Project Report

On

Development of Machine Learning model for prediction of House prices.

Submitted in the partial fulfillment for the award of

Post Graduate Diploma in Big Data Analytics (PG-DBDA)

from Know-IT ATC, CDAC ACTS, Pune

Guided by:

Mrs. Trupti Joshi

Submitted By:

Akshay Patil (220943025003)

Aniket Jore (220943025011)

Hitesh Narang (220943025020)

Sahil Chavan (220943025033)

CERTIFICATE

TO WHOMSOEVER IT MAY CONCERN

This is to certify that

Akshay Patil (220943025003)

Aniket Jore (220943025011)

Hitesh Narang (220943025020)

Sahil Chavan (220943025033)

have successfully completed their project on

Development of Machine Learning model for prediction of House prices.

Under the guidance of Ms. Trupti Joshi

ACKNOWLEDGEMENT

This project **“Development of Machine Learning model for prediction of House prices”** was a great learning experience for us and we are submitting this work to Know-IT ATC, CDAC ACTS, Pune.

We all are very glad to mention the name of **Mrs. Trupti Joshi** for her valuable guidance to work on this project. Her continuous guidance and support helped us to overcome various obstacles and intricacies during the course of project work.

We are highly grateful to **Mr. Vaibhav Inamdar** (Center Coordinator, Know-it, Pune) his guidance and support whenever necessary while doing this course Post Graduate Diploma in Big Data Analytics (PG-DBDA) through C-DAC ACTS, Pune.

Our most heartfelt thanks goes to **Mr. Shrinivas Jadhav** (Vice-President, Know-it, Pune) who provided all the required support and his kind coordination to provide all the necessities like required hardware, internet facility and extra lab hours to complete the project, throughout the course and till date, here in Know-IT ATC, CDAC ACTS, Pune.

From:

Akshay Patil (220943025003)

Aniket Jore (220943025011)

Hitesh Narang (220943025020)

Sahil Chavan (220943025033)

TABLE OF CONTENTS

ABSTRACT

The ability to estimate house prices is a crucial issue in the real estate industry with ramifications for investors, purchasers, and homeowners. Analyzing a wide variety of elements, including as location, size, age, condition, and several other qualities, is necessary to estimate the cost of a property. Using the features offered by the Housing Dataset, machine learning algorithms have been created to forecast home values. The main strategies and methods for predicting home prices, such as feature selection, data preparation, and model selection, are outlined in this abstract. The difficulties and constraints of this undertaking are also discussed, including the requirement for high-quality data and the impossibility of capturing all pertinent elements that influence home values. House price forecasting is a crucial field of study with real-world implications.

1. INTRODUCTION

The housing market is one of the most dynamic and complex sectors of the economy. Understanding the factors that influence house prices is crucial for homeowners, real estate agents, property investors, and policymakers. With the rapid advancement of machine learning techniques, it has become possible to analyze large datasets and develop accurate models for predicting house prices.

- In this report, we present a study of house price prediction using machine learning on the California housing dataset. The California housing dataset contains information on the median house prices, population, and other socio-economic factors for different neighborhoods in California. We will use this dataset to develop and evaluate machine learning models that can predict house prices based on the available features.
- The main objective of this study is to explore the performance of different machine learning algorithms for house price prediction and identify the most accurate model. We will also investigate the significance of different features in predicting house prices and explore ways to improve the model's accuracy.
- Overall, this report aims to provide insights into the use of machine learning techniques for house price prediction and the factors that influence house prices in California.

Datasets and features:

- Data used was collected from www.kaggle.com . The Dataset is of California state in USA which has median house values of a different latitudes and longitudes.
- However, overall the dataset provides a rich source of data for analyzing patterns and trends in lending behavior, and for developing and evaluating house prices.
- The main goal of the analysis is to build an accurate and robust regression model to predict the outcome of House Price. This project uses Random Forest, Decision Tree, Linear Regression, Polynomial Regression, and Gradient-Boosting.

2. SYSTEM REQUIREMENTS

Hardware Requirements:

- Platform – Windows 7 or above
- RAM – Recommended 8 GB of RAM
- Peripheral Devices – Keyboard, Monitor, Mouse
- WiFi connection with minimum 2 Mbps speed

Software Requirements:

- Language: Python 3
- Apache Spark
- Machine Learning
- Tableau
- **OS – Windows**

3. FUNCTIONAL REQUIREMENTS

1) Python 3:

❿ Python is a high-level programming language that is easy to learn and use.

❿ Python is an interpreted language, which means that code can be executed on the fly, without the need for compilation.

❿ Python is open source and free to use, with a large and active community of developers contributing to its development and maintenance.

❿ Python has a vast collection of third-party libraries and packages, such as NumPy, Pandas, Matplotlib, and Scikit-learn, among others, that make it easy to perform data analysis.

2) Apache Spark:

❿ Apache Apache Spark is a distributed computing framework designed to process large amounts of data in parallel across a cluster of computers.

