MACHINE LEARNING TECHNIQUES TO PREDICT HOUSING PRICES AND IDENTIFY FACTORS AFFECTING THEM

ADITYA KALPESHBHAI PATEL
FAWWAD AHMED MIRZA

A project report submitted in conformity with the requirements
for the degree of Master's of Science in Information Technology

Department of Mathematical and Physical Sciences
Faculty of Graduate Studies
Concordia University of Edmonton

MACHINE LEARNING TECHNIQUES TO PREDICT HOUSING PRICES AND IDENTIFY FACTORS AFFECTING THEM

ADITYA K. PATEL & FAWWAD A. MIRZA

Approved:

Rossitza Marinova, Ph.D.

Supervisor Date

Committee Member Date

Patrick Kamau, Ph.D.

Dean of Graduate Studies Date

Abstract

House prices have a big impact on the economy, and customers and real estate agents are quite concerned about the price changes. Every year, housing prices rise, which ultimately highlights the necessity for a method or plan that might forecast home prices in the future. According to research, house owners and the real estate industry frequently worry about price swings in real estate. To identify pertinent features and the most effective models to anticipate home values, a review of the literature is conducted before the creation and analysis of machine learning models. Physical attributes including the living area, plot area, location, number of bedrooms etc. affect the price of a property directly. Previous studies have been based on just a few of these variables. This paper used dataset for USA housing with different internal characteristics and applied CART with ensemble techniques (Boosting and Bagging), K-Nearest Neighbor, Local Regression and Random Forest to predict property prices of houses and their relationship with different features. The paper will validate the different machine learning techniques applied by using k-fold cross validation and RMSE to provide an optimistic view on property price prediction.

Keywords: Property Price Prediction, Real estate Forecast, Classification and Regression Trees (CART), K-Nearest Neighbor, Random Forest, House attributes, Cross Validation

Acknowledgements

I would like to extend our deepest gratitude to my supervisor Prof. Dr. Rossitza Marinova, whose constant encouragement, guidance and support contributed towards the success of this Capstone project. Her expertise in the field and patience were unmatched and helped us bring shape and meaning to our research.

I also want to specially thank Prof. Dr. Baidya N. Saha for his help and guidance and extensive knowledge on the subject of Machine Learning & Big Data. My interaction with Dr. Baidya always inspired us to strive a little harder.

It would have been impossible to continue the work without endless support from our friends, we have met some incredible people in this journey who allowed us to think outside the box and bring new ideas to our fruition.

A special thanks at the end to my Mom and Dad, my brothers, my sister-in-law and my fiancé for their unconditional love and endless support.

Fawwad Mirza

Acknowledgements

I would like to thank my supervisor Prof. Dr. Rossitza Marinova for her constant guidance, feedback, and motivation throughout the project. Their suggestions and inputs have been invaluable in shaping this project.

I would also like to thank Prof. Dr. Baidya N. Saha for his technical expertise and assistance during our course of Big Data. Without his expertise and guidance the project would have been much more challenging.

I had unwavering support from my friends throughout the process and I would say that I have met a very supportive and genius bunch of people on this journey.

Lastly, I want to thank my parents for their love and support towards my passion.

Aditya Patel

Contents

List of Figures

1 Introduction

Real estate has been a crucial component of the economic activity which has changed drastically since the last few years. The changes are not only seen in pricing dynamics but also demand, general market trends and technology adaptation. Even though many people lack any prior real estate trading knowledge, they nonetheless spend money in this market [19]. Lately there has been extra focus on energy sustainable resources and eco-friendly living as the emerging trend in the real estate industry. People are demanding greener homes, or suburbs with more trees and greenery. Incorporation of IoT is also seen as an emerging trend of the real estate sector and how it's now shaping the industry of smart homes pricing them accordingly.

This sector is increasingly becoming unpredictable and the rapid changes in the price can make it a little hard for the buyers as well as realtors to make wise decisions. The unavailability of houses in favorable areas can cause bidding wars leading to higher costs. Purchases in expensive and higher demand cities where cost of living is already higher, negotiating the mortgage loans and dealing with complex and secure financing can be extremely difficult. These issues are faced by both buyers and realtors due to the complicated legal and economic issues faced by the real estate sector. Higher demand suburbs also face extreme rivalry between buyers as well as realtors making it even more hard to get a suitable house. Economic uncertainty plays the part of undermining the customer confidence as they deal with market reluctance.

Over time, both the number of households and the growth of housing units have leveled off. However, the kind of dwelling occupancy—whether owner-occupied or renter-occupied—represents the most important shift that has taken place. The number of owner-occupied homes stagnated in the years following the financial crisis, while the number of renter-occupied homes increased. This tendency has recently changed, demonstrating that house ownership remains a key component of the American Dream as shown in Figure 1.

The purchasing power of consumers may be influenced by three key variables: income, home prices, and mortgage rates. Consumer power rises along with rising earnings. Since 2016, the median household income in the United States has increased to more than 60,000 dollars, which is among the highest salaries recently. Housing affordability decreases as mortgage rates or property prices increase. Since 2011, the average price per square foot of floor space of existing single-family homes in the US has grown yearly, reaching 143.83 dollars in 2021 [6].

1.1 External & Internal Factors

A combination of external and internal factors is involved in the process of deciding the real estate pricing including:

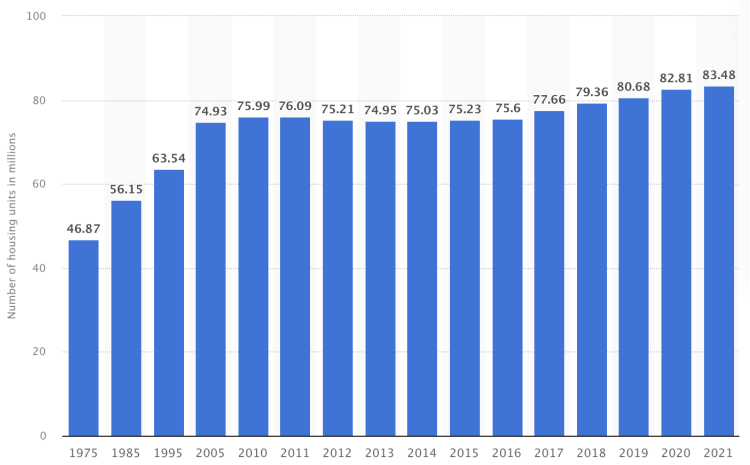


Figure 1: Number of Housing Units in USA from 1975 to 2021 [7]

- Location: The value of houses is substantially influenced by the proximity to important facilities such as hospitals, quality schools, entertainment & recreation hubs, parks and shopping malls. Prices are naturally higher in districts with lower crime rate and scenic neighborhood.
- Market Demand and Supply: Main aspect of home prices is dependent on the concept of demand and supply. Prices are generally higher in high-demand locations where there's scarcity of homes and can decrease where there are many property options available [8].
- Economic Factors: Affordability and demand are directly relied on the state of economy, standard of living and the buying power of the people including the employment levels, purchasing power, GDP growth etc.
- Interest Rates: Mortgage interest rates are known to have a strong influence on the affordability of the property for buyers as with lower interest rates the prices generally rise due to higher demand.
- Infrastructure Development: Expected infrastructure initiatives in the area such as under construction high way connections or commercial hubs around the area also impact the prices in a positive manner.
- Government Regulations and Policies: Most importantly, the rules and policies set by the government such as tax breaks or limitations on foreign ownership for buyers have an impact on the housing values and the housing market itself.

1.2 Value Added Through Personal Characteristics

Apart from external factors of housing, the value of property is directly impacted by the personal characteristics of the house such as:

- Size of the property and utilizable space, property. Number of rooms, bathrooms, finished basements and the structural design of property is really important.
- Age and condition also have a direct influence where newer and well-maintained homes tend to be placed with higher prices compared to older houses without any renovations.
- Features and amenities of the house also directly affect the price where the presence of pools, theatre, garden, high-end appliances and contemporary kitchens can increase the price of house.
- Houses in greener neighborhoods and space for home grown gardens are preferred by many. Also, a well-lit house with natural light coming in directly are usually high priced.

1.3 Deployment of Machine Learning

Understanding property values using the traditional techniques are becoming inefficient and losing accuracy with every passing day as the real estate markets continue to expand in a more complicated and dynamic way. The ability to capture large amounts of data and identify insightful patterns quickly, deployment of machine learning is nothing but a game changer in this situation. This has not only led to quick analysis but a more precise and accurate property value predictions [20].

Thorough knowledge of the price drivers, market trends and individual property characteristics is important to successfully understand the real estate industry and its trends while overcoming the obstacles. The buyers and realtors can cope with the challenging environment and make well-informed decisions about their property purchase by making use of effective and smart techniques as well as staying informed.

It has become increasingly difficult to predict property values effectively in today's changing economy. Here machine learning has proved to be a valuable resource for everyone especially real estate sector. Stakeholders can know the different variables affect the real estate pricing by utilizing machine learning algorithms to analyze the information and identify ambiguities. This has resulted in more informed real estate agents with better idea about the customer's choices and are equipped to negotiate better terms, price properties more fairly, and decide on investments with knowledge.

Machine learning techniques can be used in the real estate in a number of ways. Firstly, to predict the property prices accurately machine learning algorithms examine the past sales information, different property characteristics and other variables. These algorithms can then find intricate relationships and connections between the selected variables and end up with accurate and consistent property appraisals [11].

Important factors that have a direct impact on real estate values can be identified using data analysis techniques. Real estate brokers and sellers can learn what characteristics of a home should be highlighted to draw in potential buyers by ranking features like location, size, amenities, and age of the property based on their influence on pricing. Large datasets can easily be analyzed by machine learning algorithms to spot market trends and price changes over time. Realtors can then utilize this data to choose the ideal dates to buy or sell a property, modify price plans, and gauge the general market environment. They can also incorporate data analytics in their business to assist potential buyers with customized property suggestions based on their favorable characteristics as well as budget. This technology can recommend homes that fit the narrow criteria of a buyer by examining their past interactions and preferences subsequently saving the time and increasing client satisfaction. It also allows the realtors to compare their prices with competitors and similarly situated properties to create aggressive pricing tactics. They might establish listing prices that would correspond to market demand and maximize property turnover.

1.4 Identifying Challenges in Real Estate Investment

Data analysis can help identifying challenges and possible dangers in the real estate investments as well. Realtors can use machine learning techniques to reduce risks and make better judgements by taking several factors into consideration such as market stability, past price patterns and economic data. To enhance the attractiveness and value of projects or listings, the realtors use machine learning to discover the attributes that buyers look for in a specific area. Based on these past pricing and market patterns, machine learning algorithms are able to predict future real estate values. Prescriptive insights are useful as buyers and realtors can now strategically plan their investments and foresee market volatility. Real time market insights assist the real estate professionals to stay updated with ever changing dynamics. They can react swiftly to changes in demand as well as pricing. Agents can come up with quicker and informed decisions while saving on both time and money [22].

The real estate sector has therefore benefited greatly from the application of machine learning and data analysis techniques. Real estate agents have a competitive edge, they can improve property appraisals, and offer a more individualized and effective experience for both buyers and sellers by utilizing the power of this technology to examine the characteristics of homes that impact price.

1.5 Contributions of Research

The research work hopes to contribute a detailed analysis of how machine learning models can be applied to real estate data and its meaningful outcomes. Real Estate industry can benefit from a suitable method of forecasting the property prices and the anticipated increase with every passing year. By reviewing multiple relevant studies and datasets to chose the best machine learning models for the given dataset

we hope to bring a new perspective out. Our work would take the application of machine learning models to real estate industry and it's forecasting attributes which will give out fruitful results. Data classification and regression techniques combined with ensemble learning to model a different set of variables of physical characteristics of houses explores untapped areas of the field. We hope to overcome the gaps in the previous studies using this research as a solid work of data exploration, modelling, processing, analysis and validation. The machine learning models would be double checked for errors so they can be tested for accountability and accuracy before deciding what stands out as the best.

