S.P. Mandali's

Ramnarain Ruia Autonomous College

Matunga, Mumbai-400019

Department of Computer Science & Information Technology

Project Report

on

HOUSE PRICE PREDICTION USING MACHINE LEARNING & DEEP LEARNING WITH COMPARISON

Project Guide

Ms. MEGHA SAWANT

Project By

ISHAAN JAYENDRA BARDE
ROLL NO - 416

s

MSC Information Technology

2023-24

INDEX

ACKNOWLEDGEMENT

I would like to express my sincere gratitude to the HOD, **Ms. Megha Sawant** for giving me the opportunity to work on this topic. It would never be possible for me to take this project to this level without their innovative ideas, their relentless support and encouragement.

The success and outcome of this project required a lot of guidance and assistance from many people and I am extremely privileged to have got this all along with the completion of the project. I would like to appreciate the effort of all my classmates who assisted me whenever I needed some sort of help in various aspects of the project development, to all of you, I say thank you. It helped me increase my knowledge and skills. I will forever be grateful.

Table of Contents

ABSTRACT

Problem Statement

The accurate prediction of housing prices is of paramount importance in various real estate-related applications, such as property investment, mortgage assessment, and urban planning. This study presents a comprehensive exploration of machine learning techniques for house price prediction. Leveraging a diverse dataset comprising features like property size, location, amenities, historical pricing trends, and economic indicators, we investigate the effectiveness of different machine learning algorithms and feature engineering strategies. This study includes a wide range of methods, including deep learning, LSTM, and linear regression. To measure predictive accuracy and model robustness, we examine their performance using measures like mean absolute error (MAE), mean squared error (MSE), and R-squared (R2).

INTRODUCTION

The housing market is a cornerstone of the global economy, influencing not only the financial well-being of individuals and families but also impacting broader economic indicators. Accurate prediction of house prices is a crucial task with wide-ranging implications, from assisting potential homebuyers in making informed decisions to aiding real estate professionals, investors, and policymakers in assessing market trends and risks. In recent years, the integration of machine learning techniques into the realm of real estate has revolutionized the way we approach house price prediction. The House Price Prediction Project represents an ambitious endeavour to harness the power of machine learning in order to provide reliable and data-driven insights into the dynamic world of real estate. This project seeks to develop predictive models capable of estimating housing prices with a high degree of precision, taking into account a multitude of factors, from property characteristics and location to economic indicators and historical trends. The significance of this project lies in its potential to address critical challenges faced by various stakeholders in the housing market. For prospective homebuyers, it offers the opportunity to make well-informed decisions about their investments. Real estate professionals can benefit from more accurate pricing strategies, while investors can better gauge market opportunities and risks. Policymakers can use these models to monitor and respond to market fluctuations, thereby promoting stability and affordability in the housing sector. In this project, we will evaluate the house prices and predict them using machine learning and deep learning. We will explore and implement various machine learning algorithms, assess the impact of different features on prediction accuracy, and fine-tune our models to achieve the best results. By the end of this project, we aim to equip ourselves with a robust predictive tool that can contribute meaningfully to the understanding and management of housing markets.

LITERATURE REVIEW

Literature Review: House Price Prediction

The prediction of house prices has been a significant area of research and practical application within the realms of economics, data science, and real estate. Over the years, numerous studies have explored various methodologies, datasets, and factors influencing housing markets to enhance the accuracy and robustness of price prediction models. This literature review provides a snapshot of key findings and trends within the field of house price prediction.

1. Regression and Machine Learning Models:

Many early approaches focused on regression techniques, such as linear regression, to predict house prices based on a set of features. However, as the complexity of data increased, researchers turned to more sophisticated machine learning algorithms, including decision trees, random forests, support vector machines, and gradient boosting. These models demonstrated improved predictive power by capturing nonlinear relationships and interactions among features.

2. Feature Selection and Engineering:

Feature selection and engineering play a pivotal role in model performance. Researchers have explored the impact of various features, such as property size, location, number of bedrooms, and economic indicators like interest rates and employment rates. Additionally, advanced feature engineering techniques, like creating interaction terms or embedding geographical information, have shown potential in enhancing prediction accuracy.

3. Geographic Factors:

Geographical factors have consistently emerged as crucial predictors in house price models. Studies often incorporate variables like proximity to schools, public transportation, amenities, and crime rates. Geographic information systems (GIS) and spatial analysis techniques have been employed to capture spatial autocorrelation and regional trends, enhancing the models' spatial predictive capabilities.

4. Time Series Analysis:

Recognizing the temporal dynamics of housing markets, time series analysis has gained traction. Autoregressive integrated moving average (ARIMA) models and more advanced approaches like Long Short-Term Memory (LSTM) networks have been applied to capture the temporal dependencies and cyclical patterns inherent in real estate data.

5. Big Data and Deep Learning:

With the proliferation of big data, deep learning techniques have been introduced to house price prediction. Convolutional neural networks (CNNs) have been used to analyze property images and extract visual features, while recurrent neural networks (RNNs) have demonstrated capabilities in handling sequential data, such as time series housing market data.

6. Model Evaluation and Interpretability:

Model evaluation metrics like Mean Absolute Error (MAE), Root Mean Squared Error (RMSE), and R-squared have been commonly employed to assess model performance. Furthermore, the interpretability of models has gained attention, especially in regulatory contexts. Efforts to explain the model's decision-making process have led to the

development of techniques like feature importance analysis and SHAP (SHapley Additive exPlanations).

7. Ethical Considerations:

As algorithms play an increasingly influential role in real estate, ethical concerns have surfaced. Biases in data and models have the potential to perpetuate discriminatory practices. Researchers are investigating ways to mitigate bias and ensure fairness in house price prediction models.

8. Online Platforms and Real-World Applications:

Several online platforms and real estate companies have integrated predictive models to provide house price estimates to users. These applications often combine machine learning with user-generated data to offer personalized predictions, enhancing user engagement and decision-making.

RESEARCH METHODOLOGY

The research methodology for a house price prediction project involves a systematic approach to collecting, pre-processing, analysing, and modelling data to develop accurate prediction models. A general outline of the research methodology is used

- **Problem Definition and Data Collection:**

Clearly define the scope of your project, including the geographical area, type of properties, and target variables (e.g., sale price). Collect relevant data from reputable sources, including property listings, real estate databases, government sources, and economic indicators. This data should encompass property attributes, location information, economic factors, and any other relevant features.

- **Data Pre-Processing:**

Clean the dataset by handling missing values, duplicates, and outliers. Impute missing values using appropriate methods or consider removing records with substantial missing information. Convert categorical variables into dummy variables by pandas library to expand the understanding of the data better. Normalize, standardize or scale numerical features to ensure they have similar ranges and distributions and fit the algorithm.

- **Feature Engineering:**

Create new features that could potentially enhance predictive power, such as feature interactions, ratios, or aggregations. Incorporate

geographic features like distance to key locations (schools, hospitals, transportation) using geospatial calculations.

- **Exploratory Data Analysis (EDA):**

Perform data visualization to understand the relationships between variables, identify trends, and spot outliers. Analyse correlations between features and the target variable to identify potential predictors.

- **Feature Scaling:**

  Feature scaling is the process of normalizing the range of features in a dataset. Real-world datasets often contain features that are varying in degrees of magnitude, range, and units. Therefore, in order for machine learning models to interpret these features on the same scale, we need to perform feature scaling. The scaling methods used are:

  1. Standardization
  2. MinMax

1. **Standardization**

Standardization refers to the process of establishing and maintaining a set of consistent, commonly agreed-upon practices, rules, or guidelines within a specific industry, organization, or field of activity. The primary purpose of standardization is to ensure uniformity, safety, quality, and interoperability of products, services, and processes.The standard score of a sample x is calculated as:

$$z = (x - u) / s,$$

Where:

- X = Observation
- U = Mean
- S = Standard Deviation

2. **Min Max**

It appears that you are referring to the Min-Max scaling or normalization technique. Min-Max scaling is a data preprocessing method commonly used in statistics, data analysis, and machine learning to transform a feature's values within a specific range, typically between 0 and 1.

The Min-Max scaling formula is as follows:

X norm=X – X min/ X max – X min,

Where:

- X norm is the normalized value of the feature.
- X is the original value of the feature.
- X min is the minimum value of the feature in the dataset.
- X max is the maximum value of the feature in the dataset.

- **Model Selection:**

Choose a variety of predictive models suitable for regression tasks. Common choices include linear regression, decision trees, random forests, gradient boosting, support vector machines, and neural networks. Consider ensemble techniques to combine the strengths of multiple models.

- **Model Training and Evaluation:**

Split the dataset into training, validation, and test sets. Training data is used to train the model, validation data helps tune hyperparameters, and the test data evaluates final model performance. The data is split in 80-20 for training and testing respectfully. Linear Regression models of machine learning and deep learning are built. The model performance is

evaluated using appropriate metrics such as Mean Absolute Error (MAE), Root Mean Squared Error (RMSE), R-squared and VarScore.

- **Model Interpretation:**

Analyse feature importance to understand which variables contribute most to the predictions. Use techniques for values to provide insights into how individual features affect predictions. The performance metrics helps to evaluate how good the model works.

- **Model Comparison:**

Compare the machine learning algorithm used and the deep learning algorithm. Check which is best suited for the problem and has higher accuracy in terms of prediction. Also, name the best suited model for this type of project.

- **Continuous Improvement:**

Monitor model performance over time and retrain the model periodically to account for changing market dynamics and new data. Add new features like user input and predictions based on that to evaluate the sale price of a particular home.

RESEARCH OJECTIVES

- Exploratory Data Analysis
- Predict the house price
- Using different models in terms of minimizing the difference between predicted and actual rating
- Implementing Deep Learning Model
- Comparing the models and gaining detailed study of which is best suited model and method for the same

EXPERIMENTAL SET-UP

Setting up experiments for house price prediction involves defining the procedures for

- Data Pre-Processing
- Model Training
- Evaluation
- Comparison

DATASET: RAW HOUSING DATASET

The dataset has historical information of houses which are sold. It has information of several features such as Sale Price, Waterfront View, No. of Bedrooms/ Bathrooms, etc.

