CS7052 - Machine Learning

House Price Prediction using Machine Learning

A Comparative Study of Random Forest and Linear Regression Algorithms

By Harpreet Kaur, Midhun Mukundan and Jim Patewa

horizontal line

# 

# Table of Contents

[Table of Contents 1](#_Toc152514250)

[Abstract 3](#_Toc152514251)

[Report Objective 3](#_Toc152514252)

[Linear Regression Algorithms 3](#_Toc152514253)

[Random Forest Algorithms 3](#_Toc152514254)

[1. Introduction 4](#_Toc152514255)

[1.1 Problem Statement 4](#_Toc152514256)

[1.2 Significance of Machine Learning 5](#_Toc152514257)

[1.3 Dataset Description 6](#_Toc152514258)

[2. Design & Methodology 7](#_Toc152514259)

2.1 Data exploration and preprocessing 7

2.2 Model selection 11

2.3 Model training and hyperparameter tuning 11

[3. Testing 13](#_Toc152514264)

[3.1 Evaluation Metrics 13](#_Toc152514265)

[3.2 Results 14](#_Toc152514266)

[4. Conclusion 16](#_Toc152514271)

[4.1 Key Findings 16](#_Toc152514272)

[4.2 Implications and Applications 17](#_Toc152514273)

[4.3 Limitations and Future Research 18](#_Toc152514274)

[4.4 Personal Summary 19](#_Toc152514275)

[5. References 20](#_Toc152514276)

## Abstract

### Report Objective

This report presents a detailed study on the application of machine learning algorithms, specifically the Random Forest and Linear Regression algorithms, for the purpose of predicting house prices. The study utilises the California housing prices dataset and employs a variety of machine learning techniques such as data preprocessing, feature engineering, model training and hyperparameter tuning. The end results showcase the effectiveness of the Random Forest algorithm in comparison to Linear Regression, providing valuable insight into the effectiveness of machine learning in order to garner accurate house price predictions.

### Linear Regression Algorithms

linear regression as a supervised machine learning algorithm used for predicting a linear trend between a dependent variable and one or more independent variables. It assumes a consistent correlation and aims to find a best-fit line minimizing the difference between predicted and actual values. The algorithm originates from statistics, where a literal line is drawn to depict the relationship mathematically. Linear regression finds applications in various domains like stock market analysis, price prediction, medical research, and sales forecasting, making it a reliable method for predicting future outcomes in diverse fields.

### Random Forest Algorithms

The Random forest algorithm is an ensemble machine learning algorithm that operates through the combination of multiple independent decision trees brought together for the purpose of increasing the accuracy of predictions. The name “Random Forest” derives from its very consisting of numerous interconnected decision trees resembling that of a forest.

## 1. Introduction

### 1.1 Problem Statement

In the study that we conducted, our main objective was not only to develop a machine learning model capable of accurately predicting the houses prices in the state of California based on a variety of features. But to make a comparative analysis of two well known algorithms within the field of Machine Learning, the two algorithms being linear regression and Random Forest. In order to address this problem, we required a sample set of data with which to base our predictions upon and therefore collected a comprehensive dataset composed of the required features necessary to make predictions on the future prices of houses. Features necessitated included those such as the number of bedrooms per household, proximity to the ocean, and the age of the buildings among others. Moreover, the dataset required would need to consist of the actual selling prices of the houses, which would serve as our target variable for training the model.

Our primary focus was on training a machine learning model that could effectively analyse the provided features and thereby establish meaningful relationships with the corresponding house prices. By leveraging different algorithms and techniques namely linear regression and the random forest algorithm, we aimed to uncover patterns and correlations within the dataset that would enable accurate predictions. Once our model was trained, it would be possible to use it as a tool for predicting the prices of houses that were not part of the training dataset.

### 1.2 Significance of Machine Learning

Machine learning holds significant importance in the modern world due to its ability to process and analyse large datasets extracting meaningful patterns from them in order to make accurate predictions. Machine learning algorithms excel at not only handling complex data but also uncovering hidden insights that may not be apparent to humans. In the context of predicting house prices such as in the case of this particular study, these algorithms can be used to analyse various features of houses and thereby identify the correlations and patterns that influence their value. Thus enabling us to make accurate predictions and aid us with the forecasting of future house prices.

In essence, the significance of machine learning lies in its capacity to extract knowledge from data, make accurate predictions, and empower decision-makers. And by applying machine learning techniques to house price prediction, we aimed to exemplify this fact and demonstrate how it can be beneficial in a practical, real world scenario.

### 1.3 Dataset Description

For our study we utilised the California Housing Prices dataset, which we obtained from Kaggle. This dataset served as the cornerstone for training and evaluating our machine learning models.