1.6 Organization of Project Report

The Report is comprised of eight chapters starting with a detailed introduction. The introduction highlights the real estate situation around the globe and the USA, factors which might affect the housing prices, challenges faced by realtors and buyers. How machine learning techniques can be applied in the real estate industry and what benefits it can bring by using the right model for forecasting.

Chapter 2 revolves around the research objectives as to why we have decided to pursue this topic of research and what are our aims to achieve through this research. It mainly highlights the need of finding the suitable machine learning model for predicting prices of houses and what possible factors can have a solid relationship with housing prices.

Chapter 3, the problem statement of the research talks about the real problem faced by real estate industry which compelled us towards this thesis that is the absence of a proper workable model to identify the anticipated future property prices and how realtors and house owners can properly price their property by understanding which features have the most impact.

Chapter 4 of the research is a detailed review of literature including multiple studies on machine learning models being already used in real estate sector, how machine learning models help in predicting prices, what factors can have possible affects on housing prices and what kind of machine learning algorithms can be used for different datasets. These research papers and journals have been extracted from reliable sources and are all well researched and thorough in their subject matter.

Chapter 5 talks about project design which is entirely focused on the methodology we are going to apply in the study. This includes the dataset we are using, it's detailed characteristics and exploration of data, our approach with data pre-processing highlighting the important variables for the research and characteristics we might remove or replace as well as the selection of machine learning models. We have decided to use Classification and Regression Trees with ensemble learning for classification (boosting and bagging) and Random Forest, K-NN and Local Regression. This is then followed by the data validation techniques that we will be applying including Cross Validation and Root Mean Squared Errors.

Chapter 6 entails the details about our implementation of the models and walks through a step-by-step procedure of exploratory data analysis, classification and regression with comparison of models on the basis of their errors.

Chapter 7 is summarization of the findings of the study highlighting which model do we think is appropriate and what are the most important characteristics of a house according to our research. This is then followed by the scope of thesis in the future and how can researchers further expand this piece of work.

Chapter 8 finishes the research work by a couple of recommendations that we have listed for the particular work and to make amends and what can help achieve even better results.

2 Objectives

The data in this study is tested using a number of different machine learning techniques which is going to use a combination of pre-processing methods to increase the accuracy. The purpose of this study is to examine the accuracy of forecasting property values using Regression, K-Nearest Neighbors, and Random Forest. The goal of this research is to understand more about these approaches in machine learning and what works best for this kind of dataset. We aim to identify if any relationship exists in property prices with certain characteristics of the houses (those mentioned in the dataset) including living area, plot area, presence of basement or not and waterfront, number of bedrooms and bathrooms, number of floors, chronological age, condition and whether the property has been renovated. We also hope to understand more about these characteristics by studying their relationships with property price and how impactful they can be. Using machine learning models, we aim to classify the data and then use regression to design favorable algorithms that would help us in predicting the property prices accurately. This would give us an idea about the models which would be best suited for our kind of dataset and what exactly would prove to be useful.

3 Problem Statement

With the growing uncertainty in the real estate market around the globe especially after being hit by the pandemic in 2019/20, real estate agents and buyers face difficulties over pricing and the never-ending bidding wars. Sellers always decide pricing for their properties with little to no knowledge about how the internal and external characteristics would impact the property's value and the realtors also base it on just a few factors totally negating the rest. This paper takes into account the most important internal characteristics of a property which can help realtors and sellers understand what would really be important to set up property prices. Not only this, the buyers would have a knowledge before making their purchase and backing their bidding wars with logic. Using Machine learning techniques to study the different

characteristics which impact a property's value would not only be a help to buyers but a breakthrough in the real estate sector to further analyze the features of property on a different level.

4 Literature Review

The process of evaluating property values using a different set of techniques is known as "house price prediction." Despite the fact that there are now high property demands, we don't have enough appropriate ways that might aid in predicting home prices. Machine learning focuses on creating algorithms that can forecast future results based on previous data. Similar principles apply to the forecast of home prices. This study discussed numerous theories and previous research on this topic. The act of generating an opinion of value is a crucial tool for analyzing housing prices whether selling, buying, financing, or insuring on real estate properties [23].

The demand for real estate is growing yearly, which consequently drives up the price of real estate. In order to make educated judgments and assist the owners in setting fair home prices, buyers and brokers want to understand the factors influencing the price of a property. House price predictions may be made using a variety of MLM models, such as ANN and VR. According to Zulkifley ET al.'s [24] study from 2020, there are two main components to property price prediction: ANN is more inclined on house characteristics, and the other focuses on the machine learning model used to estimate housing values. Based on the factors that affect them, house prices may be divided into three characteristics: location, structure, and suburban condition. His research examined the factors that earlier academics had taken into account while developing various prediction models to forecast real estate values [24]. The findings indicate that when combined, SVR, ANN, and XGBoost can be effective models for predicting home prices. It was found that homes in strategic locations cost more than homes in remote areas with less amenities, such as close access to a retail center.

Researchers want to forecast property values properly while avoiding losses and ineffective price predictions. When estimating efficient house pricing for customers based on their budget and preferences and anticipating home prices, several factors must be taken into account. Many relevant models are tested and based on results most accurate models are recommended for forecasting the housing prices. The study by Truong in 2019[20], examined and assessed the performance of three distinct ML model types, including Hybrid Regression and Stacked Generalization Regression, Light Gradient Boost, Random Forest, and XGBoost. All of these models produced fair results with a specific combination of advantages and disadvantages. RFA had the lowest error and was the most effective approach for training data, although it was evidently prone to over-fitting. In terms of accuracy, XGboost and LGBM were excellent models, but they struggled with time complexity. The performance of hybrid regression or HRA was noteworthy and marginally superior to the other three. Staked GR was a complicated model that is only useful when data accuracy is impor-

tant. Both HRA and Staked GR produce acceptable results, but time complexities must be taken into account because both models contain the RFA, a very time-complex model [21]. The least time-consuming method is Stacked Generalization Regression, which also uses K-fold cross-validation.

The study by Chouthai et al. [4], in predicting housing prices applying machine learning showed regression trees to be as effective as linear regression, although polynomial regression produced results with smaller errors. While neural networks struggled to effectively use the dataset. In their study, they used a variety of machine learning techniques to forecast housing prices [4]. These included logistic regression, support vector regression, the Lasso Regression approach, and decision trees. The research includes the characteristics and data for 100 dwellings. This study produced accurate findings.

Numerous regression models combining tree-based algorithms, Support Vector Regression, and Least Square Based Linear Regression have been utilized in order to build an accurate real estate price prediction model and investigate the most effective component on the home price. A hybrid Lasso and Gradient Boosting method for predicting individual house prices was developed by Lu et al. in 2017 [14] using the Iowa state dataset that was made available on the internet for the researchers. The selling prices log is their goal variable, and their final prediction is a mix of Lasso and Gradient Boosting [14].

Ozdemir's study from 2022 studied many tree-based algorithms to create the ideal regression model for forecasting real estate values. The analysis revealed that the CatBoost model was most suited for the property dataset prediction, and it was determined that living area is the most useful of the three residential properties in predicting selling price. Out of 80 factors, the above-ground housing area, general condition, and locality are very significant [25]. Apart from their general quality, garage properties are also effective.

HPI or the house price index is to determine the increase in residential real estate prices in multiple countries, including the USA Federal Housing Finance Agency HPI, the United Kingdom Halifax HPI, UK National Statistics HPI, and Singapore's URA HPI. According to studies, they include location, concept, and physical state. The area of the house, number and size of bedrooms, kitchen and bathrooms, and garage, the presence of a yard, the plot area and the year of build of the property are physical characteristics that may be seen to have strong impact on prices [3].

According to a study by Kang [12] there are certain other characteristics which might directly impact the housing prices and can be called as the physical attributes of the house just as when was it built, the size of the building, the area and design of rooms, and other aspects which have an effect on the interior design of a house. Although ideas cover a range of advertising strategies used by developers to draw in possible investors. This shows how approachable the house is to entertainment centers, schools, markets, and highways. A house's price is directly impacted by its

locality and vicinity of the amenities [12].

Increased property value can decrease financial flexibility and lead to debt-financed spending, which might increase demand by levitating owner's income. There are usually two types of forecasting techniques for prices[1]. The first group of techniques include oil and stock price prediction, like predicting market patterns in a time-series manner. The prices of let's say plane tickets or real estate fall in the second group. Here the time series technique would be searching for a correlation between past and future rates.

This second group would use hedonic pricing based on linear regression. The study by Abigail has applied a regression model to examine Boston housing statistics and make predictions about house prices based on the dataset's attributes. It investigated the use of the random forest approach for forecasting home prices. The suggested model's performance was assessed using the Boston housing dataset, with performance measures including Mean Absolute Error, Root Mean Square (RMSE) and R2 or Coefficient of Determination [2]. The study demonstrated the capabilities of the random forest machine learning approach to have successfully forecast home values based on the factors available in the dataset.

According to research, before negotiating further, realtors, sellers, and buyers typically consider if an MRT station is close. Real estate values are significantly influenced by the age of the home, which is the second important factor. According to Li et al. (2015)[14], the earth is home to a number of seismic zones, including those in Tonga, Fiji, Japan, Indonesia, China, Turkey, and Iran. A house's collapse during an earthquake could be influenced by its age. Additionally, older homes are less resistant to damage than newer homes. Therefore, before making a purchase or renting a home, clients are recommended by dealers to find out the age of the property. Geographic coordinate information significantly affects real estate values, as claimed in a study by Melanda, 2016[15]. The transaction date and the number of convenience stores that are accessible by foot from the housing circle make up the remaining factors. These are minor than the rest in the process of estimating real estate prices [15]. In general, this study provides real estate tenants, sellers, and buyers with early knowledge on the worth of property variables.

5 Project Design

The dataset housing prices in USA contains entries of over 4,600 houses with 17 variables representing the real estate situation in 2014 published on the internet and updated on 2018. The variables present in the dataset were crucial in impacting the prices of houses directly. The important step in implementation is data collection followed by exploration (to understand the variables in dataset), pre-processing stage for data cleaning and making it suitable for the development of model.

5.1 Data Collection & Exploration

The dataset includes houses from different cities all over the USA and contains houses with bigger plot sizes and smaller living spaces. Exploratory data analysis is required prior to model construction. In order to choose the best machine learning techniques, it is necessary to first find the underlying tendencies in the data. In order to understand the features in the data and their purpose, exploration of data was done. The features include house price (in US dollars) as our dependent variable, number of floors, bedrooms & bathrooms, plot area, living area, area above ground, waterfront, presence of basement, year of construction & renovation, address. The area variables are in square ft. The house of prices range between 7.8 thousand dollars to 26.6 million dollars. Mean number of bedrooms in houses were three while the average living area was 2100 square feet and many houses were below that average. The average plot area was around 14,800 square feet almost seven times more than size of house (living) however there were less than 50% plots which exceeded 10,000 square feet. This shows that some houses are with larger plot size but smaller living area and it's not evenly distributed. Only 30 houses were located by the sea which is a very small number compared to the other 4570 houses making it a very insignificant variable. Floor area was 1830 square feet on average with a right-hand symmetry showing that many houses have below average areas.

5.2 Data Pre-Processing

Now that we have a detailed look at the variables, we know about the important features and lesser important features which do not have much impact on the dependent variable of price. All of the features involved in the dataset are not fit for analysis and some need modification or elimination.