DATA EXPLORATION & PRE-PROCESSING

I have used

- EDA
- Data Cleaning
- Visualization

- Data Exploration & Transformation

EXPLORATORY DATA ANALYSIS (EDA):

EDA or **EXPLORATORY DATA ANALYSIS** is an approach that is used to analyse the data and discover trends, patterns, or check assumptions in data with the help of statistical summaries and graphical representations. In the project I have used all the types including Univariate, Bi-variate and Multi-Variate.

Univariate Analysis – In univariate analysis, we analyse or deal with only one variable at a time.

Bi-Variate Analysis – This type of data involves analysis two different variables.

Multivariate Analysis – The data involves three or more variables, it is categorized under multivariate.

DATA CLEANING:

Data cleaning is the process of fixing or removing incorrect, corrupted, incorrectly formatted, duplicate, or incomplete data within a dataset. When combining multiple data sources, there are many opportunities for data to be duplicated or mislabelled.

VISUALIZATION:

Data visualization is the graphical representation of information and data. By using visual elements like charts, graphs, and maps, data visualization tools provide an accessible way to see and understand trends, outliers, and patterns in data. Additionally, it provides an excellent way for employees or business owners to present data to non-technical audiences without confusion. In the world of Big Data, data visualization tools and technologies are essential to analyse massive amounts of information and make data-driven decisions.

POWER BI REPORT:

Power BI is a powerful tool used by data scientist to analyse raw data and extract valuable insights. The report consists of valuable insights of the data, it adds different and unique figures to display the data. Using the interactivity on the app the user can play with all the features in the dataset and get in depth explaination.

DATA EXPLORATION AND TRANSFORMATION:

Data exploration is an approach similar to initial data analysis, whereby a data analyst uses visual exploration to understand what is in a dataset and the characteristics of the data, rather than through traditional data management systems.

Data transformation is the process of converting data from one format to another, typically from the format of a source system into the required format of a destination system. Data transformation is a component of most data integration and data management tasks, such as data wrangling and data warehousing.

MODEL TRAINING

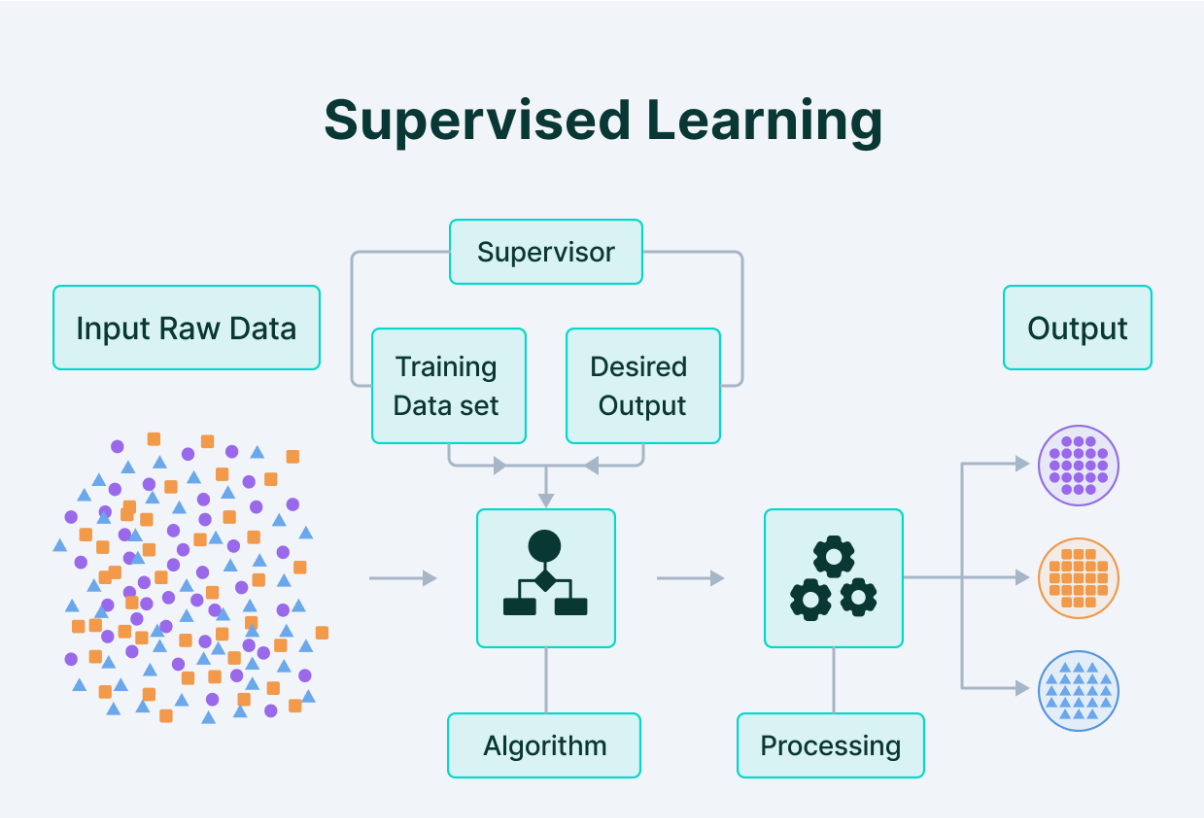
Model training in machine language is the process of feeding an ML algorithm with data to help identify and learn good values for all attributes involved. I have used

- Supervised Learning
- Deep Learning

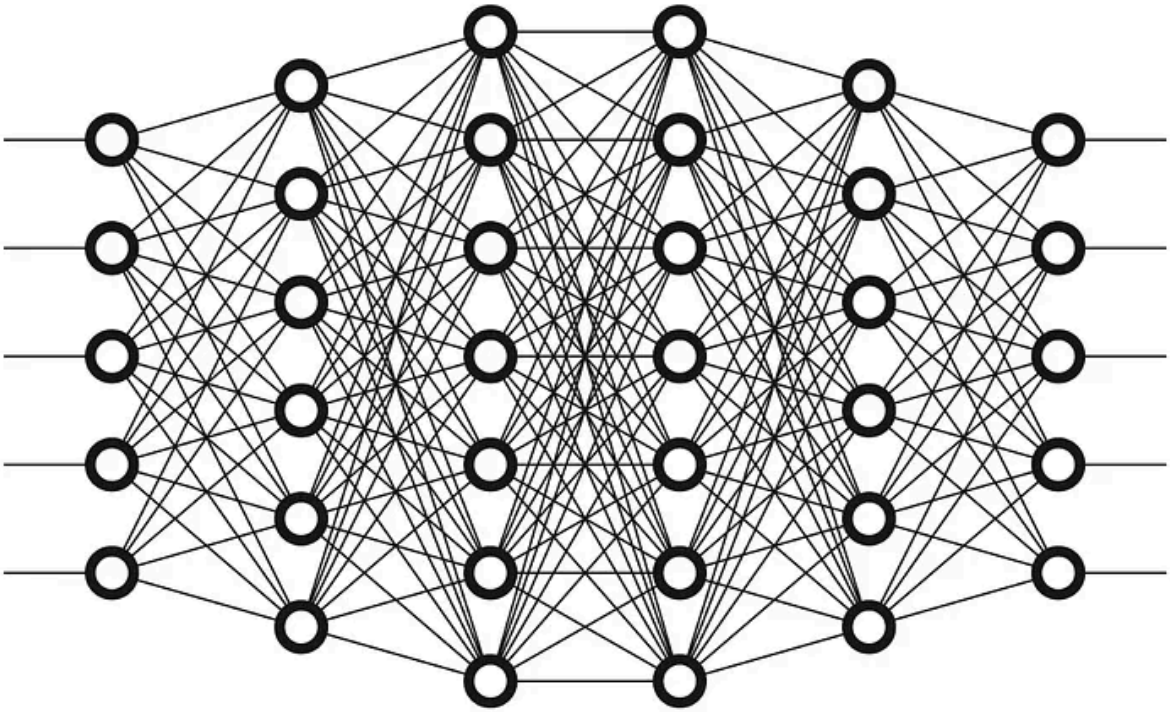
SUPERVISED LEARNING also known as supervised machine learning, is a subcategory of machine learning and artificial intelligence. It is defined by its use of labelled datasets to train algorithms that to classify data or predict outcomes accurately. As input data is fed into the model, it adjusts its weights until the model has been fitted appropriately, which occurs as part of the cross validation process. Supervised learning helps organizations solve for a variety of real-world problems at scale, such as classifying spam in a separate folder from your inbox. Supervised learning uses a training set to teach models to yield the desired output. This training dataset includes inputs and correct outputs, which allow the model to learn over time. The algorithm measures its accuracy through the loss function, adjusting until the error has been sufficiently minimized. It can be separated into two types of problems when data mining—classification and regression:

- Classification uses an algorithm to accurately assign test data into specific categories. It recognizes specific entities within the dataset and attempts to draw some conclusions on how those entities should be labelled or defined. Common classification algorithms are linear classifiers, support vector machines (SVM), decision trees, k-nearest Neighbor, and random forest, which are described in more detail below.
- Regression is used to understand the relationship between dependent and independent variables. It is commonly used to make projections, such as for sales revenue for a given

business. Linear regression, logistical regression, and polynomial regression are popular regression algorithms.



DEEP LEARNING subset of machine learning, which is essentially a neural network with three or more layers. These neural networks attempt to simulate the behaviour of the human brain—albeit far from matching its ability allowing it to “learn” from large amounts of data. While a neural network with a single layer can still make approximate predictions, additional hidden layers can help to optimize and refine for accuracy.