❿ Spark includes various libraries and APIs for processing structured and unstructured data, including Spark SQL, Spark Streaming, MLlib, and GraphX.

❿ Spark uses in-memory computing and data partitioning techniques to achieve high performance and scalability. Spark supports parallel processing and data partitioning, which enables it to scale horizontally and handle large amounts of data.

❿ Apache Spark is widely used in various industries, such as finance, healthcare, e-commerce, social media, telecommunications, etc.

3) Tableau:

❿ Tableau is a data visualization and business intelligence software that allows users to connect, analyse, and share data in a visual and interactive way.

❿ It offers a user-friendly drag-and-drop interface that enables users to create interactive dashboards, reports, and charts without the need for complex coding or programming.

❿ Tableau supports various data sources, including spreadsheets, databases, cloud services, and big data platforms, such as Hadoop and Spark.

Data Cleaning:

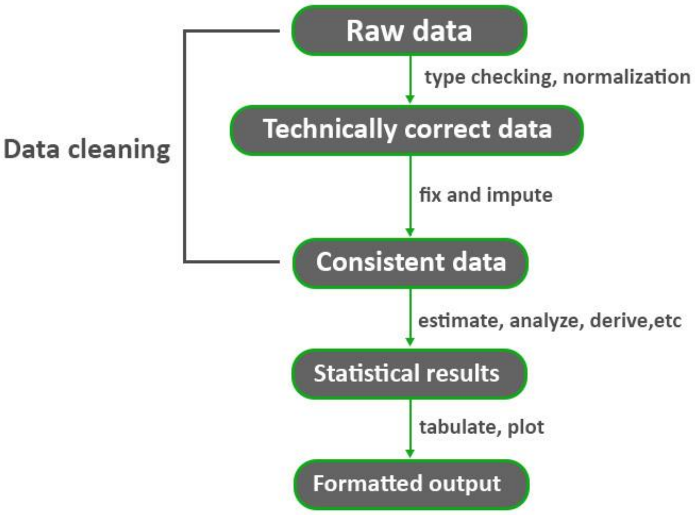


Fig: Data Cleaning Process

Data cleaning is a crucial process in Data Mining. It carries an important part in the building of a model. Data Cleaning can be regarded as the process needed, but everyone often neglects it. Data quality is the main issue in quality information management. Data quality problems occur anywhere in information systems. These problems are solved by data cleaning.

Without proper data cleaning, data analysis and modelling can lead to erroneous or biased results, which can have serious consequences for businesses and organizations.

Hence, it is a critical step in the data preparation process, as it can significantly impact the accuracy and reliability of the insights and decisions that are derived from the data. By improving the quality of data, organizations can gain a better understanding of their operations, customers, and market trends, and make more informed and effective decisions.