- Address would be removed from the database.
- Number of floors also does not contribute anything important so it was removed.
- View is removed as it makes the data redundant.
- Basement and renovation are changed to dichotomous features (where 1 means yes and 0 means no).
- Construction year was also subjected to dichotomization too with 0 taking old houses and 1 as new houses and cutoff year of old houses set at 1970.

Figure 2 will show that prices higher than 1 million clearly stand out from the rest, so we eliminate entries with price greater than 1 million dollars. After elimination of 340 of such entries we're left with 4260 entries.

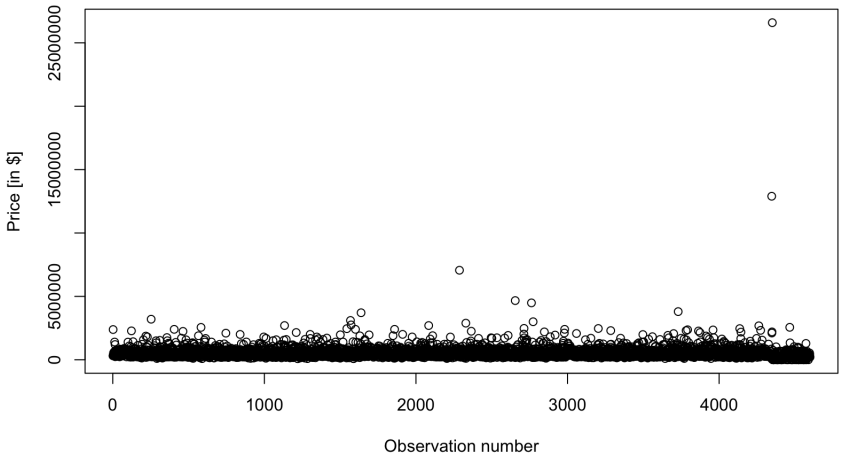


Figure 2: Number of Houses and Prices in the dataset

5.3 Model Selection

The data should be properly processed before developing the machine learning models so they can learn the patterns more quickly. We divide the dataset by splitting into train and test sections using the ratio 3:1. The training set has some 3159 entries while the test set contains 1052 entries of houses. The work is based on supervised machine learning models for both classification and regression. Based on the dataset variables, regression forecasts a continuous value. Regression issues' primary objective is to calculate a mapping function based on the input and output variables. Simply put, it also helps identify a solid relationship or correlation between the dependent and independent variables. Whereas a classification machine learning model forecasts distinct output features, such as categories, by approximating a mapping function from input variables. This results in making mapping functions in charge of forecasting and predicting the category of input variables. This makes both discrete and real-value variables be used classification process.

5.3.1 Decision Trees

The strongest and most popular technique for categorization and prediction is the decision tree. It's a tree structure that looks like a flowchart, each internal node stands in for a test on a characteristic, each branch for the result of the test, and each terminal node for the class label. An ML prediction method is called a Classification and Regression Tree (CART). It demonstrates how other values may be used to

anticipate the values of a target variable. Every branch of the decision tree has a prediction for the target variable at the end of each node and a predictor variable at the beginning of each branch. One essential decision tree technique that presents a foundation for machine learning is the classification and regression tree algorithm[18]. Leo Breiman came up with the term "CART" (Classification and Regression Trees) to refer to decision tree techniques that might be used to address classification or regression tree predictive modeling problems.

5.3.2 Random Forest

Random forest is a popular supervised machine learning model used for both classification and regression. It uses the ensemble learning techniques by using different classification and regression models up to perfect precision. This model takes the results of multiple decision trees and averages them to increase the prediction accuracy of data. Random Forest does not depend on single decision tree but it uses results from trees individually and calculates the performance based on majority of votes [**?**].

5.3.3 K-NN

For classification and regression problems, one common machine learning method is the K-Nearest Neighbor (K-NN) algorithm. It is predicated on the notion that labels or values for related data points are frequently similar. The K-NN method uses the full training dataset as a reference throughout the training phase. When producing predictions, it uses a selected distance metric, such as Euclidean distance, to determine the distance between each input data point and each training sample. When classifying data, the method chooses the K neighbors as the most popular class label out of the projected labels for the input data point. In order to forecast the value for the input data point, regression computes the average or weighted average of the target values of the K neighbors. Since it is a non-parametric regression, it makes no assumptions about the distribution of the data and hence encourages the training stage. It swiftly picks up complex target functions without forgetting what it has learned. 'K' observations with x_i nearby are taken into consideration for a given input x of training data, and the average response of those 'K' independent variables offers [9].

5.3.4 LOESS or Local Regression

In a K-NN method, the window size, or k, is an adjustable parameter that, in this case, controls the estimate's smoothness. This algorithm is pretty similar to a K-NN algorithm in that respect. K might be thought of as your bias vs. variance slider. Lower values of K will provide larger variance, whereas large values of K will produce higher bias. To fit linear or quadratic predictor functions at the centers of neighborhoods, the Loess approach employs weighted least squares. A specific percentage of

the data points must be present in each neighborhood; therefore, its radius is set accordingly. The smoothness of the predicted surface is determined by the percentage of data, or smoothing parameter, in each local area. A smooth decreasing function of the data points' distance from the neighborhood's center determines how heavily each local neighborhood's data points are weighted [10].

5.3.5 Ensemble Learning

Comparison to using just a single decision tree, the ensemble learning approach incorporates many decision trees to give out higher predicting performance. The ensemble model works on the principal that weak learners can be combined to form much stronger learners. The objective of this learning is to generate a powerful classifier combining several learners to get more precise classification results. The four types of ensemble learning techniques are voting, stacking, bagging and boosting. In this paper, the popular ensemble learning methods of boosting and bagging are applied to the experimental data and then compared on basis of their cross-validation errors [16].

5.4 Data Validation & Accuracy

Once the data has been modelled, we can now test the model's accuracy by using certain techniques to compare errors between the different machine learning models and then identify which one suits the best and has the lowest error. Accuracy score evaluates the number of correct predictions made by a model. A good model would not only have a lower error, higher accuracy but also be successful in building meaningful relationships and identifying patterns between variables.

5.4.1 K-Fold Cross Validation

A method for assessing prediction models is K-fold cross-validation. There are k folds or subgroups in the dataset. The model is first trained and assessed k number of times, with each evaluation utilizing a distinct fold as the validation set. To gauge the model's generalization performance, performance measures from each fold are averaged. This approach supports the evaluation, selection, and hyper-parameter tweaking of models and offers a more accurate indicator of a model's performance [17].

5.5 Root Mean Squared Error (RMSE)

RMSE is considered as one of the most used and popular performance measuring technique in regression models. It typically calculates the difference between the actual values and value predicted by a model. This gives out the accuracy of a model and the quality of it's prediction abilities. Lower RMSE value means that the model is more accurate and better at prediction. Ideally, the RMSE value should be

0 which is rarely the case since no model would be perfect in always predicting the accurate results [5].

5.6 Confusion Matrix

Figure 3 shows that confusion matrix or simply the error matrix is known for forecasting the performance of a classification machine learning model. It's a summary of number of correct and incorrect predicted values by the model (actual and predicted). With actual values on horizontal axis and predicted values on
vertical axis, it is divided into 4 quadrants of true positive (model predicts value

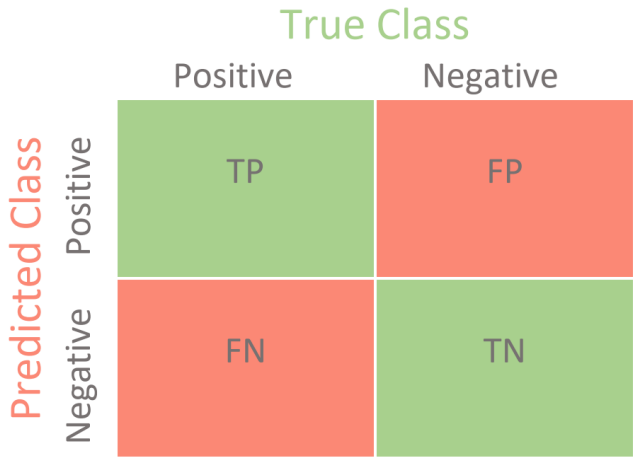


Figure 3: Confusion Matrix for Multi-Class Machine Learning Model

as positive and it's positive), false positive (model predicted a positive value which is in fact negative actually) it's also known as type 1 error, false negative (model predicted value as negative but it is actually positive) it's also known as type 2 error, and true negative (model predicts a value as negative and it's actually negative) [13].

6 Implementation & Results

We have gone through a detailed machine learning modelling starting from exploration of data and then applying the CART algorithm. This chapter would include a step-by-step process of the machine learning models implementation and their results followed by a comparison of their errors in the end.

6.1 Exploratory Data Analysis

Here in Figure 4 we can see that certain values are more distinctive than others. Some of these are really expensive homes. As a result, we decide to discard the instances where

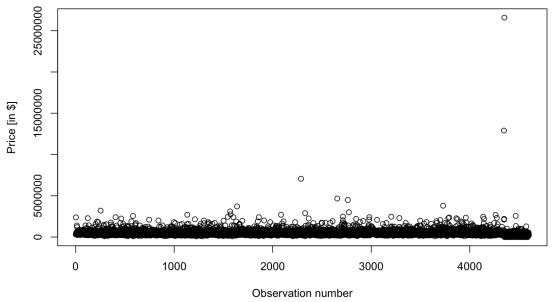


Figure 4: Number of Houses and Prices in the dataset

the cost of the home exceeds $1 million because they are exceptionally high in the context of the research question (the possible seller or buyer will not be looking of such a high-priced properties). We have 4260 cases following this method leaving 340 of such properties out.

After eliminating 340 cases, here in Figure 5 we have shown 4260 properties following the method in which we leave out any of the property where the cost crosses the benchmark of $1 million. While looking at Figure 5 we see that many of the houses are between $20,000 to $60,000 range.

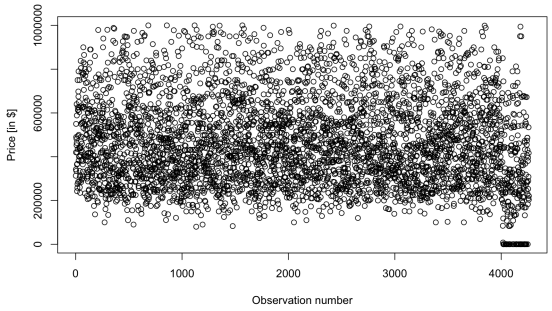


Figure 5: Houses within the Range of $1 Million

In Figuer:6 it shows that the cost of a property is 467.5 thousand dollars on average. Most of the properties are below the average value as shown in Figure 6 with a right-hand asymmetry in the distribution. Looking at this we can say that based on number a house will typically cost between $250,000 and $500,000.
The distribution in Figure 6 supports the fact that the median is below average at about $315,000.

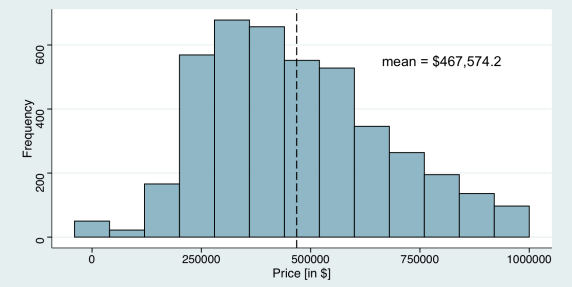


Figure 6: Average House Prices

The elimination of outliers forced the price range to be from $7,800 to $1 million. Moving on we will see how the dichotomization effects where we will divide our property features in two variants and then analyze.

In Figure 7 it is shown that property that is priced at "0" in this method is going to be considered as inexpensive, whereas a property that is priced at "1" is considered to be costly, that is, one that is priced higher than the mean. We have 2348 affordable/below average properties in our dataset. When compared to costly properties which accumulate at almost 1900. Moving on we will be investigating the factors that determine a cost of a property.

