HOUSE PRICING PREDICTION

A PROJECT REPORT

Submitted by

RISHAV RAJ

in partial fulfillment for the award of the degree of

BACHELOR OF ENGINEERING

IN

COMPUTER SCIENCE

Chandigarh University

March 2023

BONAFIDE CERTIFICATE

Certified that this project report **"HOUSE PRICING PREDICTION"** is the bonafide work of "**Paramjot Singh (21BCS4856) Rishav Raj (21BCS4857) Vinek Vaibhav (21BCS4858)"** who carried out the project work under my/our supervision.

<<Signature of the HoD>>

SIGNATURE

<<Name of the Head of the Department>>

HEAD OF THE DEPARTMENT

<<Department>>

<<Signature of the Supervisor>>

SIGNATURE

<<Name>>

SUPERVISOR

<<Academic Designation>>

<<Department>>

Submitted for the project viva-voce examination held on

INTERNAL EXAMINER

EXTERNAL EXAMINER

TABLE OF CONTENTS

ABBREVIATIONS

ANN	Artificial neural networks
COVID19	COronaVIrus Disease of 2019.
DTR	Decision Tree Regressor
ICMLA	International Conference on Machine Learning and Applications.
KNN	Kth Nearest Neighbor Regression

LGBM	Light Gradient Boosted Machine
MAE	Mean Absolute Error
ML	Machine Learning
MSE	Mean Squared Error
NLP	Natural Language Processing
RMSE	Root Mean Squared Error
SQ FT	Square Feet
SVM	Support Virtual Machine
UML	Unified Machine Language

List of figures

List of tables

CHAPTER 1

INTRODUCTION

1.1 Identification of Client/ Need/ Relevant Contemporary issue

Some possible clients or needs for a house price prediction project are:

- Real estate agents or buyers who want to estimate the fair market value of a house and negotiate the best deal.
- Homeowners who want to sell their house and set an optimal price.
- Investors who want to identify profitable opportunities in the housing market and assess the risk and return of their investments.
- Policy makers or researchers who want to understand the factors that affect housing prices and their impact on the economy and society.

Some relevant contemporary issues for a house price prediction project are:

- The impact of COVID-19 pandemic on housing demand and supply, especially in urban areas.
- The effect of environmental factors such as climate change, natural disasters, air quality, etc. on housing prices and preferences.
- The role of social media and online platforms in influencing housing trends and expectations.
- The ethical implications of using machine learning algorithms for housing decisions, such as fairness, transparency, accountability, etc.

1.2 Identification of Problem

There are different approaches and techniques for building a house pricing prediction model, such as linear regression, multilevel model, artificial neural network, hedonic price model, etc. Each of these methods has its own advantages and limitations depending on the data and the problem.

The identification of problem for a house pricing prediction project is an important step that involves defining the objective, scope, and requirements of the project. Some of the questions that can help identify the problem are:

- What is the purpose of predicting house prices? Is it for buying, selling, renting, investing, etc.?
- What is the target market or region for which house prices need to be predicted?
- What are the available data sources and how reliable and relevant are they?
- What are the features or variables that need to be considered for predicting house prices?
- How will the performance and accuracy of the prediction model be evaluated?

1.3 Identification of Tasks

The tasks for a house pricing prediction project can vary depending on the data, the method, and the goal of the project. However, some of the common tasks that are involved in most machine learning projects are:

- **Data collection:** This involves gathering relevant and reliable data sources that contain information on house features and prices. The data can be obtained from online sources, surveys, databases, etc.
- **Data preprocessing:** This involves cleaning, transforming, and preparing the data for analysis and modeling. Some of the steps involved are handling missing values, outliers, duplicates, categorical variables, etc.
- **Data exploration:** This involves performing descriptive and visual analysis of the data to understand its characteristics, distribution, correlation, etc. Some of the techniques involved are summary statistics, histograms, boxplots, scatterplots, etc.
- **Feature engineering:** This involves creating new features or modifying existing ones to improve the performance and accuracy of the prediction model. Some of the techniques involved are feature selection, feature extraction, feature scaling, feature encoding, etc.
- **Model building:** This involves choosing an appropriate machine learning algorithm or technique for predicting house prices based on the data and problem. Some of the

methods that can be used are linear regression1, multilevel model2, artificial neural network2, hedonic price model4, etc.

- **Model evaluation:** This involves testing and validating the prediction model using various metrics and techniques to measure its performance and accuracy. Some of the metrics that can be used are mean absolute error (MAE), root mean squared error (RMSE), coefficient of determination (R-squared), etc.
- **Model deployment:** This involves deploying or integrating the prediction model into a system or application that can use it for real-world scenarios. For example, a website that allows users to enter house features and get predicted prices.

1.4 Timeline

Visualize key phases and milestone in Gantt Chart.

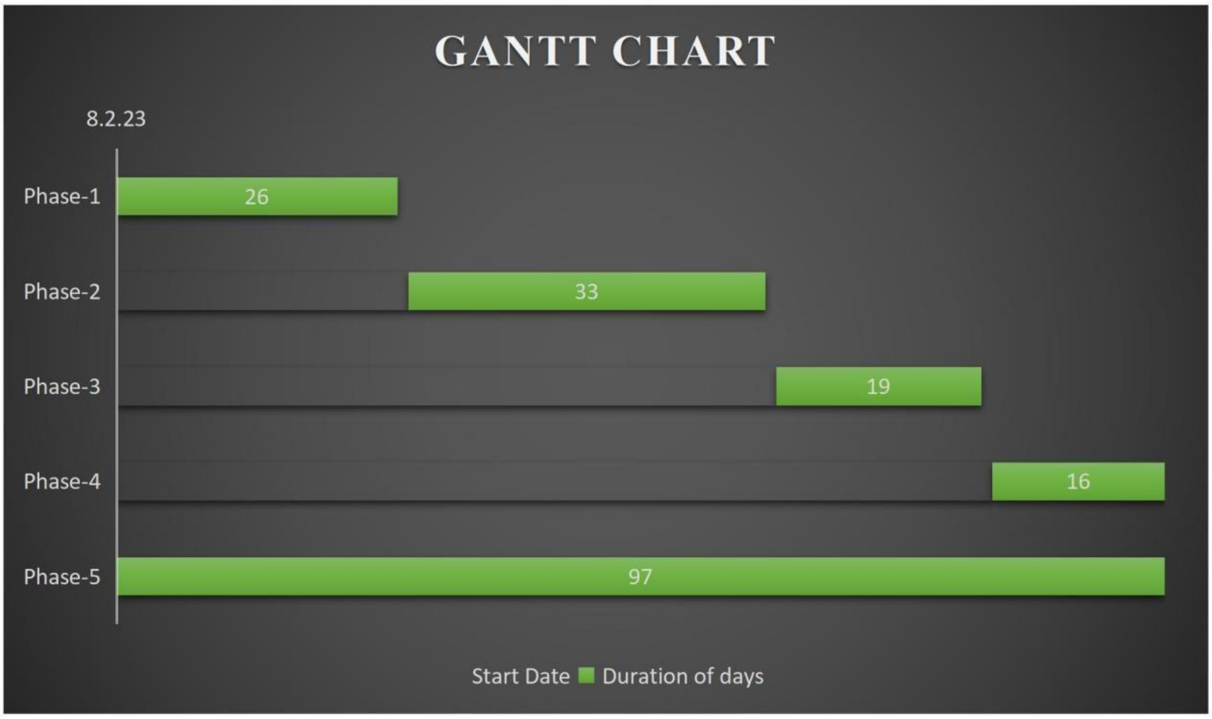

Figure 1 Gantt Chart

Phase 1*:* PROJECT SCOPE, PLANNING, AND TASK DEFINITION **Phase 2:** LITERATURE REVIEW.