4. SYSTEM ARCHITECTURE

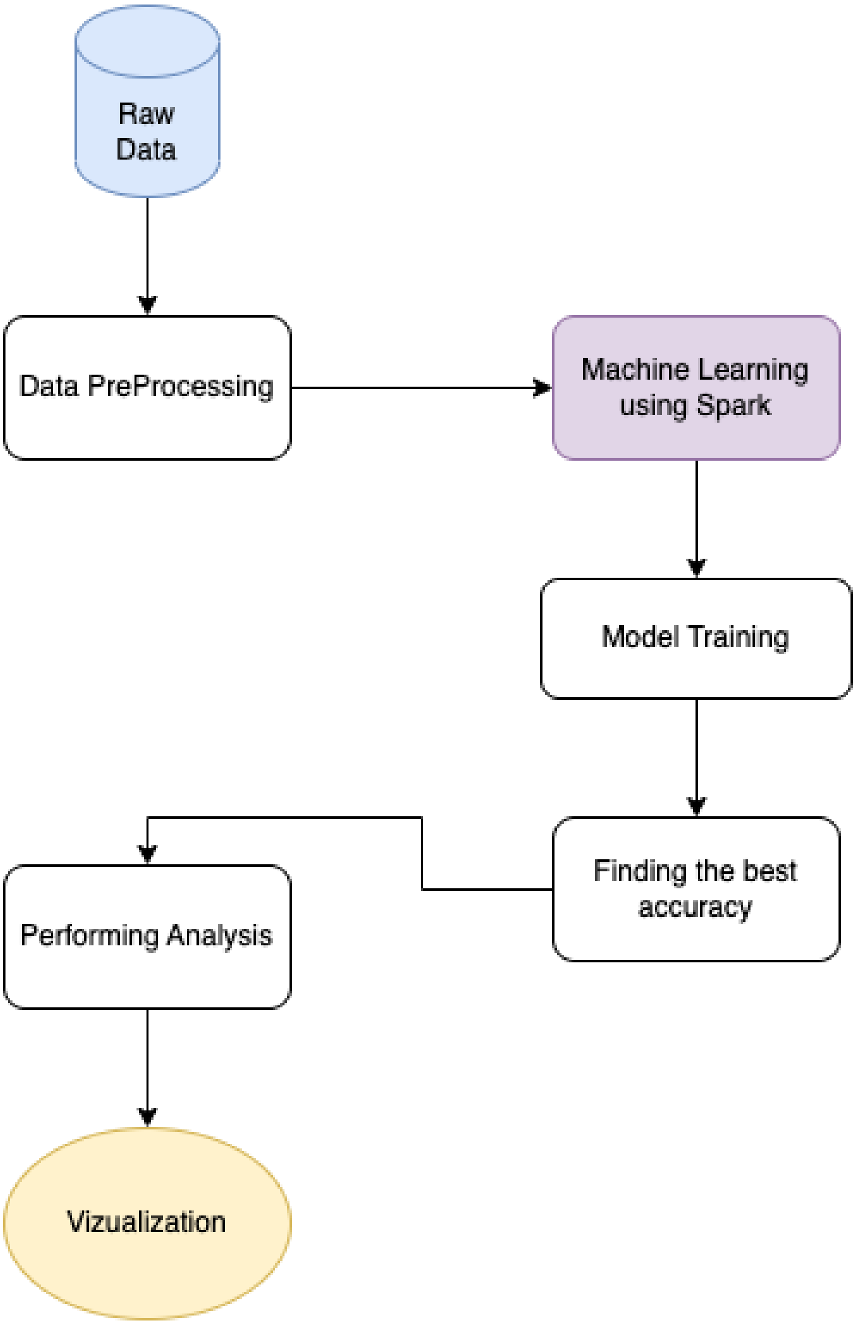


Fig: System Architecture of Prediction of House Price

5. METHODOLOGY

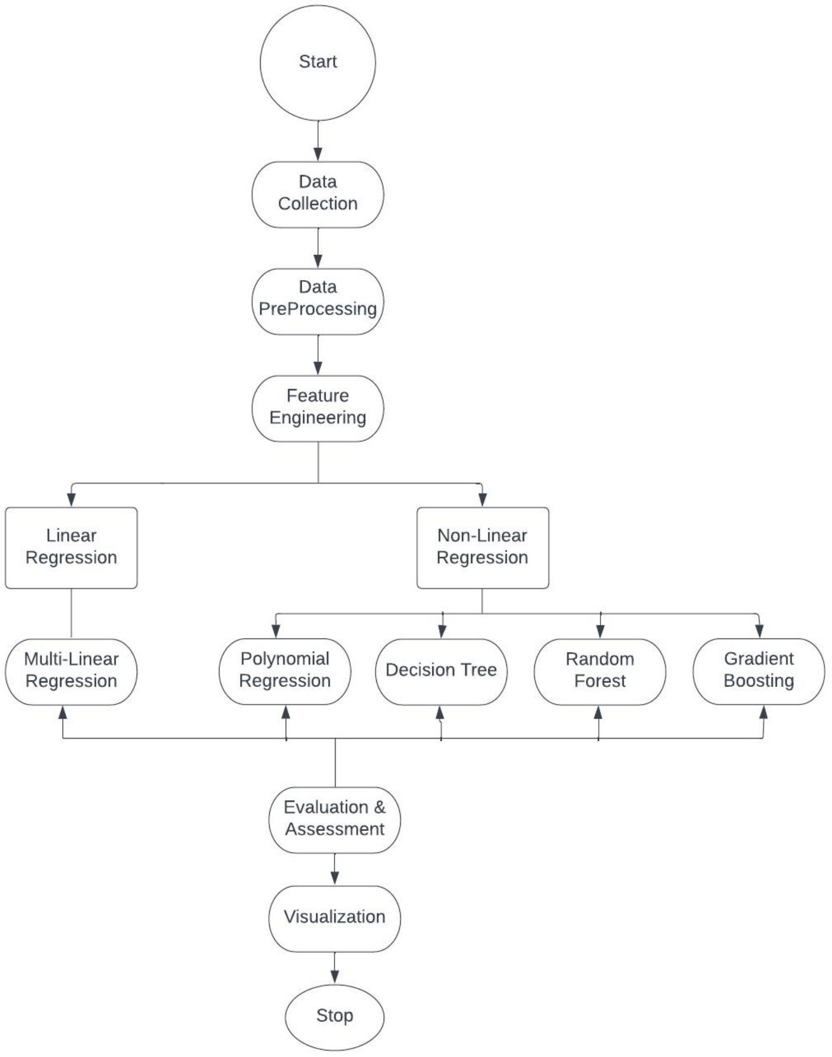


Fig: Methodology of House Price Prediction

6. MACHINE LEARNING ALGORITHMS

- Machine learning is a subfield of artificial intelligence that involves developing algorithms and models that enable computers to learn from data and make predictions or decisions without being explicitly programmed. The goal of machine learning is to enable computers to improve their performance over time by learning from experience and feedback.
- In our project, we applied various Regression Algorithms such as Random Forest, Decision Tree, Linear Regression, Polynomial Regression, and Gradient-Boosting.. After the implementation, were able to analyze the accuracy of the algorithms on our data.
- Accuracy was one of the major factors that helped to decide which model has the accurate predictions.