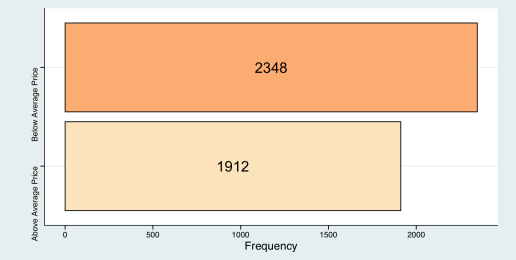


Figure 7: Above & Below Average Houses

Figure 8 shows that three bedrooms are common in about half of the properties. Four bedrooms are also common in properties showing up in nearly every fourth property. Properties with one, five, or more bedrooms are not very common. The feature gives a mean of 3.347 showing that on average a property has more than 3 bedrooms.

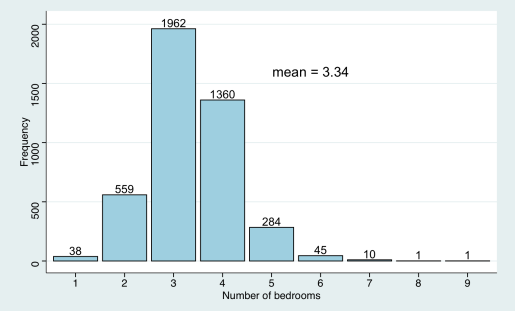


Figure 8: Average Number of Bedrooms

Figure 9 shows the Floor Area which also has a very clear right-hand asymmetry like most of the graphs mentioned. Most of the properties are smaller than the 2000 square feet whereas the average here is a little over 1800 square feet making it less than almost 300 square feet of the House Area.

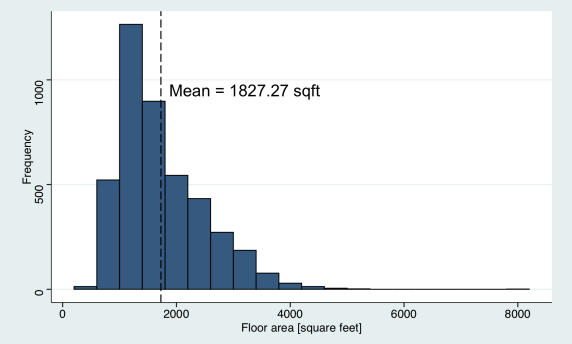


Figure 9: Average Square Feet (Floor Area)

Figure 10, the property is often less than 15,000 square feet as projected in Figure 10, or more than 7 times the size of the typical home. Majority of the plots are less than 10000 square feet. The graph here explaining the Plot Area which includes House and Floor area also has a substantial right-hand asymmetry.

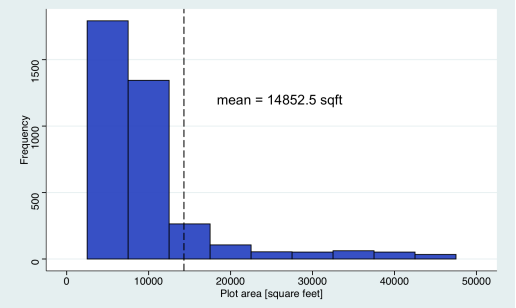


Figure 10: Average Square Feet (Plot Area)

There are just 18 properties that can be labelled as seaside residences shown in Figure 11, or 0.42% of the overall properties in our dataset. The remaining properties are spread across different cities.

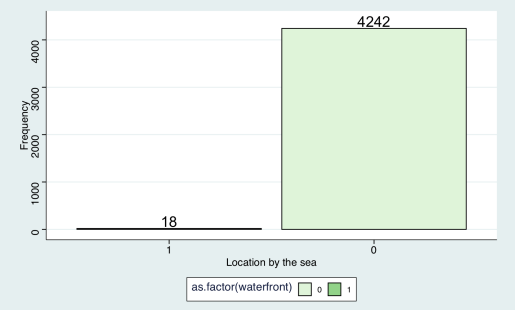


Figure 11: Properties by the Sea

In Figure 12 a scale from one to five determines the condition of the properties, half of them were evaluated at 3. 381 properties to be exact, or less than 9% of all the properties were rated at 5 receiving the highest rating available. Evaluations 1 and 2 were not at all common showing that most of the properties were in good condition.

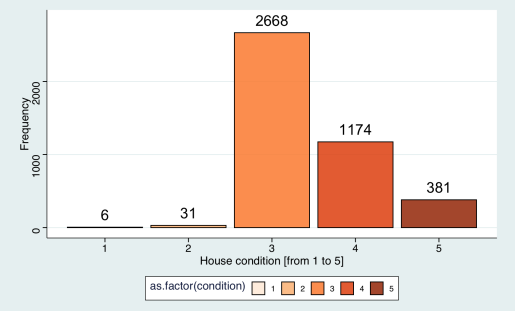


Figure 12: Rating of Properties Condition

As per Figure 13, 1740 houses, or a little over than 40% of all properties, have gone through renovation. Which means that most of the properties have never been renovated.

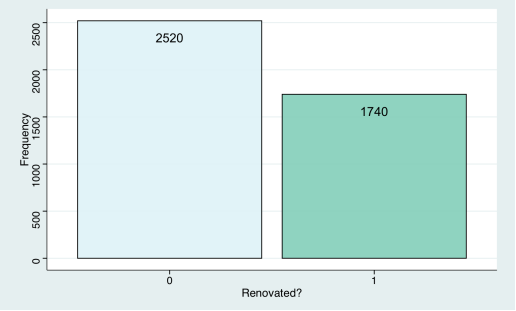


Figure 13: Number of Properties being Renovated

In Figure 14 we have divided properties by making 1970 as a benchmark year and distributing them as older and newer by analyzing that how many properties were built before and after 1970 shown in Figure 14. A lot of properties are more than 50

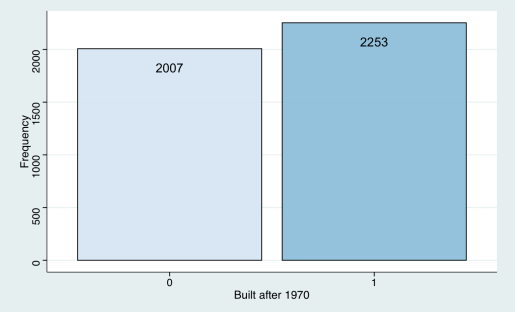


Figure 14: Build Before and After 1970

years old (2007 to be exact), making them relatively old. A total of 2253 properties, or 52.8% of all properties, were constructed after the year 1970.

The final feature is shown in Figure 15 where we will be using dichotomization is going to specify if the property contains a basement or not. Around 38% of the properties does include a basement. A total of 2612 properties (or a little over 60%) doesn't have a basement or an underground place.

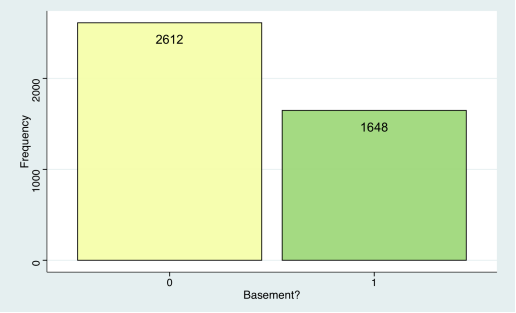


Figure 15: Properties Containing Basement

In Figure 16, the correlation plot shows that the surface of the property (House Area) and the Floor Area of the property are strongly correlated. As a result, we will be putting the regression line on a scatterplot.

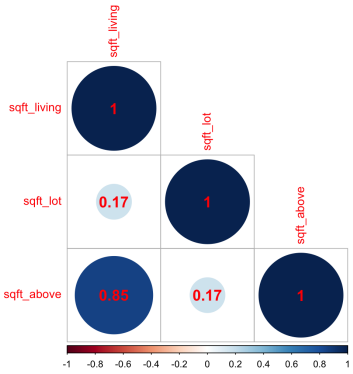


Figure 16: Co-relation of Main Factors

In Figure 17, we have taken both of the main features from the previous correlation plot and here it demonstrates that both the features have almost the same/equal values (which are located along a line that passes through the coordinate system). Although a small dispersion is shown through the line.

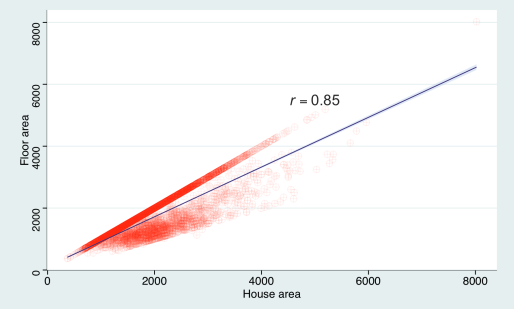


Figure 17: Scatterplot of Main Factors

Therefore, we determine to pick one of the two mentioned feature and discard the other. The feature to stand out is the Floor Area.

The Chi Square test results in Figure 18 shows that the pair of variables are not strongly related to each other and a moderate relationship exists between the age of the house and whether it has been renovated, so we don't eliminate any further variables.

```
0.000 0.022 0.004 0.343 0.059 0.164
```

Figure 18: Chi-Square Results

The data modeling section started with dividing the dataset into train and test with a 3:1 split respectively as shown in Figure 19.

```
Number of rows in the training set: 3195
Number of rows in the test set: 1065
```

Figure 19: Training & Test dataset

6.2 Classification

We come to classification of data using decision trees first with seven main variables chosen to be basement, build, bedrooms, condition, renovation, living area and plot area.

It is evident in Figure 20 that the tree created in first attempt is substantial and trees further optimization. The idea of choosing the best decision tree would be creating the smallest possible where the k folds cross validation error is just 1 standard deviation greater than the minimum error.

```
Classification tree:
rpart(formula = as.character(price2) ~ ., data = train, control = rpart.control(cp = 0,
    xval = 10))

Variables actually used in tree construction:
[1] basement    bedrooms    build       condition   renovated   sqft_living sqft_lot

Root node error: 1435/3195 = 0.44914

n= 3195

           CP nsplit rel error  xerror     xstd
1  0.36933798      0   1.00000 1.00000 0.019593
2  0.01300813      1   0.63066 0.64878 0.017899
3  0.00905923      4   0.59164 0.62369 0.017688
4  0.00441347      5   0.58258 0.60279 0.017502
5  0.00313589     11   0.54704 0.58467 0.017333
6  0.00209059     13   0.54077 0.57561 0.017246
7  0.00185830     30   0.50105 0.57770 0.017266
8  0.00139373     33   0.49547 0.58467 0.017333
9  0.00116144     66   0.44530 0.60488 0.017521
10 0.00104530     69   0.44181 0.61672 0.017627
11 0.00092915     71   0.43972 0.62091 0.017664
12 0.00069686     77   0.43415 0.62300 0.017682
13 0.00055749     94   0.42160 0.63902 0.017819
14 0.00046458     99   0.41882 0.64530 0.017870
15 0.00034843    105   0.41603 0.64530 0.017870
16 0.00027875    119   0.41045 0.64878 0.017899
17 0.00000000    124   0.40906 0.66341 0.018015
```

Figure 20: Single Tree Classification

After being able to successfully create a much smaller tree as shown in Figure 21 which was able to meet the prior conditions, we'll then move on to envision our model created.

```
Classification tree:
rpart(formula = as.character(price2) ~ ., data = train, control = rpart.control(cp = 0,
    xval = 10))

Variables actually used in tree construction:
[1] build       sqft_living sqft_lot

Root node error: 1435/3195 = 0.44914

n= 3195

         CP nsplit rel error  xerror     xstd
1 0.3693380      0   1.00000 1.00000 0.019593
2 0.0130081      1   0.63066 0.64878 0.017899
3 0.0090592      4   0.59164 0.62369 0.017688
4 0.0044135      5   0.58258 0.60279 0.017502
```

Figure 21: Single Tree Classification Optimization

As shown in Figure 22, the highest node shows that living area is the variable with most impactful values compared to other features. This was further visualized by using the weights function.