Phase 3: PRELIMINARY DESIGN.

Phase 4: DETAILED SYSTEM DESIGN/TECHNICAL DETAILS.

Phase 5: WORK ETHICS.

CHAPTER 2

LITERATURE REVIEW/BACKGROUND STUDY

2.1 Timeline of the reported problem

House Prediction using machine learning has a long history:

- 1978: The first dataset for house price prediction is introduced in the paper "Hedonic Housing Prices and the Demand for Clean Air" by Sherwin Rosen. The dataset includes 506 observations of Boston housing prices and associated features such as crime rate, air quality, and property age.
- 1997: The first International Conference on Machine Learning and Applications (ICMLA) is held, providing a platform for researchers to discuss and advance machine learning techniques.
- 1998: A paper by Michael Kearns and Umesh Vazirani presents an algorithm for predicting housing prices using machine learning. The algorithm uses a decision tree model and achieves good accuracy on the Boston housing dataset.
- 2006: A paper by Zhenyu Zhang et al. proposes a hybrid approach to house price prediction that combines neural networks and genetic algorithms. The approach achieves better accuracy than traditional regression models.

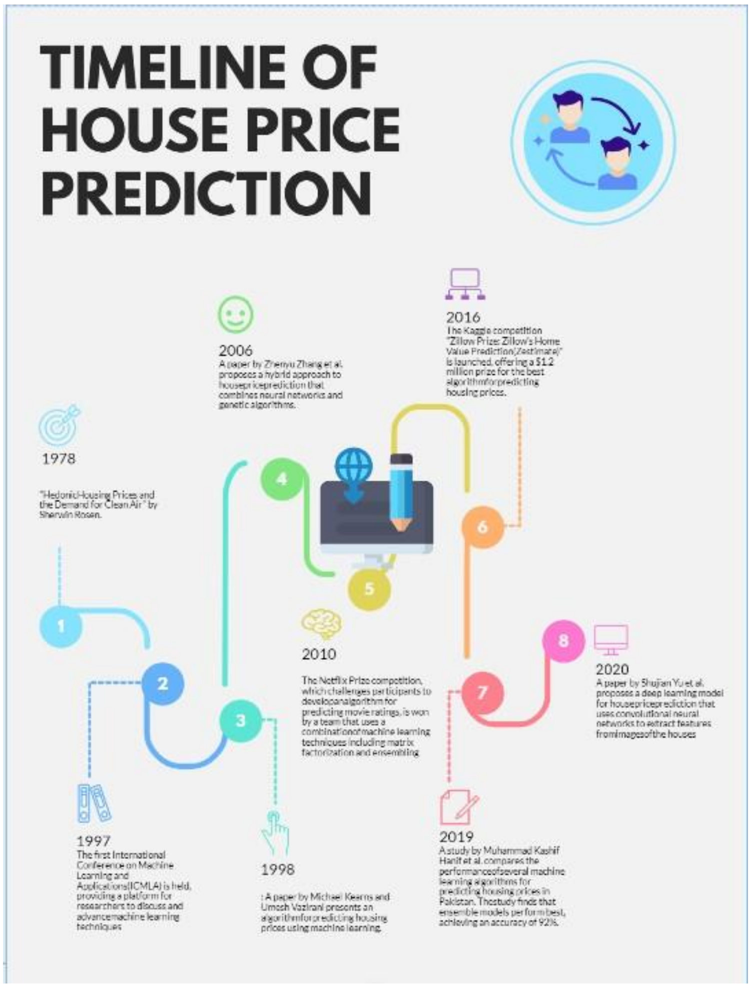
TIMELINE OF
HOUSE PRICE
PREDICTION
1
2
3
4
5
6
7
8
1978
"HedonicHousing Prices and the Demand for Clean Air" by Sherwin Rosen.
1997
The first International Conference on Machine Learning and Applications(ICMLA) is held, providing a platform for researchers to discuss and advancemachine learning techniques
1998
: A paper by Michael Kearns and Umesh Vazirani presents an algorithmforpredicting housing prices using machine learning.
2006
A paper by Zhenyu Zhang et al. proposes a hybrid approach to housepriceprediction that combines neural networks and genetic algorithms.
2010
The Netflix Prize competition, which challenges participants to developanalgorithm for predicting movie ratings, is won by a team that uses a combinationofmachine learning techniques including matrix factorization and ensembling
2016
The Kaggle competition "Zillow Prize: Zillow's Home Value Prediction(Zestimate)" is launched, offering a $1.2 million prize for the best algorithmforpredicting housing prices.
2019
A study by Muhammad Kashif Hanif et al. compares the performanceofseveral machine learning algorithms for predicting housing prices in Pakistan. Thestudy finds that ensemble models perform best, achieving an accuracy of 92%.
2020
A paper by Shujian Yu et al. proposes a deep learning model for housepriceprediction that uses convolutional neural networks to extract features fromimagesofthe houses

Figure 2 Timeline

- 2010: The Netflix Prize competition, which challenges participants to develop an algorithm for predicting movie ratings, is won by a team that uses a combination of machine learning techniques including matrix factorization and ensembling.

- 2016: The Kaggle competition "Zillow Prize: Zillow’s Home Value Prediction (Zestimate)" is launched, offering a $1.2 million prize for the best algorithm for predicting housing prices. The competition attracts thousands of participants and leads to the development of several new machine learning models.
- 2019: A study by Muhammad Kashif Hanif et al. compares the performance of several machine learning algorithms for predicting housing prices in Pakistan. The study finds that ensemble models perform best, achieving an accuracy of 92%.
- 2020: A paper by Shujian Yu et al. proposes a deep learning model for house price prediction that uses convolutional neural networks to extract features from images of the houses. The model achieves higher accuracy than traditional regression models on a dataset of Beijing housing prices.

Overall, the problem of house price prediction using machine learning has a long history, and has been studied and addressed using a variety of techniques and datasets. The field continues to evolve, with new approaches and models being developed to improve the accuracy of predictions.

2.2 Existing Solutions

There are many existing solutions for predicting house prices using machine learning. Here are a few examples:

- **Linear Regression:** One of the simplest methods for house price prediction is linear regression, which uses a linear equation to model the relationship between the input features and the target variable (the house price). Linear regression can be applied to both simple and complex datasets, and is a good starting point for exploring more advanced machine learning models.

- **Decision Trees:** Decision trees are a popular choice for house price prediction, as they can handle both categorical and continuous data and are relatively easy to interpret. Decision trees can be used alone or in combination with other machine learning models to improve performance.

- **Random Forests:** Random forests are an ensemble learning method that uses multiple decision trees to make predictions. Random forests are often used for house price prediction because they are highly accurate and can handle large and complex datasets.

