Southampton Solent University

FACULTY OF Business, Law, and Digital Technologies

Combining Machine Learning models to predict House Prices

An experimental study of machine learning and forecasting methods applied to California housing data.

Master's Thesis

Author: ISAAC AKE

Supervisor: Shadi Eltanani

MSC thesis submitted in fulfillment of the requirements for the degree of MSc. Artificial intelligence and Data Science at Solent University Southampton.

9th September 2022

DECLARATION OF AUTHORSHIP

This thesis is being submitted in partial fulfillment of the requirements of Southampton Solent University for the degree of Master of Artificial intelligence and Data Science

Acknowledgments

I want to thank Shadi Eltanani, who served as my primary supervisor, for guiding me through this project, Throughout the research effort, Shadi provided ongoing support and was always willing and happy to assist in any manner he could. Additionally, I want to express my gratitude to my friends and family for their encouragement and insightful comments on the study.

Table of Contents

Abstract

House price valuation is critical when it comes to real estate selections. Mortgage lenders and house buyers and sellers utilize valuation estimations to gauge risk and the reasonableness of an asking price. In this thesis, models for predicting changes in property prices are examined, as well as the factors that drive those changes.

In real estate listings, a set list of numerical characteristics is used to describe the property. Spatial data, or location information, is a very important part of predicting home prices because the same house can sell for very different amounts in different places. We will investigate how each of these types of data affects how well we can predict home prices. We'll show how each type of data helps models that predict prices. This thesis makes a big contribution to making a software artifact by showing how a software implementation can be used to improve an algorithm for predicting house prices.

The housing market could use a better mechanism for projecting house values. The housing price projection can help a house seller, buyer, or real estate broker make more educated decisions. The regression method can still be beneficial because of its speed and ability to give generally accurate predictions. Regarding the prediction of property prices, we conclude that machine learning is a potential, alternative technique.

We have used the California House Price Prediction data for the current study. The purpose of this study is to create a model of home prices by making use of the information that has been provided to produce predictions about the median house value in the state of California. The corresponding dataset for the project includes 20,640 municipalities in total.

In the current study, to predict the median house price, appropriate regression models where applied which include appropriate parametric regression like simple linear regression model, ridge, lasso regression model as well as other regression models like Decision Tree Regression and Random Forest Regression Model. The Random Forest model proved to be the most appropriate model giving the highest value of the R-square and the minimum Root Mean Square Error Value. Hence, it can be concluded that the Random Forest Model is the most appropriate model for this dataset and should be used for predicting housing prices.

Chapter one

1.0 Introduction

Because of the wide variety of prices and the resulting impact on the economy, this thesis aims to make predictions about the house price index. For the past few decades, the housing market has received a lot of academic attention. Real estate is so popular because it provides both a place to live and a source of income, according to Shiller (2005). The economy relies heavily on the housing industry, which is also interwoven with the financial sector. A decline in the housing market frequently has substantial impacts, and this decline has the potential to trigger economic crises and recessions. Over the course of history, several nations have experienced substantial price shifts, which were typically preceded by an all-time high and then followed by a precipitous decline. Price fluctuations have substantial implications on household welfare as well as business cycles and financial stability. As a result of these effects, indicators for financial regulatory bodies, central banks, and other economic stakeholders could be generated from price changes Bork & Mller, (2016)

Machine learning is a subfield of artificial intelligence (AI) that extracts information from data via the use of various technological tools and algorithmic processes. The field of big data is one in which machine learning techniques can be applied, as it would be extremely difficult to manually examine such enormous amounts of data. In the field of computer science, machine learning is an approach that, rather than focusing primarily on mathematical answers to problems, looks for algorithmic ones. The accuracy of several regression algorithms will be assessed by comparing their ability to accurately estimate home price changes to the number of unknown variables that influence the value to be anticipated. Home prices are affected by the features of each individual property.

Houses can have a wide range of features and pricing points depending on where they are located. For example, the price of a large house may be higher if it is in a wealthy, desirable area rather than a poor one. A wide range of pre-processing methods will be applied to the data collected from the experiment to improve the accuracy of the predictions that are produced. Some

variables will also be added to the local dataset so that further research may be conducted on the connection between these factors and the price at which the item is sold.

The data from the California Census have been released by the United States Census Bureau. These data include ten different sorts of metrics for each block group in California, such as the population, median income, and median home price, among other things. Additionally, the dataset acts as an input for project scoping and tries to identify the requirements, both functional and nonfunctional, that are necessary for it.

The purpose of this study is to create a model of home prices by making use of the information that has been provided to produce predictions about the median house value in the state of California. This model should be able to learn from the data and be able to anticipate the median home price in any district given those features, given all the other factors that have been taken into consideration. The United States Census Bureau disseminates sample data for the smallest feasible geographic regions, which are referred to as districts or block groups (a block group typically has a population of 600 to 3,000 people). The corresponding dataset for the project includes 20,640 municipalities in total. In the current study, to predict the median house price, we have used appropriate regression models which include appropriate parametric regression like simple linear regression model as well as other regression models like Decision Tree Regression and Random Forest Regression Model.

In the current study, it will be tried to create appropriate regression models to predict the median house prices along with some Exploratory Data Analysis done on the data to get an understanding of what the dataset looks like. In the report, the problem statement and objective section (related to describing the overall aim of the analysis and study to be done), Review of Literature Section (related to the review of different relevant kinds of literature), Methodology Section (Discussion of different methods used in the analysis – especially the regression methods), Results and Analysis Section (Section giving details of all the Exploratory Data Analysis and Regression) are presented.

1.1 Background

Estimating a market price might be difficult. Any item's pricing can alter based on conditions. Market strength, craftsmanship, advertising, and brand awareness affect the price. The "hedonic model" for predicting home prices is based on an economic theory proposed by Lancaster in 1966. According to this theory, every home is made up of a variety of characteristics that consumers purchase. Adding or removing any of these characteristics will fundamentally change the product being purchased and any price prediction model. affordability, and supply-and-demand factors (Rosen, 1974).

To estimate the buying price of a home, we must first create a prediction model and study the descriptive components that might be utilized for this purpose. This thesis examines the role that machine learning plays in this process.Since 2005, rising interest rates have caused the U.S. housing market to slow down a lot. A big investment bank called Lehman Brothers Holdings was hit especially hard and had to go bankrupt in 2008. This caused house prices to drop quickly, which, along with the subprime mortgage crisis, slowed the economy even more and lowered the value of assets. This, in turn, led to a drop in the value of the global housing market, which set off a global disaster (Park & Kwon Bae, 2015). So, economists started to pay more attention to spotting possible weaknesses that could threaten economic stability.

After the worldwide financial crisis of 2008, the real estate market remained in a downward trend for several years, notably in major cities, all the way up until the end of 2011. Since 2012, the housing market has been on an upward trend, which may be attributed to the diminishing supply of homes, the growing demand for homes, and the inevitable rise in home prices. Once again, this spurred economists and market analysts to focus their attention on developing more accurate forecasting models in order to shield the economy against threats that could trigger economic downturns (Park & Kwon Bae, 2015).

1.2 Research Relevance

Very little research has been done on the home price forecast model specifically to address the issue by employing machine learning techniques and developing application software for user interaction. This is even though these two things have been the focus of most of the research efforts.

Predicting the price of a market is not a new subject in the real estate market. Valuing a piece of real estate is important to a lot of people who are involved in or affected by the real estate market. Before a customer buys a house, the bank must figure out how much it costs. Agents in real estate need to set the right price so that the seller can get the most money possible. Lastly, private people want to know how much their house, or a possible new house is worth so they can decide when to sell or buy with more information.

Machine learning and big data may be used by businesses and organizations to extract value and knowledge from data. It is possible that the use of machine learning and big data will enhance the performance of these actors. When it comes to predicting real estate prices, machine learning can help banks, real estate brokers, and individuals make more informed decisions.

To produce accurate predictions of prices in the real estate market, it is necessary to overcome the hurdles posed by these opportunities. The data that will be used to produce the price forecasts must already exist, and the methodologies that will be used to make the price predictions will need to be tweaked. To realize the full potential of machine learning when applied to the real estate market, these obstacles need to be overcome.

1.3 Research Question

The primary purpose of this thesis is to evaluate how well machine learning methods perform in comparison to one another. The relevance and motivation for this study are the two primary reasons why this study was undertaken. The following is the key research question that we have posed for ourselves:

- Utilizing the data that was supplied, construct a model of housing prices that can accurately estimate the median value of a home in the state of California.

Despite the likelihood that the prediction mechanism needs computationally costly procedures, users should be able to promptly acquire forecasts utilizing the internet application. As a result, we plan to implement this application on both the client and the server side. It is the responsibility of the client side to handle the forecast visualization's that show the geographical variance. A widely used platform is necessary for us to reach the widest possible audience with our application. On the other hand, the server should take care of the prediction system, which includes the training of the dataset and the management of queries from the client. It should be built on a server with sufficient computational power and implementation libraries loaded.

Chapter Two

2.0 Review of Literature

In the current section, some related works relevant to fitting relevant models to California Housing Data have been discussed. The thesis adds to a modest but expanding body of research on machine learning in the housing market. There is a wide range of literature on predicting house prices, including conventional regression and autoregressive models. Only relevant machine learning literature is presented in this area.

It's crucial to determine a home's value accurately. The evaluation procedure is crucial for lending organizations when making loan choices. To avoid losing money on defaulted loans, these institutions must make sure the property values they are evaluating are proportionate to the loan amount. During the loan application process, lenders frequently want an inspection and appraisal. Numerous angles have been taken into consideration during the lengthy study of home valuation.

Considering the housing data available for Fairfax County, Virginia, Park, and Bae (2015) tackled the problem of predicting house prices. The authors created a house price classification model using machine learning techniques like Naive Bayesian, AdaBoost, and RIPPER to address the issue. Their disclosed findings showed that the RIPPER algorithm outperforms all other models. The authors of the study indicated that in addition to expanding the size of their small dataset, future research should consider a property's appraised worth and property taxes. To create a more reliable model and improve the body of existing literature, our current research took into account those recommendations and included all three components for the Volusia County dataset.

The fall of the home price bubble, which started the financial crisis in the United States in 2007 (Martin, 2011; Baker, 2008; Acharya & Richardson, 2009), may help to understand the relevance of anticipating changes in house prices. These findings highlight the importance of spotting the early warning indications of significant economic changes developing over time and show the significant impacts of housing market volatility or shocks on actual economic activity. This shows the disparities in trends across various American cities as certain cities saw exponential growth in their housing values throughout this time, while

other cities observed steady or hardly any price movements (Ferreira & Gyourko, 2012). House price forecasting is influenced and directly influenced by other key economic variables. The ability to predict housing prices has a direct impact on other important economic phenomena and economic systems

In a different study, Gu et al. (2011) investigated the issue of housing prices with the intention of predicting a house price model. To solve the model, a mix of support vector machines and genetic algorithms was suggested. The model was applied to a housing dataset that was gathered in China between 1993 and 2002. The results demonstrated that the G-SVM technique was more effective than the grey model. The issue with the U.S. real estate house price index was also addressed by Plakandaras et al. (2015). In this study, a brand-new hybrid forecasting technique that combines Support Vector Regression (SVR) and Ensemble Empirical Mode Decomposition (EEMD) was developed (SVR). Random Walk (RW), Bayesian Vector Autoregressive, and Bayesian Autoregressive are used to compare the produced solutions of their suggested model.