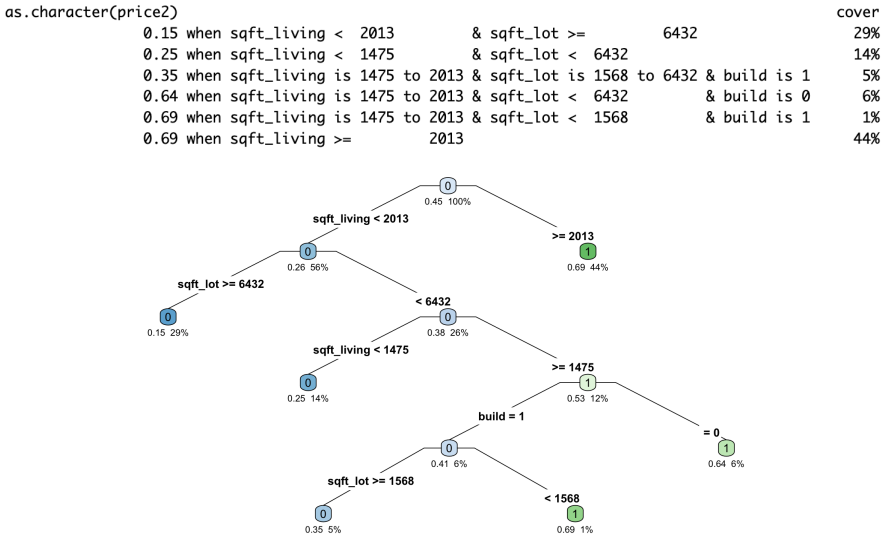


Figure 22: Single Tree Classification Optimization

The Figure 23 further supports the above claim that living area is the most prominent variable in predicting property prices in the classification tree model followed by the number of rooms and the plot area. The most unimportant variable came out to be presence by the sea or simply waterfront.

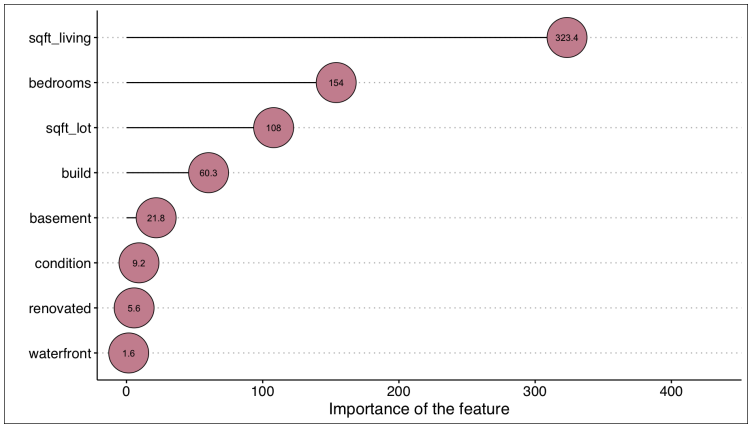


Figure 23: Single Tree Important Variables

Now we know what were the important characteristics according to this model, it's now time to check the viability of model by using it on a test set.

Prediction error in Figure 24 shows that the model is better suited to predicting property type than randomly classifying. The confusion matrix above identified that 263 (103+160) almost 26.7% of total properties were incorrectly classified. 160 properties which were

```
Prediction error on the Test is:  0.2469
```

Real	Prediction 0	1
0	428	160
1	103	374

Figure 24: Prediction Error Single Tree

actually cheaper were predicted to be expensive and 103 happen to be expensive in reality but were predicted to be cheap. The model was comparatively simple and quick to run so we now move on our two ensemble models of tree aggregation of boosting and bagging.

6.2.1 Bagging

Bagging or bootstrap aggregation is used to decrease the variance of a single decision tree classifier. The model works as follows.
Bagging operates as:

1. Take n number of samples from the first data that has been bootstrapped.
2. Draw a decision tree for every sample that was bagged (bootstrapped).
3. Calculate the mean of individual tree's projection and a final model.

Here in Figure 25 we have used 30 trees apriori to bagging tree classification.

```
n= 3195

node), split, n, loss, yval, (yprob)
      * denotes terminal node

 1) root 3195 1466 0 (0.5411581 0.4588419)
   2) sqft_living< 2005 1799  480 0 (0.7331851 0.2668149)
     4) sqft_lot>=6431.5 943  126 0 (0.8663839 0.1336161) *
     5) sqft_lot< 6431.5 856  354 0 (0.5864486 0.4135514)
      10) sqft_living< 1236 250   44 0 (0.8240000 0.1760000) *
      11) sqft_living>=1236 606  296 1 (0.4884488 0.5115512)
        22) build>=0.5 302  123 0 (0.5927152 0.4072848) *
        23) build< 0.5 304  117 1 (0.3848684 0.6151316) *
   3) sqft_living>=2005 1396  410 1 (0.2936963 0.7063037) *
```

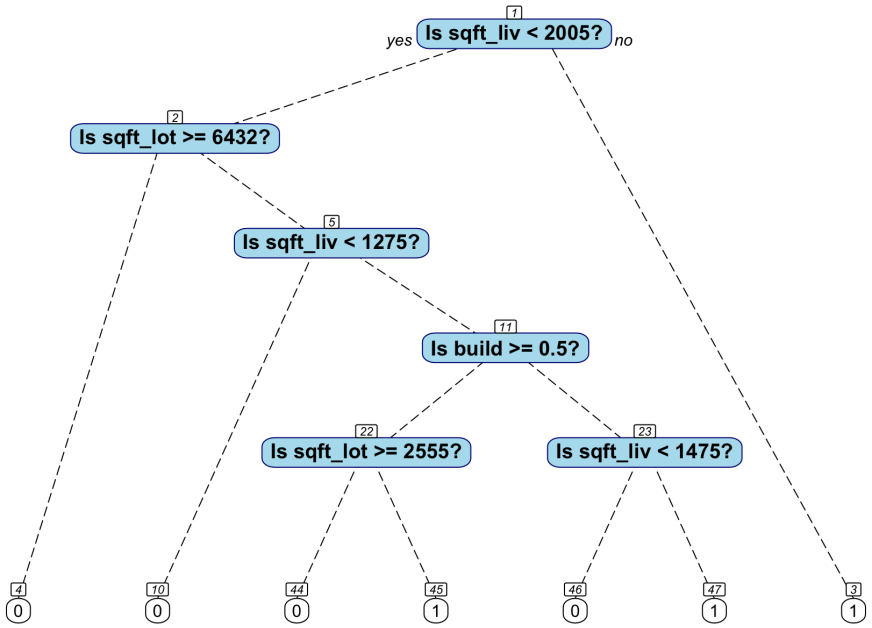


Figure 25: Bagging Classification Tree

The model is run and visualized and it's already smaller compared to the previous classification tree without clipping so we then move on the vote process.

Values in yellow shown in Figure 26 (164 and 161) shows that the values were incorrectly classified. There were rarely any instances where the numbers of both classes matched. However, half of them got correct prediction on the thirty trees.

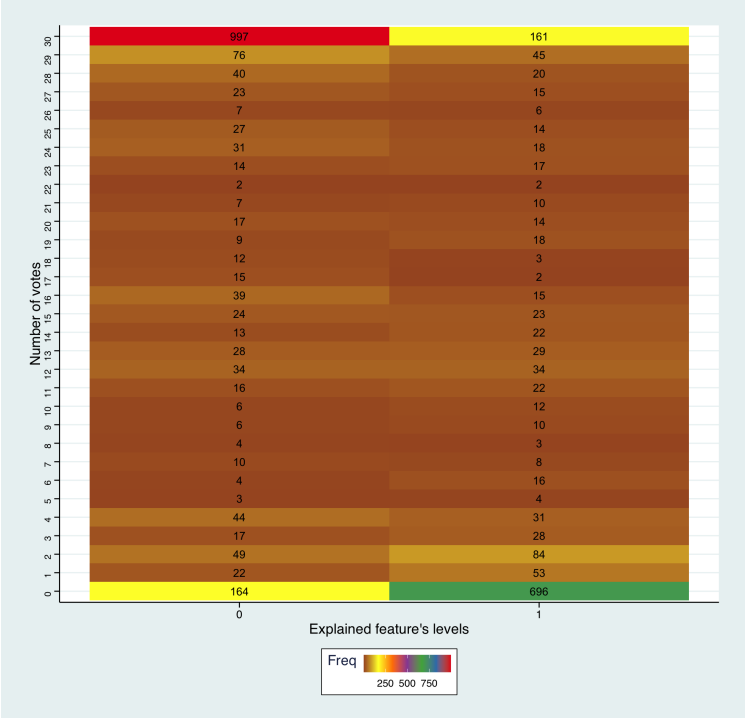


Figure 26: Bagging Classification Votes

As we lay out the weights of this model in Figure 27, we again see living area stand out from the rest of variables and has the highest impact and plot area and build

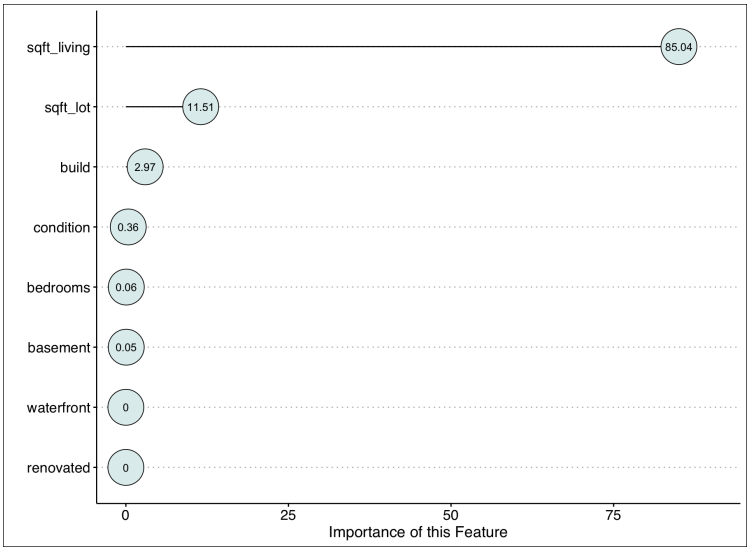


Figure 27: Bagging Classification Important Variables

follow second and third. However, unlike before the bedrooms are considered to be least important which was surprising.
Now moving to the quality check of the model by using Cross Validation.

In Figure 28, the K-fold cross validation gave out the prediction error to be 0.267 where k=10. This also meant that almost 73.3% of the properties were accurately classified.

```
Prediction Error from Cross Validation in Bagging is:  0.2673
```

Figure 28: Prediction Error Bagging

The prediction error on test set using the 10-fold cross validation came out to be 0.246 (as shown in Figure 29) meaning that 75.4% of it.

```
Prediction Error calculated on the Test Set of Bagging is:  0.246

     Prediction
Real   0   1
   0 451 137
   1 125 352
```

Figure 29: Prediction Error Bagging (Test)

Both the errors are close and thus it negates the idea of the model being over trained. The results of confusion matrix show that (262 out 1065) were incorrectly classified. And 125 were classified as false negative whereas 137 came out to be false positive.

Figure 30 shows that the error is fairly constant just after a few attempts as we visualize on the graph how the number of trees impact the classification error. The ideal model would be the one consisting 30 trees.