The California Housing Prices dataset provided us with a rich and diverse collection of data with which to work with, and ensured that our machine learning models were trained on a representative real world sample of the housing market. The dataset which is made up of information taken from the 1990 California census served as a valuable resource in our study. It enabled us to explore the relationships between various features and house prices, ultimately leading to improved accuracy in our predictions.

The California Housing Prices dataset in and of itself is made up of a comprehensive accumulation of data points that provide valuable insights into the various factors influencing house prices within California. Its compositional makeup includes the location of the houses, their size, the number of rooms, and their proximity to the ocean.

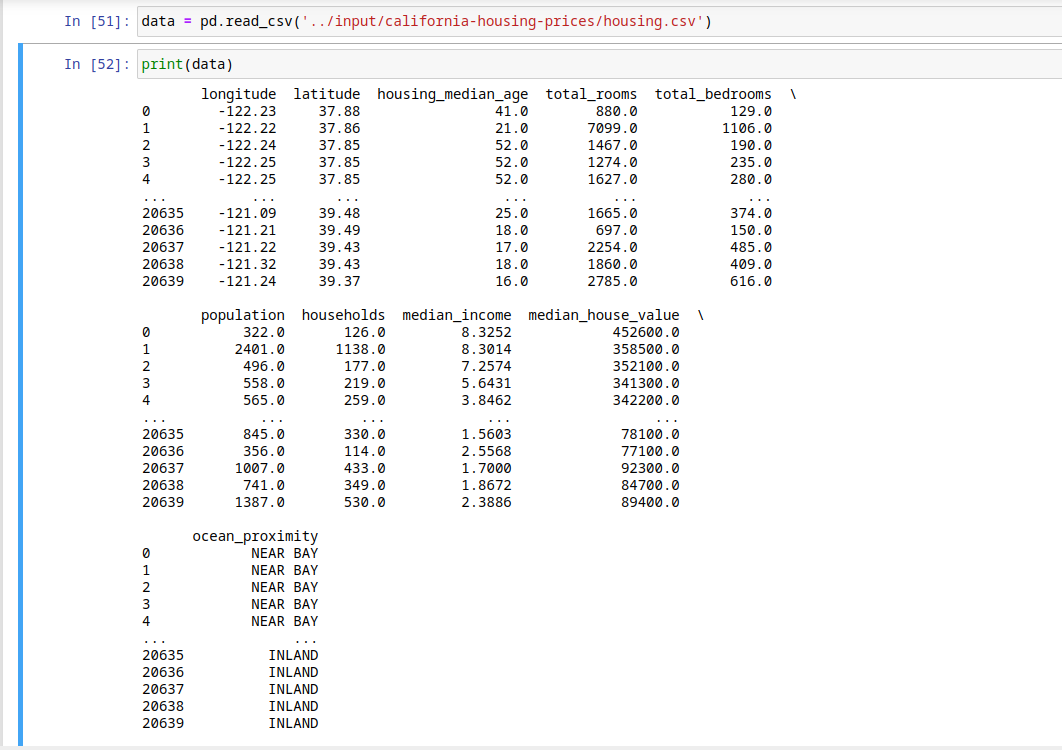
Using this dataset, we were able to train our machine learning models to learn the relationships between the features with which it was made up of and calculate the correlation between each and every feature and the overall house prices. In the end allowing us to forecast and make accurate predictions about the cost of houses based on their characteristics.

## 2. Design & Methodology

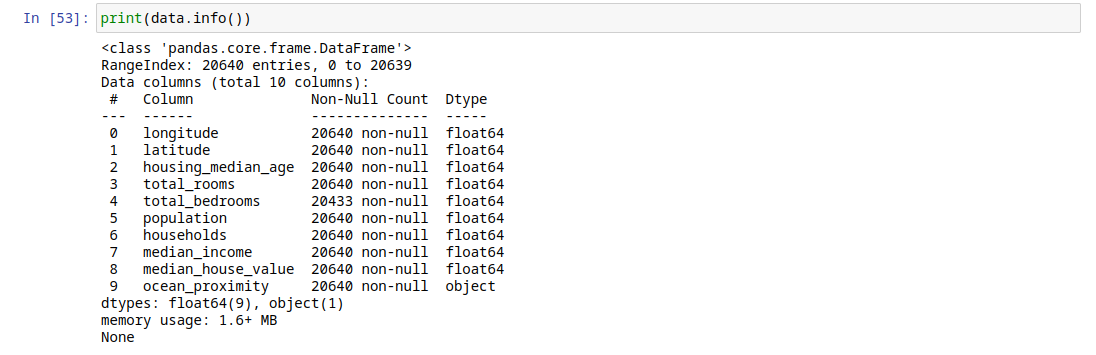
### 

### 2.1 Data Exploration & Preprocessing

After collecting the data, we conducted essential preprocessing to ready it for machine learning models. To enhance predictive accuracy, we addressed missing values by removing corresponding rows. For robust model evaluation, we split the dataset into training and testing sets, ensuring model generalization to unseen data. Employing one-hot encoding handled categorical features, making them suitable for numerical input. These preprocessing steps ensured a clean, appropriately split dataset, instilling confidence in the model's reliability and accurate predictions based on provided metrics.