1. Linear Regression

Linear regression is a statistical method used to model the relationship between a dependent variable and one or more independent variables. The goal is to find the best-fit line or hyper plane that minimizes the distance between the predicted values and the actual values.

Pros:

- Linear regression is a simple and easy-to-understand method that requires little technical knowledge.
- It can be used to identify the strength and direction of the relationship between variables.
- It is useful for predicting the value of a dependent variable when the values of the independent variables are known. .

Cons:

- It assumes that the error terms are normally distributed and have equal variances, which may not be true in some cases.
- It can be sensitive to outliers, which can affect the accuracy of the predictions.
- It can be affected by multi-collinearity, which occurs when two or more independent variables are highly correlated with each other.

2. Random Forest:

Random forest is a machine learning algorithm that is used for classification, regression, and feature selection tasks. It is an ensemble method that combines multiple decision trees, where each tree is trained on a subset of the training data and a subset of the input features.

Pros:

- It is a highly accurate and powerful machine learning algorithm that can perform well on a wide range of classification and regression tasks.
- It can handle both categorical and continuous input variables, and it can detect and handle interactions between variables.

Cons:

- It may not perform well on small datasets or with rare or unseen classes, which may require more specialized techniques or models.
- It may not be suitable for online or real-time prediction tasks, which require faster and more lightweight models or techniques.

3. Decision Tree Regressor:

Decision Tree is a supervised learning algorithm which can be used for solving both classification and regression problems. It can solve problems for both categorical and numerical data. Decision Tree regression builds a tree-like structure in which each internal node represents the "test" for an attribute, each branch represent the result of the test, and each leaf node represents the final decision or result. A decision tree is constructed starting from the root node/parent node (dataset), which splits into left and right child nodes (subsets of dataset). These child nodes are further divided into their children node, and themselves become the parent node of those nodes

Pros:

- Decision tree regression is a simple and easy-to-understand method that requires little technical knowledge.
- It can handle non-linear relationships between the features and the target variable.
- It can handle missing data and outliers effectively.
- It can automatically select the most important features for predicting the target variable.
- It can be used for both classification and regression problems.

Cons:

- Decision tree regression is a simple and easy-to-understand method that requires little technical knowledge.
- It can handle non-linear relationships between the features and the target variable.
- It can handle missing data and outliers effectively.
- It can automatically select the most important features for predicting the target variable.
- It can be used for both classification and regression problems.

4. Gradient Boosting:

Gradient boosting is a machine learning algorithm that is used for supervised learning tasks, such as classification and regression. It is an ensemble method that combines multiple weak learners, typically decision trees, into a strong learner that can make accurate predictions on new data.

Pros:

- It is highly accurate and can achieve state-of-the-art performance on a wide range of tasks, especially with large datasets and complex models.
- It can provide a measure of feature importance, which can be used for feature selection and interpretation.
- It can be parallelized and distributed, which can improve its performance on large datasets.

Cons:

- It can be computationally intensive and time-consuming to train and tune, especially for large datasets and complex models.

5. Polynomial Regression

Polynomial Regression is a type of regression which models the non-linear dataset using a linear model. It is similar to multiple linear regressions, but it fits a non-linear curve between the value of x and corresponding conditional values of y. suppose there is a dataset which consists of data points which are present in a non-linear fashion, so for such case, linear regression will not best fit to those data points. To cover such data points, we need Polynomial regression. In Polynomial regression, the original features are transformed into polynomial features of given degree and then modelled using a linear model.

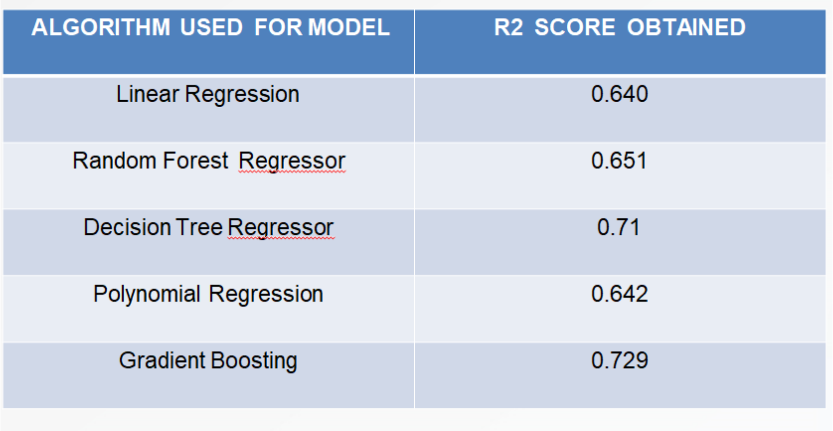
Pros:

- Polynomial regression can capture non-linear relationships between the independent and dependent variables, which linear regression cannot.

- It can provide a good fit to the data if the relationship is non-linear and the degree of the polynomial is chosen appropriately.

