

**College of Natural and Computational**

**Science**

**Department of Computer Science**

**COURSE TITLE:- Itroduction To Artificial Intelligence**

**COURSE CODE: CoSc3112**

Title of the Report:- house price prediction using machine learning

**Submitted to Ins:- Lencho Desalegn**

GROUP 1 MEMBER

|  |  |  |
| --- | --- | --- |
| NO | NAME | ID\_NO |
| 1 | YAIKOB TESGERA | UGR/15498/14 |
| 2 | BILISE GEMECHU | Ugr/13500/14 |
| 3 | GURMESA BORU | UGR/13998/14 |
| 4 | RAMATO KAMAL | UGR/14738/14 |
| 5 | MUDIN SHAFI | UGR/14507/14 |

Sub-date:-22/06/2024

JINKA,ETHIOPIA

Table of Contents: Page

Abstract\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_III

Chapter one\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_ 1

1.1. Introduction\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_ 2

1.2. General Objectives\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_ 2

1.2.1 Specific objectives\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_ 2

1.3. Scope\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_ 2

2. Statements of the Problem\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_ 2

3. Methodology\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_ 3

3.1 Research design\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_ 3

3.2 Data collection\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_ 3

3.3 Data preprocessing\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_ 3

3.4 Testing and training\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_ 4

4. Results and analaysis\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_ 4

5. Discussion\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_ 5

6. Conclusion\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_ 6

11. Resources\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_ 6

12. References\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_ 6

List of table

Table 1. Model outputs for trains\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_ 5

Table 2. Model outputs for test\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_ 5

Appendex-1\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_ 7

Appendex -2\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_ 7

**Abstract**

Our project consists of what and how the house price model works with the assistance of machine learning technique using scikit-learn and which datasets we will be using in our proposed model. This project provides us an overview on how to predict house prices using various machine learning models with the help of different python libraries. This proposed model considers as the most accurate model used for calculating the house price and provides a most accurate prediction. Predicting the price of a house helps for determine the selling price of the house in a particular region and it help people to find the correct time to buy a home. In this task on house price prediction using machine learning, our task is to use data to create a machine learning model to predict house prices in the given region. We will implement a linear regression algorithm on our dataset. by using real world data entities, we are going to predict the price of the house in that area. For better results we require data pre-processing units to improve the efficiency of the model. for this project we are using supervised learning, which is a part of machine learning. We have to go through different attributes of the dataset.

**Chapter 1**

**1.1 Introduction**

One of the basic requirements of livelihood in the recent world is to buy a house of your own. Predicting house prices is a crucial task in real estate, as it helps buyers and sellers make informed decisions. Traditional methods of house price prediction rely on manual analysis and may not be accurate. Machine learning models, on the other hand, can analyze large datasets and identify complex patterns, making them more accurate. the price of the house may depend on various factors. real estate agents and many who are involved in selling the house want a price tag on the house which would be the real worth of buying the house. The prediction of the price of the house is often very hard for the inexperienced. The residence price is a factor that issues a lot of citizens either wealthy or white-collar magnificence as possible by no means decide or gauge the value of a residence primarily based totally on place or places of work accessible. Purchasing a residence is the best and unique desire of a own circle of relatives because this expends the whole thing of their funding budget and every now and then cover them under loan. It is a hard challenge for expecting the correct value of residence price. This proposed version could make viable who are expecting the precise costs of house. This model of Linear regression in machine learning takes the internal factor of house valuation dependencies like area, number of bedrooms, locality etc. and external factors like air pollution and crime rates. This Linear regression in machine learning gives the output of price of the house with more accuracy.

Here in this project, we are going to use linear regression algorithm (a supervised learning approach) in machine learning to build a predictive model for estimation of the house

price for real estate customers. In this project we are going to build the Machine learning model by using python programming and other python tools like NumPy, pandas, matplotlib etc. We are also using scikit-learn library in our approach of this project. For going further into this project, we prepare dataset which consists of features like numbers of bedrooms, area of house, locality, etc.

**1.2 General objectives**

The General objective of the study is to develop a house price prediction model for all stakeholders who have a stake in the selling and buying process to pass effective decisions.

**1.2.1 Specific objectives**

Our project aims to achieve the following objectives:

* Implement Ridge Regression, Lasso Regression, and XGBoost Regressor algorithms to predict house prices.
* Evaluate and compare the performance of these models using metrics such as Root Mean Squared Error (RMSE), Mean Squared Error (MSE), and Mean Absolute Error (MAE).
* Provide insights into the significant features influencing house prices and how different algorithms handle these features.