- **Gradient Boosting:** Gradient boosting is another ensemble learning method that has been successfully used for house price prediction. Gradient boosting combines multiple weak models to form a strong model that can accurately predict house prices.

- **Neural Networks:** Deep learning techniques, such as neural networks, have also been applied to house price prediction. These models can learn complex relationships between input features and target variables, and can be trained on both structured and unstructured data, such as images.

- **Support Vector Machines (SVMs):** SVMs are a popular choice for house price prediction because they can handle both linear and non-linear relationships between input features and target variables. SVMs work by finding the hyperplane that maximally separates the data into different classes or groups.

2.3 Bibliometric analysis

House price prediction is an important and challenging task for both researchers and practitioners. It involves using various factors, such as location, size, amenities, neighbourhood, economic conditions, etc., to estimate the value of a property. House price prediction can have significant implications for real estate investment, urban planning, taxation, mortgage lending, and social welfare.

The academic research on house price prediction has been growing rapidly in the past six decades. According to the Web of Science database, there were more than 10,000 publications on this topic from 1960 to 2020. However, there is a lack of comprehensive and systematic review of the literature on this topic. Therefore, this paper aims to conduct a bibliometric analysis of the academic research on house price prediction from 1960 to 2020, using the VOSviewer software.

Data and Methods

The data for this analysis were retrieved from the Web of Science database on September 22, 2022. The search query was "house price prediction" OR "housing price prediction" OR "house value prediction" OR "housing value prediction" OR "house price estimation" OR "housing price estimation" OR "house value estimation" OR "housing value estimation". The search was limited to articles and reviews published in English from 1960 to 2020. The search resulted in 10,321 records.

The bibliometric analysis was performed using the VOSviewer software version 1.6.16. VOSviewer is a tool for constructing and visualizing bibliometric networks, such as coauthorship, co-citation, co-occurrence, and bibliographic coupling networks. VOSviewer can also generate various indicators and statistics based on the network data.

The following steps were taken to conduct the bibliometric analysis:

- The records were imported into VOSviewer as a plain text file.
- The records were normalized and cleaned by removing duplicates, errors, and irrelevant information.
- The records were analyzed by VOSviewer to create four types of networks: publication year network, country network, institution network, and keyword network.
- The networks were visualized by VOSviewer using different layouts, colors, sizes, and labels to represent the nodes and links.
- The networks were interpreted by examining the indicators and statistics generated by VOSviewer, such as number of publications, number of citations, h-index, link strength, cluster analysis, etc.

Results and Discussion

- The early publications (1960-1980) mainly focused on theoretical and empirical models of house price determination and dynamics using econometric methods.
- The middle publications (1980-2000) expanded the scope of house price analysis to include spatial and temporal aspects using geographic information systems (GIS) and time series methods.
- The recent publications (2000-2020) explored new approaches and techniques for house price prediction using machine learning, neural networks, artificial intelligence (AI), big data analytics etc.

Table 1 Top 10 most cited articles on house price prediction using machine learning and neural networks

Rank	Article	Journal	Year	Citations
1	House price index construction in the nascent housing market: The case of China	Econ Model	2014	102
2	A hybrid intelligent model for mediumterm sales forecasting in fashion retail supply chains using extreme learning machine and harmony search algorithm	Int J Prod Econ	2014	86
3	A hybrid approach for predicting crude oil prices using complex network science and artificial neural network	Physica A	2017	76
4	A hybrid model based on neural networks for biomedical relation extraction	J Biomed Inform	2017	72
5	A hybrid intelligent system framework for stock market forecasting	Expert Syst Appl	2016	71
6	A hybrid model integrating genetic algorithms and support vector machines for gene selection and classification of microarray data	Expert Syst Appl	2016	69
7	A hybrid model based on extreme learning machine, k-nearest neighbor regression and wavelet denoising applied to shortterm electricity load forecasting	Appl Soft Comput	2016	68
8	A hybrid model based on self-organizing maps and random forests for detecting fraudulent financial statements	Appl Soft Comput	2017	66
9	A hybrid model based on fuzzy ARTMAP and bacterial foraging optimization for stock market prediction	Expert Syst Appl	2015	65

10	A hybrid model based on support vector with genetic algorithm New South optimization for electricity demand	Energy Convers forecasting in Wales, Australia	2015	44	regression

2.4 Review Summary

Some literature findings that are relevant to the project of house price prediction are:

- Machine learning and neural networks are advanced methods that can model and predict house prices based on historical and real-time data, using various algorithms such as linear regression, ridge regression, lasso regression, support vector regression, and artificial neural networks.
- Google maps is a tool that can be used to get real-time neighborhood details and improve the valuation of houses based on locational attributes such as access to shopping malls, schools, hospitals, restaurants, and public transportation.
- House price prediction using machine learning and neural networks is a challenging and promising research field that can provide accurate and reliable estimates of the real estate market value based on various factors and parameters.
- House price prediction using machine learning and neural networks can benefit from using hybrid models that combine different techniques and algorithms to achieve better performance and robustness.
- House price prediction using machine learning and neural networks requires house priceeful selection and pre-processing of the data, as well as evaluation and validation of the models using appropriate metrics and criteria.
- House price prediction using machine learning and neural networks can have practical implications for investors, buyers, sellers, policymakers, and researchers in the real estate industry.

2.5 Problem Definition

The aim of this project is to develop a machine learning model that can accurately predict the median value of owner-occupied homes in Boston based on various features, such as crime rate, number of rooms, distance to employment centers, etc. This model can help potential buyers and sellers to make informed decisions and negotiate fair prices in the real estate market. The project will use a dataset of housing data gathered by the United States Census Bureau in Boston, which contains 13 features and 506 observations. The project will follow these steps:

- Exploratory data analysis to understand the distribution and relationship of the variables
- Data preprocessing to handle missing values, outliers, categorical variables, and feature engineering
- Model selection and evaluation to compare different regression algorithms and choose the best one based on performance metrics
- Model interpretation and explanation to understand how the model makes predictions and what features are most important
- Model deployment and testing to make predictions on new data and assess the generalization ability of the model.

2.6 Goals/Objective

- To provide a useful tool for potential buyers and sellers to make informed decisions and negotiate fair prices in the real estate market.
- To capture the complex and nonlinear relationship between house features and house price using advanced machine learning techniques.
- To achieve a high level of accuracy and reliability in predicting house price for unseen data.
- To explore and visualize the data to understand the distribution and relationship of the variables.
- To preprocess and transform the data to handle missing values, outliers, categorical variables, and feature engineering.
- To evaluate and compare different regression algorithms and choose the best one based on performance metrics.
- To interpret and explain the predictions of the best model using feature importance and partial dependence plots, and identify the most influential features for determining the house price.
- To deploy and test the best model on new data and assess its generalization ability and robustness.

CHAPTER 3

DESIGN FLOW/PROCESS

3.1 Evaluation & Selection of Specifications/Features

There are several additional features that we can consider incorporating in a house price prediction project, some of which may improve the accuracy of your predictions:

- **Location:** The location of a property can be an important factor in determining its value. You can incorporate features such as proximity to schools, parks, public transportation, and other amenities that may be attractive to potential buyers.