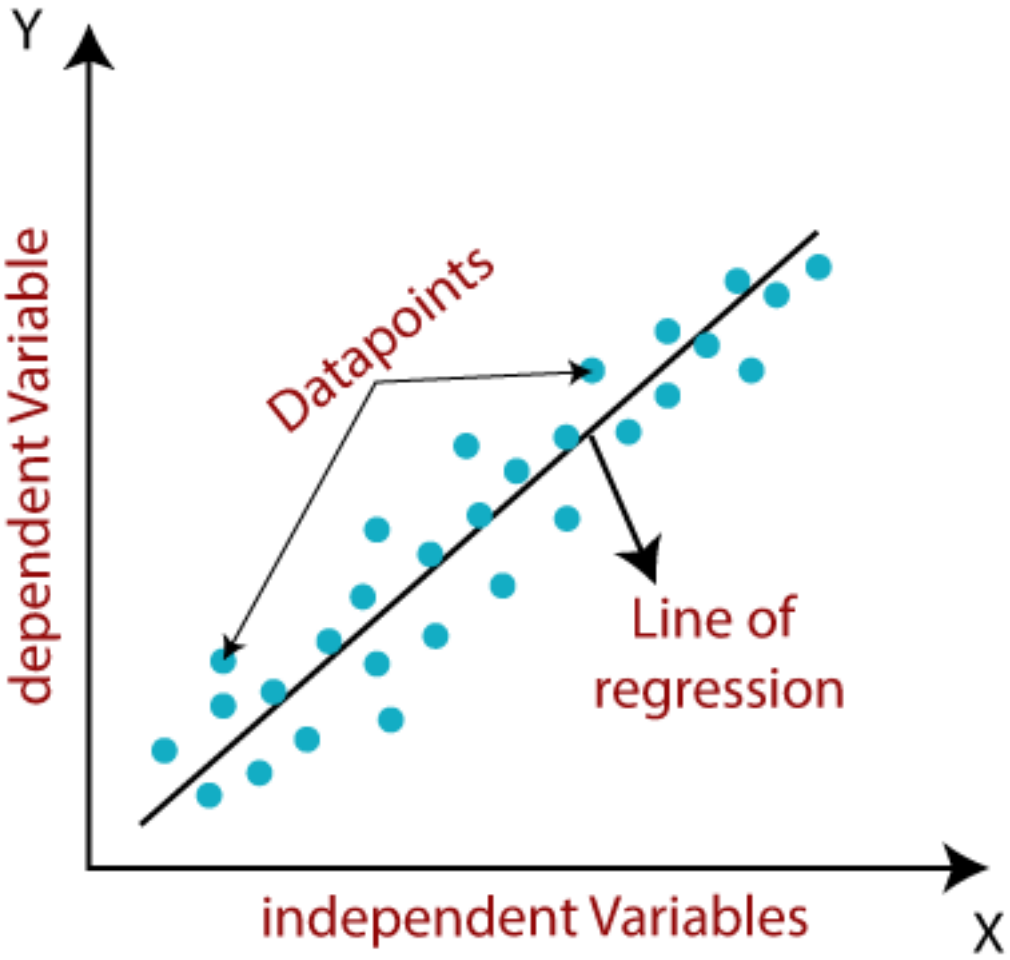
It drives many artificial intelligence (AI) applications and services that improve automation, performing analytical and physical tasks without human intervention. Deep learning technology lies behind everyday products and services (such as digital assistants, voice-enabled TV remotes, and credit card fraud detection) as well as emerging technologies (such as self-driving cars).

MODELS USED:

- Linear Regression
- LSTM

LINEAR REGRESSION

Linear regression analysis is used to predict the value of a variable based on the value of another variable.



The variable you want to predict is called the dependent variable. The variable you are using to predict the other variable's value is called the independent variable. There are two main types of linear regression:

1. Simple Linear Regression
2. Multiple Linear Regression

Simple Linear Regression

In simple linear regression, there is one independent variable (X) and one dependent variable (Y). The relationship between X and Y is modelled using a straight line equation: $Y = \beta0 + \beta1 * X + \varepsilon$.

$\beta0$ is the intercept, $\beta1$ is the slope, and ε represents the error term (residuals). The goal is to find the best-fitting line that minimizes the sum of squared residuals (least squares method).

Artificial Intelligence

Artificial intelligence is the simulation of human intelligence processes by machines, especially computer systems. Specific applications of AI include expert systems, natural language processing, speech recognition and machine vision.

As the hype around AI has accelerated, vendors have been scrambling to promote how their products and services use AI. Often what they refer to as AI is simply one component of AI, such as machine learning. AI requires a foundation of specialized hardware and software for writing and training machine learning algorithms. No one programming language is synonymous with AI, but a few, including Python, R and Java, are popular.

In general, AI systems work by ingesting large amounts of labeled training data, analyzing the data for correlations and patterns, and using these patterns to make predictions about future states. In this way, a chatbot that is fed examples of text chats can learn to produce lifelike exchanges with people, or an image recognition tool can learn to identify and describe objects in images by reviewing millions of examples.

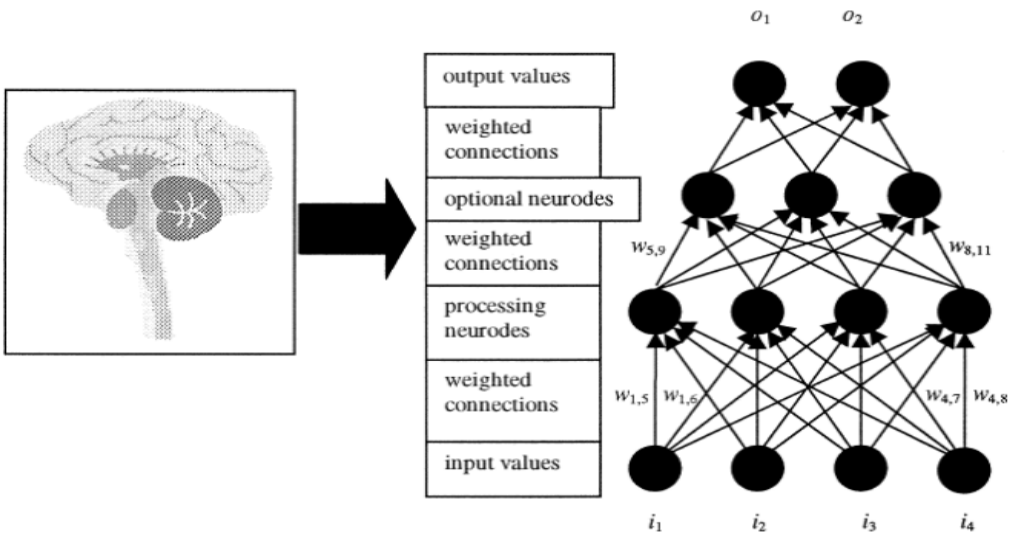
AI programming focuses on three cognitive skills: learning, reasoning and self-correction.

Learning processes. This aspect of AI programming focuses on acquiring data and creating rules for how to turn the data into actionable information. The rules, which are called algorithms, provide computing

devices with step-by-step instructions for how to complete a specific task.

Artificial Neural Network

Artificial neural networks (ANNs), usually simply called neural networks (NNs), are computing systems inspired by the biological neural networks that constitute animal brains.



An ANN is based on a collection of connected units or nodes called artificial neurons, which loosely model the neurons in a biological brain. Each connection, like the synapses in a biological brain, can transmit a signal to other neurons. An artificial neuron receives a signal then processes it and can signal neurons connected to it. The "signal" at a connection is a real number, and the output of each neuron is computed by some non-linear function of the sum of its inputs.

The connections are called edges. Neurons and edges typically have a weight that adjusts as learning proceeds. The weight increases or decreases the strength of the signal at a connection. Neurons may have

a threshold such that a signal is sent only if the aggregate signal crosses that threshold. Typically, neurons are aggregated into layers. Different layers may perform different transformations on their inputs. Signals travel from the first layer (the input layer), to the last layer (the output layer), possibly after traversing the layers multiple times.

All of the weight-adjusted input values to a processing element are then aggregated using a vector to scalar function such as summation (i.e., $y = \Sigma w_{ij} x_i$), averaging, input maximum, or mode value to produce a single input value to the neurode. Once the input value is calculated, the processing element then uses a transfer function to produce its output (and consequently the input signals for the next processing layer). The transfer function transforms the neurode's input value.

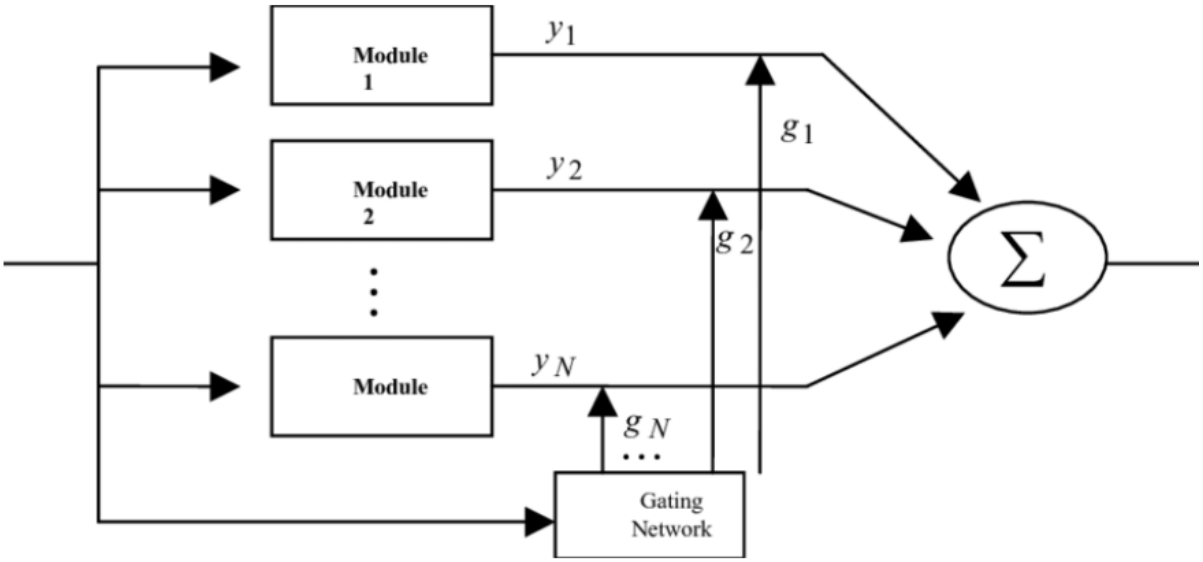
Typically this transformation involves the use of a sigmoid, hyperbolic-tangent, or other nonlinear function. The process is repeated between layers of processing elements until a final output value, o_n, or vector of values is produced by the neural network.