Cons:

- Polynomial regression can be sensitive to the choice of degree of the polynomial. A high degree polynomial can lead to over fitting, while a low degree polynomial can lead to under fitting.
- It can be affected by outliers, which can distort the fit of the polynomial equation.



ALGORITHM USED FOR MODEL	R2 SCORE OBTAINED
Linear Regression	0.640
Random Forest Regressor	0.651
Decision Tree Regressor	0.71
Polynomial Regression	0.642
Gradient Boosting	0.729

Fig R2 Score of different ML models

7. DATA VISUALIZATION AND REPRESENTATION

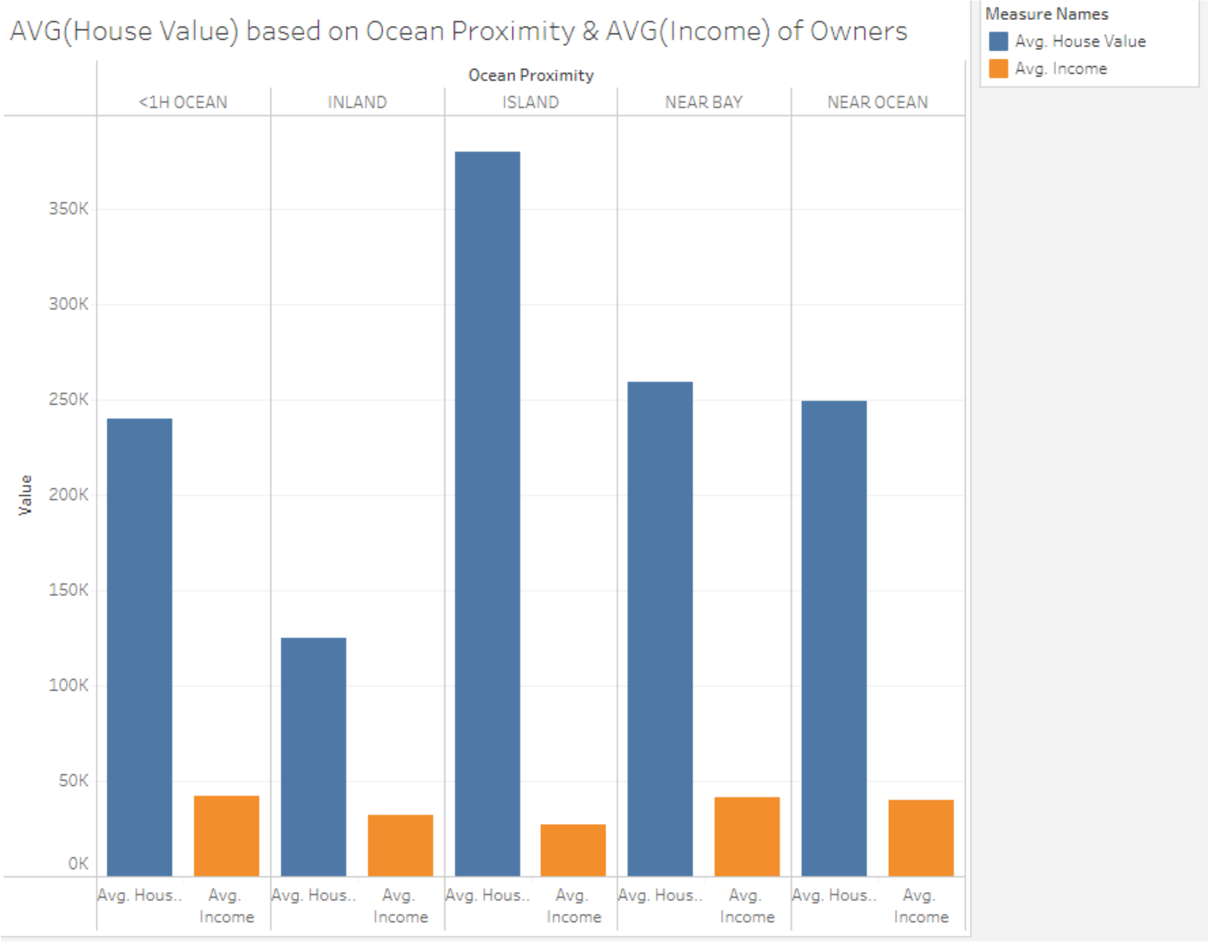
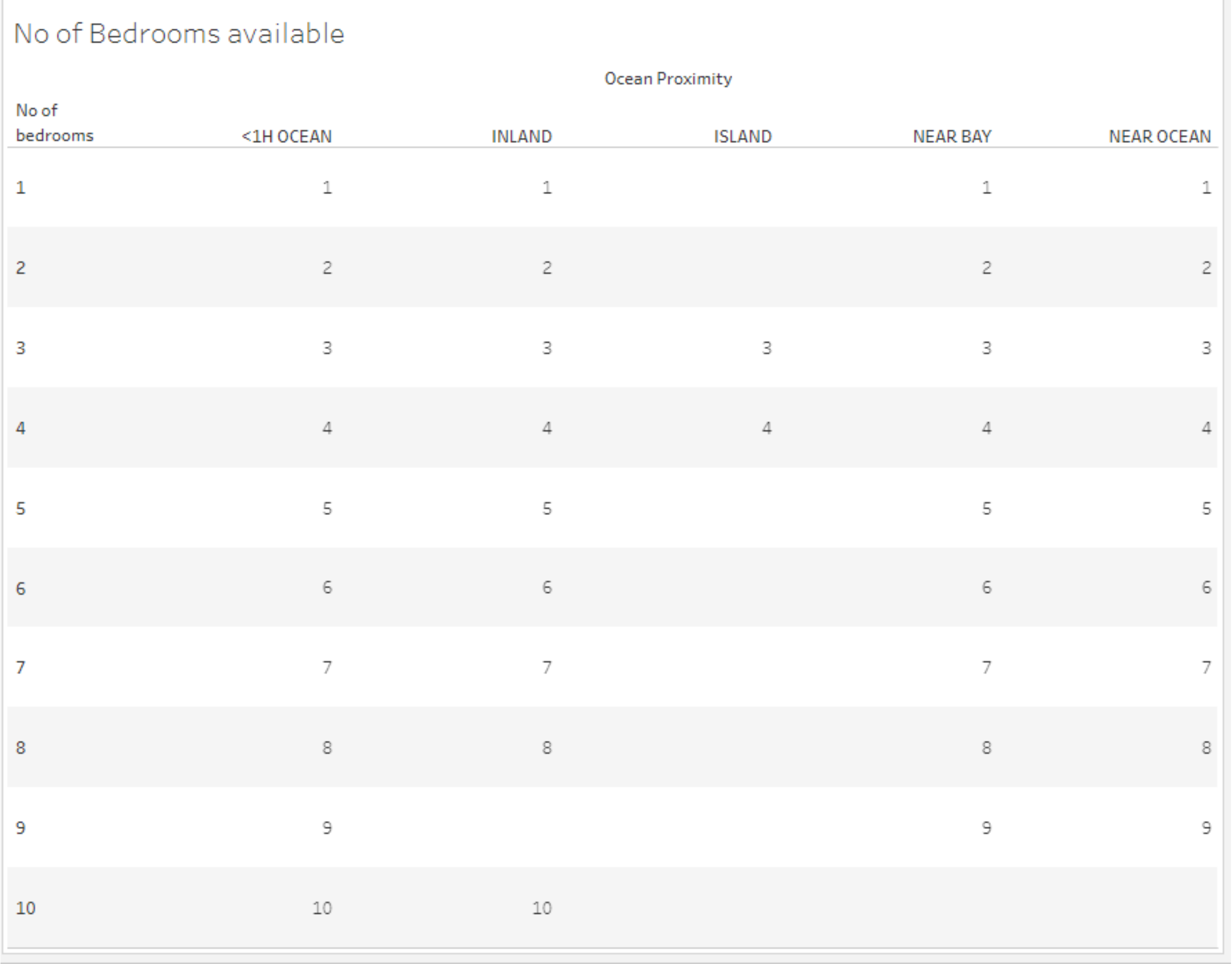