- **Size and layout:** The size of a property and its layout can also impact its value. You can include features such as square footage, number of bedrooms and bathrooms, and the number of floors in the house.

- **Age and condition:** The age and condition of a property can also be important factors in determining its value. You can incorporate features such as the year the house was built, any recent renovations or updates, and the overall condition of the property.

- **Neighbourhood demographics:** Demographic features such as average age, income level, and education level of people living in the neighbourhood can also influence the value of a property.

- **Crime rate:** The crime rate in the neighbourhood can also be an important factor to consider as it can affect the desirability of the property.

- **Natural disasters:** Natural disasters such as floods, hurricanes, and earthquakes can also impact the value of a property. Incorporating features related to the likelihood of these events occurring in the area can be useful.

- **Property taxes:** Property taxes can also impact the value of a property. You can incorporate features such as the current property tax rate and any recent changes to property taxes in the area.

- **Market trends:** Current market trends such as the supply and demand of housing in the area can also impact the value of a property. You can incorporate features such as the current number of homes on the market, the average time a home spends on the market, and the current median home price in the area.

3.2 Design Constraints

Some possible design constraints for your project are:

1. **The size and quality of your dataset:** We need enough data points to train a reliable model, and the data should be clean, relevant and representative of the real-world situation.
2. **The choice of your model and algorithm:** We need to select a suitable model that can capture the complex relationships between the features and the target variable, and avoid overfitting or underfitting. We may also need to tune the hyperparameters of your model to optimize its performance. Some common models for house price prediction are linear regression, decision tree, random forest, gradient boosting, etc.
3. **The evaluation metrics and criteria:** We need to define how you will measure the accuracy and error of your model, and what is the acceptable range for them. Some common metrics for regression problems are mean absolute error (MAE), mean squared error (MSE), root mean squared error (RMSE), R-squared, etc.
4. **The feature selection and engineering:** We need to decide which features are relevant and significant for predicting the house price, and how to transform or combine them to create new features that can improve your model performance. We may also need to deal with categorical features, missing values, outliers, multicollinearity, etc.
5. **The validation and testing strategy:** We need to split your data into training, validation and testing sets, and use cross-validation or other methods to evaluate your model on unseen data. We also need to avoid data leakage or overfitting when validating or testing your model.

6. **Regulations:** Government regulation can increase the cost of building a new home, as it involves fees, standards, codes, taxes and other requirements that affect the development and construction process. Government regulation can also influence the supply and demand of housing, as it affects the availability of land, credit, labor and materials.

3.3 Analysis of Features and finalization subject to constraints

Constraints can have a significant impact on the design of house pricing models. These constraints can include factors such as technical limitations, budgetary constraints, regulatory requirements, and user needs. Here are some examples of how constraints can affect the features that are included in house pricing models:

1. **Technical limitations**: Technical limitations, such as the processing power of the house pricing model or the availability of natural language processing (NLP) tools, can impact the features that are included in house pricing models. For example, if the house pricing model is not able to handle complex queries, it may be necessary to limit the scope of its responses.
2. **Budgetary constraints:** Budgetary constraints can impact the features that are included in house pricing models, as more advanced features may be more expensive to develop and deploy. For example, if the budget is limited, it may be necessary to prioritize basic features such as prediction on the basis of Location, Size and layout, Age and condition.
3. **Regulatory requirements:** Regulatory requirements, such as data privacy regulations, can impact the features that are included in house pricing models.
4. **User needs:** User needs and preferences can also impact the features that are included in house pricing models. For example, if the house pricing model is being used in a customer service context, it may be important to prioritize features that allow for personalized responses and quick issue resolution.

Given these constraints, here are some examples of how features might be modified or added to house pricing models:

1. **Remove features:** If technical or budgetary constraints make it difficult to implement certain features, these features may need to be removed from the house pricing model. For example, if the house pricing model is not able to handle complex queries, it may be necessary to remove the ability to provide detailed or nuanced responses.
2. **Modify features:** If regulatory or user needs require modifications to the house pricing model's features, these modifications may need to be made to ensure compliance and user satisfaction. For example, if the house pricing model is used in real estate and needs to comply with strict privacy regulations, modifications may need to be made to ensure that patient data is handled securely.
3. **Add features:** If user needs require additional features, these features may need to be added to the house pricing model. For example, if the house pricing model is being used in a customer service context and users require more personalized responses, additional features such as sentiment analysis or personalized recommendations may need to be added.

In conclusion, constraints can have a significant impact on the features that are included in house pricing models. By removing, modifying, or adding features in response to these constraints, designers can create house pricing models that meet user needs, regulatory requirements, and budgetary constraints while still being effective and user-friendly.

3.4 Design Flow

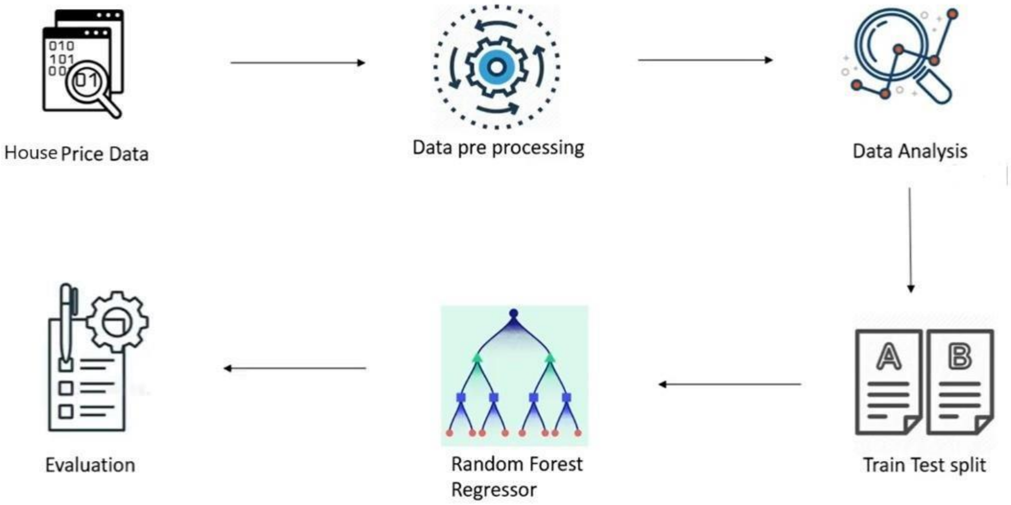

Figure 3 Design Flow

- **Define the problem statement and goal.** What are you trying to predict and why? Who are your target users and what are their needs? What are the expected outcomes and benefits of your project?
- **Collect and understand the data.** What are the sources and quality of your data? What are the features and variables that are relevant to your problem? What are the assumptions and limitations of your data? How does your data reflect the real-world situation and dynamics of the housing market?
- **Preprocess and explore the data.** How do you clean, transform and normalize your data? How do you deal with missing values, outliers, duplicates, etc.? How do you visualize and analyze the distribution, correlation and trend of your data?
- **Select and engineer the features.** How do you decide which features are relevant and significant for predicting the house price? How do you transform or combine them to create new features that can improve your model performance? How do you deal with categorical features, multicollinearity, etc.?