Below is a screenshot of us loading the dataset into Jupyter Notebook, and thereafter printing the data within the dataset.

Below is a screenshot of us running the data.info() pandas function in order to see if any null values existed within the dataset. From the output of this function we observed that the total\_bedrooms column consisted of 207 null values, which could lead to inaccurate predictions as a result of data inconsistencies.



### 

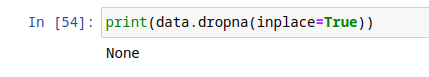
### 

### 

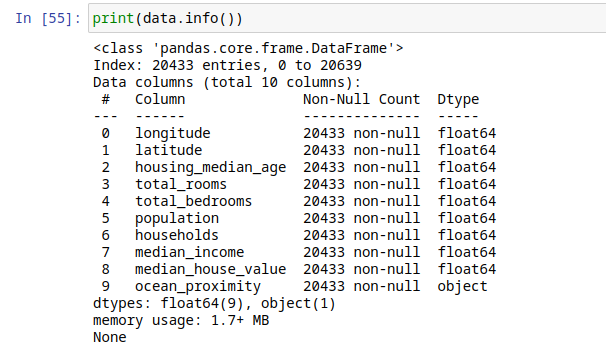
### 

### 

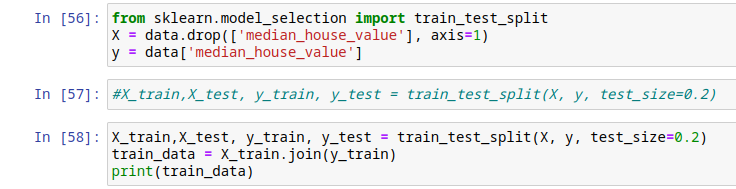
Below is a screenshot of us running removing the null values from our dataset, thereby creating a more consistent and clean dataset with which to work from.



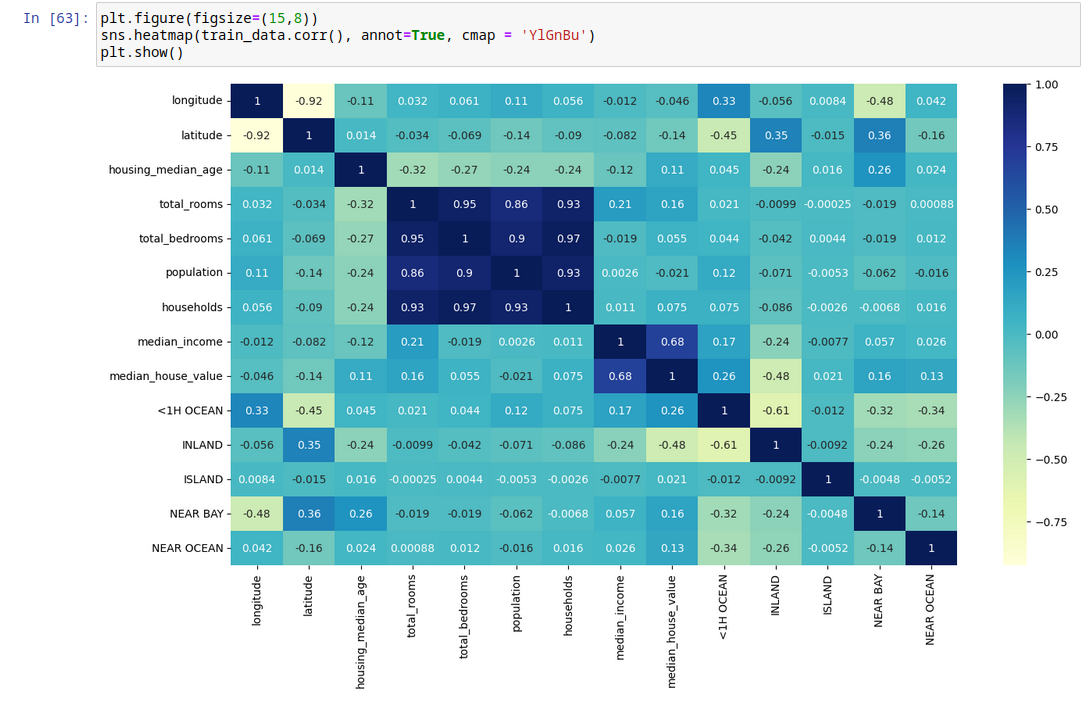
Below is a screenshot of the result of us removing the null values;