Theoretically, to simulate the asynchronous activity of the human nervous system, the processing elements of the artificial neural network should also be activated with the weighted input signal in an asynchronous manner. Most software and hardware implementations of artificial neural networks, however, implement a more discretized approach that guarantees that each processing element is activated once for each presentation of a vector of input values.

There are several Artificial neural network models they are as follows:

Modular Neural Networks

In this type of neural network, many independent networks contribute to the results collectively. There are many sub-tasks performed and constructed by each of these neural networks. This provides a set of inputs that are unique when compared with other neural networks. There is no signal exchange or interaction between these neural networks to accomplish any task.



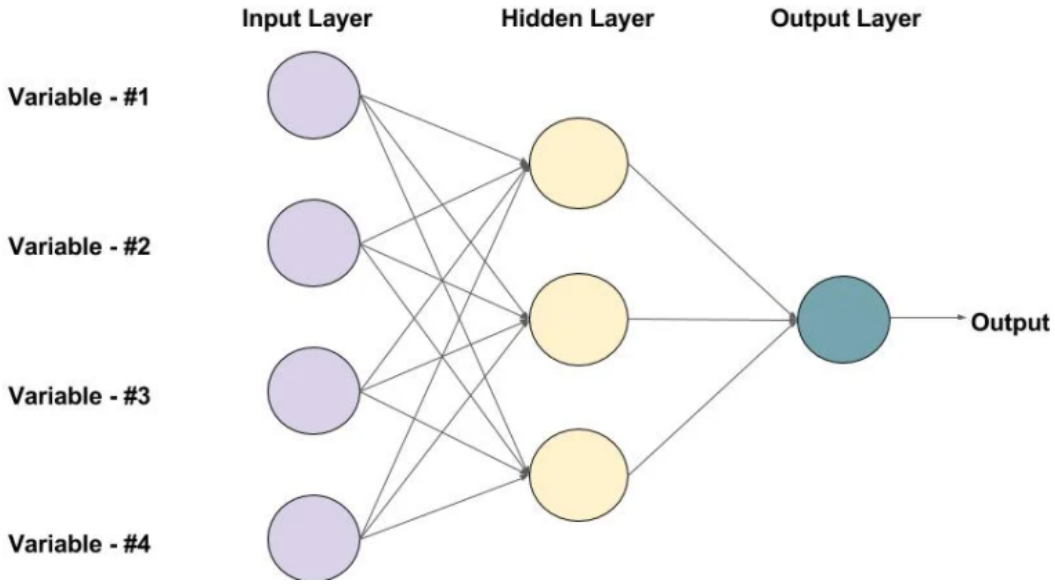
Architecture of a modular neural network.

The complexity of a problem is easily reduced while solving problems by these modular networks because they completely break down the sizeable computational process into small components. The computation speed also gets improved when the number of connections is broken down and reduces the need for interaction of the neural networks with each other.

The total time of processing will also depend on the involvement of neurons in the computation of results and how many neurons are involved in the process. Modular Neural Networks (MNNs) is one of the fastest-growing areas of Artificial Intelligence.

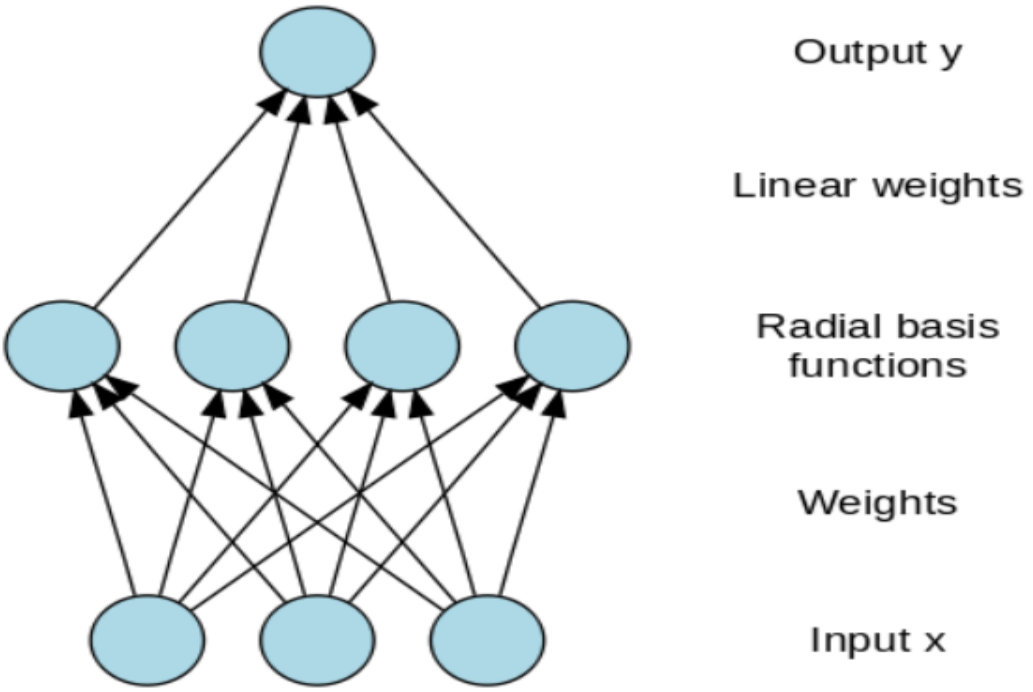
Feedforward Neural Network – Artificial Neuron

The information in the neural network travels in one direction and is the purest form of an Artificial Neural Network. This kind of neural network can have hidden layers and data enter through input nodes and exit through output nodes. Classifying activation function is used in this neural network. There is no backpropagation, and only the front propagated wave is allowed.



There are many applications of Feedforward neural networks, such as speech recognition and computer vision. It is easier to maintain these types of Neural Networks and also has excellent responsiveness to noisy data.

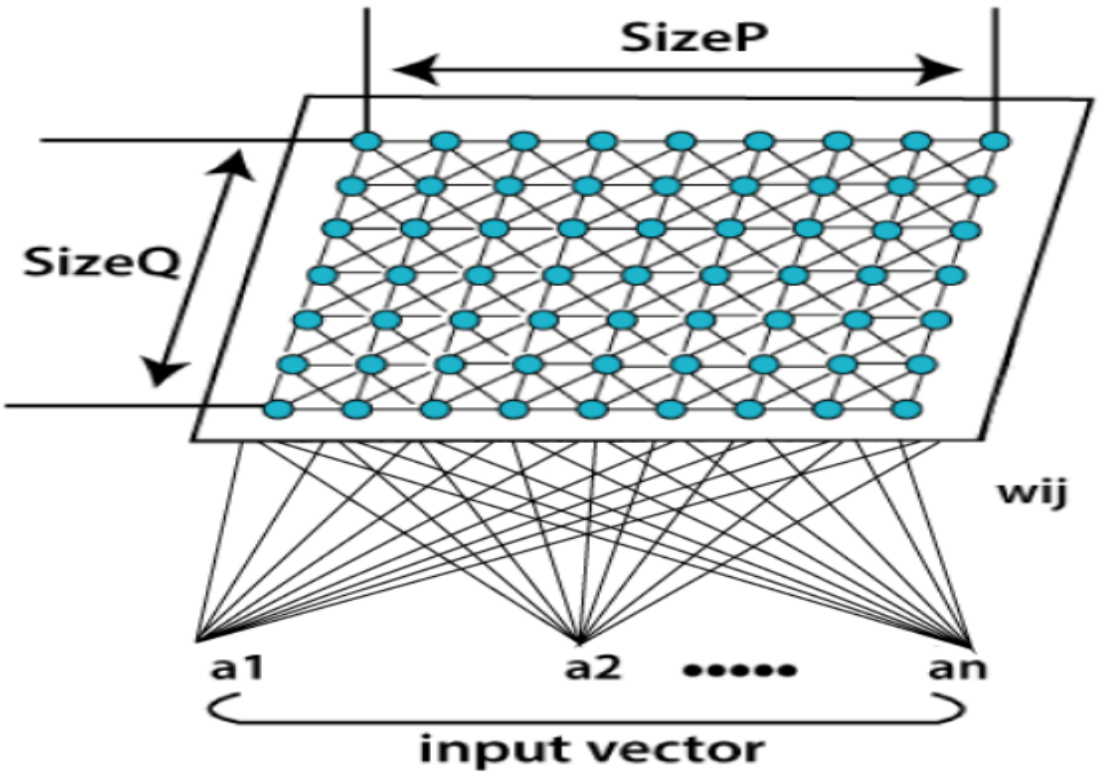
Radial basis function Neural Network



There are two layers in the functions of RBF. These are used to consider the distance of a centre with respect to the point. In the first layer, features in the inner layer are united with the Radial Basis Function. In the next step, the output from this layer is considered for computing the same output in the next iteration. One of the applications of Radial Basis function can be seen in Power Restoration Systems. There is a need to restore the power as reliably and quickly as possible after a blackout.

Kohonen Self Organizing Neural Network

In this neural network, vectors are input to a discrete map from an arbitrary dimension. Training data of an organization is created by training the map. There might be one or two dimensions on the map. The weight of the neurons may change that depends on the value.