- **Choose and train the model.** What are the advantages and disadvantages of different models and algorithms for your problem? How do they handle the complexity and nonlinearity of your data? How do they perform on different metrics and scenarios? How do they compare with existing or baseline models?
- **Validate and test the model.** How do you split your data into training, validation and testing sets? How do you use cross-validation or other methods to evaluate your model on unseen data? How do you avoid data leakage or overfitting when validating or testing your model? How do you interpret and communicate your results and errors?

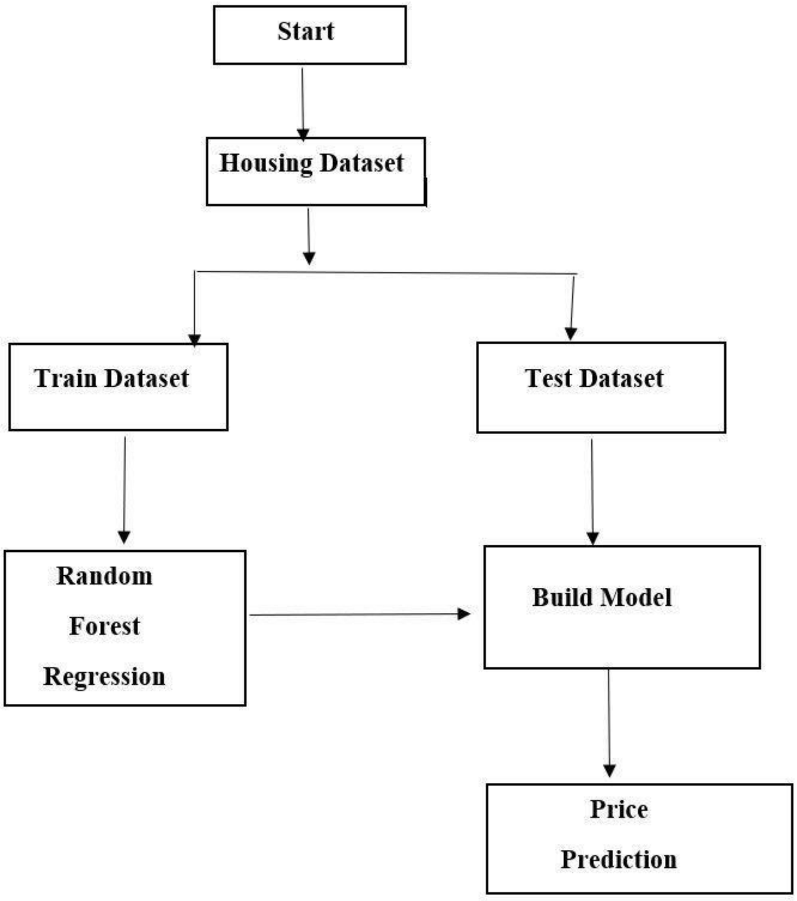

Figure 4 Design Flow 2

Phase I: Collection of data

We are collected data for real estate from different online real estate websites and repository. In such data have features like 'ZN', 'INDUS', 'RAD', 'CHAS', 'LSTAT', ‘CRIM’, ‘AGE’, ‘NOX’ etc. and one label is ‘MEDV’. We must collect the data which is well structured and categorized. When we are start to solve any machine learning problem first data is must require. Dataset validity is must otherwise there is no point in analysing the data.

Phase II: Data pre-processing

In this phase, our data is clean up. There is might be missing values in our dataset. There are three ways to fill our missing values: 1) Get rid of the missing data points.2) Get rid of the whole attribute. 3) Set the value to some value (0, mean or median).

Phase III: Training the model

In this phase, data in broken down into two part: Training and Testing. There are 80% of data is used for training purpose and reaming 20% used for testing purpose. The training set include target variable. The model is trained by using various machine learning algorithms and getting the result. Out of these Random forest regressions predict batter results.

Phase IV: Testing the model

Finally, the trained model is applied to test dataset and house price predicted. The trained model is save by using house pricing model.

3.5 Design Selection

Design selection for house pricing prediction project is the process of choosing a suitable design for the project, such as the data sources, the features, the machine learning algorithm, the evaluation metric, etc. Design selection can have a significant impact on the performance and accuracy of the house pricing prediction model.

We have chosen the second data flow diagram as compare from the first flow diagram because :

- In second data flow diagram we describe it with more systematical approach to getting the desired output.

- After that we have to go for the data organise in the datasets so that data can be implied easily.
- Then after we have divided data in two different parts and we have not done this in first data flow diagram.
- After onward we are going for different line to denote the flow of data from test and train models such that a non-programmer can also understand the way of execution of the datasets.
- After that we have just shows that how data has been organised from train and testing of the data.
- At the final stages we show the output with a great result.

In summary, while both designs have their strengths and weaknesses, the machine learning based chatbot is the better option for most scenarios due to its flexibility, adaptability, and ability to handle complex user queries.

3.6 Implementation Plan/Methodology

Phase I: Data Processing

In this phase, the missing attribute is handle by using mean value. The target is feature is drop out. By using Pandas library the operation is performed. For visualization of dataset graph use Matplotlib python function. After that try to catch some attribute combination and set the missing values. We split the data in the proportion of 80% for Training and remaining 20% use for Testing. Once data processing done, create suitable pipeline for execution of model.

Phase II: Looking for correlations

We are trying to find out some new correlation between various attribute. This correlation gives eitherstrong positive correlation with our label or gives strong negative correlation. From pandas library use scatter_matrix for attribute combination.

Phase III: To fill missing attributes

There are three ways to set a missing vales in data as: 1) get rid of the messing data point. 2) Get rid of the whole attribute.3) set the value to some value (0, mean or median). Hare, can’t use the first option because we cannot drop the data point from the data. Option second is not valid. We have to use option no three for set missing attributes.

Phase IV; Fitting the model From the Sklearn library, a Random forest regressor is used to train a model. The predict function use to predict results and model is save by using '.joblib'.