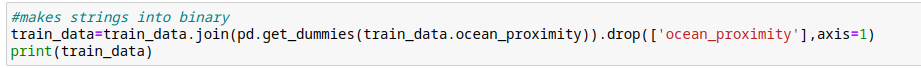
Below is a screenshot of us separating the columns into two distinct variables for training purposes, and specifying the percentage of the data will be used for training and also how much data will be used for testing purposes.



Below is a screenshot of us implementing a correlation heatmap in order to help us analyse the strength of relationships between variables. From this we observed that median household income had a strong correlation with the median house value and consequently would signify that the income of a household would therefore be a good predictor of a house's value.



Below is a screenshot of us converting the ocean proximity features from strings into binary values in order to ensure that they can be correctly identified.



### 2.2 Model Selection

As can be inferred from the title of the report both the Random Forest and Linear Regression algorithms were naturally selected for use in training our model and for comparison to observe which algorithm produced the most accurate predictions. By selecting both Random Forest and Linear Regression, we aimed to evaluate the trade-off between model complexity and interpretability. Random Forest can capture intricate relationships and interactions, while Linear Regression provides a simpler and more interpretable model. This, allowing us to understand the potential benefits of employing a more complex algorithm like Random Forest in comparison to a simpler algorithm like Linear Regression.

### 2.3 Model Training and Hyperparameter Tuning

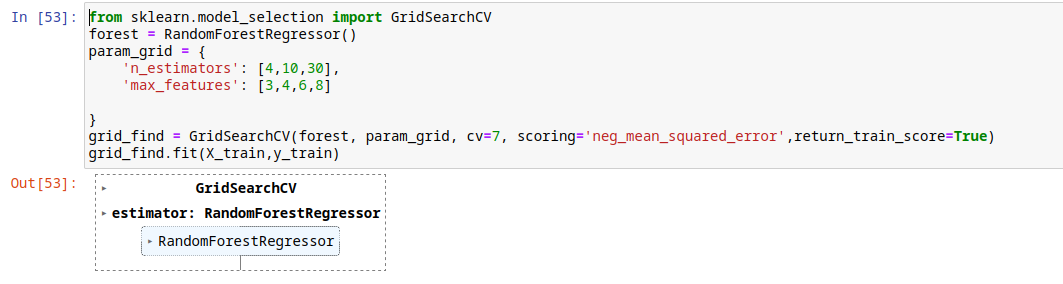
The selected models were trained using the aforementioned California Housing Prices dataset with which we obtained from Kaggle. However, due to the simplicity of the Linear Regression algorithm we opted only to carry out hyperparameter tuning on the Random Forest model.

We employed hyperparameter tuning techniques, such as grid search with cross-validation, to optimise the models' performance. Parameters such as the number of estimators and maximum depth were tuned to achieve the best results.

To systematically explore different combinations of hyperparameters, we utilised grid search with cross-validation. Grid search involves specifying a range of values for each hyperparameter, with the technique exhaustively evaluating all possible combinations. Moreover, Cross-validation helps to assess the performance of the models for different hyperparameter settings.

Through the process of hyperparameter tuning, we aimed to find the optimal combination of hyperparameters that maximised the models' performance. By carefully tuning the hyperparameters, we ensured that the models were optimised to provide the best possible performance for predicting house prices based on the given features.

Below is a screenshot of us carrying out Random Forest hyperparameter tuning on the Random Forest model.



## Below is a screenshot of us carrying out training on the Linear Regression model.

## 3. Testing

### 

### 3.1 Evaluation Metrics

The predictive accuracy of the models was evaluated using R-squared (R^2): R-squared is a statistical metric that indicates the proportion of the variance in the target variable (house prices) that can be explained by the model's predictions. It ranges from 0 to 1, with a higher value indicating a better fit. A value of 1 means the model perfectly predicts the target variable, while a value of 0 suggests that the model's predictions are no better than randomly guessing.