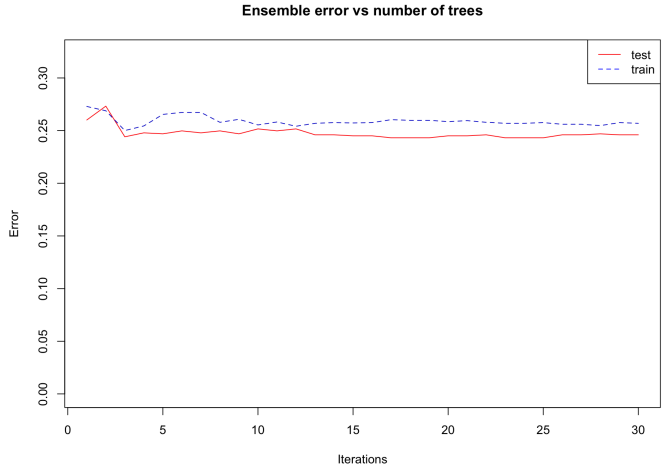


Figure 30: Bagging Error vs Iteration

6.2.2 Boosting

Boosting combines several weak learners into a single better model, similar to bagging. Contrary to bagging, boosting instructs these slow learners in a step-wise manner, with each learner building on the errors of the last. Depending on the size of errors weights are presented in the model. Like bagging, we do 30 trees model. However, unlike bagging, the trees are not very independent and are affected by the previous models.

It is seen in Figure 31 that, living area on the highest node shows its importance in this dataset and how impactful it is compared to other variables. As we visualize the results with weights distribution we can further analyze.

```
n= 3195

node), split, n, loss, yval, (yprob)
      * denotes terminal node

 1) root 3195 1503 0 (0.5295775 0.4704225)
   2) sqft_living< 1695 1221  307 0 (0.7485667 0.2514333) *
   3) sqft_living>=1695 1974  778 1 (0.3941236 0.6058764)
     6) sqft_living< 2705 1469  661 1 (0.4499660 0.5500340)
      12) sqft_lot>=5448.5 1058  528 0 (0.5009452 0.4990548)
        24) bedrooms>=3.5 575  252 0 (0.5617391 0.4382609)
          48) sqft_living< 2185 283   97 0 (0.6572438 0.3427562) *
          49) sqft_living>=2185 292  137 1 (0.4691781 0.5308219)
            98) sqft_lot< 8629 114   48 0 (0.5789474 0.4210526) *
            99) sqft_lot>=8629 178   71 1 (0.3988764 0.6011236) *
        25) bedrooms< 3.5 483  207 1 (0.4285714 0.5714286) *
      13) sqft_lot< 5448.5 411  131 1 (0.3187348 0.6812652) *
     7) sqft_living>=2705 505  117 1 (0.2316832 0.7683168) *
```

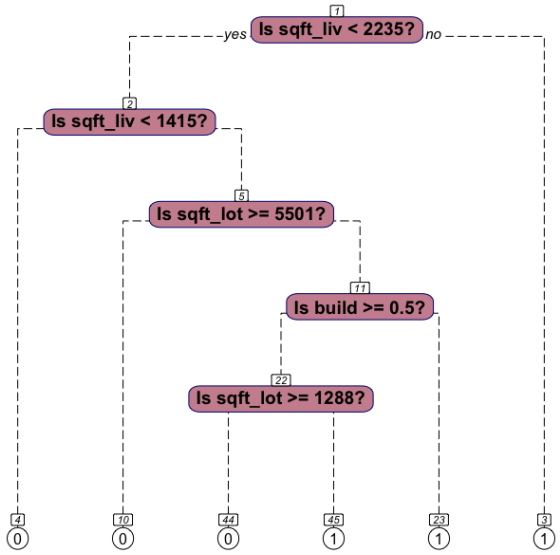


Figure 31: Boosting Classification Tree

We can understand how the weights had been distributed shown in Figure 32.

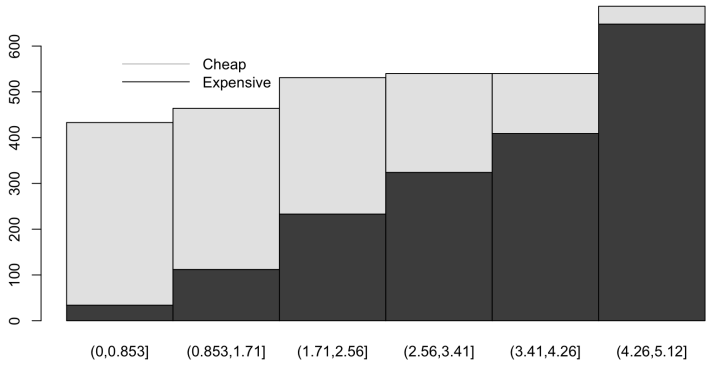


Figure 32: Weights for Cheap vs Expensive

The problem is evident that the votes are for the wrong class and the trees have incorrectly classified the houses.

The importance of feature shown in Figure 33 for this model indicates that living area is considered to be the most important characteristic for deciding property prices followed by plot area and age.

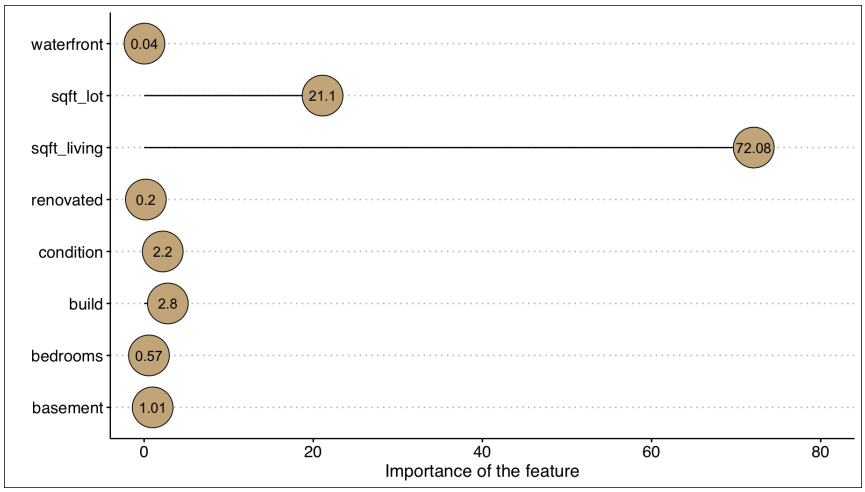


Figure 33: Boosting Important Variables

However, it has also given less importance to renovation which seems unsuitable since build and renovation would be impactful in reality. Presence by water is considered as the most unsought characteristic because it does not have much impact on the price according to this model.

Figure 34, the model is now tested for accuracy using the k-fold cross validation where the prediction error turns out to be 0.258 when k=10 and is somewhat similar to the previous model.

```
i:  1 Mon Jul 31 17:02:48 2023
i:  2 Mon Jul 31 17:02:51 2023
i:  3 Mon Jul 31 17:02:55 2023
i:  4 Mon Jul 31 17:02:58 2023
i:  5 Mon Jul 31 17:03:01 2023
i:  6 Mon Jul 31 17:03:05 2023
i:  7 Mon Jul 31 17:03:08 2023
i:  8 Mon Jul 31 17:03:12 2023
i:  9 Mon Jul 31 17:03:15 2023
i:  10 Mon Jul 31 17:03:19 2023

Prediction Error from Cross Validation in Boosting is:  0.2579
```

Figure 34: Boosting Prediction Error

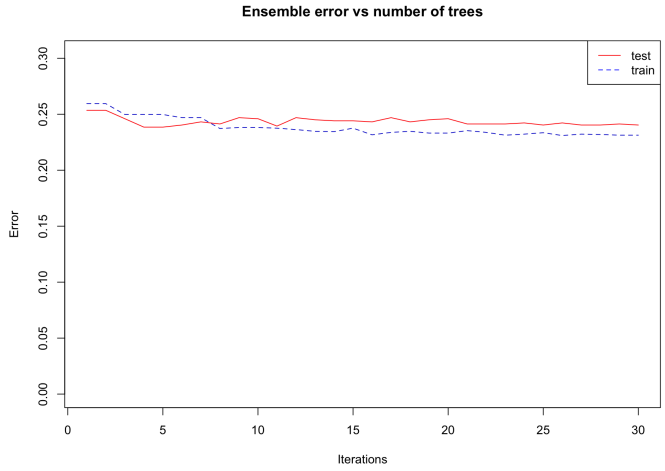
The error is then checked (in Figure 35) on the test set and it came out to be 0.24 and we have 76% correctly classified cases in this scenario. It is again in line with the k-fold validation error and very similar to the bagging error. The confusion error matrix shows less error compared to bagging with only 256 wrongly classified cases out of 1065.

```
Prediction Error calculated on the Test Set of Boosting is:  0.2404

     Prediction
Real   0   1
   0 484 104
   1 152 325
```

Figure 35: Boosting Error Matrix (Test)

Figure 36 helped us see how the error is changed based on the number of trees, the error is seen to be stabilized after 10 iterations on both test and train sets so we decided to build model with 10 trees and then check the prediction error on test set. This gave an even lower error (0.2376) and can replace the 30-tree model.




```
Prediction Error calculated on the Test Set of Boosting2 is:  0.2376
```

Figure 36: Boosting Error vs Iteration

6.2.3 Random Forest

As the technique uses random selection of subset of the data then building several decision trees on it to call it as random forest classification, we move on to that. Here every tree acts an individual source of knowledge about the data and draws its way of classification for the dataset. We start with a model of 1000 trees which are small so they don't need any pruning.

We can see in Figure 37 that just after a few attempts the error fell drastically and stabilizes and so we build a new model with number of Trees=100.

	all	0	1
sqft_living	0.2501	0.2145	0.2933
sqft_lot	0.1065	0.1915	0.0021
bedrooms	0.0608	0.0702	0.0492
waterfront	0.0555	0.0711	0.0363
condition	0.0529	0.0779	0.0220
build	0.0288	0.0209	0.0385
basement	0.0228	0.0413	0.0001
renovated	0.0084	0.0135	0.0021

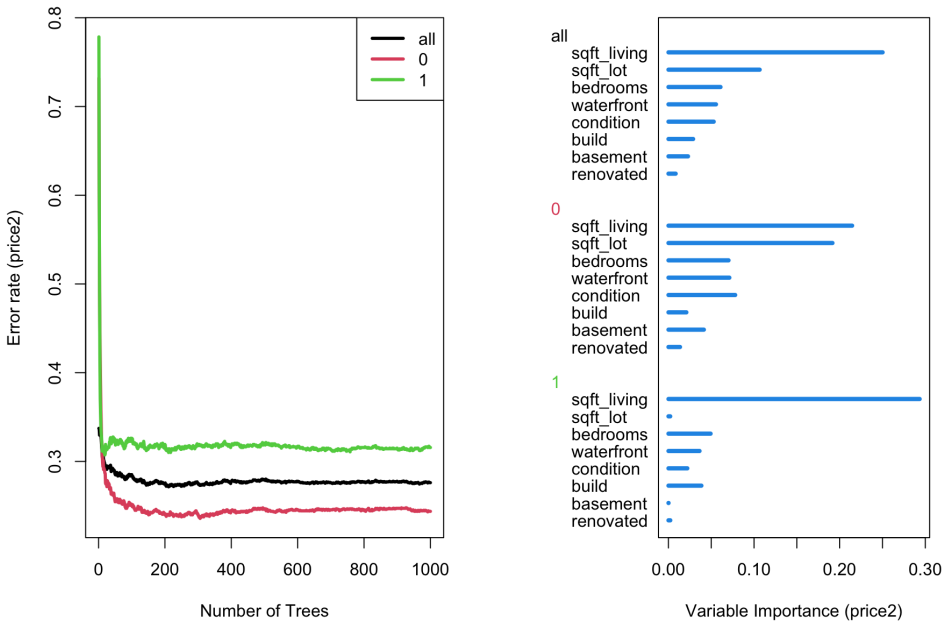


Figure 37: Random Forest Classification (1000 Trees)