According to previous studies (Plakandaras, Gupta, Gogas, & Papadimitriou, 2015; Glaeser, & Nathanson, 2017; Granziera & Kozicki, 2015), using supervised machine learning approaches to estimate home values is effective. Machine learning is a relatively new method used in economics that aims to lower the level of uncertainty in problems involving prediction (Mullainathan & Spiess, 2017). Numerous forecasting models have been developed in recent years, providing insightful data on the housing market.

Another study examined the behavior of financial markets and attempted to forecast the daily bitcoin exchange rate (Mallqui & Fernandes, 2019). To address the difficulty of predicting the exchange rate of bitcoin, the authors suggested a machine learning-based approach. The solutions to cutting-edge articles in this broad field of study were used to validate the conclusions of this challenge. Spanish day-ahead electricity price prediction was examined by Dáz et al. in 2019.

A regression tree-based strategy has been suggested as a solution to the issue. Additionally, the dataset for this challenge, in particular the model variables, were taken from publicly available records on energy use. The model's findings provide a respectable level of accuracy for price formation prediction, and they support the use of non-linear analysis to forecast prices using independent variables.

Cao et. al. (2018) is an important study on the California Housing Price Data. As per the authors, there are a variety of elements that impact housing prices.

The present model for predicting home prices often falls under the category of what is known as a single predictor model. As per the author, this model's accuracy in making predictions is not optimal, and the over-fitting problem frequently occurs as a result of the noise in the data. An ensemble learning-based housing price prediction model that takes into account a number of different predictors is what this research suggests as a solution to these problems. Extra trees, random forest, GBDT, and XGB are the algorithms that have been chosen to serve as benchmarks for the purpose of determining how successful the suggested model is. According to the authors, the dataset that was used is the California house price information that is accessible online. The findings from the study indicate that the suggested technique has the potential to increase both the accuracy of predictions and the stability of those predictions in comparison to the other four single prediction models.

As per Wu (2020), over the course of the past few years, the price of real estate has been steadily climbing, and this trend, which is connected to the worries of both individuals and the larger community and has emerged as a major topic of discussion in recent times, is one of the most pressing issues facing people today. Because of this, it is of the highest need to develop precise projections about the price of real estate. Using data on house prices in the state of California, the goal of this research is to provide an answer to the question of how to estimate the average annual sales price of houses in the state of California by considering several different parameters.

The data are used to gather the primary distribution of housing expenses, and linear and lasso regression are used to analyze the factors that have an influence on those prices. The square footage of the floor, the number of rooms, the proportion of residents with low incomes, and the educational possibilities offered in neighboring communities are all elements that may have an effect. Research into the factors that affect the value of homes is not only very important but also necessary if one is serious about fixing the myriad of problems that are now besetting the real estate industry. According to the authors, they will not be able to manage the price of commercial housing in an efficient and reasonable manner unless they have a complete grasp of the micro mechanism of price creation in the

housing market as well as the factors that influence housing pricing. The authors opined that, without this knowledge, they will not be able to achieve the goal of efficient and reasonable price management.

As per Montero et. al. (2018), predictions of property values are becoming an increasingly common topic of discussion in the scholarly economics literature. In the past, hedonic models that are a-spatially linear, also known as intrinsically linear, were often used for the purpose of making forecasts about house values. However, research has demonstrated that geographical effects are an important factor in determining property prices. Both parametric and semi-parametric variations of spatial hedonic models are examined in depth over the course of this piece of writing.

As per Komagome-Towne et. al. (2016), using visualizations, the major objective of this work is to investigate the distinctions and parallels that exist across several pricing models for single-family houses in California. Specifically, this investigation will focus on the housing market in California using the California Housing Price Prediction Dataset. To carry out this research, the authors searched through Redfin.com and collected information on 5,142 unique listings. The sample consists of all single-family homes, townhouses, and condominiums that were sold in California over the five-year period starting in October 2012 and ending in October 2015 across the whole state. This span of time begins and ends with the months of October.

The authors have utilized, python, a programming language and software environment developed for statistical calculation, to handle and analyze data relevant to real estate transactions. To arrive at accurate estimates about the prices at which future transactions would take place is the ultimate objective. To choose a method for prediction, many regression models are analyzed and compared to one another to locate one that is a suitable fit. Methods such as multiple linear regression, k-nearest neighbors, tree-based methods (including decision trees, bagging, and random forests), and techniques for nonlinear regression are some of the ones that are described in this article. Other methods that are described include nonlinear regression techniques (splines and generalized additive models).

We compare and assess the performance of these different methods by first looking at the median proportion of inaccurate predictions. When making a prediction, it is important to

consider a variety of factors, some of which are commonly reported statistics. These statistics include the interior square footage, the size of the lot, the number of bedrooms, the number of bathrooms, the year the home was constructed, and the date it was sold. An examination of the relationship between location and property prices in the Pasadena area has been carried out using heat maps that have been superimposed on top of maps of the area. According to the data, the generalized additive models seem to have the best performance.

Additionally, the direction and amplitude of the relationships between the predictors and home prices were the same in all the models and techniques, suggesting that they were consistent. This is further evidence that the models and procedures are consistent. For example, when all other factors remain the same, the number of bedrooms has a startlingly negative impact on the price of a home. On the other hand, the interior square footage, the size of the lot, and the number of bathrooms all have a positive association with house pricing. It has been determined that the square footage, in combination with the geographical coordinates, is the single most important component and that it is accountable for a significant portion of the variation in price. According to the authors, the best median figure for the percent of error is around 10 percent, which suggests that our projections are often off by 10 percent from the amount that is actually paid. These findings provide insights into the factors that influence house values. Additionally, these findings provide sellers with the practical assistance they need to market their homes at reasonable prices. Additionally, these findings assist purchasers in preventing themselves from overpaying for their homes.

According to Ghatnekar et. al. (2021), these days, Machine Learning is becoming more important in a wide range of industries, including medical diagnosis and trading on the stock market. In a manner like the work that came before it, the authors have focused most of our efforts in this investigation on a specific application of machine learning. In this research, the author’s utilized machine learning methods to construct prediction forms for home prices based on actual data of real estate in California.

These forms were developed using data from the state of California. This article starts out with a general introduction to the learning algorithms that were used during the duration of the study on comparing property prices. Following that comes a description of a variety of different regression models, followed by a discussion of the parameter values that have

been optimized for each model. At the beginning of the study, a variety of Machine Learning models' results in terms of quantitative measures are compared. Next, it employs SHAP values to assess the relevance of several qualities that impact the worth of homes in Boston and California. These factors include location, square footage, and the number of bedrooms and bathrooms. In the end, it arrives at its conclusion concerning real estate pricing by the application of economic logic. The outcomes of the forecasts serve as the foundation for this article.

As per Neloy et. al. (2019), The amount of money needed to pay for apartment rent is dependent on a wide range of distinct factors. This study's objective is to explore several aspects of residential property to arrive at an accurate estimate of the rent that would be necessary monthly to occupy the property. To achieve the goal, a prediction model that is constructed via ensemble learning will need to be created. The authors made use of the Housing Price Prediction dataset for this purpose. This dataset includes information on the monthly rent as well as the numerous facilities that are offered in apartments that are situated in the city of California. The results not only shed light on the several types of categorical values that influence the machine learning models, but they also indicate the degree to which the cost of renting an apartment may be reliably anticipated and calculated. The identification of the factors that go into the calculation of the total cost of renting an apartment in California is another aim of this study. The authors made use of the Advanced Regression Techniques (ART) to increase the accuracy of our prediction by comparing them to the numerous aspects of an apartment in order to construct a model that is suitable. This allows us to create a model that meets our needs.

Chapter Three

3.0 METHODOLOGY

The purpose of this thesis is to employ machine learning techniques and competitive regression models to estimate the median house price value of housing in the district. Fig. 1 depicts the research framework for the home price prediction problem. This system's five key building blocks are data collection, data preparation, feature processing, model training, and model evaluation. Each block in the diagram is explained in detail in the following subsections.

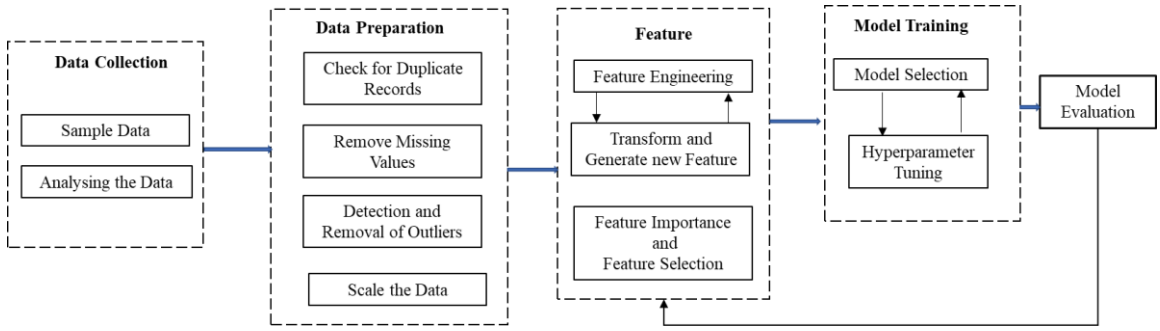


Fig. 1. Research methodology for the housing price issue

The No Free Lunch Theorem says that algorithms act differently when they are used in the same situations. This study looks at how well different algorithms, such as the linear, Lasso, Ridge, Random Forest, and Decision trees algorithms, can predict house prices. So, the goal of this study is to learn more about how regression techniques are used in machine learning.

Processing the datasets that have been provided is also required to improve performance. Because every house possesses specific characteristics that are utilized to determine an estimate of its price, this is accomplished by first determining the characteristics that are required, and then applying one of the selection procedures to get rid of the factors that

are not acceptable. As a result of the fact that these characteristics may or may not be present in all houses, they may not have the same influence on the pricing of houses, which results in results that are misleading.

3.1 Introduction to Machine Learning

This section talks about machine learning and related topics. It explains how to choose a model and how to measure how well machine learning algorithms work. Also, a brief history of how machine learning has been used will help show the strengths, challenges, and opportunities of this technology.

The terms ‘machine learning and 'artificial intelligence are frequently used interchangeably, however machine learning is a subfield/type of AI. Common synonyms for machine learning include predictive analytics and predictive modelling. Arthur Samuel, an American computer scientist, coined the phrase "machine learning" in 1959 to describe "a computer's ability to learn without being expressly programmed. Machine learning algorithms take and evaluate input data to predict output values that are within an acceptable range. As new data is fed into these algorithms, they learn and adjust their operations to improve performance, gaining 'intelligence' in the process.

