices indicating that when these factors increase the houseprice decrease. It is difficult to analyse the real estate market compared with goods market because the real estate market has more variables to examine. In addition, inflation has a positive impact on house prices, indicating that when inflation increases the house pricesincrease.

The empirical results show that the crime rate, repo, lending, and deposit rates have a weak negative correlation with the sale price, indicating that when these factors increase the house price decrease. In addition, the empirical result shows that inflation and year have weak positive correlation indicating that when these factors increase the house prices increase.

The theoretical and empirical results are similar, where both have proved that crimes, repo, lending, and deposit rates have a negative influence on the house prices. However, the only difference between the theoretical results and empirical

results is that in the empirical results the unemployment rate has a positive correlation with the sale price, where it is a negative correlation in the theoretical results.

7. Conclusion

The study shows a comparison between the regression algorithms and artificial neural networkwhen predicting house prices in Ames, Iowa, United States and Malmö, Sweden. The results were promising for the public data due to it being rich with features and having strong correlation, whereas the local data gave a worse outcome when the same pre-processing strategy was implemented due to it being in a different shape compared with the public data interms of the number of features and the correlation strength.

Hence, the local data needs more features to be added preferably with a strong correlation withthe house price. However, ANN gave the best RMSE score, and Lasso got the best R2 score overall. The final results of this study showed that Lasso makes better prediction compared toother used algorithms.

Crime, deposit, lending, and repo rates have a weak negative influence on house prices, whereas inflation and year have a weak positive influence.

The results answer the research questions as follows:

- *Question 1 – Which machine learning algorithm performs better and has the most accurate result in house price prediction? And why?*

  Lasso made the best performance overall when both R2 and RMSE scores are taking into consideration. It has achieved the best performance due to its L1 norm regularisation for assigning zero weights to the insignificant features.

- *Question 2 – What are the factors that have affected house prices in Malmö over the years?*

  The number of crimes, repo, lending, and deposit rates has a weak correlation with thehouse prices. Which means there are lower likelihood relationships between these factors and sale price. However, when these factors increase the house price decrease. Besides, inflation and year have changed the house prices positively, which means when these factors increase, the house price increase.

7.1. Future Work

Future work on this study could be divided into seven main areas to improve the result evenfurther. Which can be done by:

- The used pre-processing methods do help in the prediction accuracy. However,experimenting with different combinations of pre-processing methods to achieve better prediction accuracy.
- Make use of the available features and if they could be combined as binning featureshas shown that the data got improved.
- Training the datasets with different regression methods such as Elastic net regressionthat combines both L1 and L2 norms. In order to expand the comparison and check the performance.
- The correlation has shown the association in the local data. Thus, attempting to enhance the local data is required to make rich with features that vary and can provide a strongcorrelation relationship.
- The factors that have been studied in this study has a weak correlation with the sale price. Hence, by adding more factors to the local dataset that affect the house price, such as GDP, average income, and the population. In order to increase the number of factors that have an impact on house prices. This could also lead to a better finding forquestion 1 and 2.
- The results of this study have shown that ANN is prone to overfitting. However, ANN still a strong algorithm that has a lot of options that could, with the right methods, provide a better prediction accuracy. ANN has a lot of possibilities that could lead to a different output. For instance, experimenting with the model when using combinations of layers and neurons over several iterations in order to find what fits the algorithm.
- ANN model was designed using feed-forward architecture. The model could make use of applying the proper back-propagation method to reduce the weight between neuronsand give a better training performance resulting in better prediction accuracy.

8. Bibliography

[1] Mansi Bosamia,'Positive and Negative Impacts of Information and Communication Technology in our Everyday Life', International Conference on "Disiplinary and Interdisciplinary Approaches to Knowledge Creation in Higher Education: CANADA and INDIA (GENESIS 2013).

[2] Rahul Reddy Nadikattu, 'THE EMERGING ROLE OF ARTIFICIAL INTELLIGENCE IN THE MODERN SOCIETY', International Journal Of Creative Research Thoughts, 2016.