It is not appropriate in predicting expensive properties even though the error keeps reducing in the training dataset with the addition of trees shown in Figure 38. It is comparatively easier to observe inexpensive houses but for this particular model, living area is again the most important and distinct variable. Presence of basement and renovation do not hold much importance according to this model.

	all	0	1
sqft_living	0.2511	0.2167	0.2927
sqft_lot	0.1023	0.1811	0.0056
waterfront	0.0573	0.0733	0.0376
bedrooms	0.0563	0.0584	0.0536
condition	0.0495	0.0728	0.0210
build	0.0314	0.0215	0.0436
basement	0.0237	0.0410	0.0024
renovated	0.0068	0.0107	0.0019

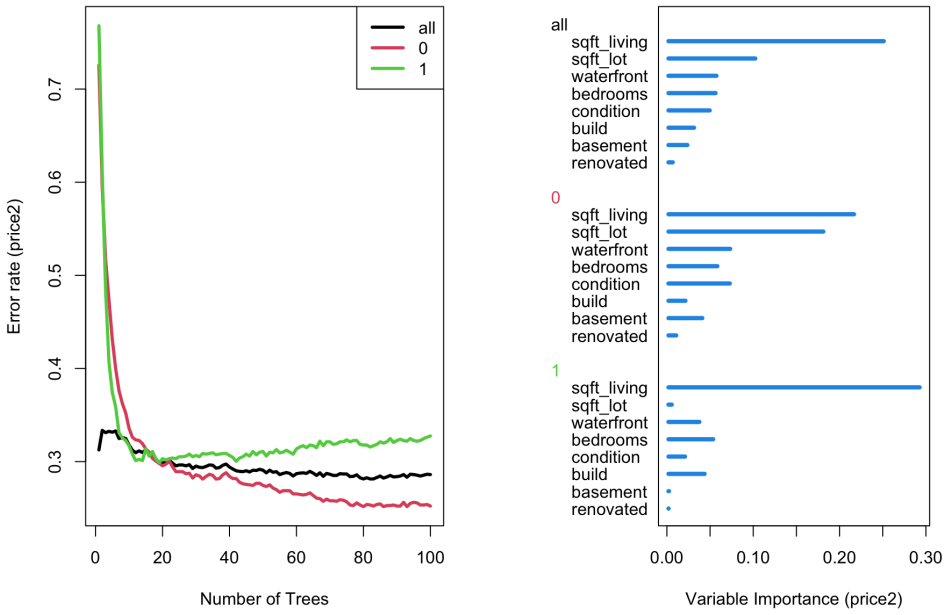


Figure 38: Random Forest Classification (100 Trees)

We check the error on train set which comes out to be 0.285 the average error is 28.5% and the area under the ROC curve is 0.79 as shown in Figure 39.

```
Sample size of test (predict) data: 3195
              Number of grow trees: 100
Average no. of grow terminal nodes: 682.78
       Total no. of grow variables: 8
     Resampling used to grow trees: swr
  Resample size used to grow trees: 3195
                          Analysis: RF-C
                            Family: class
                  Imbalanced ratio: 1.2265
                       Brier score: 0.18763715
           Normalized Brier score: 0.75054859
                               AUC: 0.79352055
                            PR-AUC: 0.75229446
                            G-mean: 0.70910299
        Requested performance error: 0.28607199, 0.25227273, 0.32752613

Confusion matrix:

         predicted
observed    0   1 class.error
       0 1321 439      0.2494
       1  473 962      0.3296

         Misclassification error: 0.285446
```

Figure 39: Random Forest Prediction Error

Figure 40 tells that similar to the previous models when looked at the important characteristics we observe that living area again stands out to be distinct.

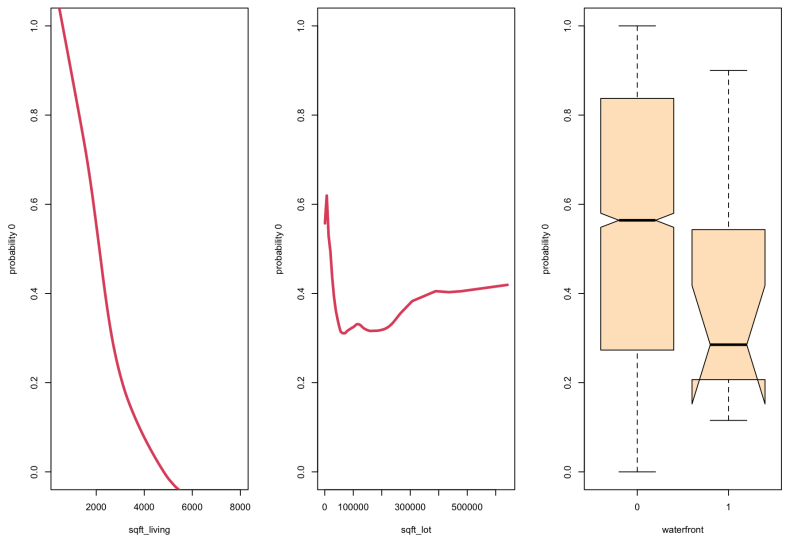


Figure 40: Random Forest Important Variables

As the living area increases, the property price increases and classifies it into an expensive one. As with plot area, the probability first decreases and then starts

increasing with greater area plots. Presence of waterfront also shows an impact.

The final step is to now analyze the model's performance and accuracy on the test set. The random forest model performed worse than the boosting model we just evaluated giving an error of 0.257 while the average error comes at 25.7% as shown in Figure 41.

```
Sample size of test (predict) data: 1065
              Number of grow trees: 100
Average no. of grow terminal nodes: 682.78
       Total no. of grow variables: 8
     Resampling used to grow trees: swr
  Resample size used to grow trees: 3195
                          Analysis: RF-C
                            Family: class
                 Imbalanced ratio: 1.2327
                      Brier score: 0.18368871
          Normalized Brier score: 0.73475483
                               AUC: 0.79926179
                            PR-AUC: 0.75458349
                            G-mean: 0.74022001
        Requested performance error: 0.25633803, 0.23129252, 0.28721174

Confusion matrix:

         predicted
observed   0   1 class.error
       0 452 136      0.2313
       1 138 339      0.2893

         Misclassification error: 0.257277
```

Figure 41: Random Forest Prediction Error (Test)

We will be outlining our results here categorically. We were able to develop and study four of the models which were: Single classification tree, random forest boosting, and bagging (all of which were based on classification trees). We are aware that an error on the test set will decide which model to apply in practice. These are the outcomes:

Boosting: *0.2376*
Bagging: *0.246*
Single Tree: *0.2469*
Random Forest: *0.2572*

The second boosting model gave us the smallest error and random forest gave us the largest. If we differentiate between the most suitable and the worst model, there is a 1.96% of gap.

6.3 Regression

We now turn to the regression modelling of data. We will describe the continuous feature, the value of the property in USD. We will be working on the same features to explaining and will be technically tweaking them a little.

Once again in Figure 42, we only keep the features chosen as predictors in the set and divide our dataset into train and test sets with a ratio of 3:1 respectively.

```
Number of rows in the training set: 3195
Number of rows in the test set: 1065
```

Figure 42: Training & Test dataset

6.3.1 K-Nearest Neighbour

The K-Nearest Neighbour method will be the first technique we employ. This approach entails finding a number of items that closely resemble the instance being studied, then estimating their values. To begin, we will develop a basic model with a single neighbour.

Here in Figure 43, error on the test set came out to be 222084.3, by changing the neighbors we also change quality of the model.

	knn_predl	test_price
1	326000	485000
2	950000	455000
3	925000	383962
4	335000	780000
5	333000	87500

```
Root Mean Squared Error (RMSE) Test is: 222084.3
```

Figure 43: K-NN RMSE

The most appropriate value for k would be decided after looking for the error on different set of neighbors.

In Figure 44, the RMSE value is seen to go down with the increase in the number of neighbors so choose the model with the lowest RMSE value and K=50 which is 163499.

	Number_of_neighbors	RMSE
1	2	193567.1
2	3	184277.9
3	5	176461.8
4	8	171642.2
5	10	169337.5
6	20	164371.5
7	50	163499.2
8	100	166155.0

Figure 44: K-NN Lowest RMSE

6.3.2 Random Forest

Similar to random forest classification. We move to using this model for the regression of data. We start in a similar fashion of creating 1000 trees of small size.

It can be observed by the graph in Figure 45 that after addition of trees after a certain stage which can lead to over-fitting, the error reduces very slowly so we take number of trees as 100.

	Importance	Relative Imp
sqft_living	94147895247	1.0000
waterfront	39645987922	0.4211
condition	20172527494	0.2143
sqft_lot	19056871971	0.2024
bedrooms	18403971350	0.1955
build	8526901462	0.0906
basement	3510632566	0.0373
renovated	866376917	0.0092

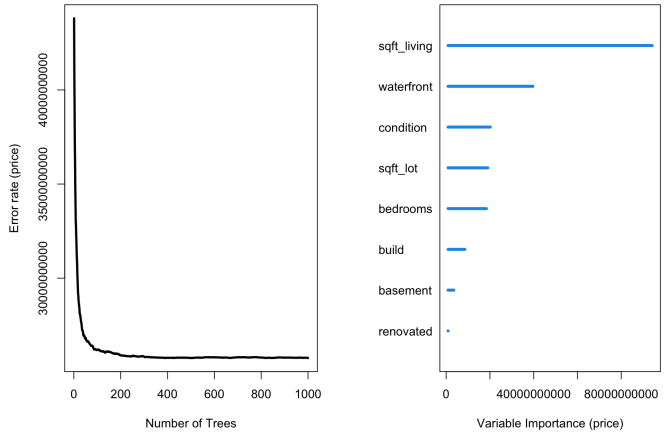


Figure 45: Regression Random Forest (1000 Trees)

Figure 46 shows comparatively better results with stabilization of error as the trees number reach to 100. The three most important variables came out to be living area, presence of water and bedrooms number. Renovation of the house came out to be much insignificant.