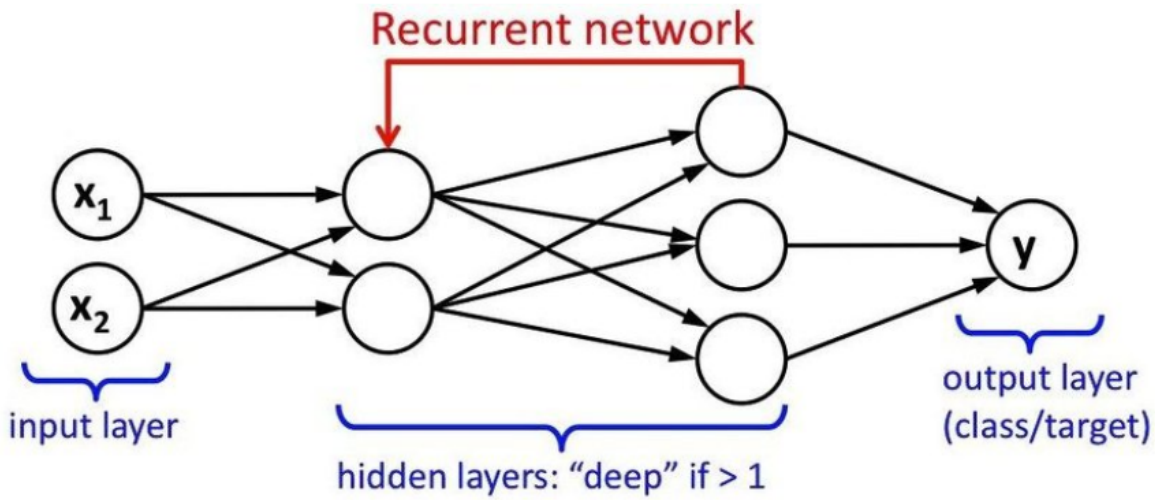
The neuron's location will not change while training the map and will stay constant. Input vector and small weight are given to every neuron value in the first phase of the self-organization process. A winning neuron is a neuron that is closest to the point. Other neurons will also start to move towards the point along with the winning neuron in the second phase.

The winning neuron will have the least distance, and Euclidean distance is used to calculate the distance between neurons and the point. Each neuron represents each kind of cluster, and the clustering of all the points will happen through the iterations.

One of the main applications Kohonen Neural Network is to recognize the data patterns. It is also used in the medical analysis to classify diseases with higher accuracy. Data are clustered into different categories after analysing the trends in the data.

Recurrent Neural Network(RNN)

The principle of Recurrent Neural Network is to feedback the output of a layer back to the input again. This principle helps to predict the outcome of the layer. In the Computation process, Each neuron will act as a memory cell. The neuron will retain some information as it goes to the next time step.



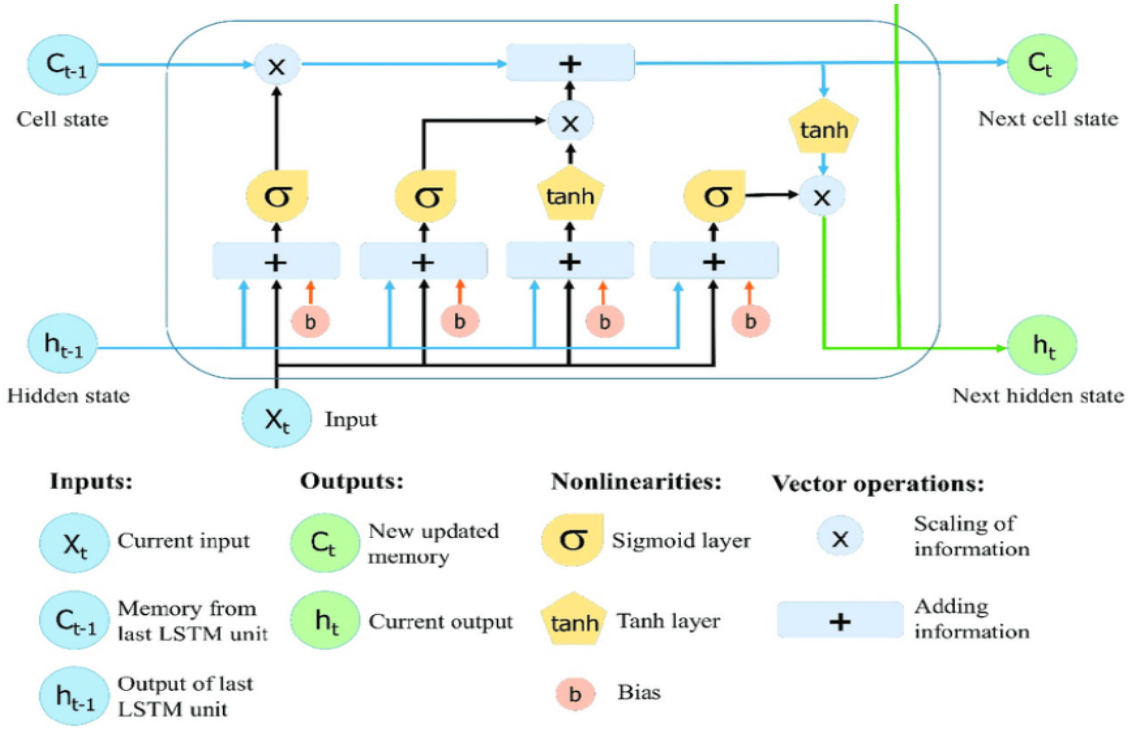
It is called a recurrent neural network process. The data to be used later will be remembered and work for the next step will go on in the process. The prediction will improve by error correction. In error correction, some changes are made to create the right prediction output. The learning rate is the rate of how fast the network can make the correct prediction from the wrong prediction.

There is much application of Recurrent Neural Networks, and one of them is the model of converting text to speech. The recurrent neural network was designed for supervised learning without any requirement of teaching signal.

Long / Short Term Memory

Schmidhuber and Hochreiter in 1997 built a neural network which is called long short term memory networks (LSTMs). Its main goal is to

remember things for a long time in a memory cell that is explicitly defined. Previous values are stored in the memory cell unless told to forget the values by “forget gate”.



New stuff is added through the “input gate” to the memory cell, and it is passed to the next hidden state from the cell along the vectors which is decided by the “output gate”. Composition of primitive music, writing like Shakespeare, or learning complex sequences are some of the applications of LSTMs.

MODEL EVALUATION

Model evaluation is the process of using different evaluation metrics to understand a machine learning model's performance, as well as its

strengths and weaknesses. Model evaluation is important to assess the efficacy of a model during initial research phases, and it also plays a role in model monitoring. The evaluations metrics used are:

i. MSE
ii. MAE
iii. RMSE
iv. VarScore

- MSE
  MSE stands for "Mean Squared Error." It is a common metric used to measure the average squared difference between the predicted and actual values in a regression problem. MSE is a measure of the model's accuracy and is particularly useful in evaluating the performance of regression models, where the goal is to predict a continuous numeric value. Here's how it's calculated

- MAE
  MAE stands for "Mean Absolute Error." It is a common metric used to evaluate the performance of a machine learning model, particularly in regression problems. MAE measures the average absolute difference between the predicted values and the actual values in the dataset. It provides a straightforward and interpretable way to understand how well a model is performing.

- RMSE
  RMSE stands for "Root Mean Squared Error." It is a commonly used metric for evaluating the performance of a machine learning model, especially in regression problems. RMSE measures the square root of the average of the squared differences between the predicted values and the actual values in a dataset. This metric combines the advantages of the Mean Squared Error (MSE) with a square root operation to provide a measure of error that is in the same units as the target variable. The RMSE is a non-negative value, and like the MSE, smaller values indicate a better model fit

to the data. The primary advantage of RMSE over MSE is that it is in the same units as the target variable, making it more interpretable.

COMPARISON STUDY

Comparative studies are investigations to analyse and evaluate, with quantitative and qualitative methods, a phenomenon and/or facts among different areas, subjects, and/or objects to detect similarities and/or differences. In this I have compared the supervised machine learning algorithms with deep learning algorithms. The machine learning model is developed using the sklearn library and the deep learning model with the popular Keras library.

TOOLS

The libraries used for this project are as follows:

- Pandas
- Matplotlib
- NumPy
- Sklearn
- Keras
- Google Colab
- Python
- Neural Network

Pandas

Pandas is a widely used open-source data manipulation and analysis library for Python. It provides easy-to-use data structures and functions for working with structured data, making it a popular choice for data scientists and analysts.

Matplotlib

Matplotlib is a popular and widely-used data visualization library in Python. It provides a flexible and comprehensive framework for creating high-quality static, animated, or interactive plots and charts. Matplotlib is often used in data analysis, scientific research, and data visualization tasks.

NumPy

NumPy, which stands for "Numerical Python," is a fundamental and widely-used open-source library in Python for numerical and scientific computing. It provides support for working with large, multi-dimensional arrays and matrices, as well as a collection of mathematical functions to operate on these arrays efficiently.

Sklearn Library:

Scikit-Learn, often abbreviated as sklearn, is a widely used and open-source machine learning library in Python. It provides a robust set of tools for various machine learning tasks, including classification, regression, clustering, dimensionality reduction, model selection, and more. Scikit-Learn is built on top of other popular Python libraries such as NumPy, SciPy, and Matplotlib and offers a consistent and user-friendly interface for working with machine learning models and data.

Keras Library:

Keras is an open-source deep learning framework that is designed to be user-friendly, modular, and highly customizable. It was developed as part of the TensorFlow project but later became an independent high-level neural networks API. Keras makes it easier for researchers and developers to build and experiment with deep learning models.

Google Collab Software

Colab is a free Jupyter notebook environment that runs entirely in the cloud. Most importantly, it does not require a setup and the notebooks that you create can be simultaneously edited by your team members - just the way you edit documents in Google Docs. Colab allows anybody to write and execute arbitrary python code through the browser, and is especially well suited to machine learning, data analysis and education.