**1.3 Scope**

Our project focuses on predicting house prices using machine learning models trained on a dataset containing various attributes such as property characteristics and neighborhood demographics. The study's scope includes:

* Data collection from a reliable real estate database.
* Data preprocessing steps to clean, transform, and prepare the dataset for analysis.
* Model development and evaluation using Ridge Regression, Lasso Regression, and XGBoost Regressor.
* Performance evaluation based on metrics like Root Mean Squared Error (RMSE), Mean Squared Error (MSE), and Mean Absolute Error (MAE).

**2. Problem Statement**

The asking price and general description are frequently presented independently from the generic and standardized real estate attributes. These qualities may be easily compared across the entire spectrum of potential houses because they are given separately and in a systematic manner. House sellers might list a summary of all the key aspects of the house in the description because every house also has distinctive elements, such as a particular view or style of washbasin. Potential purchasers can take into account all provided real estate features, but owing to the great diversity, it is almost not possible to provide an automatic comparison of all variables. This also applies in the opposite direction house sellers must evaluate the worth based on the attributes of the house in relation to the current market price of comparable houses. It is difficult to determine a fair market price due to the variety of features. Housing prices are a significant indicator of the health of the economy, and both buyers and sellers are keenly interested in price points.

**3. Methodology**

In our project, the House Prediction dataset is imported from Kaggle in comma separated values (csv) format. The dataset is analyzed with the help of pandas, numpy and scikit-learn. Tableau is used as a data visualization tool. After drawing insights from the dataset with the help of tableau, we identify the important factors i.e. factors majorly affecting the change in prices. The factors adding insignificant values to the overall result are omitted. The dataset is divided into two parts - training set and testing set. The various machine learning models are trained with the help of the training set. The testing set is then used to check the performance of all the machine learning models. Accuracy score is calculated. Root Mean Square Error of all the models is calculated. In the final step the model with the least RMSE (Root Mean Square Error) value is used for predicting house prices. To estimate housing values in this study, we used a number of well-known machine learning methods. Ridge Regression, Lasso Regression and XGB Regressorn were some of the methods we used in our project.

**3.1 Research design**

Our research aims to develop a robust predictive model for house prices using machine learning algorithms: Ridge Regression, Lasso Regression, and XGBoost Regressor. The framework involves comprehensive data collection, meticulous preprocessing, rigorous model selection, and thorough evaluation to ensure accurate predictions.

**3.2 Data collection**

Data was collected from a reputable real estate database, encompassing various attributes such as property features (e.g., size, location, number of rooms), transaction details (e.g., sale price, date of sale), and neighborhood characteristics (e.g., crime rates, school quality). The dataset was curated to include a diverse range of properties and geographical locations, ensuring the model's applicability across different market segments.

**3.3 Data preprocessing**

1. **Data Cleaning**: We performed initial data cleaning to handle missing values and outliers using statistical techniques such as mean imputation or median replacement. Outliers were identified and treated appropriately to prevent them from skewing the model's predictions.
2. **Feature Engineering**: Feature selection and transformation were crucial steps to enhance model performance:

* **Numerical Features**: Scaled using methods like Min-Max scaling or Standard scaling to normalize their ranges. Number of numerical features are: 37
* **Categorical Features**: Encoded using techniques such as one-hot encoding or label encoding to convert categorical variables into a format suitable for machine learning algorithms. Number of categorical features are: 43

**3.4 Testing and Training**

1. **Training Process**: Each selected model (Ridge Regression, Lasso Regression, XGBoost Regressor) was trained on the training dataset. Hyperparameter tuning was conducted using techniques like grid search or randomized search to optimize model performance.
2. **Testing Process**: After training, models were evaluated on the test dataset to measure their ability to generalize to unseen data. Performance metrics were compared across the Ridge, Lasso, and XGBoost models to identify the most effective approach for predicting house prices.



**4. Results and Analaysis**

To use various machine learning algorithms for solving this problem. Ridge Regression, Lasso Regression and XGB Regressorn. XGB Regressorn achieves a high accuracy score and a low root mean squared error (RMSE) value of 0.00139 is scored. This suggests that the XGB Regressorn model captures the underlying patterns and relation-ships in the data effectively, resulting in accurate predictions of house prices. Similarly, Lasso Regression achieves a commendable accuracy score of 0.887 and a reasonably low RMSE value of 0.088 for train and RMSE 0.1173 for test. XGB- Regressorn is a boosting algorithm that builds an ensemble of weak learners iteratively. It has the ability to handle complex feature interactions and can effectively capture non-linear relationships, resulting in accurate predictions. The regularization techniques employed in XGBoost help prevent overfitting and improve generalization performance.