	Importance	Relative Imp
sqft_living	94426394541	1.0000
waterfront	43871160137	0.4646
bedrooms	20940779352	0.2218
condition	20665230810	0.2189
sqft_lot	19092764838	0.2022
build	8205123513	0.0869
basement	4089445788	0.0433
renovated	1008777861	0.0107

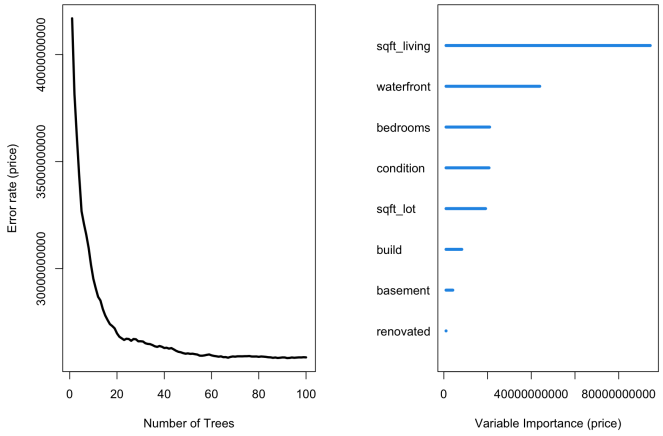


Figure 46: Regression Random Forest (100 Trees)

With the increase in house prices (in Figure 47), the living area is seen to increase similar to what we saw in classification. Area of the plot is seen to increase as with the prices increase but after certain point, with larger and larger plot areas the price is seen to go down.

```
Sample size of test (predict) data: 3195
              Number of grow trees: 100
Average no. of grow terminal nodes: 650.77
       Total no. of grow variables: 8
     Resampling used to grow trees: swr
  Resample size used to grow trees: 3195
                          Analysis: RF-R
                            Family: regr
                         R squared: 0.37763184
      Requested performance error: 25846487727.8735
```

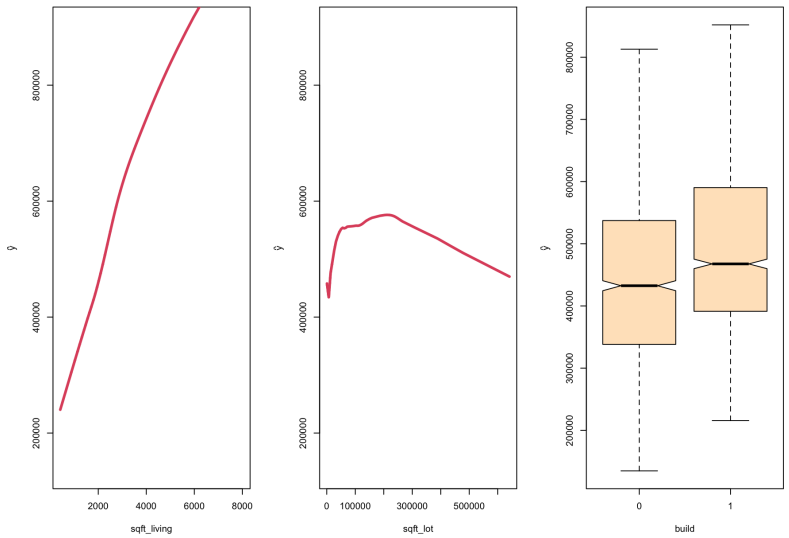


Figure 47: Regression Random Forest Error & Important Variables

Houses built after the year 1970 are seen to have relatively higher prices compared to older houses.

RMSE of this model in Figure 48 came out to be 157740 which is 5759 less than that of K-NN model so it can be considered a little better than the previous mode.

```
Sample size of test (predict) data: 1065
              Number of grow trees: 100
Average no. of grow terminal nodes: 650.77
       Total no. of grow variables: 8
     Resampling used to grow trees: swr
  Resample size used to grow trees: 3195
                          Analysis: RF-R
                            Family: regr
                         R squared: 0.38672636
      Requested performance error: 24882033443.4115

Root Mean Squared Error (RMSE) Test is: 157740
```

Figure 48: Regression Random Forest RMSE

6.3.3 LOESS - Local Regression

By using Loess regression using the Loess() function to a numerical vector, one may smooth it out and predict the Y locally (i.e., the learned values of Xs). Using the span option, which has a value between 0 and 1, it is possible to regulate the neighborhood's size. It regulates the level of smoothing. The fitted curve is hence smoother the higher the amount of span. We have chosen four variables which have showed significance in earlier models including living area, plot area, number of bedrooms and the built year.

With this model however, the RMSE value came out to be greater than in the previous regression models (Figure 49).

```
Call:
loess(formula = price ~ sqft_living + sqft_lot + bedrooms + build,
    data = train2, span = 2)

Number of Observations: 3195
Equivalent Number of Parameters: 14.52
Residual Standard Error: 164900
Trace of smoother matrix: 14.52  (exact)

Control settings:
  span     :  2
  degree   :  2
  family   :  gaussian
  surface  :  interpolate         cell = 0.2
  normalize:  TRUE
 parametric:  FALSE FALSE FALSE FALSE
drop.square:  FALSE FALSE FALSE FALSE
       Root Mean Squared Error (RMSE) Test is: 159569
```

Figure 49: Regression LOESS RMSE

Similar to the classification models we were able to sort the 3 models used in regression based on their errors and are listed as:

Random Forest (*RMSE=157740*)
Loess Regressor (*RMSE=159569*)
K-Nearest Neighbor (*RMSE=163499*)

7 Conclusion

While running the various models, one feature emerged with the greatest importance when it comes to the price of a house. Be it any of the classification methods used, square feet living was given most importance. When mapped onto a weighted graph, square feet living carried greatest importance. In single tree classification modes, the least weighted feature was the location of the house by the water. This model also gave rise to a larger number of false positives and false negatives in the data. Random Forest gave rise to the highest error of 25.7% during the classification process but indicated that square feet living is again the one with greatest importance. The model with the lowest error in classification of data was boosting (23.7%).

In regression however, Random Forest showed the most influential variable on the price is the living space of the house. With the living space increasing, the house starts to be classifying into expensive houses. An over-fitting problem occurs because by adding new trees, the error starts to decrease gradually leading to an over-fitting problem. There is a different situation with the plot area. Initially, the price for a house increases with increasing plot size, but at 20,000 square feet, the value of the house begins to slowly decline with more and more space. Renovation is a debatable feature. New houses which are built after 1970 are on average slightly more expensive than old houses.

The best model in regression was Random Forest but it did not turn out to be the best classification model. In the case of classification, it is possible to use other models (e.g., neural network) or improve the existing model. The Random Forest model showed the least variation when predicting the price of a house based on several features.

Future Work

This study can allow real estate developers and realtors to understand what features that are presented in the data can impact the value of the property Machine Learning Algorithms can help identify the important target features for clients and then appropriately price the house to match the clients price point as well as requirements. The study can not only help sellers but also buyers. This study can be used as a foundation for a variety of research in the future, including real estate market forecasting, and stock price prediction.

The dataset used in this study was limited and had a limited number of variables to analyze the predication of property rates. Due to time constraints involved and complexity of different models, only three models were run and compared. Also, the data was relevant to the US market, for analysis of different geographical areas and the factors affecting their property prices, different datasets can be used to generate a simple yet all-rounder machine learning model for real estate price prediction.

8 Recommendation

Future studies can examine other predictive models, such as neural networks, Lasso, and Ridge regressions, whose performance has been demonstrated previously. this can put stress on future engineering and incorporate user-defined characteristics acquired by mixing a subset of current attributes. This textual tabular information may be combined with visual aspects of the houses, such as photographs of the interiors and exteriors, to create a more robust, innovative house price forecast in the future. Finally, other macroeconomic variables such as the price of gold, the stock market index, property taxes, and the appraised value of a property can impact the housing market; taking these into account can assist in constructing home price prediction models that properly anticipate house values.

9 Author Contributions

The afore-mentioned Authors **Fawwad Ahmed Mirza (F.M.)** and **Aditya Kalpeshbhai Patel (A.P.)** collaborated for this capstone project.

Task	Author	Author
Conceptualization	F.M.	A.P.
Introduction & Literature Review	F.M.	
Methodology	F.M.	
Software		A.P.
Programming & Validation		A.P.
Formal Analysis	F.M.	A.P.
Investigation	F.M	A.P.
Resources	F.M.	
Data Curation		A.P.
Writing Original Draft Preparation	F.M.	A.P.
Writing Review & Editing	F.M.	
Visualization	F.M.	

References

[1] Fresh Books Accounting. Accounting forecasting techniques and tips for small businesses.

[2] Abigail Adetunji, Ajala Funmilola Alaba, Ajala, Ololade Oyewo, Yetunde Akande, Gbenle Oluwadara, and Akande Oluwatobi. House price prediction using random forest machine learning technique. volume Vol. 199, Feb 2022.

[3] Adyan Alfiyatin, Ruth Ema, Hilman Taufiq, and Wayan Mahmudy. Modeling house price prediction using regression analysis and particle swarm optimization case study. *International Journal of Advanced Computer Science and Applications, Malang, East Java, Indonesia*, Vol. 8, Jan 2017.

[4] Sanveed Amate Prayag Adhikaari Vijay Kukre Atharva Chouthai, Mohammed Athar Rangila. House price prediction using machine learning. Vol. 3, Mar 2019.

[5] Tianfeng Chai and R.R. Draxler. Root mean square error (RMSE) or mean absolute error (MAE)?– Arguments against avoiding RMSE in the literature. *Geoscientific Model Development*, Vol. 7:Pg – 1247–1250, June 2014.

[6] Statista Research Department. Average price per square foot of floor space in new single-family houses in the united states from 2000 to 2021.

[7] Statista Research Department. Number of housing units in the united states from 1975 to 2021(in millions).

[8] Bin Geng, Haijun Bao, and Ying Liang. A study of the effect of a high-speed rail station on spatial variations in housing price based on the hedonic model. *Habitat International*, Vol. 49, Oct 2015.

[9] Rinkaj Goyal, Pravin Chandra, and Yogesh Singh. Suitability of knn regression in the development of interaction based software fault prediction models. *IERI Procedia*, Vol. 6:Pg – 15–21, Dec 2014.

[10] William Jacoby. Loess: a nonparametric, graphical tool for depicting relationships between variables. *Electoral Studies*, Vol. 19:Pg – 577–613, Dec 2000.

[11] Ruben Jaen. Data mining: An empirical application in real estate valuation. pages Pg – 314–317, Jan 2002.

[12] Yuhao Kang, Fan Zhang, Wenzhe Peng, Song Gao, Jinmeng Rao, Fábio Duarte, and Carlo Ratti. Understanding house price appreciation using multi-source big geo-data and machine learning. *Land Use Policy*, Vol. 111, July 2020.

[13] Zohreh Karimi. Confusion matrix. Oct, 2021.

[14] Sifei Lu, Zengxiang Li, Zheng Qin, Xulei Yang, and Rick Goh. A hybrid regression technique for house prices prediction. pages Pg – 319–323, Dec 2017.

[15] Edson Melanda, Andrew Hunter, and Michael Barry. Identification of locational influence on real property values using data mining methods. *Cybergeo*, Vol. 2016, Feb 2016.

[16] Fürnkranz J. et al. Plaia A., Buscemi S. Comparing boosting and bagging for decision trees of rankings. pages Pg – 78–99, 2022.

[17] Tang L. Liu H. Refaeilzadeh, P. Cross-Validation. *In: LIU, L., ÖZSU, M.T. (eds) Encyclopedia of Database Systems. Springer, Boston, MA*, 2009.

[18] Shweta Sharma. Classification and regression trees: The use and significance of trees in analytics. *Journal on Recent Innovation in Cloud Computing, Virtualization Web Applications Vol. 5, Issue 1*, Feb, 2022.

[19] Neelam Shinde and Kiran Gawande. Survey on predicting property price. pages Pg – 2–7, Oct 2018.

[20] Quang Truong, Minh Nguyen, Hy Dang, and Bo Mei. Housing price prediction via improved machine learning techniques. *Procedia Computer Science*, Vol. 174:Pg – 433–442, Jan 2020.

[21] Quang Truong, Minh Nguyen, Hy Dang, and Bo Mei. Housing price prediction via improved machine learning techniques. *Procedia Computer Science*, Vol. 174:Pg – 433–442, Jan 2020.

[22] Gulsum Uzut and Selim Buyrukoglu. Prediction of real estate prices with data mining algorithms. *Euroasia Journal of Mathematics Engineering Natural and Medical Sciences*, Vol. 7:Pg – 77–84, May 2020.

[23] Yun Zhao, Girija Chetty, and Dat Tran. "deep learning with xgboost for real estate appraisal". *IEEE Symposium Series on Computational Intelligence (SSCI), Xiamen, China*, pages Pg – 1396–1401, 2019.

[24] Nor Zulkifley, Shuzlina Rahman, Ubaidullah Nor Hasbiah, and Ismail Ibrahim. House price prediction using a machine learning model: A survey of literature. *International Journal of Modern Education and Computer Science*, Vol. 12:Pg – 46–54, Dec 2020.

[25] Ozancan Özdemir. House price prediction using machine learning: A case in iowa. Feb 2022.