Python

Python is an interpreted high-level general-purpose programming language. Its design philosophy emphasizes code readability with its use of significant indentation. Its language constructs as well as its object-oriented approach aim to help programmers write clear, logical code for small and large-scale projects.

Python is dynamically-typed and garbage-collected. It supports multiple programming paradigms, including structured (particularly, procedural), object-oriented and functional programming. It is often described as a "batteries included" language due to its comprehensive standard library.

Guido van Rossum began working on Python in the late 1980s, as a successor to the ABC programming language, and first released it in 1991 as Python 0.9.0. Python 2.0 was released in 2000 and introduced new features, such as list comprehensions and a garbage collection system using reference counting. Python 3.0 was released in 2008 and was a major revision of the language that is not completely backward-compatible. Python 2 was discontinued with version 2.7.18

in 2020.

Neural network

Neural networks, also known as artificial neural networks (ANNs) or simulated neural networks (SNNs), are a subset of machine learning and are at the heart of deep learning algorithms. Their name and structure are inspired by the human brain, mimicking the way that biological neurons signal to one another.

Artificial neural networks (ANNs) are comprised of a node layers, containing an input layer, one or more hidden layers, and an output layer. Each node, or artificial neuron, connects to another and has an associated weight and threshold. If the output of any individual node is above the specified threshold value, that node is activated, sending data to the next layer of the network. Otherwise, no data is passed along to the next layer of the network.

IMPLEMENTATION

DATASET

https://www.kaggle.com/datasets/harlfoxem/housesalesprediction

Data Exploration and Pre-Processing

LINK:
https://colab.research.google.com/drive/1lOl9KdVyJhDPl1XSRI2umFSAUqf5K_tD?usp=drive_link

House Price Prediction (Linear Regression)

LINK:
https://colab.research.google.com/drive/1chH1mAVqrDDU1wCaEJEifVJDorK_cU-n?usp=drive_link

House Price Prediction Deep Learning Linear Regression Model

LINK:
https://colab.research.google.com/drive/1P38tDsQKIMhtUzIirG5QM5yxhwNuO9Pk?usp=drive_link

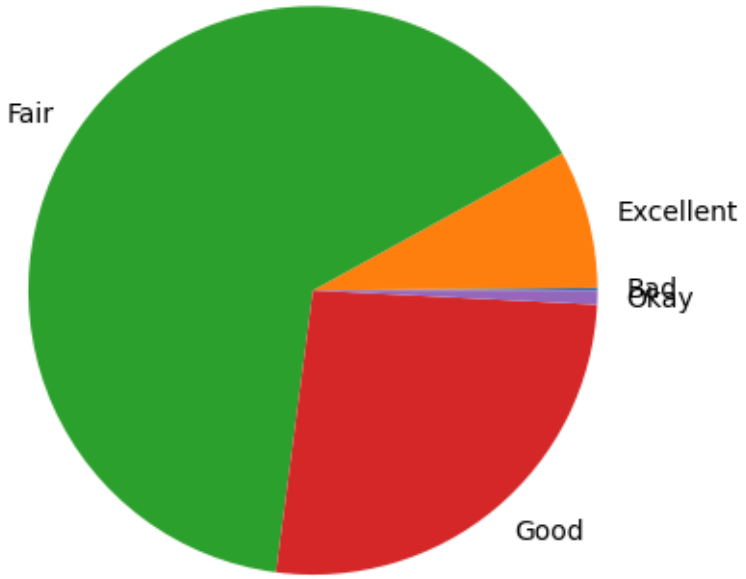
House Price Prediction Deep Learning With LSTM

LINK:
https://colab.research.google.com/drive/1wcbtghcwZ8ZnQs0IgShXT_VC5TJzTXTU?usp=drive_link