Fig: AVG (House Value) based on Ocean Proximity & AVG(Income) of Owners (Side-by-Side Bar Chart)



No of Bedrooms available

	Ocean Proximity				
No of bedrooms	<1H OCEAN	INLAND	ISLAND	NEAR BAY	NEAR OCEAN
1	1	1		1	1
2	2	2		2	2
3	3	3	3	3	3
4	4	4	4	4	4
5	5	5		5	5
6	6	6		6	6
7	7	7		7	7
8	8	8		8	8
9	9			9	9
10	10	10			

Fig. Number of Bedrooms available in a House based on different Ocean Proximity (Text Table)

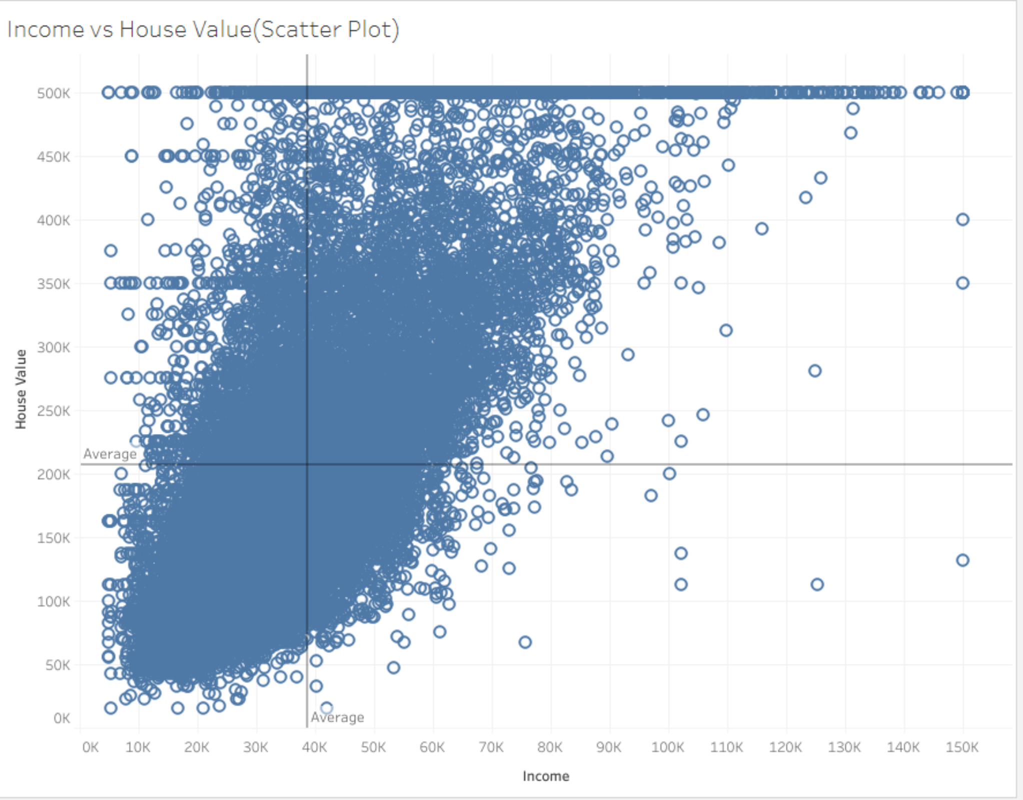


Fig. Income vs House Value (Scatter Plot)

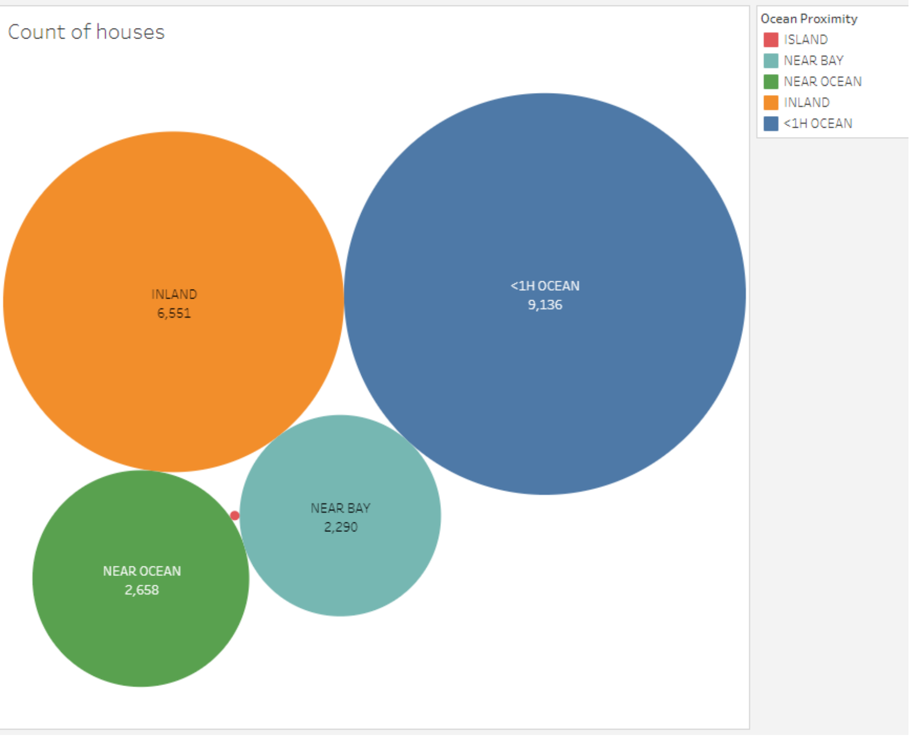
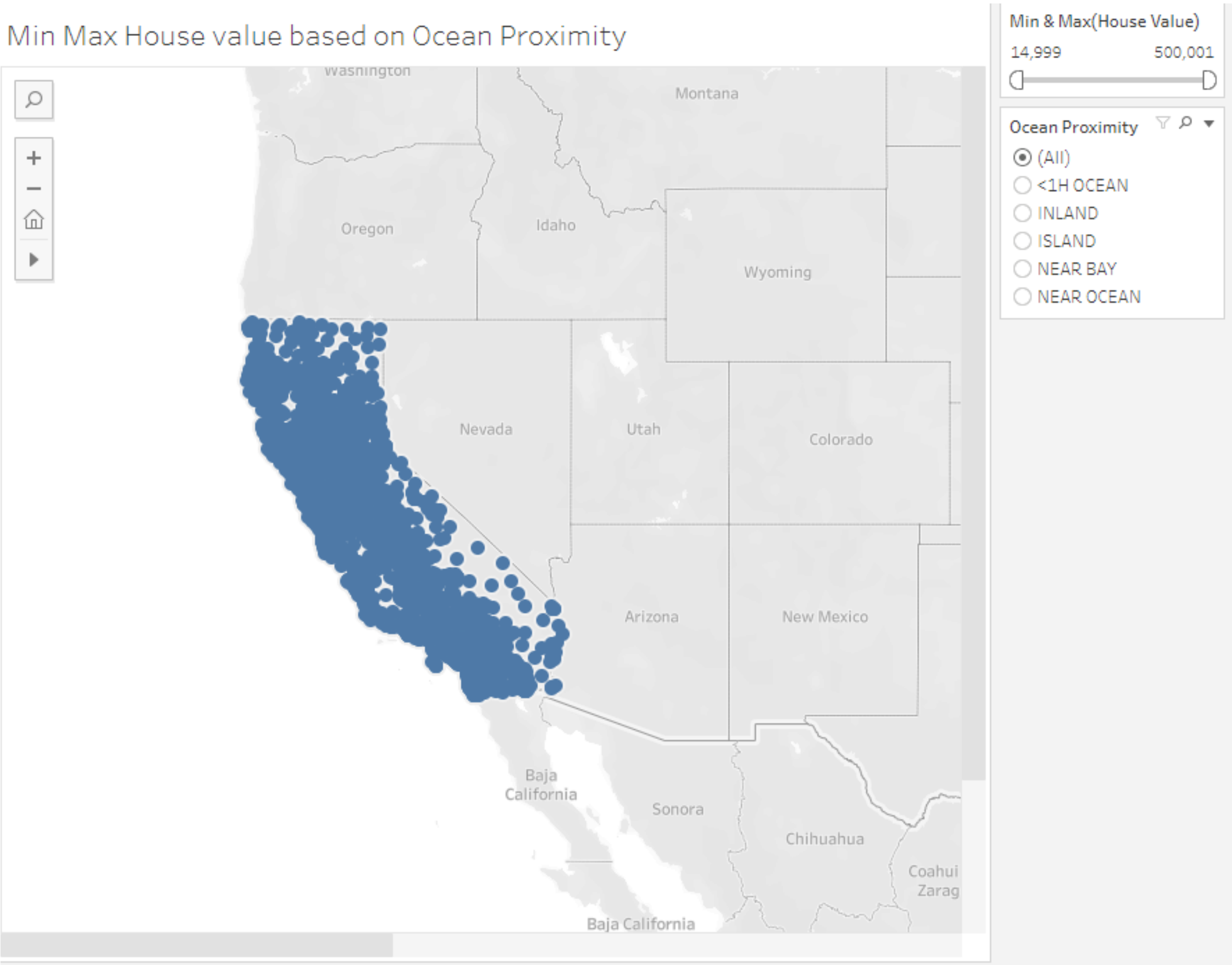