### 

### 3.2 Results

#### Linear Regression

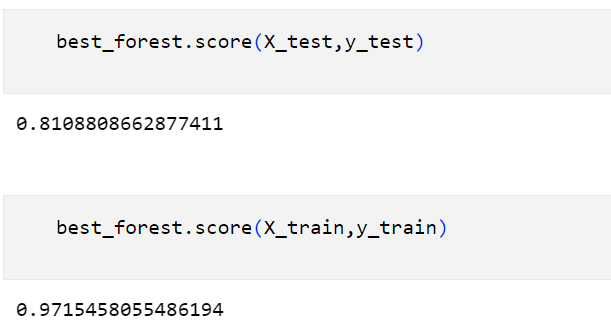
Below is a screenshot of the test results obtained from the Linear Regression model.

### 

#### 

#### Random Forest

Below is a screenshot of the test results obtained from the Random Forest.



This was our 1st test and train score which was highly overfitting so after tuning some parameters we got a better score which is: -

## 

## 4. Conclusion

### 4.1 Key Findings

In our study, the Random Forest model surpassed the predictive accuracy of the Linear Regression model for house price prediction. Leveraging an ensemble of decision trees, Random Forest captured non-linear relationships and complex interactions, leading to more accurate predictions. It effectively understood intricate relationships between house prices and features like location, size, and number of rooms. In contrast, Linear Regression's linear assumption couldn't adequately capture non-linear relationships and the diverse variables in the dataset, despite providing a simpler and interpretable model.

These findings suggest that when dealing with house price prediction, the Random Forest model is a more suitable choice in comparison to Linear Regression due to its ability to handle non-linear relationships and capture complex interactions. However, ultimately the choice of model depends on the specific requirements of the task and the trade-off between predictive accuracy and interpretability.

### 

### 

### 4.2 Implications and Applications

House price prediction can have significant applications in the corporate world. Hypothetically, insights gained from our study could be leveraged by various real estate stakeholders within the California area, including real estate agents, property buyers, and sellers, to make more informed decisions Although this may be the case it is important to consider that the dataset was taken from a 1990 census (over 30 years ago) and as a result might well not be wholly applicable to today's market.

Real Estate Agents: Real estate agents can utilise accurate house price predictions to advise their clients on pricing strategies. By understanding the factors that influence house prices and using predictive models, agents can provide more accurate and competitive listing prices, leading to faster sales and higher client satisfaction. Additionally, agents can use these predictions to identify undervalued properties or potential investment opportunities for their clients.

Property Buyers: Accurate house price predictions assist property buyers in making informed decisions regarding their purchasing budget. By having reliable predictions, buyers can assess whether a property is overpriced or underpriced, enabling them to negotiate better deals. This information can also assist buyers with the identification of properties that might align with their budget and investment goals.

### 4.3 Limitations and Future Research

Our study had limitations that we recognize for potential improvements in future prediction models. Relying on a single dataset, incorporating additional diverse datasets from different sources could enhance model generalizability. Although we focused on comparing Random Forest and Linear Regression, other algorithms exist. However, our priority was developing a practical, accurate model within a university semester. The dynamic nature of housing markets, influenced by factors like natural disasters not in our dataset, suggests the need for time analysis techniques in future research. Despite these limitations, our study aimed at a comparative analysis of Linear Regression and Random Forest, not producing a model for precise short-term predictions.

### 4.4 Personal Summary

## 1. Midhun Mukundan

My personal summary is that I had the sole responsbility of the coding part and the idea of doing house price prediction as Machine Learning coursework was a group idea. Before coming to this we have tried Emotion detection but we haven’t got the desired output. Even though house price prediction is relatively a simple project we have used our creativity by predicting the accuracy using 2 to 3 models as well as adding new features into the dataset. We have stumbled upon some obstacles like not getting the desired accuracy using linear regression so tried changing some parameters but no change so finally come up with an idea of using a complex model and decided to use random forest. Initially there was trouble setting up the random forest model like getting highly overfitting model but by doing some research we had sorted it out and finally got the desired accuracy.

## 5. References

1. “About Linear Regression.” n.d. IBM. Accessed December 2, 2023. <https://www.ibm.com/topics/linear-regression>.
2. Müller, Andreas C., and Sarah Guido. Introduction to Machine Learning with Python : A Guide for Data Scientists, O'Reilly Media, Incorporated, 2016. ProQuest Ebook Central, <https://ebookcentral.proquest.com/lib/londonmet/detail.action?docID=4698164>.
3. <https://www.machinelearningnuggets.com/decision-trees-and-random-forests//>
4. <https://www.kaggle.com/datasets/camnugent/california-housing-prices?rvi=1>
5. <https://vitalflux.com/boston-housing-dataset-linear-regression-predicting-house-prices//>
6. <https://mageluer.github.io/blog/python-machine-learning/>