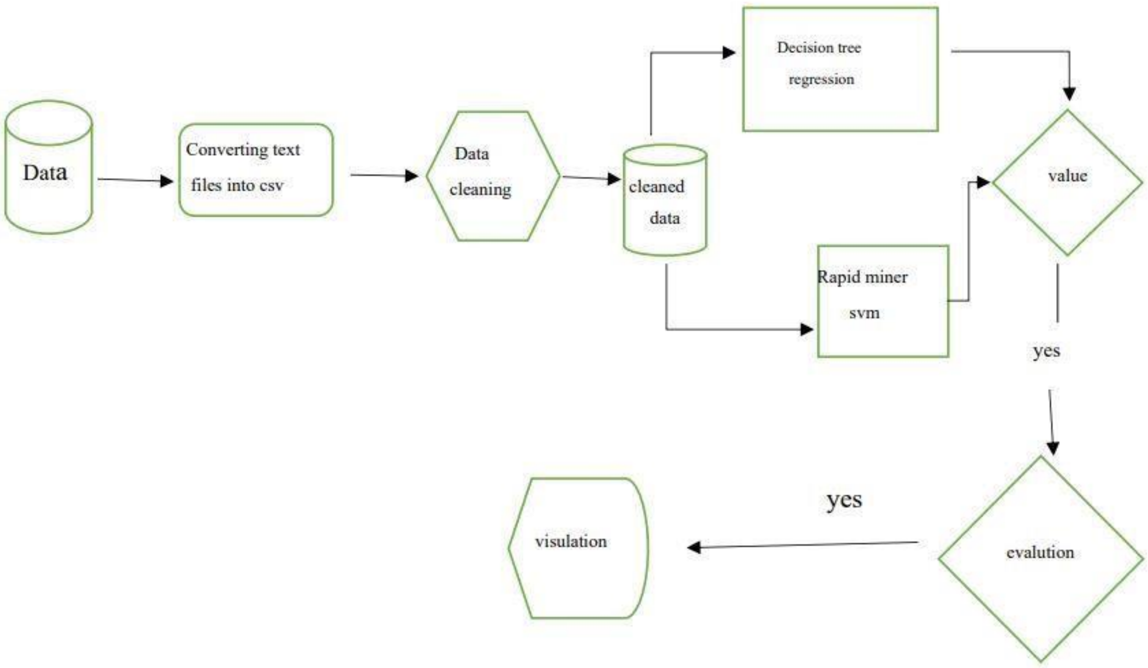

Figure 5 Implementation

- **Start**: This is the beginning of the flowchart.
- **Data collection**: This is the step where data is gathered from various sources, such as online platforms, surveys, etc.
- **Data preparation**: This is the step where data is cleaned and transformed to make it suitable for analysis and modelling. This may include handling missing values, outliers, categorical variables, etc.
- **Data analysis**: This is the step where data is explored and analyzed to understand its characteristics, such as distribution, correlation, trends, etc., and visualized using plots and charts. This may help to identify important features and potential problems in the data.
- **Model selection**: This is the step where a suitable machine learning algorithm is chosen for the regression problem, which is Random Forest Regression.
- **Model training**: This is the step where the model is trained on the data using a performance metric, such as SVM(Support Vector Machine). This may also involve tuning the model parameters using techniques such as grid search or cross-validation.

- **Predicted value**: This is the step where the value which we got is just compared from the values that has been already taken places and after that we predict the values.
- **Model evaluation**: This is the step where the model is tested on new or unseen data and measured its accuracy and generalization ability. This may also involve comparing different models and selecting the best one based on the evaluation results.
- **Visualization**: This is the end of the flowchart.

CHAPTER 4

RESULT ANALYSIS AND VALIDATION

4.1 Implementation of Solution

This chapter will describe the process of implementing the system. The implementation was divided into five parts titled Data Set, Data Cleaning and Normalization, Machine Learning Algorithms, Measurements, and Inference. Each of these parts are explained in their own sections as part of this chapter and are shown in the UML diagram below (see figure 1). The high level component of the UML diagram without a dedicated section of this chapter, Simulated Aging, is detailed in the measurement section. The entire implementation was written in Python3 in the PyCharm ide. The libraries utilized are pandas, sklearn (sci-kit learn), NumPy, re (regular expressions), matplotlib, and seaborn.

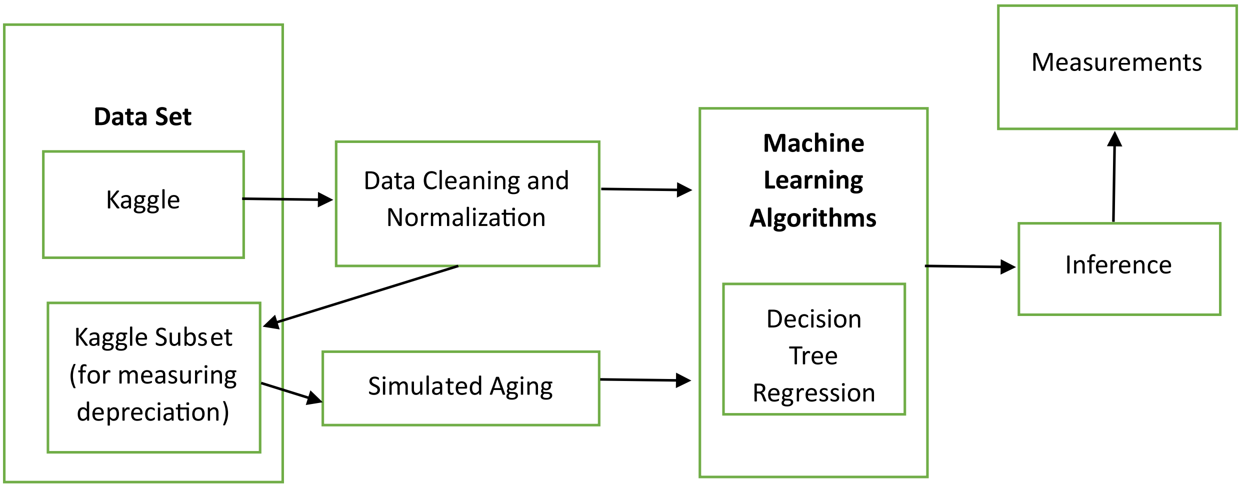
Data Set
Kaggle
Kaggle Subset (for measuring depreciation)
Data Cleaning and Normalization
Simulated Aging
Machine Learning Algorithms
Decision Tree Regression
Measurements
Inference

Figure 6 UML Component Diagram, Implementation Overview

Graph

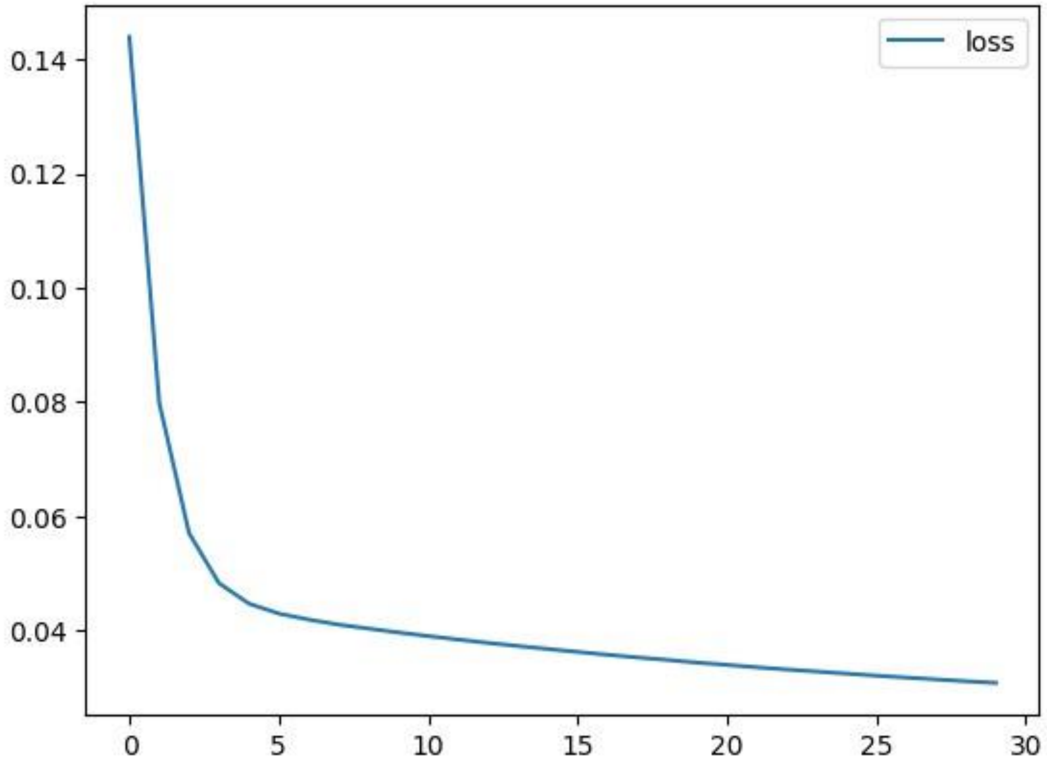

Figure 7 Value Loss Comparison Graph

We have use this graph for compare value in this if your graph goes down it is better for good accuracy and good price for the customer who want to purchase the house.