As a part of AI, machine learning includes a wide range of advanced methods that let computers find and predict patterns in data, even if they don't know anything about the data ahead of time. These methods use math, statistics, and computer science. In contrast to traditional algorithms, these techniques are designed to "learn" from the data given and adjust the model parameters and variable weights without further human intervention.

During the last few decades, a wide array of unique algorithms for machine learning have been developed. It is vital to classify various methods and applications because there is such a wide variety of both types. According to Russell et al. (2010), there are three primary "learning kinds," which are supervised learning, unsupervised learning, and reinforcement learning.

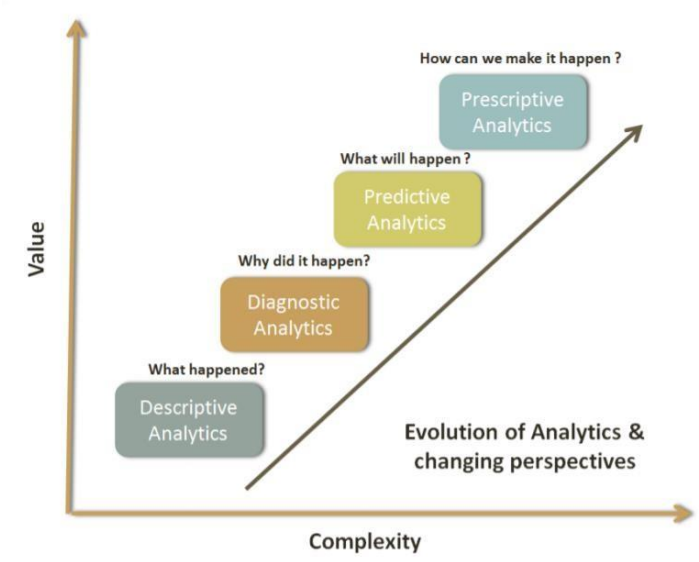
Supervised learning is a type of learning in which the algorithm is given input and output data, and an error function is used to map inputs to outputs. This function is the most important part of learning because it helps you get closer to the real function. Learning

means that the algorithm's parameters are changed, and the weights of the input variables are changed. Regression and classification algorithms are both types of supervised approaches. Regressions use continuous variables as input to get a numerical vector as an output, while classification models label the continuous variables as input.

Unsupervised learning is a type of machine learning technique that is part of a larger category of learning programs that are designed to solve challenges with inadequate information. Without having any prior knowledge of the results, these algorithms devise rules that are determined by the connections between the data points.

The term "reinforcement learning" refers to a style of learning in which the primary focus is on acquiring knowledge by way of positive and negative reinforcement in the context of an ever-changing setting. Those behaviours that contribute to the successful completion of the objective are rewarded, while those that will not do so are penalized. After receiving adequate training, the algorithm selects the most effective sequence of behaviours that will lead to rewards to attain the highest possible level of performance.

The image below highlights the usefulness of modern analytic approaches, such as machine learning, which enable research to progress beyond describing past occurrences and explaining their causes to predicting future events. By continuously learning from historical data, these analytic methods can even be utilised to determine the optimal method for achieving the intended outcome.



Even though machine learning has been getting a lot of attention lately, many of these techniques have been used in science for decades. One reason why it's getting more and more popular is that there's more and more computer power to support these complicated processes.

Another reason is that the internet of things (IoT) and big data are getting a lot of attention right now. Gartner Research (2018b) describes the Internet of Things as "the network of physical objects that contain embedded technology to communicate and sense or interact with their internal states or the outside environment." In other words, IoT connects everything to everything, including houses, smartphones, cars, factories, and so on.

Furthermore, with the rise of social media and smartphones, we've seen a significant increase in the amount of data generated, which is known as big data. Variability, veracity, veracity, value, and variability are all characteristics that Gandomi and Haider (2015) consider to be big data characteristics. Machine learning, on the other hand, is attracted by the data's volume, accuracy, and monetary value. Detecting new patterns in an existing big amount of data is what we mean when we say that the system receives new data frequently enough to derive value from it.

The methods that are used in machine learning are diverse and can be used to a variety of contexts where there is data available. High-frequency trading, playing chess, and driverless driving are just few of the current applications of artificial intelligence. Other applications of machine learning are already serving as personal assistants in a lot of different people's homes. Among these are Amazon's Alexa and Apple's Siri, both of which are voice recognition programs, as well as efficient spam filters built into users' email accounts.

However, Baer and Kamalnath (2017) point out some of the difficulties that machine learning practitioners are having to deal with. It's important to note that machine learning is not an easy operation, but rather a complex piece of software. Even though the technological side is quite complicated, many of the solutions have already been implemented. However, the danger rests in the user's usage of the algorithms as a black box, rather than the algorithms themselves. Algorithms have flaws that can lead to inaccurate or even incorrect forecasts if they are used without considering their defects and limitations.

Moreover, aside from mistakes made by people, biases can also be caused by how often the environment changes. This is because past data might not be the best way to predict what will happen in these areas in the future. Problems with performance could be caused by patterns in data that have never been seen. Negative interest rate policies are an example of something that no one expected. So, algorithm engineers and data scientists need to keep an eye on how their software is changing and make changes to it often to reduce biases caused by statistics, people, and other things.

Positively, machine learning algorithms base their conclusions on established mathematical or statistical notions and are not biased. Once the appropriate machine learning model has been built and deployed, it delivers significant time savings because computers can rapidly process various inputs while continuously improving performance.

3.2 Machine Learning Framework

In the subject of machine learning, there are numerous algorithms that may be used to a variety of issues. The performance of these algorithms may be comparable, but they may differ in terms of computational costs, training time, or assumptions. Thus, the most difficult aspect is determining which machine learning technique is most suited to handle the problem at hand. As data requirements and availability vary widely in real-world applications, it is more common than not for scenarios to be unique.

This thesis proposes a framework to save research time to facilitate the process of selecting the best algorithms. This approach begins by introducing Chapman et alcross-industry.'s standard method for data mining (CRISP-DM) (2000). Before presenting a more specialized framework, this broad strategy tries to improve comprehension of the data mining and machine learning process. The latter outlines a process flow for supervised regression algorithms, the most appropriate machine learning technique for predicting individual house prices.

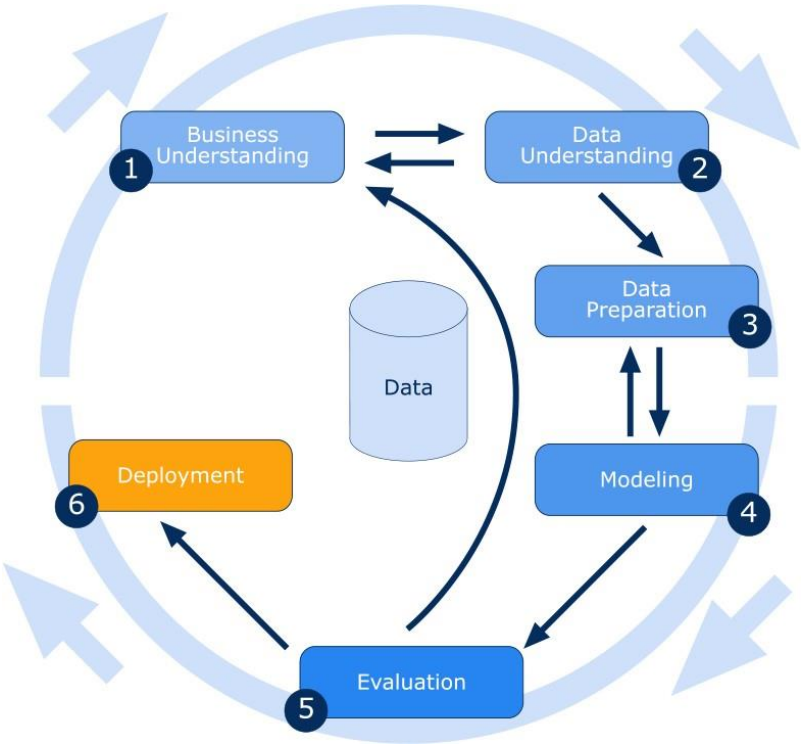
3.3 Machine Learning Setup

Using the open programming language Python, the empirical investigation applies data from the actual world to five machine learning algorithms. Five machine learning algorithms are

employed in the thesis. linear, lasso, ridge, and random forest. because each of these methods involves a repeating set of machine learning procedures. This completes the first phase in the machine learning process, which is the choosing of an algorithm.

3.4 General Approach

As seen in the Figure below, finding the best potential answer to a problem through iterative problem solving is not an easy task in data science. At the beginning of the data science process, the researcher must be aware of the business and its ramifications. Thus, a sound research topic can be formulated to guide the project's direction and confine its scope. As a result of a well-defined research issue, a machine learning algorithm can be given specific instructions on how to carry out its work.



Chapman et al. describe the stages of the CRISP-DM process (2000).

As soon as the researcher knows what the problem is, the next step is to deal with the data. This process involves more than just knowing what the data is, how it works, and what it's worth. Before data can be analyzed, it must be collected. So, once the goal of the research

is clear, it is important to investigate data sets about the topic. Even though the quality of the data is important, the main goal is to get a broad picture of all the possible data sources and features. After looking into these options, it's important to investigate the datasets and see if they can help reach the research goal on their own or together. Depending on the results of the search for data, a project may need more data or may be possible with the data that is already available.

Preparing or pre-processing data after it has been gathered is a common term for this process. The expression "garbage in, trash out," used frequently in the information technology business, sums up the need for good data quality. According to this theory, even if the data is well-maintained, any machine learning algorithm will struggle to find patterns. Adding irrelevant variables to analysis is an example that may be found in practically any real-world endeavor.

Consequently, the algorithm may pick up on patterns that aren't real but exist because of a surplus of data. These variables may cause algorithms to overlook patterns in the data, which may have a negative impact on forecast performance. Machine learning techniques rely on accurate data to make solid predictions, so maintaining excellent data quality is critical.

3.5 Simple Linear Regression

Regression models, which include describing the relationship between variables by "fitting a line to the data that has been gathered," may be used to explain the link between the variables (Zou et. al., 2003). The line that is used in linear regression models is straight, whereas the line that is used in logistic and nonlinear regression models is curved. Linear regression models employ the straight line which is used to find out the relationship among the variables. Regression makes it feasible to estimate what occurs to a dependent variable in response to changes in one or more independent variables. This may be done by comparing the two sets of variables in a dataset.

Utilizing a technique known as simple linear regression, one can get an estimate of the relationship that exists between two quantitative factors. One may use simple linear regression if he/she is interested in finding out any of the following things:

- The degree to which the values of two variables are tightly connected to one another (e.g., the relationship between rainfall and soil erosion).
- The value of the variable that is reliant on the independent variable when that variable is at a certain value (e.g. the amount of soil erosion at a certain level of rainfall).

The assumption that is made by linear regression is as follows:

- A straight line is the line of best fit through the points in the data, indicating that the connection between the independent variable and the dependent variable is linear (rather than a curve or some sort of grouping factor).