[3] Ravi Manne and Sneha C Kantheti, 'Application of Artificial Intelligence in Healthcare: Changes and Challenges', Current Journal of Applied Science and Technology, 2021, DOI:10.9734/CJAST/2021/v40i631320

[4] Woubishet Zewdu Taffese, 'A Survey on Application of Artificial Intelligence in Real Estate Industry', 3rd International Conference on Artificial Intelligence in Engineering andTechnology, 2006.

[5] Ferreira, F. G. D. C., Gandomi, A. H., & Cardoso, R. T. N. 'Artificial Intelligence Applied to Stock Market Trading: A Review', IEEE, 2021, doi:10.1109/access.2021.3058133.

[6] Anandharajan, T. R. V., Hariharan, G. A., Vignajeth, K. K., Jijendiran, R., & Kushmita. (2016).' Weather Monitoring Using Artificial Intelligence'. 2016 2nd International Conference on Computational Intelligence and Networks (CINE). doi:10.1109/cine.2016.26.

[7] Tong, W., Hussain, A., Bo, W. X., & Maharjan, S., ' Artificial Intelligence for Vehicle-to-Everything: a Survey', IEEE, 2019 ,doi:10.1109/access.2019.2891073.

[8] Sumit Das, Aritra dey, Akash Paul and NAbamita Roy, 'Applications of Artificial Intelligence in Machine Learning: Review and Prospects', International Journal of Computer Applications, 2015, DOI:10.5120/20182-2402.

[9] Avneet Pannu, 'Artificial Intelligence and its Application in Different Areas', International Journal of Engineering and Innovative Technology (IJEIT), Volume 4, Issue 10, April 2015.

[10] John A. Bullinaria, 'IAI : The Roots, Goals and Sub-fields of AI', 2005. https://www.cs.bham.ac.uk/~jxb/IAI/w2.pdf

[11] Bahia, I. S. (2013). A Data Mining Model by Using ANN for Predicting Real Estate Market: Comparative Study. International Journal of Intelligence Science,03(04), 162- 169. doi:10.4236/ijis.2013.34017.

[12] Mu, J., Wu, F., & Zhang, A. (2014). Housing Value Forecasting Based on Machine Learning Methods. Abstract and Applied Analysis, Volume 2014 (2014), Article ID 648047, 7 pages. Retrieved April 2017, from https://www.hindawi.com/journals/aaa/2014/648047/.

[13] Eli Beracha, Ben T Gilbert, Tyler Kjorstad, Kiplan womack, "On the Relation between Local Amenities and House Price Dynamics", Journal of Real estate Economics, Aug. 2016.

[14] Stephen Law, "Defining Street-based Local Area and measuring its effect on house price using a hedonic price approach: The case study of Metropolitan London", Cities, vol. 60, Part A, pp. 166–179, Feb. 2017.

[15] Binbin Lu, Martin Charlton, Paul Harris & A. Stewart Fotheringham, "Geographically weighted regression with a non-Euclidean distance metric: a case study using hedonic house price data", International Journal of Geographical Information Science, pp. 660-681, Jan 2014.

[16] Marco Helbich, Wolfgang Brunauer, Eric Vaz, Peter Nijkamp, "Spatial Heterogeneity in Hedonic House Price Models: The Case of Austria", Urban Studies, vol. 51, Issue 2, Feb. 2014.

[17] Sean Holly, M. Hashem Pesarana, Takashi Yamagata, "A spatiotemporal model of house prices in the USA", Journal of Econometrics, vol. 158, Issue 1, pp. 160–173, Sep. 2010.

[18] Joep Steegmans, Wolter Hassink, "Financial position and house price determination: An empirical study of income and wealth effects", Journal of Housing Economics, vol. 36, pp. 8-24, June 2017.

[19]. Annina S, Mahima SD, Ramesh B. An Overview of Machine Learning and its Applications. International Journal of Electrical Sciences & Engineering (IJESE). 2015 January; I(1): 22-24.