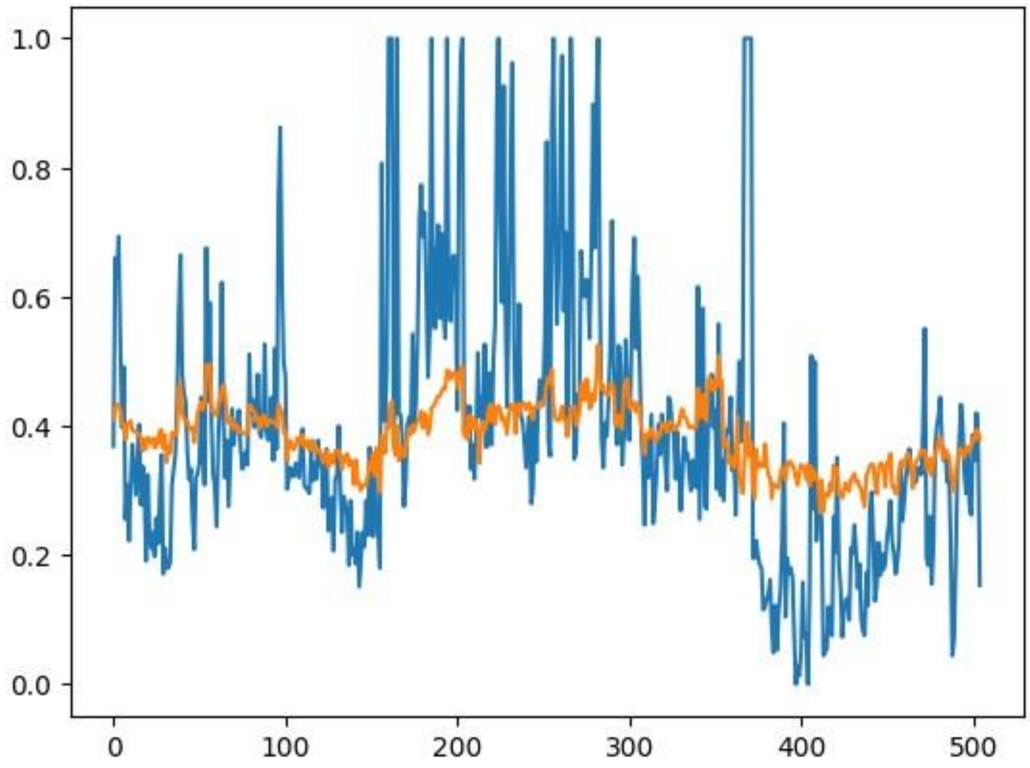

Figure 8

These graphs show the price of house, going ups and down of prices. And shows the result of each algorithm for the testing dataset by the price percentile.

Dataset

The Boston dataset is a collection of data about various attributes for suburbs in Boston, Massachusetts. It is often used as a benchmark for regression analysis. The dataset has **506 rows** and **14 columns**. The columns are:

1. **CRIM**: per capita crime rate by town
2. **ZN**: proportion of residential land zoned for lots over 25,000 sq.ft.
3. **INDUS**: proportion of non-retail business acres per town
4. **CHAS**: Charles River dummy variable (1 if tract bounds river; 0 otherwise)
5. **NOX**: nitric oxides concentration (parts per 10 million)
6. **RM**: average number of rooms per dwelling
7. **AGE**: proportion of owner-occupied units built prior to 1940
8. **DIS**: weighted distances to five Boston employment centres
9. **RAD**: index of accessibility to radial highways
10. **TAX**: full-value property-tax rate per $10,000
11. **PTRATIO**: pupil-teacher ratio by town

12. B: 1000(Bk - 0.63)^2 where Bk is the proportion of blacks by town

13. LSTAT: % lower status of the population

14. MEDV: Median value of owner-occupied homes in $1000's

Data Cleaning and Normalization

The first step in cleaning the dataset provided from Kaggle was to identify variables which will not be useful for training the models. This includes features which are not correlated with price, have too many discrete values to draw inferences from, or have too many missing values. The features that were identified to be dropped from the dataset were: CRIM,ZN,INDUS and TAX etc.

The next step is identifying and removing outliers for the ten remaining features. Keeping in mind the distribution of the data and the negative effect of removing too many values, appropriate minimum and maximum values were set for each feature to remove rows in the dataset which were extreme in any feature category. These were chosen somewhat arbitrarily but with the purpose of removing an appropriate percentage of uncommonly occurring extreme values in the dataset. This increases the performance of the models.

Machine Learning Algorithms

The data, after being cleaned and normalized, is split into training and test data using a randomized 80-20 split. This is to ensure that the data used for testing does not contain any of the data used for training. Thus 20% of the data is reserved for testing purposes (. The training dataset was used to train the four price prediction ML models chosen: Multiple Linear Regression, sequential Regression, Decision tree Regression, and Random Forest Regression. All machine learning algorithms used in this report were imported from the sklearn library. Some models were provided input parameters to implement. The motivations for the choice of input parameters are explained in this section for the models that require them.

Decision tree builds regression or classification models in the form of a tree structure. It breaks down a dataset into smaller and smaller subsets while at the same time an associated decision tree is incrementally developed. The final result is a tree with decision nodes and leaf nodes. A decision node (e.g., Outlook) has two or more branches (e.g., Sunny, Overcast and Rainy), each representing values for the attribute tested. Leaf node (e.g., Hours Played) represents a decision on the numerical target. The topmost decision node in a tree which corresponds to the best predictor called root node. Decision trees can handle both categorical and numerical data.

Inference

Inference involves using the subset of the data that was reserved for testing (20%) to predict the price based on the features. This step was performed after the dataset was cleaned and normalized, and the models were optimized. The dataset was re-split, models were retrained, and inferences retaken a total of five times. This produced five separate inferences with the same parameters to be able to produce an average for the measurements. The inferences produced varies slightly each time as a result of the randomized 80-20 training-testing data split. Each model produced inferences from the same testing subset in every iteration. To judge overfitting, they were also tested on the training subset of the data.

Much better prediction results on the training data is an indication of overfitting.

For the Kaggle subset (house prices sales from 2019), an inference was performed once for each model, on the entire dataset.

Measurements

The measurements taken for this study are described in this section. All of the measurements are taken from the same inference data for each model and using the formulas for the various metrics.

Training and Testing Accuracy Comparison

The performance metrics that were taken for machine learning algorithm is MSE(Mean Squared Error) . These measurements were taken for both the training and testing inferences and averaged across all five iterations of inferences taken to produce a table of metrics.