3.6 Decision Tree Regression

A regression or classification model may be constructed using a decision tree, which presents the data in the form of a tree hierarchy. It takes a dataset and divides it up into smaller and smaller sections while simultaneously developing an associated decision tree in an incremental fashion (Xu et. al., 2005). The completed task produces a tree that has decision nodes as well as leaf nodes. A decision node, such as Outlook, may have two or more branches, such as Sunny, Overcast, and Rainy, with each branch reflecting a different possible value for the analysed property.

A choice on the numerical objective is represented by each leaf node, such as "Hours Played." The root node is the decision node at the very top of a tree, and it corresponds to the predictor with the highest accuracy. Data of either a categorical or numerical kind may be processed using decision trees.

J. R. Quinlan was the one who first created the core methodology known as ID3, which is used for the construction of decision trees. This method does not do any backtracking as it conducts a top-down, greedy search through the space of probable branching options. By switching the Information Gain variable for the Standard Deviation Reduction variable in the ID3 method's decision tree design, it is feasible to create a decision tree for regression to be used in statistical analysis.

A decision tree may be formed in a top-down approach starting at a root node by first partitioning the data into subsets that include instances with values that are most similar to one another (homogenous). The standard deviation is a useful tool that may be used to determine the degree to which a numerical sample is homogeneous. If the numerical sample is completely stable throughout, the standard deviation will end up being equal to 0 in this scenario.

3.7 Random Forest Regression

Random Forest Regression is a supervised learning technique that utilizes the ensemble learning approach to perform regression (Segal, 2004). The ensemble learning method is a methodology that combines predictions generated by multiple machine learning algorithms to produce a forecast that is more accurate than a single model's prediction.

During the training phase of a Random Forest operation, several decision trees are built. After this phase is complete, the Random Forest generates the mean of the classes as the prediction for all the trees. To have a better understanding of how the Random Forest algorithm works, let's go over each step one at a time:

1. Select k data points at random from the larger set that served as the basis for the training.
2. Construct a decision tree with these k data points serving as the nodes and branches, respectively.
3. Repeat steps 1 and 2 until one has built the number of trees that you specified in step N. This will take many iterations.
4. In the case of a new data point, one should first instruct all of your N-tree trees to make a prediction about the value of y for the data point in question, and then he/she should assign the new data point to the average of all of the predictions made regarding y. This will ensure that the value of y is as accurate as possible.

The Random Forest Regression model has both robustness and accuracy in its analysis. It is often capable of achieving outstanding outcomes in a broad range of scenarios, including ones that entail non-linear interactions. On the other hand, there are a few drawbacks, such as the fact that there is no interpretability, that overfitting is an easy possibility, and that one must choose how many trees to put in the model.

3.8 Lasso Regression

Lasso (Tibshirani, 1996) is a linear model that is widely used for the feature reduction process. This is accomplished by setting the coefficients of some uninteresting attributes to zero in the model. Lasso's target function can be defined as a linear model with a regularisation component added. Lasso's objective function can be defined mathematically as a linear model with an additional regularisation element.

The Lasso objective function is $min\ w\ 1\ 2nsamples\ ||Xw - y||\ 2\ 2 + \alpha\|w\|1$.

Minimizing the least square penalty is the goal here with $\alpha\|w\|1$ added, where α is a constant and $\|w\|1$ is the $l1$ -norm of the coefficient vector.

3.9 Ridge Regression

In 1962 Hoerl created the Ridge Regression, an L2-norm regularised regression method. It is an estimating method for managing collinearity without removing any variables from the regression model. Multicollinearity is a common problem in multiple linear regression, causing least square estimation to be unbiased and its variances to be far off the true value. Ridge Regression, by introducing a small amount of bias to the regression model, reduces the standard errors and shrinks the least square coefficients in the direction of the parameter space.

Ridge formula is $R = Min\ (sum\ of\ squared\ residuals + \alpha * slope\ 2\ )$. Where $Min\ (sum\ of\ squared\ residuals)$ is the Least Squared Error, and $\alpha * slope\ 2$ is the penalty term that Ridge adds to the Least Squared Error.

The sum of squared residuals is minimized when Least Squared Error is used to calculate parameter values. Ridge, on the other hand, reduces the sum of squared residuals while determining the parameter values. It also includes a punishment term, which specifies both the length of the slope and the severity of the penalty. Additionally, the slope asymptotically approaches zero when the is increased. The cross-validation approach is used to determine, much like with Lasso. Ridge thereby makes the prediction less sensitive by reducing parameters and reducing variance.

Chapter Four

4.0 Dataset Understanding

The California house price dataset was used to validate our methods in this study. The US Census Bureau has released Census Data for California, which has 20640 records. The sample dataset contains 10 distinct metrics for each Californian block group, such as population, median income, and median housing price. The median house value attribute of the dataset will be predicted utilizing the various features as independent variables

The dataset contains the following fields or variables –

- Longitude
- Latitude
- Housing Age Median Total Rooms
- Count of Bedrooms
- Population
- Households
- Ocean proximity
- Median Income Median House Value

In this challenge, the median house value must be anticipated.

Data cleaning and analysis are covered in the first section according to EDA and data cleansing. ipynb.

The second is machine learning model training, which is covered in Training Machine Learning Algorithms. ipynb.

EDA and Data Cleaning

I have performed the following data changes and exploratory data analysis.

- Adding fresh features

A thing added to something else to complete or enhance it

- Taking out anomalies

Irregularity recognition is the distinguishing proof of intriguing occasions, things, or perceptions which are dubious because they vary altogether from standard ways of behaving or designs. Abnormalities in information are likewise called standard deviations, anomalies, clamor, curiosities, and special cases.

- modifying distorted features

A mutilated design can, by definition, be defined as far as its deviations from a known 'parent' structure and has a space-bunch balance that is a subgroup of the evenness of the parent. Ordinarily, twists lower balance and increment primary intricacy.

- Multicollinearity testing

Multicollinearity happens when autonomous factors in the relapse model are exceptionally connected to one another. It makes it hard to decipher of model and furthermore makes an overfitting issue. A typical suspicion individual test prior to choosing the factors for the relapse model.

4.1 DATA PREPROCESSING

When data is pre-processed, it is cleaned up and ready to be used by machine learning algorithms. Data pre-processing techniques are used on data that is raw and can't be analyzed yet. Using different methods, data records were changed into a format that can be used for machine learning analysis.

We cleared the dataset of invalid and missing values before doing an iterative analysis of the data. Data Wrangling and Data Munging are related words used in the Data Science field; data wrangling/data munging are strategies used to turn raw data into a format that is best for using the data. In our instance, we transformed textual information like Ocean proximity into binary variables, this technique resulted in better analytical results

Model development requires data pre-processing. Evaluation indicators have varying

degrees and units. This affects data analysis results. Data standardization (normalization) is needed to compare data indications and reduce level differences.

4.2 Exploratory Data Analysis

First, the dataset has been loaded in Python and a sample dataset is shown.

	longitude	latitude	housing_median_age	total_rooms	total_bedrooms	population	households	median_income	median_house_value	ocean_proximity
0	-122.23	37.88	41	880	129.0	322	126	8.3252	452600	NEAR BAY
1	-122.22	37.86	21	7099	1106.0	2401	1138	8.3014	358500	NEAR BAY
2	-122.24	37.85	52	1467	190.0	496	177	7.2574	352100	NEAR BAY
3	-122.25	37.85	52	1274	235.0	558	219	5.6431	341300	NEAR BAY
4	-122.25	37.85	52	1627	280.0	565	259	3.8462	342200	NEAR BAY

Figure 1 – Sample Dataset

Next, the field types of each of the variables are displayed.

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 20640 entries, 0 to 20639
Data columns (total 10 columns):
 #   Column              Non-Null Count  Dtype
---  ------              --------------  -----
 0   longitude           20640 non-null  float64
 1   latitude            20640 non-null  float64
 2   housing_median_age  20640 non-null  int64
 3   total_rooms         20640 non-null  int64
 4   total_bedrooms      20433 non-null  float64
 5   population          20640 non-null  int64
 6   households          20640 non-null  int64
 7   median_income       20640 non-null  float64
 8   median_house_value  20640 non-null  int64
 9   ocean_proximity     20640 non-null  object
dtypes: float64(4), int64(5), object(1)
memory usage: 1.6+ MB
```

Figure 1 – Field Types of the Dataset

Next, the summary statistics of the dataset have been calculated and it is displayed in Figure 3.

	longitude	latitude	housing_median_age	total_rooms	total_bedrooms	population	households	median_income	median_house_value	o
count	20640.000000	20640.000000	20640.000000	20640.000000	20640.000000	20640.000000	20640.000000	20640.000000	20640.000000	
mean	-119.569704	35.631861	28.639486	2635.763081	535.284351	1425.476744	499.539680	3.870671	206855.816909	
std	2.003532	2.135952	12.585558	2181.615252	420.053240	1132.462122	382.329753	1.899822	115395.615874	
min	-124.350000	32.540000	1.000000	2.000000	1.000000	3.000000	1.000000	0.499900	14999.000000	
25%	-121.800000	33.930000	18.000000	1447.750000	292.000000	787.000000	280.000000	2.563400	119600.000000	
50%	-118.490000	34.260000	29.000000	2127.000000	431.000000	1166.000000	409.000000	3.534800	179700.000000	
75%	-118.010000	37.710000	37.000000	3148.000000	643.250000	1725.000000	605.000000	4.743250	264725.000000	
max	-114.310000	41.950000	52.000000	39320.000000	6445.000000	35682.000000	6082.000000	15.000100	500001.000000	

Figure 3 – Summary Statistics

Next, we are checking whether there are any null values in the dataset or not.

```
longitude               0
latitude                0
housing_median_age      0
total_rooms             0
total_bedrooms        207
population              0
households              0
median_income           0
median_house_value      0
ocean_proximity         0
dtype: int64
```

Figure 4 – Null Value Detection

It is seen that there are 207 null values in the total_bedrooms and then we have imputed the data using the mode.