Fig: Total Number of Houses available according to different Ocean Proximity (Bubble Chart)



Min Max House value based on Ocean Proximity

Min & Max(House Value)

14,999 500,001

Ocean Proximity

● (All)

○ <1H OCEAN

○ INLAND

○ ISLAND

○ NEAR BAY

○ NEAR OCEAN

Fig. Min & Max House Value based and the Geographical Location of the houses based on Latitude & Longitude (Map)

8. CONCLUSION AND FUTURE SCOPE

- In conclusion, predicting house prices based on features such as ocean proximity, number of bedrooms, area, population, and median income of people can be an effective way to estimate the value of a property. The features mentioned in the analysis are important factors that can affect the demand and supply of housing in a given area.
- Ocean proximity can be a significant predictor of house prices, as houses closer to the ocean tend to be in higher demand and can command higher prices. The number of bedrooms and area of a property can also impact its value, as larger properties and those with more bedrooms tend to be more desirable for families and can command higher prices.
- Overall, the combination of these features can provide a robust basis for predicting house prices. By leveraging machine learning algorithms, it is possible to create accurate models that take into account these important features and provide reliable estimates of the value of a property.

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

1. The used pre-processing methods do help in the prediction accuracy. However, experimenting with different combinations of pre-processing methods to achieve better prediction accuracy.
2. Using Hyper-Parameter-Tuning for current existing algorithms for enhanced accuracy.
3. We can use the challenges and best practices for deploying machine learning models in production environments. We can explore different deployment options like containerization, server less computing, and API-based deployment.

References

1. California Housing

   https://www.kaggle.com/datasets/camnugent/california-housing-prices

2. Zikopoulos, P., Eaton, C., deRoos, D., Deutsch, J., Lapis, G., & Brown, R. (2012). Understanding Big Data: Analytics for Enterprise Class Hadoop and Streaming Data. McGraw-Hill.
3. Spark documentation: https://spark.apache.org/docs/latest/
4. Python documentation: https://docs.python.org/3/
5. "Machine Learning using Python" by Prof. U Dinesh Kumar, IIM Bangalore.
6. . Annina S, Mahima SD, Ramesh B. An Overview of Machine Learning and its Applications. International Journal of Electrical Sciences & Engineering (IJESE). 2015 January; I(1): 22-24.