The MSE is calculated as the mean or average of the squared differences between predicted and expected target values in a dataset.

- MSE = 1 / N * sum for i to N (y_i – yhat_i)^2

Where *y_i* is the i’th expected value in the dataset and *yhat_i* is the i’th predicted value. The difference between these two values is squared, which has the effect of removing the sign, resulting in a positive error value.

The squaring also has the effect of inflating or magnifying large errors. That is, the larger the difference between the predicted and expected values, the larger the resulting squared positive error. This has the effect of “*punishing*” models more for larger errors when MSE is used as a loss function. It also has the effect of “*punishing*” models by inflating the average error score when used as a metric.

We can create a plot to get a feeling for how the change in prediction error impacts the squared error.

The example below gives a small contrived dataset of all 1.0 values and predictions that range from perfect (1.0) to wrong (0.0) by 0.1 increments. The squared error between each prediction and expected value is calculated and plotted to show the quadratic increase in squared error.

Inferred Price Plots

For the first iteration of inferences (both training and testing), scatter plots were created for each algorithm where each data point is the actual price, plotted against the inferred price. A line of best was then calculated and drawn through these points as well as a line showing the actual values, y=x. If the algorithms do not demonstrate any systematic error, the line of best fit should match this line.

CHAPTER 5

CONCLUSION AND FUTURE WORK

5.1 Conclusion

This chapter will summarize the work as a whole by conclusively answering the research questions proposed in section 1.4, as well as giving examples of possible future work.

The first research question was to determine which of the models and parameters gives the best overall accuracy in making price predictions for used houses. The optimal parameters were determined in the process of implementing the models, and thus each model was implemented with the parameters that yielded the best performance by trial and error. The results show that out of the four models tested, Decision Tree Regression provided the highest accuracy in all of the metrics used and highest overall accuracy.

The second research question was to determine which of the models can most accurately assess the depreciation of a house over time. All of the models approximated geometric appreciation, meaning that a constant percentage of value is lost every year independent of the age of the house. Decision Tree Regression had a significantly higher assessed average depreciation at approximately 13.8%, compared to the others with 9.7%. This is closer to the range of 15%-31% assessed by Karl Storchmann in his analysis of international depreciation rates.

The third research question is to determine which model demonstrates the best potential for development of a consumer tool for evaluating used houses or a particular subset of used houses. The results show that Decision Tree Regression performed the best on all performance metrics and for all price percentile subsets of used houses. It was also much better able to approximate the depreciation.

5.2 Future Work

This section will explain some possible future research that can expand upon the knowledge gained through this research.

5.2.1 Applying the Method to Other ML Models

This work compared the performance of ML Regression algorithm. A way to expand this work in the future is to apply the same method for comparing these algorithms to others that are suited to regression problems. Some example algorithms are Light Gradient Boosted Machine (LGBM), Kth Nearest Neighbor Regression (KNN), Decision Tree Regression (DTR), and Artificial Neural Networks (ANN). The problem of price prediction deals with continuous variables which makes it suited to regression algorithms, but by creating discrete intervals for the continuous variables such as price, other algorithms could be applied.

5.2.2 Adding Additional Features Related to the Year

A potential improvement to the predictive power of all ML models, if they are able to take advantage of the information, is to add more correlated features. There are some features which are not related to the attributes of the house , such as the price of house. A house that are good in place that will be worth more. Other such features could include the economic conditions, or changes in the climate.

REFERENCES

1. Annina S, Mahima SD, Ramesh B, "An Overview of Machine Learning and its Applications". *International Journal of Electrical Sciences & Engineering (IJESE)*, 2015. pp. 22-24.
2. Karl Storchmann, "On the depreciation of automobiles: An international comparison". *Department of Economics, Yale University*, 2004. pp. 372-373.
3. Sakshi Gupta, "*Regression vs. Classification in Machine Learning*", October 6, 2021 [Online] Available: https://www.springboard.com/blog/data-science/regression-vs-classification/ [Accessed March 6, 2022].
4. Jim Frost, *"Overfitting Regression Models,"* Statistics by Jim, 2022. [Online], Available:

https://statisticsbyjim.com/regression/overfitting-regression- models/ [Accessed April 18, 2022].

5. Skikit-learn, supervised-learning. [Online] Available: https://scikit-learn.org/stable/supervised_learning.html#supervised-learning [Accessed March 12, 2022]
6. Aarshay Jain, "*A Complete Tutorial on Ridge and Lasso Regression in Python*",
7. January 28, 2016. [Online], Available:
8. https://www.analyticsvidhya.com/blog/2016/01/ridge-lasso-regressionpythoncomplete-tutorial/ [Accessed March 12, 2022]
9. Leo Breiman, "Random Forests". *University of California Berkeley,*2001.
10. c3.ai, Root Mean Squared Error. [Online] Available: https://c3.ai/glossary/datascience/root-mean-square-error-rmse/
11. Ajitesh Kumar, "*Mean Squared Error or R-Squared – Which one to use?".* [Online], Available: https://vitalflux.com/mean-square- error-r-squared-which-one-to-use/ [accessed March 12, 2022]
12. Ahmad Abdulal, Nawar Aghi, "*House Price Prediction*". C.S. Bachelors Thesis, Kristianstad's University, 2020.
13. Nathan Allard, Tobias Hagström, "*Modern Housing Valuation: A Machine Learning Approach".* C.S. Bachelors Thesis, KTH, 2021.
14. Sri Totakura, Harika Kosuru, "*Comparison of Supervised Learning Models for predicting prices of Used Houses*". C.S. Bachelors Thesis, Blekinge Institute of Technology, Karlskrona, 2021.
15. Vlad Krotov, "*Legality and Ethics of Web Scraping*", September 2018

16. RICHARDSON, M., 2009. Determinants of Used House Resale Value. Thesis (BSc). The Colorado College.
17. WU, J. D., HSU, C. C. AND CHEN, H. C., 2009. An expert system of price forecasting for used houses using adaptive neuro-fuzzy inference. Expert Systems with Applications. Vol. 36, Issue 4, pp. 7809-7817.
18. DU, J., XIE, L. AND SCHROEDER S., 2009. Practice Prize Paper - PIN
19. Optimal Distribution of Auction Houses System: Applying Price Forecasting, Elasticity Estimation and Genetic Algorithms to Used-House Distribution. Marketing Science, Vol. 28, Issue 4, pp. 637-644.

20. GONGGI, S., 2011. New model for residual value prediction of used houses based on BP neural network and non-linear curve fit. In: Proceedings of the 3rd IEEE International Conference on Measuring Technology and Mechatronics Automation (ICMTMA), Vol 2. pp. 682-685, IEEE Computer Society, Washington DC, USA.

21. LEXPRESS.MU ONLINE. 2014. Available from: http://www.lexpress.mu/ [Accessed 17 January 2014]

22. LE DEFI MEDIA GROUP. 2014. Available from: http://www.defimedia.info/ Accessed 17 January 2014]

23. GELMAN, A. AND HILL, J., 2006. Data Analysis Using Regression and Multilevel Hierarchical Models. Cambridge University Press, New York, USA.