```
longitude               0
latitude                0
housing_median_age      0
total_rooms             0
total_bedrooms          0
population              0
households              0
median_income           0
median_house_value      0
ocean_proximity         0
dtype: int64
```

Figure 5 – Null Value Detection after Imputation

As we can see from figure 5 is that we are not finding any null value after imputation.

Next, it is tried to see the distributions of the different variables available in the dataset.

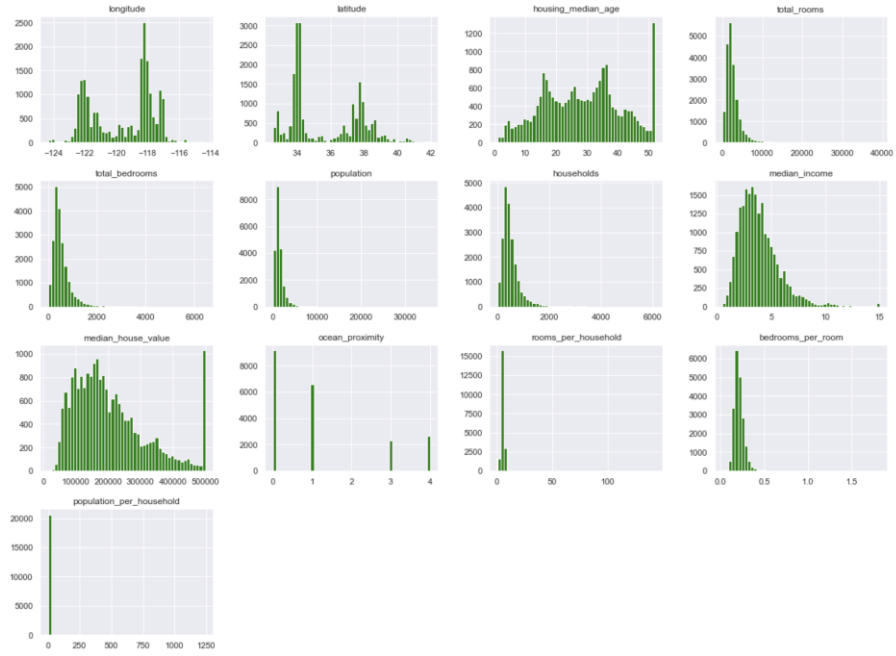


Figure 6 – Histograms of Different Variables

The histograms of the different variables are seen. It is seen from the graphs that not of the variables except bedrooms_per_room are normally distributed and in some cases, they have multi-modal distributions.

The correlated study has been performed and it is seen that none of the variable pairs are highly correlated.

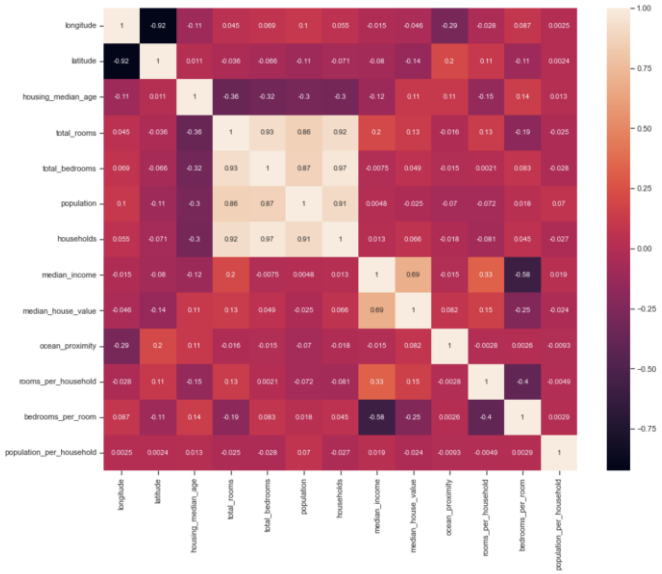


Figure 7 – Correlation Analysis

Next, it is tried to see whether there are any outliers in the median_house_value field. It is seen that there are some outliers in the dataset.

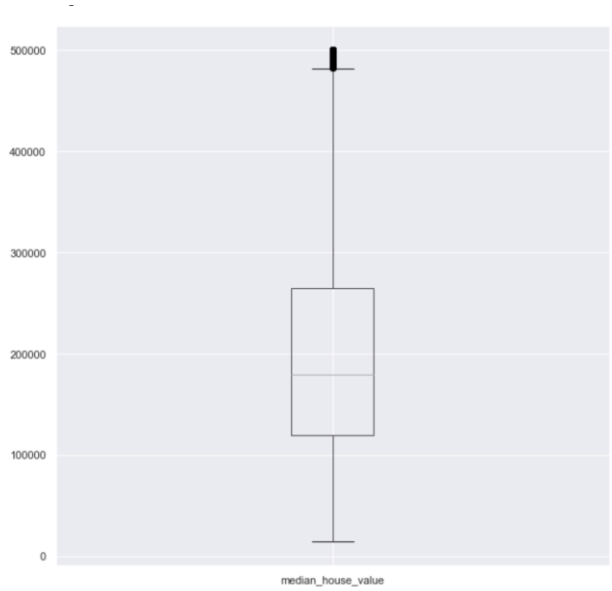


Figure 8 – Box-plot of Median House Price

We are re-creating the box-plot of Median House Price in connection to the ocean proximity.

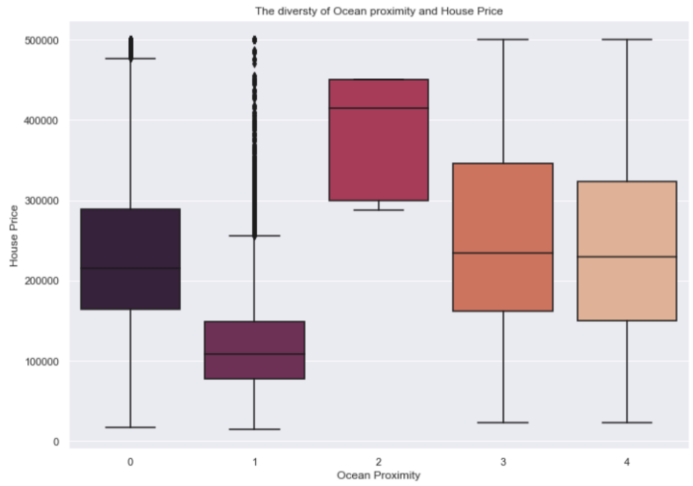


Figure 9 – Boxplot of Median House Price Vs Ocean Proximity

Here, it is seen that for ocean-proximity value one and zero there are many outliers. It means, if the property is located nearby the ocean, there is a tendency for the house prices to touch exceptionally high, however, as and when it moves away from the ocean, the prices come to certain range.

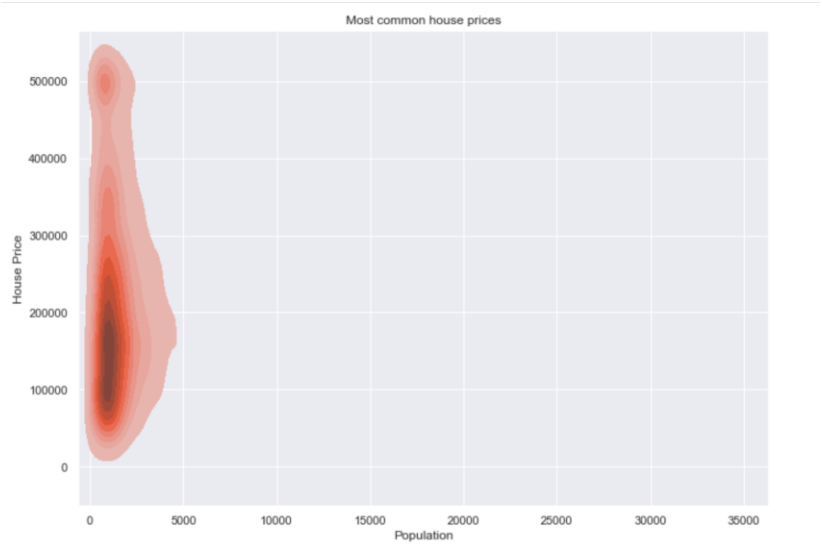


Figure 10 – Most Common House Prices

We are seeing that most common house prices are spread across the areas with population size less than 5000.

We have tried to plot the latitude longitude of the areas with the population size and we are getting the Figure 11.

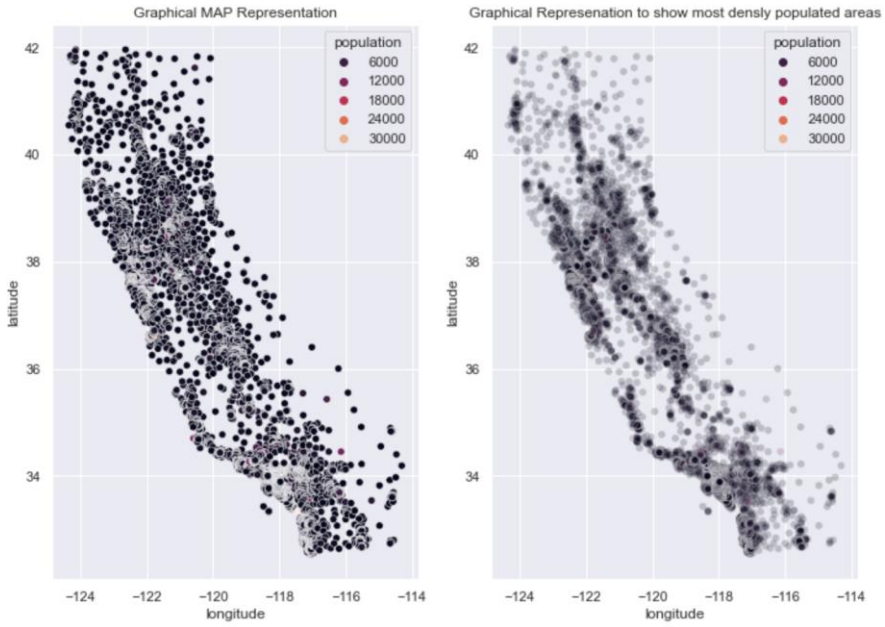


Figure 11 – Graphical Map Representation of the Latitude Longitude with respect to Population Size

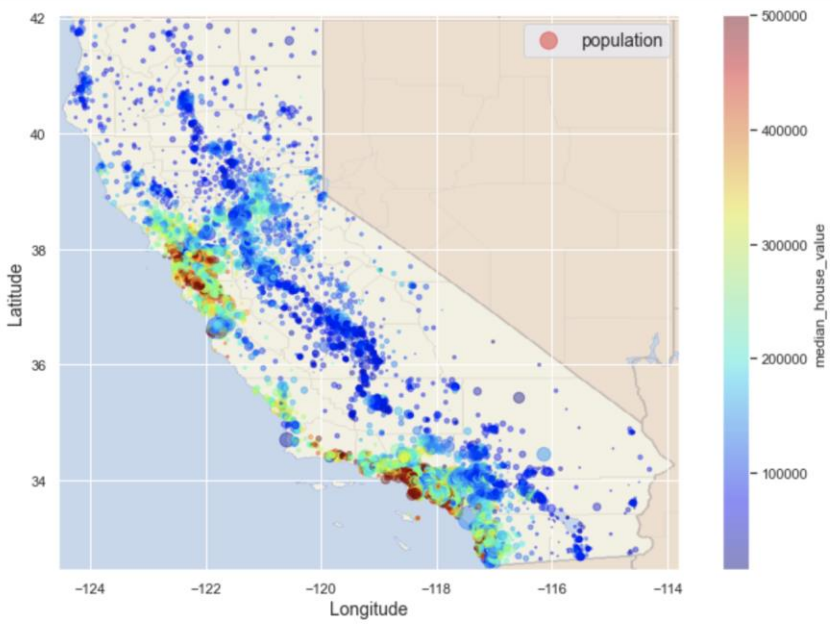


Figure 12 – Graphical Map Representation of the Latitude Longitude with respect to Population Size

It has been tried to plot median income Vs the median housing price in the Figure 13.

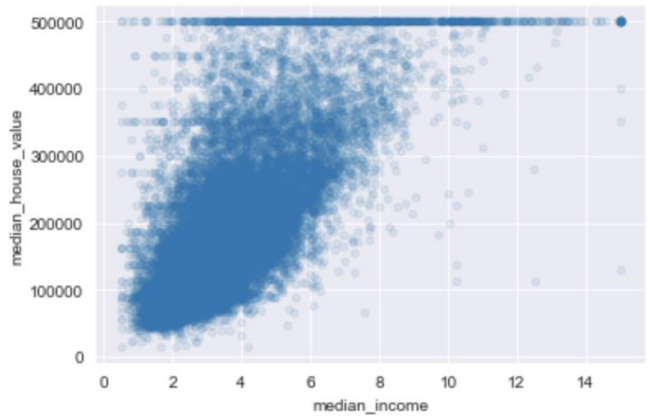


Figure 13 – Median Income Vs Median Housing Price

Here, it is seen that median income is somewhat related to the median housing price and when the income is increasing the median house price is also increasing to certain extent.

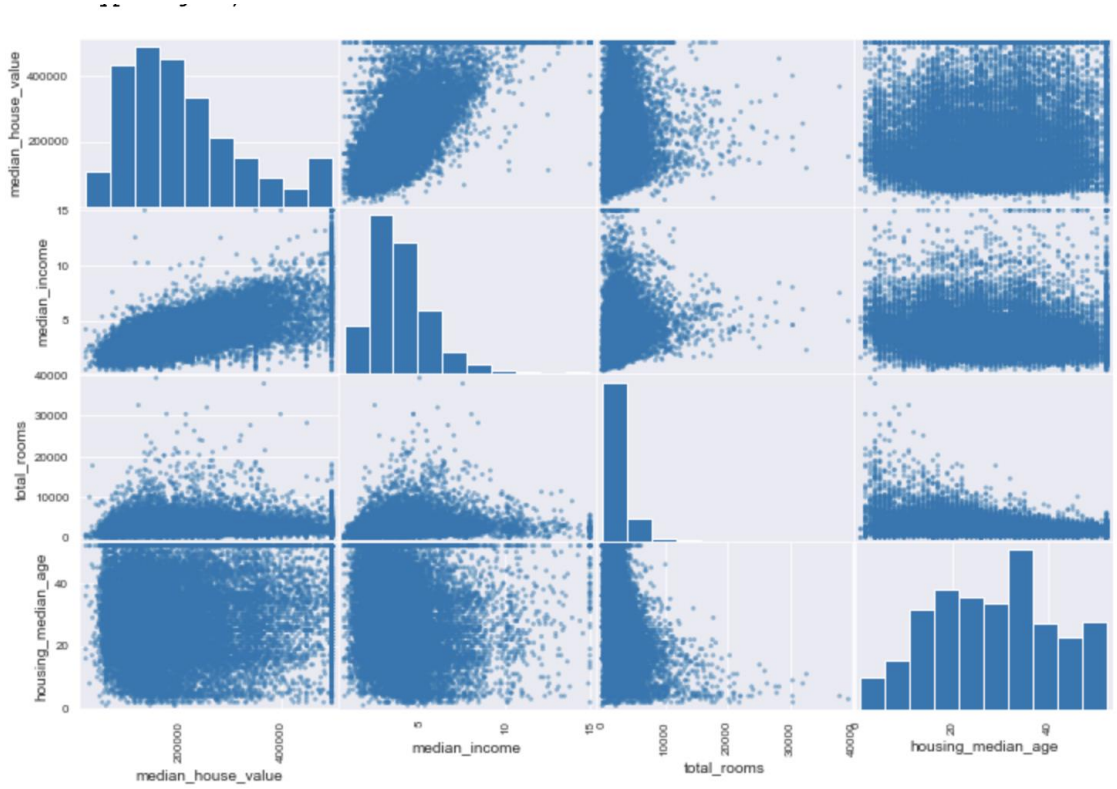


Figure 14 – Scatter Plots

We are seeing here that only median income has some kind of positive linear relationship with the median house prices; however, other variables don't have such relationships directly.

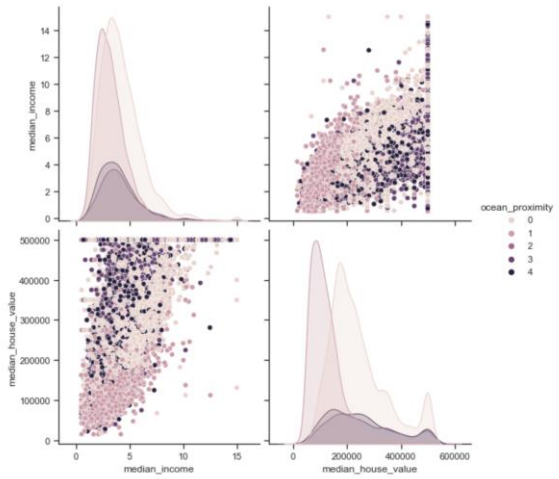


Figure 15 – Pair-Plots

Chapter Five

5.0 Model Building and Results

Next, it has been tried to develop the linear regression models to predict the median house price. After the pre-processing method, we divide the dataset into training, validation, and testing subsets after pre-processing it. Each partitioned dataset is further subdivided into dependent and independent variables, denoted by the letters X and Y.

First, we are fitting the same for the linear regression. The results obtained from the same is presented below.