[20]. Fonti V. Feature Selection using LASSO. VU Amsterdam Research Paper in Business Analytics. 2017 Mars: p. 1-25.

[21]. David HW, William GM. No Free Lunch Theorems for Optimisation. IEEE TRANSACTIONS ON EVOLUTIONARY COMPUTATION. 1997 April; I(1): 67-82.

[22]. Svensk Mäklarstatistik. [Online].; 2020. Available from: www.maklarstatistik.se.

[23]. Uyanık GK GN. A study on multiple linear regression analysis. Procedia-Social and Behavioral Sciences. 2013 Dec ; 106(1): 234-240.

[24]. Peter JB, Bo L. Regularization in Statistics. Sociedad de Estad´ıstica e Investigaci´on Operativa. 2006; XV(2): 271-344.

[25]. R T. Regression shrinkage and selection via the lasso. Journal of the Royal Statistical Society: Series B (Methodological). 1996 January; 58(1): 267-288.

[26]. Friedman J, Hastie T, Tibshirani R. The elements of statistical learning. New York: Springer series in statistics. 2001.

[27]. Clark AE, Troskie CG. Ridge Regression – A Simulation Study. Communications in Statistics - Simulation and. 2006: p. 605-619.

[28]. Yahya WB OJ. A note on ridge regression modeling. Electronic Journal of Applied Statistical Analysis. 2014 Oct : p. 343-361.

9. Appendixes

Appendix A

A list of all features in the public data placed in Ames, Iowa, United States:

ID	*Feature*	*Description*
1	*id*	*Identifies an entry*
2	*MSSubClass*	*Identifies the type of dwelling involved in the sale*
3	*MSZoning*	*Identifies the general zoning classification of the sale*
4	*LotFrontage*	*Linear feet of street connected to property*
5	*LotArea*	*Type of road access to property*
6	*Street*	*Type of alley access to property*
7	*Alley*	*General shape of property*
8	*LotShape*	*Flatness of the property*
9	*LandContour*	*LandContour*
10	*Utilities*	*Type of utilities available*
11	*LotConfig*	*Lot configuration*
12	*LandSlope*	*Slope of property*
13	*Neighborhood*	*Physical locations within city limits*
14	*Condition1*	*Proximity to various conditions*
15	*Condition2*	*Proximity to various conditions (if more than one is present)*
16	*BldgType*	*Type of dwelling*
17	*HouseStyle*	*Style of dwelling*
18	*OverallQual*	*Rates the overall material and finish of the house*
19	*OverallCond*	*Rates the overall condition of the house*
20	*YearBuilt*	*Original construction date*
21	*YearRemodAdd*	*Remodel date (same as construction date if no remodelling or additions)*
22	*RoofStyle*	*Type of roof*
23	*RoofMatl*	*Roof material*
24	*Exterior1st*	*Exterior covering on house*
25	*Exterior2nd*	*Exterior covering on house (if more than one material)*

26	*MasVnrType*	*Masonry veneer type*
27	*MasVnrArea*	*Masonry veneer area in square feet*
28	*ExterQual*	*Evaluates the quality of the material on the exterior*
29	*ExterCond*	*Evaluates the present condition of the material on the exterior*
30	*Foundation*	*Type of foundation*
31	*BsmtQual*	*Evaluates the height of the basement*
32	*BsmtCond*	*Evaluates the general condition of the basement*
33	*BsmtExposure*	*Basement exposure*
34	*BsmtFinType1*	*Refers to walkout or garden level walls*
35	*BsmtFinSF1*	*Type 1 finished square feet*
36	*BsmtFinType2*	*Rating of basement finished area (if multiple types)*
37	*BsmtFinSF2*	*Type 2 finished square feet*
38	*BsmtUnfSF*	*Unfinished square feet of basement area*
39	*TotalBsmtSF*	*--*
40	*Heating*	*Type of heating*
41	*HeatingQC*	*Heating quality and condition*
42	*CentralAir*	*Central air conditioning*
43	*Electrical*	*Electrical system*
44	*1stFlrSF*	*First Floor square feet*
45	*2ndFlrSF*	*Second floor square feet*
46	*LowQualFinSF*	*Low quality finished square feet (all floors)*
47	*GrLivArea*	*Above grade (ground) living area square feet*
48	*BsmtFullBath*	*Basement full bathrooms*
59	*BsmtHalfBath*	*Basement half bathrooms*
50	*FullBath*	*Full bathrooms above grade*
51	*HalfBath*	*Half baths above grade*
52	*BedroomAbvGr*	*Bedrooms above grade (does NOT include basement bedrooms)*
53	*KitchenAbvGr*	*Kitchens above grade*
54	*KitchenQual*	*Kitchen quality*
55	*TotRmsAbvGrd*	*Total rooms above grade (does not include bathrooms)*