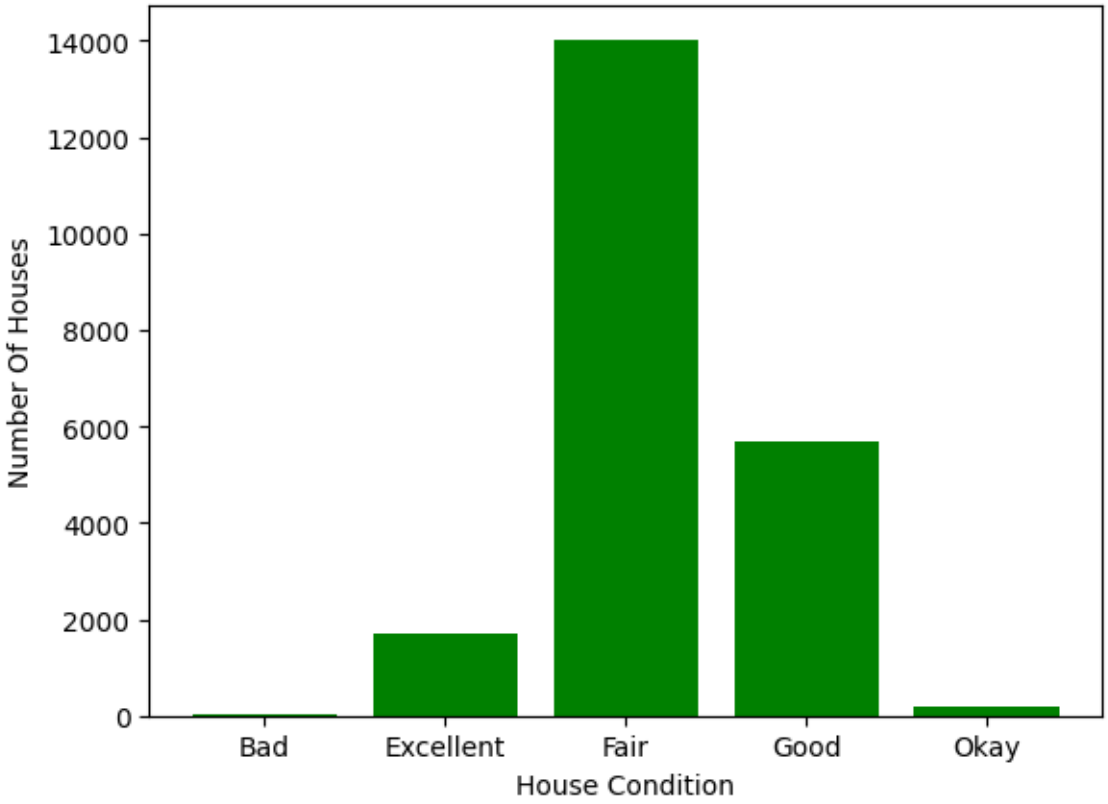
RESULT

Data Visualization Using Python

i. Condition Of The House

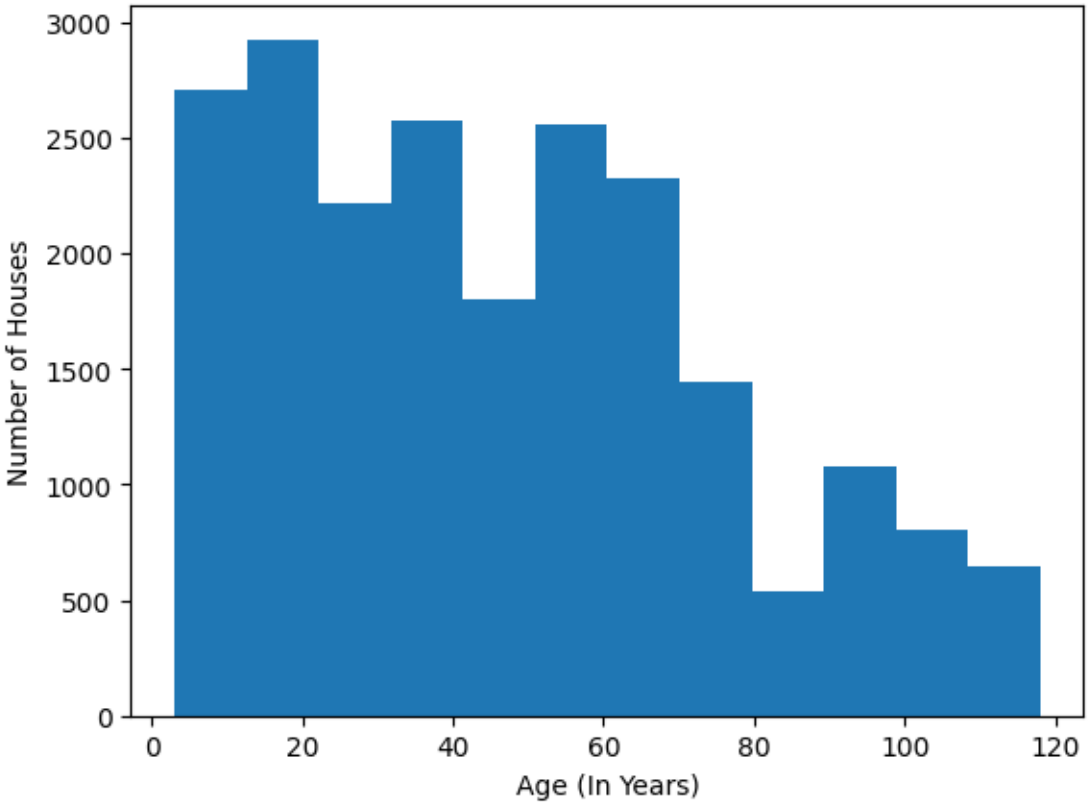


ii. Number Of House By House Condition

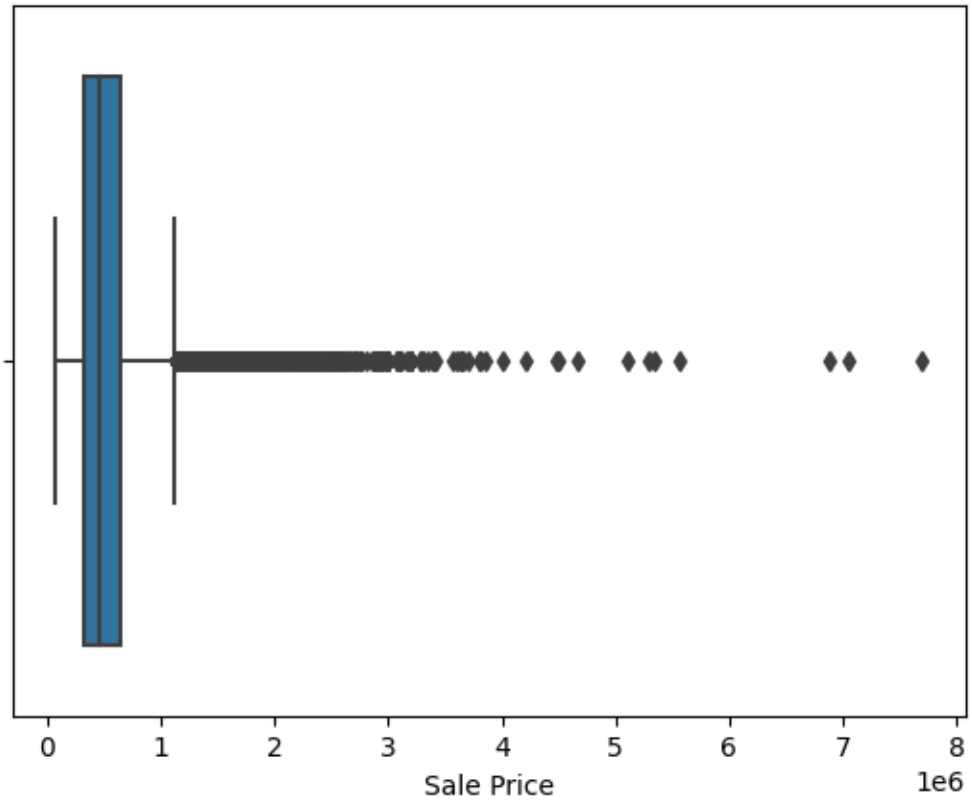


iii. Selling Price Vs Area (Scatter Plot)

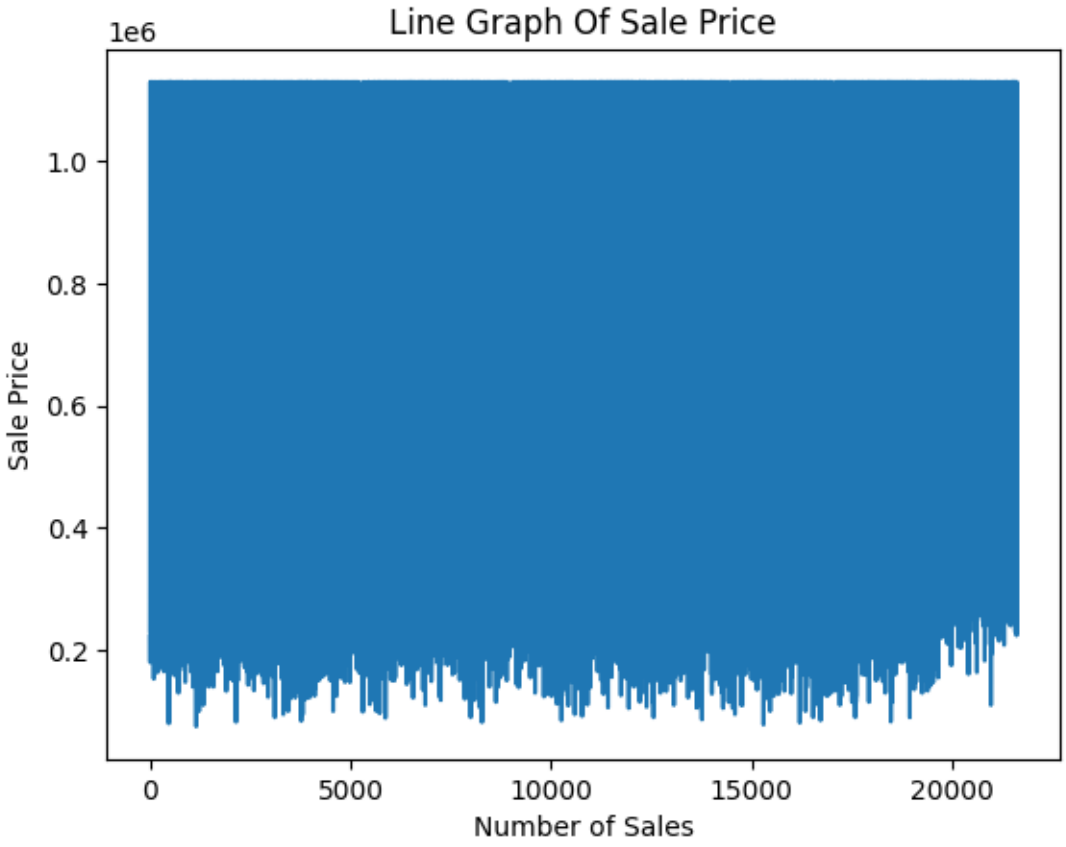
iv. Number Of Houses By Age



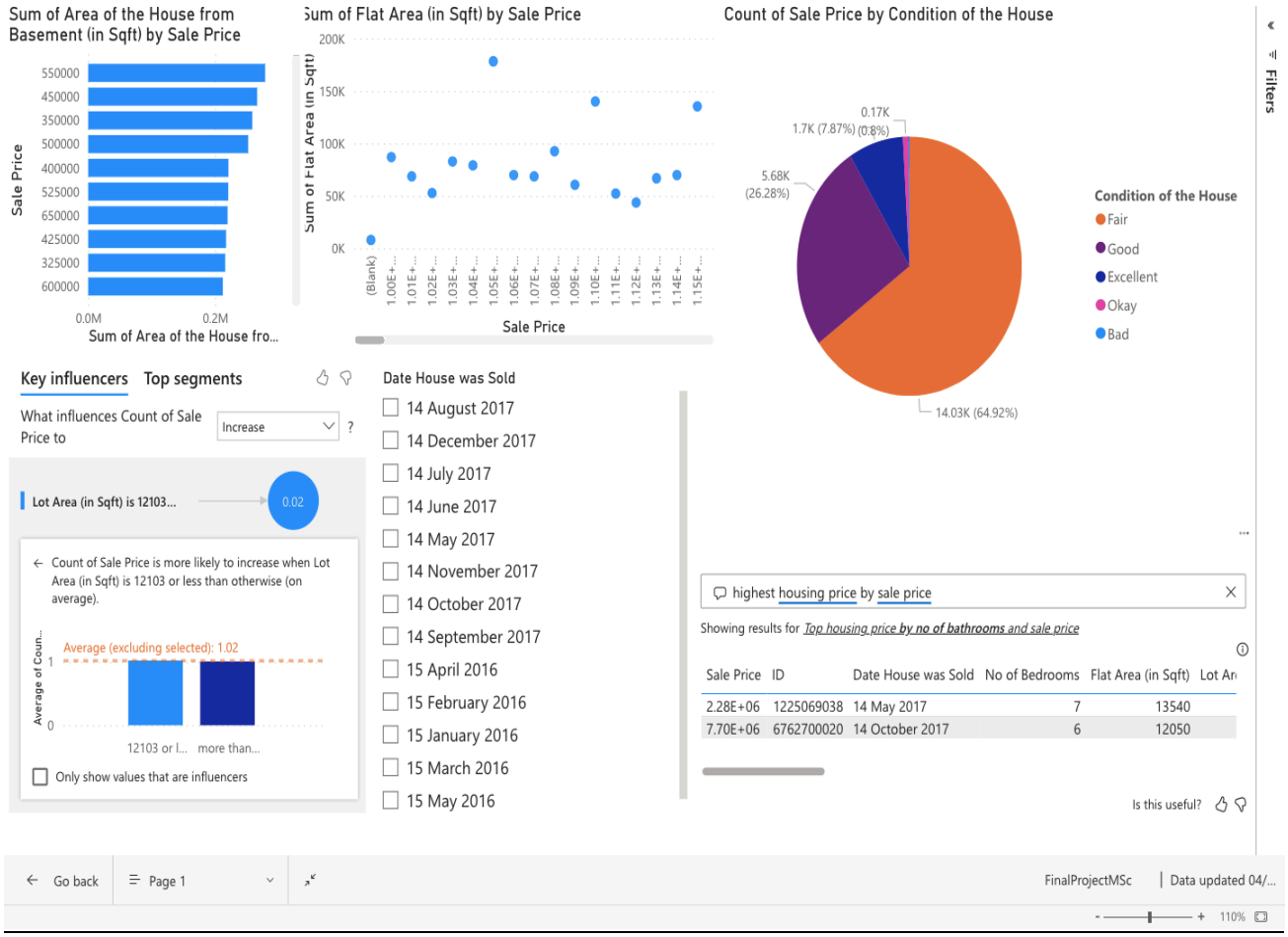
v. Sale Price Box Plot



vi. Line Graph Of Sale Price



POWER BI REPORT:



Accuracy & Evaluation Metrics

Three models were built. The first one is built on linear regression algorithm using sklearn library.

The accuracy is

Evaluations performances are

```
from sklearn import metrics
print('MAE:', metrics.mean_absolute_error(Y_test, predictions))
print('MSE:', metrics.mean_squared_error(Y_test, predictions))
print('RMSE:', np.sqrt(metrics.mean_squared_error(Y_test, predictions)))
print('VarScore:',metrics.explained_variance_score(Y_test,predictions))
```

```
MAE: 72629.05460799548
MSE: 9743084513.775206
RMSE: 98707.06415335837
VarScore: 0.8462105355624405
```

The second model was built on the same algorithm based on a deep learning model using the Keras library.

```
from sklearn import metrics
print('MAE:', metrics.mean_absolute_error(Y_test, predictions))
print('MSE:', metrics.mean_squared_error(Y_test, predictions))
print('RMSE:', np.sqrt(metrics.mean_squared_error(Y_test, predictions)))
print('VarScore:',metrics.explained_variance_score(Y_test,predictions))
```

```
MAE: 72629.05460799548
MSE: 9743084513.775206
RMSE: 98707.06415335837
VarScore: 0.8462105355624405
```

The third model was also a deep learning model but the algorithm used was LSTM or Long-Short Term Memory.