Table 1. Model outputs for trains

|  |  |  |  |
| --- | --- | --- | --- |
| No | model | RMSE | MAE |
| 1 | . Ridge Regression | 0.2736 | 0.2029 |
| 2 | Lasso Regression | 0.088 | 0.0617 |
| 3 | XGB Regressorn | 0.088 | 0.0091 |

Table 2. Model outputs for test

|  |  |  |  |
| --- | --- | --- | --- |
| No | model | RMSE | MAE |
| 1 | Ridge Regression | 0.2919 | 0.2196 |
| 2 | Lasso Regression | 0.1173 | 0.0852 |
| 3 | XGB Regressorn | 0.1225 | 0.0904 |

**5. Discussion**

In interpreting these findings, it's evident that the selected features and model architecture effectively capture the underlying patterns in housing data. For instance, features such as location, property size, and neighborhood characteristics play significant roles in determining house prices. By understanding these relationships, stakeholders in the real estate industry can make informed decisions regarding property valuation and investment strategies. Practically, our findings have several implications for various stakeholders. Real estate agents and property developers can leverage our model to streamline pricing strategies and optimize listing prices based on accurate market insights. Home buyers and investors benefit from improved transparency and reliability in property valuation, facilitating informed decision-making processes. Theoretically, our study contributes to the growing body of literature on applying machine learning to real estate economics, highlighting the efficacy of predictive modeling in complex market environments. Despite its strengths, our study has several limitations that warrant acknowledgment. Firstly, the quality and completeness of data sources may vary, potentially introducing biases or inaccuracies into our model predictions. Furthermore, the generalizability of our findings may be constrained by regional variations or specific market conditions not fully captured in our dataset. Understanding these limitations is crucial for interpreting the results prudently and refining future research endeavors in real estate prediction modeling.

**6. Conclusion**

The primary objective of developing a reliable predictive model for house prices was success-fully met. By leveraging machine learning techniques, we not only achieved competitive per-formance metrics but also deepened our understanding of the factors driving housing market trends. The model's ability to generalize well to unseen data underscores its utility in practical applications within the real estate industry. moving forward, further research can explore several avenues to enhance the predictive accuracy and applicability of house price prediction models. exploring advanced machine learning algorithms or ensemble methods may yield optimizations in model performance. Additionally, longitudinal studies could investigate the model's stability over time and its adaptability to evolving market conditions.this study contributes to the field of real estate economics by demonstrating the effectiveness of machine learning in predicting house prices. By combining rigorous data analysis with advanced modeling techniques, we provide practical insights that can empower industry professionals and stakeholders. After preparing the dataset we gain results for train to use XGB Regressor rootMean Squared Error: 0.0013954439, Mean Squared Error: 1.947263800459433e-06, Mean Absolute Error: 0.0009145243300365053 and for Test Result of Root Mean Squared Error: 0.122528794296656, Mean Squared Error: 0.015013305431792273, Mean Absolute Error: 0.09042832502988225 of data for testing the model. Our findings underscore the potential of machine learning to revolutionize real estate valuation practices and efficiency in property markets.

**7. Resources**

We use many development software and hardware tools in the process of our project. The software tools that we used are such as:-

* **python** with Anaconda navigator and **SciKitLearn** machine learning library.
* we used ***Scikitlearns*** to evaluate the performance of the model by calculating
* different performance metrics.
* **Jupiter Notebook**: web-based interactive computing notebook environment for run python codes and used to edit and microsoft word for edit docummentation.

The hardware tools that we used are processor Intel(R) Core(TM) i7-8700 CPU @ 3.20GHz 3.19 GH, desktop installed RAM 7.86 GB and above, Flash Drive forbackup our data.

**Reference**

* Harrison, D. and D. L. Rubinfeld, 1978. Hedonic Housing Prices and the Demand for Clean Air. J. Environ. Econ. Management.
* House price prediction using a hedonic price model vs an artificial neural network. American Journal of Applied Sciences. Limsombunchai, Christopher Gan, and Minsoo Lee. 3:193–201

**APPENDIX – 1**: list of table

Model outputs for trains

Model outputs for test

**APPENDIX – 2**: Abbreviations

MSE:- Mean Squared Error

MAE:- Mean Absolute Error

RMSE:- Root Mean Squared Error

CPU:- Central Processing Unit

RAM:- Random Access memory

CSV:- comma separated values