56	*Functional*	*Home functionality (Assume typical unless deductions are warranted)*
57	*Fireplaces*	*Number of fireplaces*
58	*FireplaceQu*	*Fireplace quality*
59	*GarageType*	*Garage location*
60	*GarageYrBlt*	*Year garage was built*
61	*GarageFinish*	*Interior finish of the garage*
62	*GarageCars*	*Size of garage in car capacity*
63	*GarageArea*	*Size of garage in square feet*
64	*GarageQual*	*Garage quality*
65	*GarageCond*	*Garage condition*
66	*PavedDrive*	*Paved driveway*
67	*WoodDeckSF*	*Wood deck area in square feet*
68	*OpenPorchSF*	*Open porch area in square feet*
69	*EnclosedPorch*	*Enclosed porch area in square feet*
70	*3SsnPorch*	*Three season porch area in square feet*
71	*ScreenPorch*	*Screen porch area in square feet*
72	*PoolArea*	*Pool area in square feet*
73	*PoolQC*	*Pool quality*
74	*Fence*	*Fence quality*
75	*MiscFeature*	*Miscellaneous feature not covered in other categories*
76	*MiscVal*	*Value of miscellaneous feature*
77	*MoSold*	*Month Sold (MM)*
78	*YrSold*	*Year Sold (YYYY)*
79	*SaleType*	*Type of sale*
80	*SaleCondition*	*Condition of sale*
81	*SalePrice*	*Transaction price*

Appendix B

A list of all features in the local dataset placed in Malmö, Sweden:

ID	*Feature*	*Description*
1	*year*	*Year of sale*
2	*month*	*Month of sale*
3	*contract_date*	*Transaction contract date*
4	*contract_price*	*Transaction contract price in SEK*
5	*municipality_ikf*	*Municipality's code*
6	*municipality*	*Municipality's name*
7	*Formatted_address*	*Formatted street address*
8	*Route_name*	*Route name*
9	*street_number*	*Address street number*
10	*postal_town*	*Postal town*
11	*type_of_housing*	*Normalised type of housing*
12	*housing_category*	*Normalised housing category*
13	*housing_tenure*	*Normalised housing tenure*

Appendix C

R2, RMSE scores and prediction accuracy when using RMSPROP optimiser for ANN for the dataset:

Table reference	R2	RMSE
Table 10	-0.1163	1289115.125
Table 11	0.078	0.117

Table 3. R2 and RMSE scores with the optimiser RMSPROP

Table 3 includes the old R2 and RMSE scores when the ANN model had the optimiserRMSPROP in table 10 and 11

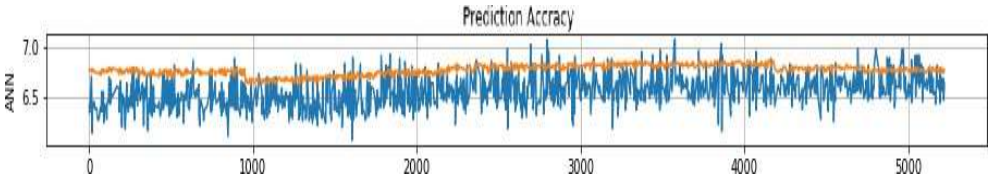


Figure 13. Prediction accuracy with the optimiser RMSPROP

Appendix D

An observation of how eliminating the outliers affect the ANN model. ANN model has been affected negatively after eliminating a total of 104 outliers that been detected with the IsolationForest library. Thus, we tested the model performance after removing outliers gradually. The results are as in the table below.