```
In [14]: linreg=LinearRegression()
         linreg.fit(x_train,y_train)

Out[14]: LinearRegression()

In [15]: y_predict = linreg.predict(x_test)

         print(sqrt(mean_squared_error(y_test,y_predict)))
         print((r2_score(y_test,y_predict)))

         0.6055289178542321
         0.6277832896594451

In [49]: cross_val_score(linreg, x_train,y_train, cv=10)

Out[49]: array([0.61119266, 0.65098956, 0.64614242, 0.6443398 , 0.60190141,
                0.64425674, 0.65599585, 0.64130878, 0.62758149, 0.62776261])

In [50]: reg_score = r2_score(linreg.predict(x_test),y_test)
         reg_score

Out[50]: 0.4497189084032044
```

Figure 16 – Linear Regression Results

According to the results obtained, the R-square value for the linear regression model fitted is coming to be around 0.45 which is not a very good fit. The mean square error for the model is coming to be 0.61.

The coefficient for the model is as follows in Figure 17 in order.

```
In [51]: linreg.coef_

Out[51]: array([-0.73979315, -0.78527293,  0.1266255 , -0.12973751,  0.2913666 ,
                -0.38598704,  0.2628795 ,  0.6614392 ,  0.00541676])
```

Figure 17 – Coefficients of the Linear Regression Fit

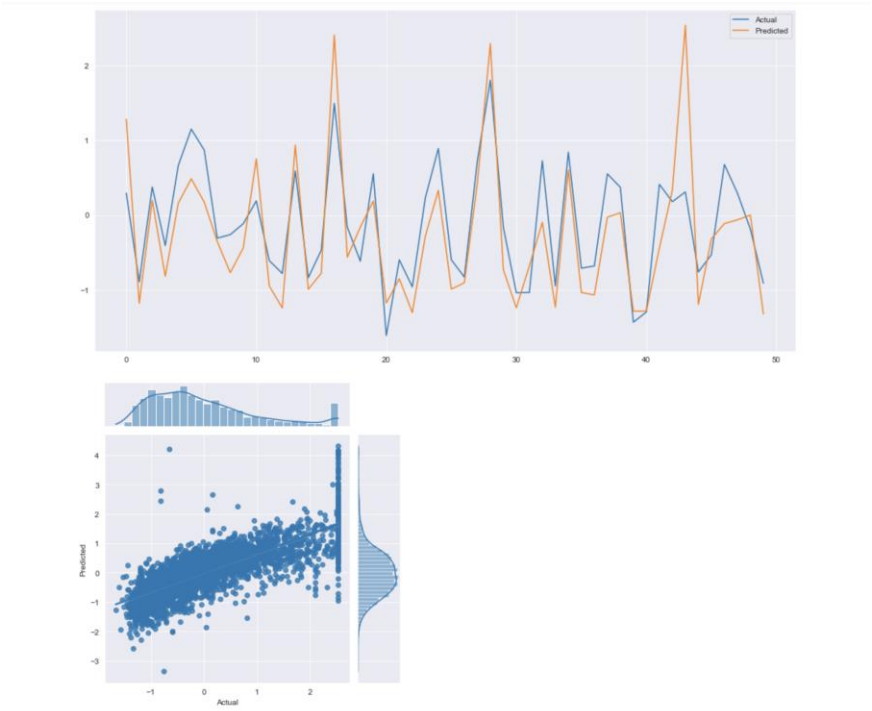


Figure 18 – Predicted Vs Actual in Linear Regression

Some plots have been created to show predicted Vs actual values (Figure 18).

We have used the Lasso Regression also and the we have got the following results.

```
lassoreg=Lasso(alpha=0.001,normalize=True)
lassoreg.fit(x_train,y_train)
lasso_pred = lassoreg.predict(x_test)
r2_lasso = r2_score(y_test, lasso_pred)
rmse_lasso = np.sqrt(mean_squared_error(y_test, lasso_pred))
print("R^2 Score: " + str(r2_lasso))
print("RMSE Score: " + str(rmse_lasso))
```

```
R^2 Score: 0.4747534206169961
RMSE Score: 0.719314096707071
```

Figure 19 – Lasso Regression Results

We are getting the Lasso Regression results as follows.

R-square value = 0.47

RMSE value = 0.72

The results of the Lasso regression are showing comparatively better than the Linear Regression.

The Lasso regression fit is shown below.

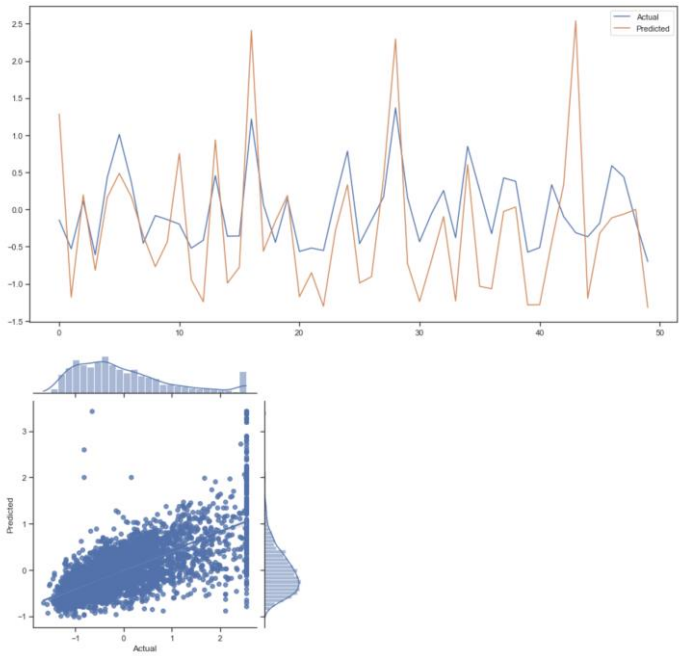


Figure 20 – Lasso Regression Fit

Next, we have used the Decision Tree regression and obtained the following results for the same.