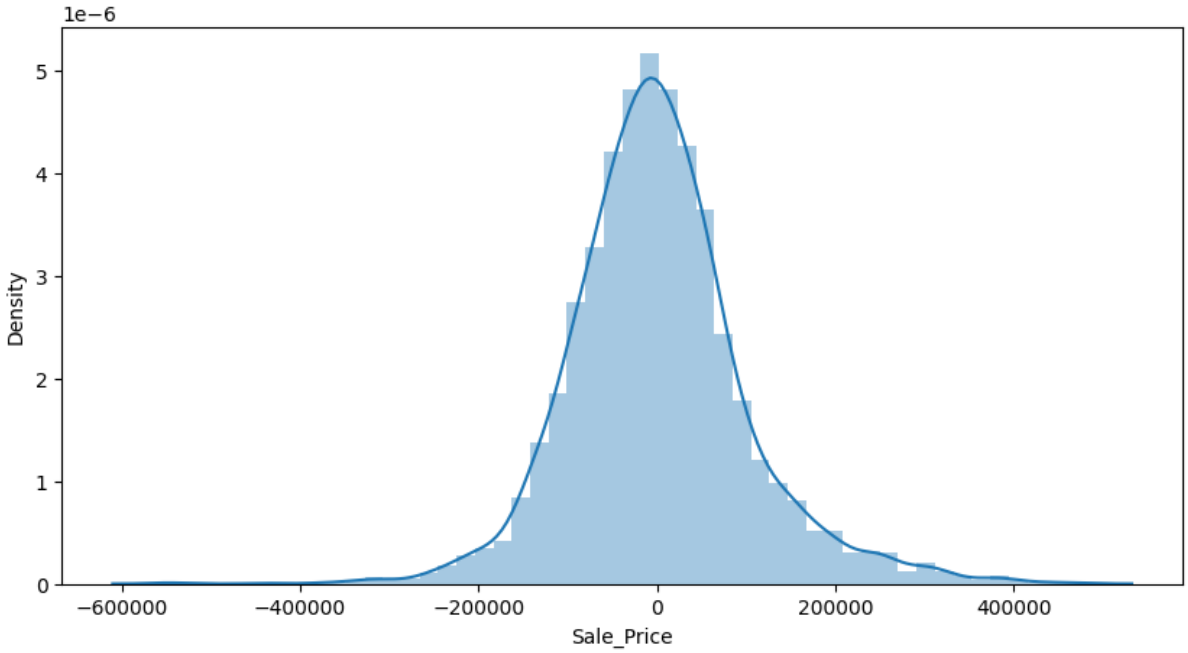
```
from sklearn import metrics

print('MAE:', metrics.mean_absolute_error(Y_test, predictions))
print('MSE:', metrics.mean_squared_error(Y_test, predictions))
print('RMSE:', np.sqrt(metrics.mean_squared_error(Y_test, predictions)))
```

```
MAE: 513564.37983732385
MSE: 327613974101.02686
RMSE: 572375.7280851686
```

Residual Plot

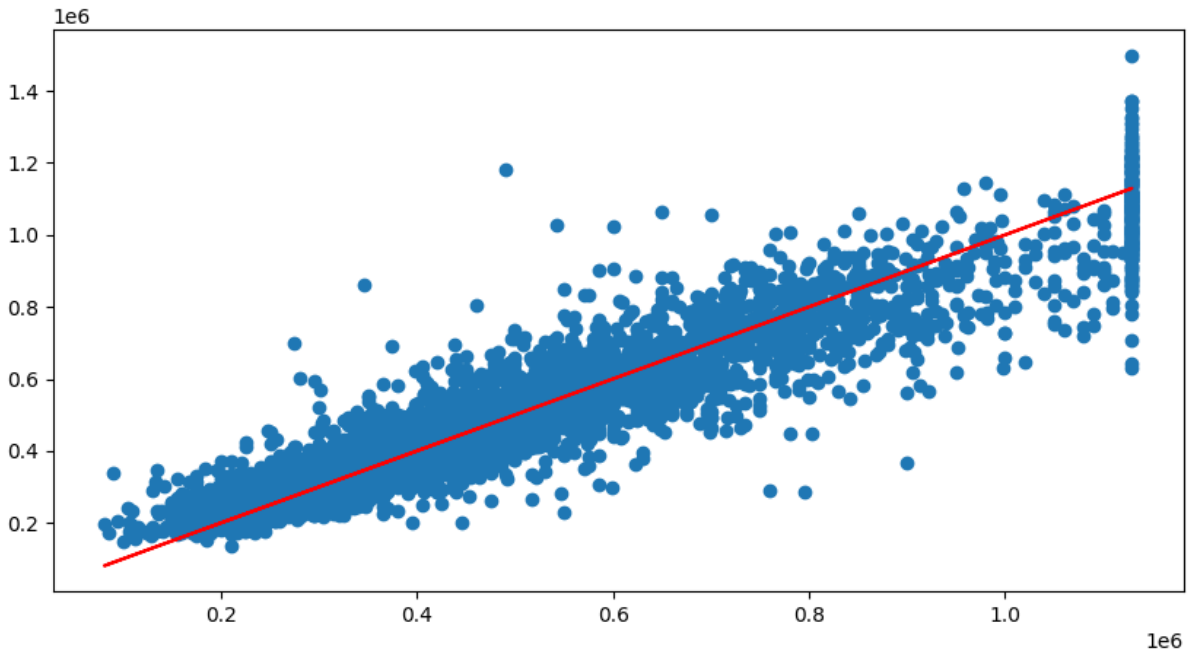
A residual plot is a graphical representation of the residuals, which are the differences between the observed (actual) values and the predicted values in a regression analysis. These plots are often used to assess the goodness of fit of a regression model and to check whether the assumptions of linear regression are met.

Sklearn Model



Scatter Plot:

A scatter plot is a graphical representation of individual data points in a two-dimensional space, with each point representing a single observation. Scatter plots are used to visualize the relationship between two variables and can reveal patterns, trends, correlations, and outliers in the data. They are particularly useful for exploring and understanding the distribution and association between variables.


1e6
1.4
1.2
1.0
0.8
0.6
0.4
0.2
0.2
0.4
0.6
0.8
1.0
1e6

CONCLUSION

In conclusion, our house price prediction model is a valuable tool for both homebuyers and real estate professionals. It provides accurate and data-driven estimates of property values, enhancing decision-making in the real estate market.

To conclude the technical part both the algorithms work great with the problem statement but the deep learning algorithms are better in terms of accuracy. This does not mean that the standard machine learning model doesn't have good accuracy. For example: A data scientist has conducted a study for comparison of two libraries and the machine learning model's accuracy is significantly low compared to the deep learning model

```
Model: Keras Regression

Mean Absolute Error(MAE): 96667.89
Mean Squared Error(MSE): 24912134897.75
Root Mean Squared Error(RMSE): 157835.78
Variance score: 80.84

**********************************

Model: Multiple Linear Regression

Mean Absolute Error(MAE): 124516.17
Mean Squared Error(MSE):39763621927.16
Root Mean Squared Error(RMSE):199408.18
Variance score: 69.42
```

Results: Keras Reg. vs Multiple Linear Reg.

Research Findings

	Keras Regression	Multiple Linear Regression
MAE	96667.89	124516.17
MSE	24912134897.75	379763621927.16
RMSE	157835.78	199408.18
Var Score	80.84	69.42

My Findings

	Keras Regression (Deep Learning Model)	Multiple Linear Regression
MAE	61648.44204433711	72629.05460799548
MSE	7608210587.887718	9743084513.775206
RMSE	87225.05711025771	98707.06415335837
Var Score	0.8794348529026528	0.8462105355624405

As give above, according to this the Keras deep learning model outperforms the sklearn model. This is because the data is not pre-processed properly. Henceforth, the dataset in this project is handled and is processed for better working model.

The Power BI dashboard and reports are also a powerful tool for data analysis and visualization. The have an edge on python analysis and visualisation because of its simplicity and interactive system. You can basically just place the data and play around with different features to gain maximum and valuable insights.

Future Enhancement

For future improvements, we can add a user interface, a feature where the model can predict the sale price by user input and more. By that, we have successfully achieved our project objectives, and we look forward to further developments and enhancements in the future. This project exemplifies the power of data-driven insights and predictive modelling in solving real-world problems. We hope that it will serve as a valuable resource for anyone looking to navigate the complex world of real estate transactions.

REFERENCES

House Price Prediction Using LSTM" by Keyu Zhang and Zhiqiang Wei (2018)

This paper investigates the application of Long Short-Term Memory (LSTM) networks in predicting housing prices, focusing on the ability of LSTM to capture temporal dependencies in time series data.

LINK: https://arxiv.org/abs/1709.08432

House Price Prediction with Machine Learning Techniques: A Review and Comparative Study by Wei Sun et al. (2020)

This study provides a comprehensive review of various machine learning techniques applied to house price prediction. It includes a comparative analysis of different algorithms and their performance.

LINK: https://www.researchgate.net/publication/325435801_House_Prices_Prediction_with_Machine_Learning_Algorithms

Deep Learning Model for House Price Prediction Using Heterogeneous Data Analysis Along With Joint Self-Attention Mechanism

This paper studies and discusses the house price predictions, uses different data analysis techniques and implements deep learning algorithms

LINK: https://ieeexplore.ieee.org/document/9395585

House Prices Prediction Using Deep Learning

The goal of this statistical analysis is to help us understand the relationship between house features and how these variables are used to predict house price.

LINK: https://towardsdatascience.com/house-prices-prediction-using-deep-learning-dea265cc3154