Number of removed outliers	R2	RMSE
1	0.6425	0.0495
2	0.5141	0.0831
3	0.6377	0.0629
4	0.4547	0.0546
5	0.8171	0.0460
10	0.7261	0.0517
15	0.7696	0.0569
20	0.8022	0.0505
25	0.7774	0.0517
30	0.7255	0.0561
35	0.7924	0.0503
40	0.8277	0.0507
45	0.8135	0.0518
50	0.6042	0.0701
55	0.6252	0.0841
60	0.6197	0.0813
65	0.6997	0.0752
70	0.5732	0.0594
75	0.6417	0.0563
80	0.6785	0.0574
85	0.6625	0.0835
90	0.5301	0.0909
99	0.5252	0.0462
104	0.4411	0.0882

Table 4. R2 and RMSE scores after removing the outliers gradually

To ease the observation, we plotted the values as in the figure below. From the figure, we observe that the R2 score behaves differently on multiple phases. From 1 to 50 outliers, the score behaves the same, then it drops. From 50 to 85 outliers, the score has different behaviour.The score drops furthermore when removing the rest of detected outliers.

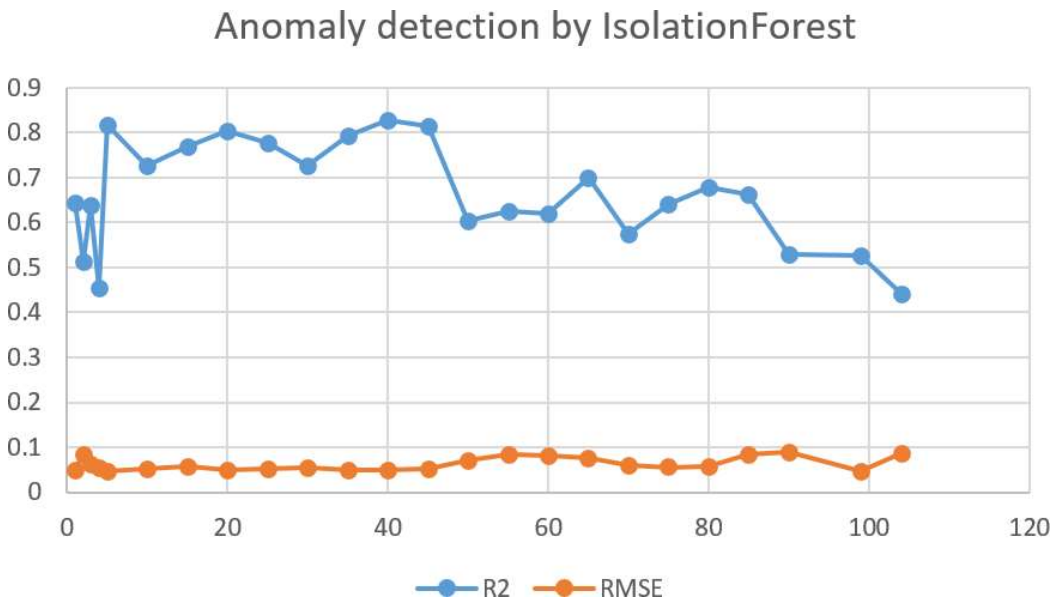


Figure 14. R2 and RMSE after eliminating outliers gradually

The anomaly detection has been performed using IsolationForest library, which works by measuring the path length from the root to a node and give it a score that will show if it normalor abnormal. We assume some of the outliers detected by this library are a part of an importantsection in the data that when taken away, the R2 gets affected negatively. On the other hand, RMSE behaves similarly throughout the elimination process.