```
predictions = tree_reg.predict(x_test)
lin_mse = mean_squared_error(y_test,predictions)
lin_rmse = np.sqrt(lin_mse)
print('rmse value is : ',lin_rmse)
```

```
rmse value is :  0.5934191390048503
```

Figure 21 – Decision Tree Regression Fit Results

Here, we are seeing that RMSE value is 0.59 which is better than the Lasso regression and hence, it can be concluded that Decision Tree is providing better fit than both Linear Regression and the Lasso Regression.

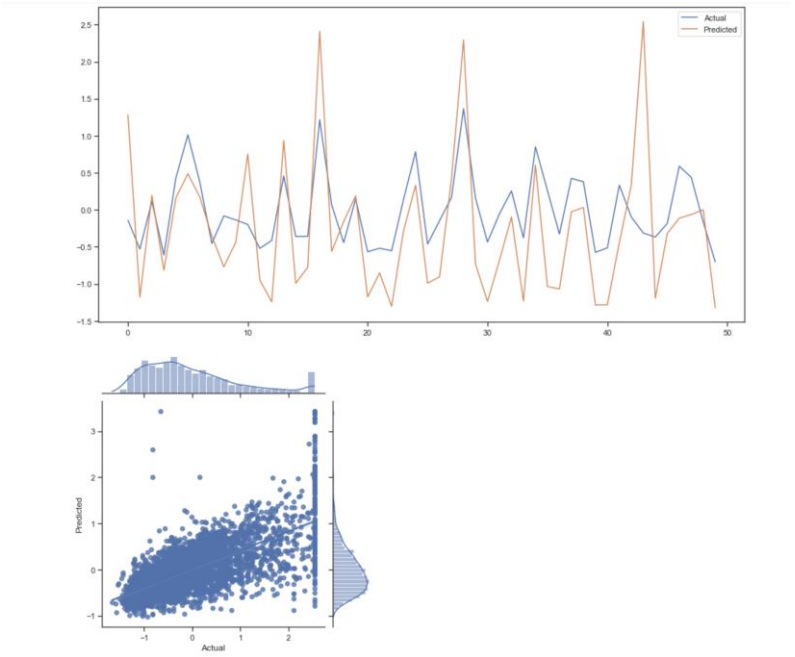


Figure 22 – Decision Tree Regression Fit

Next, we are trying to fit the Random Forest Regression Model.

```
Out[43]: RandomForestRegressor()

In [94]: rand_predictions = forest_reg.predict(x_test)
         lin_mse = mean_squared_error(y_test,rand_predictions)
         lin_rmse = np.sqrt(lin_mse)
         print('rmse value is : ',lin_rmse)

         rmse value is :  0.42535094886031216

In [95]: r2_score(forest_reg.predict(x_test),y_test)
Out[95]: 0.768832420463045
```

Figure 23 – Random Forest Model Fitting

From the Random Forest regression model, we are seeing that it is giving 0.42 as the RMSE value and 0.77 as the R-square value. This is the highest R-square value and the lease RMSE value we are getting among all the competitive models. Hence, it can be concluded that the Random Forest model is giving the best model in the current case.

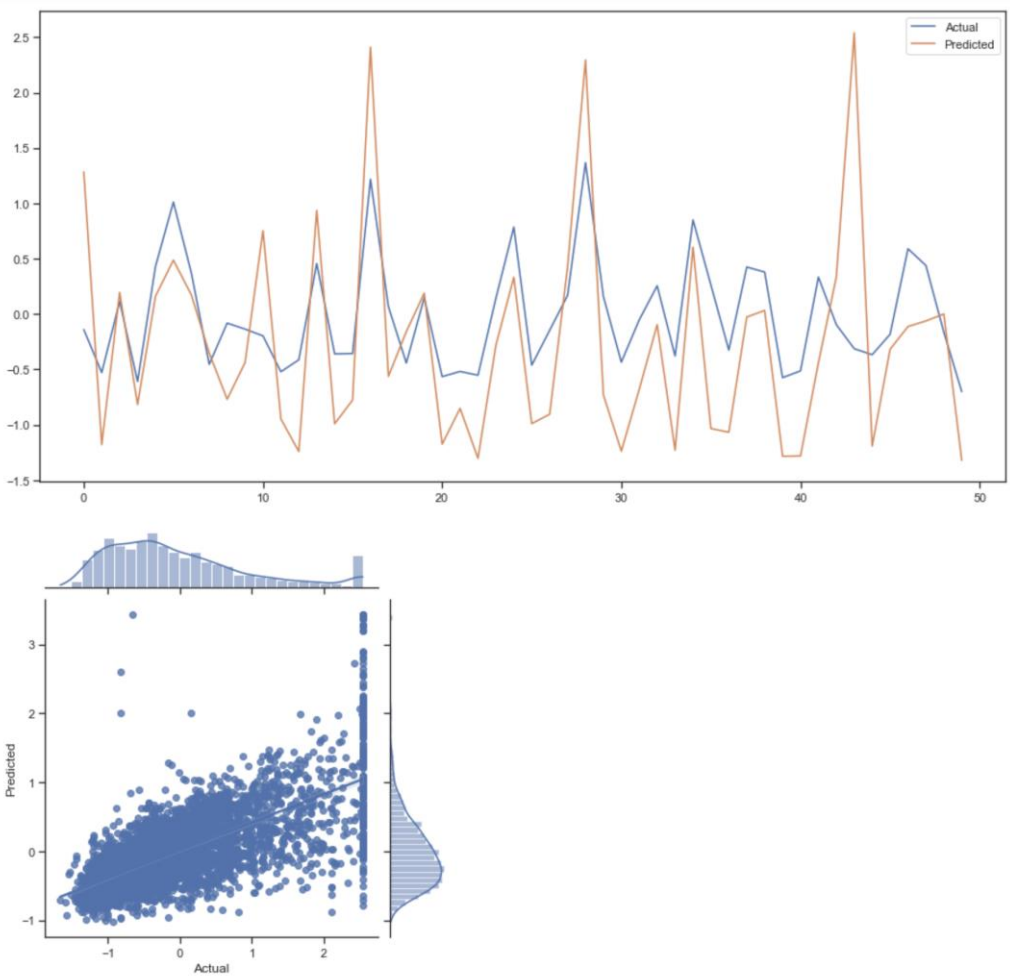


Figure 24 – Random Forest Model Fitting Results

Hence, it can be concluded that the Random Forest is providing the best model for fitting the data.

Ridge Regression results

Lastly, I would be implementing ridge regression model.

```
ridgereg=Ridge(alpha=0.001,normalize=True)
ridgereg.fit(x_train,y_train)
ridge_pred = ridgereg.predict(x_test)
r2_ridge = r2_score(y_test , ridge_pred)
rmse_ridge = np.sqrt(mean_squared_error(y_test, ridge_pred))
print("R^2 Score: " + str(r2_ridge))
print("RMSE Score: " + str(rmse_ridge))
```

```
R^2 Score: 0.6278267747704707
RMSE Score: 0.605493545636119
```

From the above analysis, the ridge regression model has an RMSE value of 0.61 and 0.62 as its R-squared value, this model had the second-best metric out of all the models implemented.

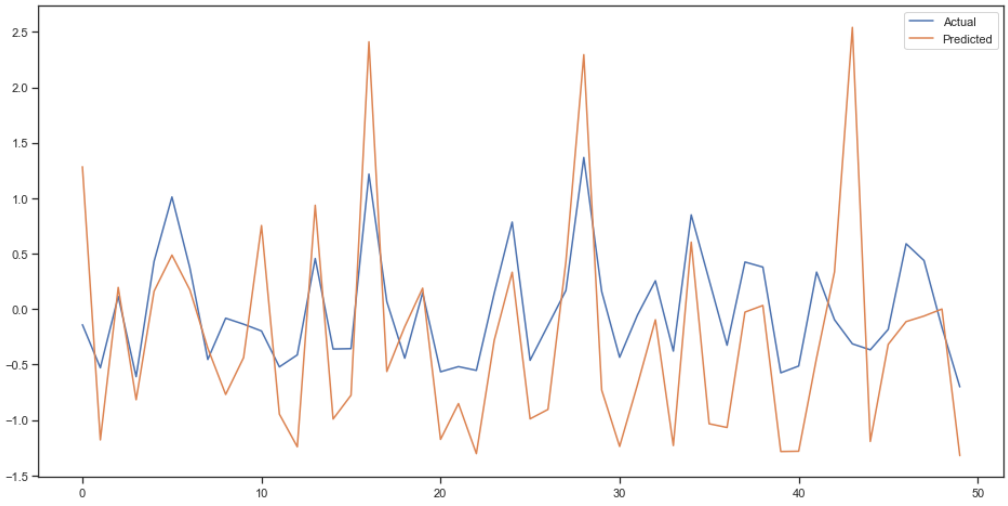


Figure 25

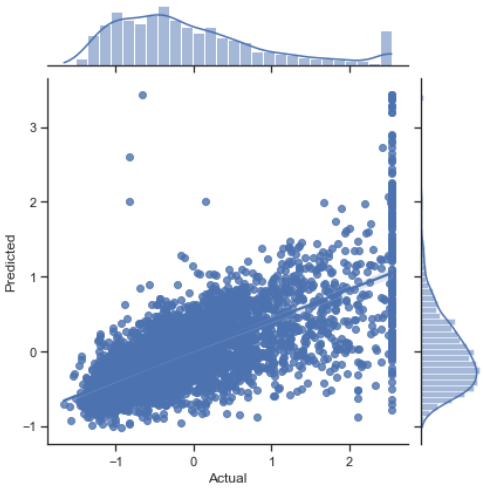


Figure 26

5.1 PERFORMANCE METRICS

When developing machine learning models, it's important to keep track of how well our model performs. We will be able to compare the model to other models or a benchmark if we wish to adjust. The RMSE (Root Mean Squared Error) and the R squared are performance evaluation matrices used to evaluate regression models.

Root mean square error – This metric is the square root of the average of the squared differences between the predicted and the actual value of the variable. The distance between data points and the regression line is measured by residuals. The RMSE is a measure of how evenly these residuals are distributed. In other words, it shows how the data is clustered around the best-fit line. The root mean square is denoted by this formula below

$$RMSE = \sqrt{\frac{\sum_{i=1}^{n}\left(P_i - O_i\right)^2}{n}}$$

The R squared – This metric indicated the distance between the predicted and actual valueson the regression line. It explains the fitting of the model

METHOD	MSE	R 2 (SQUARED)
LR	0.6055	0.6278
Lasso Regression	0.7193	0.4747
Ridge Regression	0.6055	0.6278
Decision Tree regression	0.5934	0.6425
Random Forest regression	0.4254	0.7688

Figure 27 shows the Comparison of regression methods performance

Chapter six

6.0 SOFTWARE IMPLEMENTATION AND SOCIETAL IMPACT

This project is all about how easy it is for the end-user to interact and understand how much a housecosts. Streamilit is a very useful tool that makes it easy to make web apps that work quickly for data analysis, machine learning, and for easy deployment. This working machine learning app lets the user change the input parameters and get a prediction from the model.

Python is used to design the application with a beautiful user interface, the Streamlit package is installed using the pip install command. Therefore, to run the Streamlit application we use the following command, streamlit run <someprogram.py>.

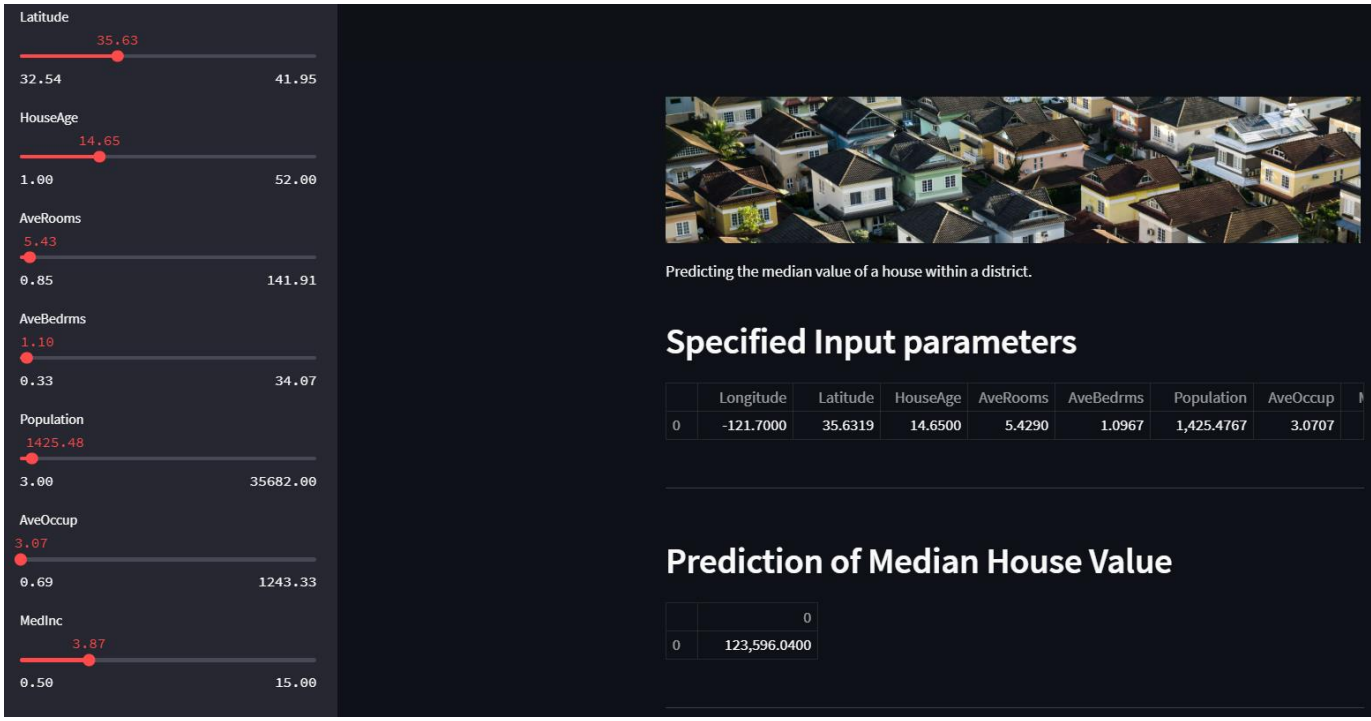


Fig 28 shows the median house value, based on different queries made on the system.

The results gotten are from the different parameters imputed by the user. The model used is the random forest model to effectively predict the median house value.

Chapter 7

7.0 Conclusion

In the current study, on the California dataset, it has been tried to fit different types of regression models which include Simple Linear Regression Model, Lasso Regression Model, Decision Tree Model, and the Random Forest Model. We are finding the Random Forest model as the most appropriate model giving the highest values of the R-square and the minimum Root Mean Square Error Value. Hence, it can be concluded that the Random Forest Model is the most appropriate model for this dataset and should be used for predicting the housing price.

For starters, our research has demonstrated that advanced machine learning algorithms such as LR, RF, and Lasso are viable tools for real estate researchers to use in predicting house prices. However, we must keep in mind that these machine learning tools have limitations of their own. Because there are frequently numerous viable features from which researchers can select and include in models, careful feature selection is required. To summarize, the application of machine learning in real estate research is still in its early stages.

We hope that our research has contributed not only to the advancement of property evaluation methodology and empirical findings, but also to the presentation of an alternative approach to house value valuation. A future study may include combining additional property transaction data from a larger geographical area with other attributes or conducting research on other property types besides house building.

7.1 Limitations

Here, we have used the regression models like Linear, Decision Tree, and Random Forest Regression, however, the results might have been better if we have worked on the K-nearest neighbor regression. One thing that has been seen is that for general data Random Forest is giving better results (least RMSE value), however, for the standardized data, ordinary linear regression is proving the better results. This needs to be rechecked with

some other data to understand whether the pattern exists in a similar manner or not. As we could work with only one dataset, it was not possible for us to recheck this with some other dataset.

7.2 Future Works

Our major research issue has been addressed and answered in this thesis; but the very act of providing a solution to this question invariably gives rise to further questions. The findings of our research and the outcomes of our analysis have sparked several new topics of interest, which we will now discuss. In this section, we will make some suggestions about how future work could be built upon the findings presented in this thesis, as well as how the body of literature on the topics of machine learning, big data, forecasting, and price prediction could be expanded.

According to our findings, machine learning and forecasting cannot be merged into a single model without additional research. This raises the question of what model combination techniques can be used to generate a more precise prediction output. Future research may study alternative techniques for integrating models with the same degree of precision. Modern machine learning combines strategies to produce more accurate output rules. We feel that research into the facets of model combining is essential for maximizing the potential of machine learning.

To determine which model to utilize for the ultimate price forecast, future research and development in this area may experiment with machine learning techniques. Consider a classification method that, given a set of house characteristics, identifies the most appropriate machine learning model for use as a predictor of that specific house type. As an alternative, this classifier could consider forecasting models.

Studying the impact of data diversity is an exciting avenue for future research in the field of machine learning and real estate price forecasting. It may be fascinating to investigate the capabilities and potential of machine learning by enhancing the dataset used for training with additional types of data, deemed to be linked to the real estate market.

References

Cao, B. and Yang, B., 2018. Research on ensemsble learning-based housing price prediction model. Big Geospatial Data and Data Science, 1(1), pp.1-8.

Ghatnekar, A. and Shanbhag, A.D., 2021, December. Explainable, Multi-Region Price Prediction. In 2021 International Conference on Electrical, Computer and Energy Technologies (ICECET) (pp. 1-7). IEEE.

Komagome-Towne, A., 2016. Models and visualizations for housing price prediction. Faculty of California State Polytechnic University, Pomona.

Montero, J.M., Mínguez, R. and Fernández-Avilés, G., 2018. Housing price prediction: parametric versus semi-parametric spatial hedonic models. Journal of Geographical Systems, 20(1), pp.27-35.

Neloy, A.A., Haque, H.S. and Ul Islam, M.M., 2019, February. Ensemble learning based rental apartment price prediction model by categorical features factoring. In Proceedings of the 2019 11th International conference on machine learning and computing (pp. 350-356).

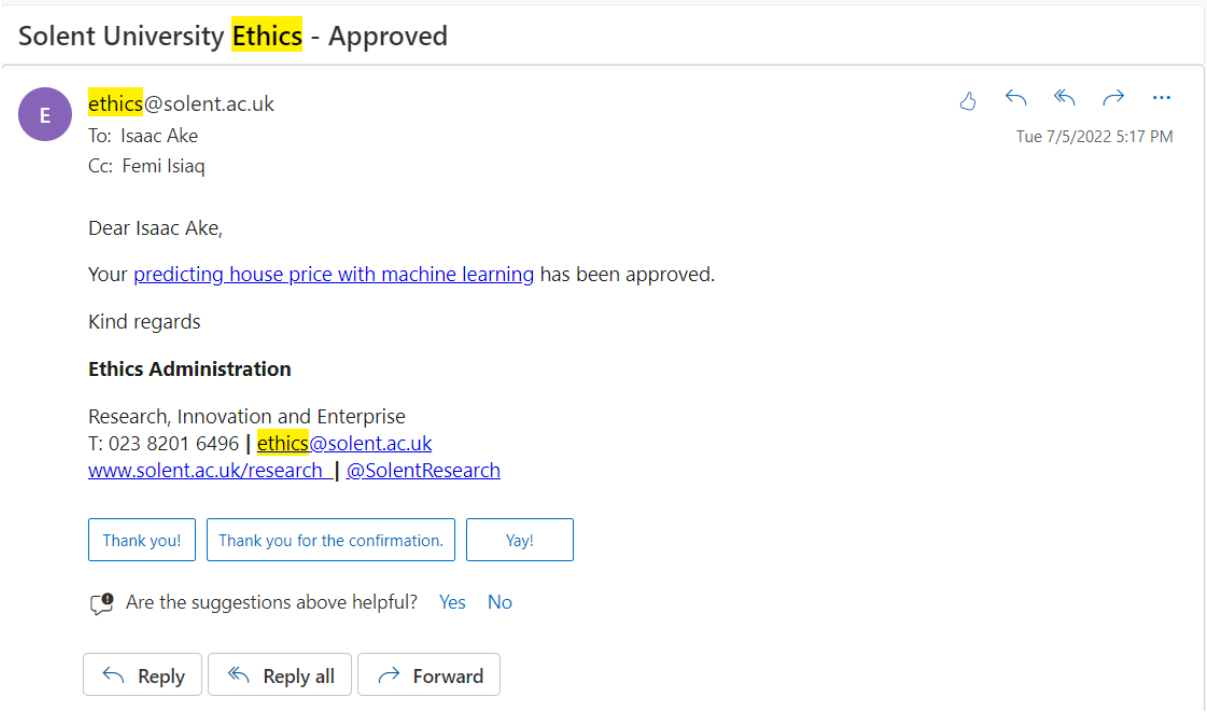
Segal, M.R., 2004. Machine learning benchmarks and random forest regression.

Wu, Z., 2020. Prediction of California house price based on multiple linear regression. Academic Journal of Engineering and Technology Science, 3(7).

Xu, M., Watanachaturaporn, P., Varshney, P.K. and Arora, M.K., 2005. Decision tree regression for soft classification of remote sensing data. Remote Sensing of Environment, 97(3), pp.322-336.

Zou, K.H., Tuncali, K. and Silverman, S.G., 2003. Correlation and simple linear regression. Radiology, 227(3), pp.617-628.

APPENDIX A: ETHICS APPLICATION.



Solent University Ethics - Approved

E

ethics@solent.ac.uk

To: Isaac Ake

Cc: Femi Isiaq

Tue 7/5/2022 5:17 PM

Dear Isaac Ake,

Your predicting house price with machine learning has been approved.

Kind regards

Ethics Administration

Research, Innovation and Enterprise
T: 023 8201 6496 | ethics@solent.ac.uk
www.solent.ac.uk/research | @SolentResearch

Thank you! | Thank you for the confirmation. | Yay!

Are the suggestions above helpful? Yes No

Reply | Reply all | Forward