Final HPP Capstone Report

MBA (Jain University)

TABLE OF CONTENTS

List of Graphs and Tables

CHAPTER 1

INTRODUCTION AND BACKGROUND

INTRODUCTION AND BACKGROUND

Executive Summary:

In the real estate sector, predicting house prices is a critical application of machine learning and data analysis. To accurately anticipate the future selling price of the houses, it makes use of previous property data, economic factors, and various features. To help buyers, sellers, investors, and real estate professionals make wise decisions, house price forecast aims to give them information.

Data collection is compiling a thorough dataset that contains details about a property, such as its size, location, number of bedrooms, and bathrooms, as well as past sales data and other pertinent information. To train and test machine learning models, you need this dataset. Data cleansing and preparation to take away anomalies, missing values, and discrepancies. It is also possible to use feature engineering to produce additional variables that can increase prediction accuracy.

Feature selection aids in decreasing dimensionality and enhancing model effectiveness. Model selection is the process of selecting the best regression or machine learning algorithm for a given prediction job. Model training helps the selected model by dividing the dataset into training and testing sets. the model learns the connections between the input features and the target pricing.

Using measures like Mean Absolute Error (MAE), Mean Squared Error (MSE), and Root Mean Squared Error (RMSE), one can evaluate a model's performance. To ensure resilience, cross-validation techniques may be used.

When a model is developed to a suitable level, it can be used to forecast home prices. It can be included in applications or websites for real estate to give users quick estimates of property values.

Currently, it is hard to predict the price of a house. There are a lot of variables that affect its price. People are having a hard time knowing the housing price when each house has different conditions. Generally, people want to buy houses that are worth the price. However, they do not know how much a specific condition of the house affects its price. For the sellers, they do not want to sell their houses at a lower-than-average cost. So, they need to be able to accurately predict housing prices so that they can get or give the best price.

When any person/business wants to sell or buy a house, they always face this kind of issue as they don't know the price which they should offer. Due to this, they might be offering too low or too high for the property. Therefore, we can analyze the available data on the properties in the area and can predict the price. We need to find out how these attributes influence house prices. The right pricing is a very important aspect of selling a house. It is very important to understand what the factors are and how they influence house prices. The objective is to predict the right price of the house based on the attributes.

A house's value is simply more than location and square footage. Like the features that make up a person, an educated party would want to know all aspects that give a house its value. For example, you want to sell a house and you don't know the price which you can take—it can't be too low or too high. To find a house price you usually try to find similar properties in your neighborhood and based on the gathered data you will try to assess your house price.

Build a model that will predict the house price when required features are passed to the model. So we will
Find out the significant features from the given features dataset that affect the house price the most.

Build the best feasible model to predict the house price with a 95% confidence level.

As people don't know the features/aspects which commute property price, we can provide them with House Buying and Selling guiding services in the area so they can buy or sell their property with the most suitable price tag and they won't lose their hard-earned money by offering low price or keep waiting for buyers by putting high prices

1.2 Introduction and Background:

House price prediction is also known as the real estate market. The real estate market plays an important role in the economy and people's lives. For many people and families, buying houses is the main thing in their financial decisions. It is also an investment opportunity for those people who looking to capitalize on property.

The real estate market is a multifaceted task that involves using different data sources, statistical techniques, and machine learning algorithms future market value of properties. This predictive modeling serves different types of assistance for homebuyers in making informed decisions to aid real estate professionals and investors in optimizing their portfolios.

The factors influencing house prices are included and can range from location and property size to economic trends, interest rates, and local facilities. Predicting house prices is not only challenging but also valuable.

Exact predictions can empower stakeholders with the knowledge needed to make strategic decisions, whether choosing the right time to buy or sell the property or identifying the right investment opportunities.

In recent years, advancements in data analytics and machine learning have changed the field of house price prediction. Data scientists and real estate professionals now have access to vast amounts of data which can be advantageous to develop sophisticated predictive models. These models can help in the evolving landscape of real estate.

Let's suppose we want to make a data science project on a house price prediction of a company. But before we make a model on this data we have to analyze all the information that is present across the dataset like as what is the price of the house, what the price they are getting, What is the area of the house, and the living measures. These all steps of analyzing and modifying the data come under EDA.

Exploratory Data Analysis (EDA) is an approach that is used to analyze the data and discover trends, and patterns, or check assumptions with the help of statistical summaries and graphical representations.

The main goal of the project is to find accurate predictions of the houses/ properties for the upcoming years. Here is the step-by-step process involved

1. Requirement Gathering – We have to gather the information and extract the main information from it.
2. Normalizing the data
3. Detecting Outliers in the data
4. Analysis and visualization using data

Types of EDA

Depending on the number of columns we can divide EDA into two types.

1. **Univariate Analysis –** In univariate analysis, we analyze or deal with only one variable at a time. The analysis of univariate data is thus the simplest form of

analysis since the information deals with only one quantity that changes. It does not deal with causes or relationships and the main purpose of the analysis is to describe the data find patterns that exist within it.

2. **Bi-variate Analysis –** This type of data involves two different variables. The analysis of this type of data deals with causes and relationships and the analysis is done to find out the relationship between the two variables.
3. **Multivariate Analysis –** When the data involves three or more variables, it is categorized under multivariate.

Depending on the type of analysis we can also subcategorize EDA into parts.

1. **Non-graphical analysis-** In non-graphical analysis, we analyze data using statistical tools like mean median or mode or skewness
2. **Graphical analysis –** In graphical analysis, we use visualization charts to visualize trends and patterns in the data

Data Encoding

There are some models like linear Regression which does not work with categorical dataset in that case we should try to encode categorical dataset into the numerical column. we can use different methods for encoding like label encoding or One-hot encoding. Pandas and Sklearn provide different functions for encoding in our case we will use the Label Encoding function from sklearn to encode.

In this article, we will understand EDA with the help of an example dataset. We will use Python language for this purpose. In this dataset, we used Pandas, NumPy, matplotlib, seaborn, and open datasets libraries. Then loading the dataset into a data frame and reading the dataset using pandas, view the columns and rows of the data, perform descriptive statistics to know better about the features inside the dataset, write the observations, find the missing values and duplicate rows. Discovering the anomalies in the given set and removing those anomalies. Univariate visualization of each field in the raw dataset, with summary statistics. Bivariate visualizations and summary statistics allow you to assess the relationship between each variable in the dataset and the target variable you‘re looking at. Predictive models, such as linear regression, use statistics and data to predict outcomes.

Plotting the graphs with different attributes of the dataset and analyzing the given dataset. Then Use the algorithms of regression to understand which is a better fit for the data set in house price prediction using model matrix i.e., Mean Squared error, Mean absolute error, Root Mean squared error, R-Squared. Analyze these model matrix for all algorithms in the form of a table then identify the best fit

Problem statement:

Accurately estimating a house's worth in the real estate market is a complex task that extends beyond just its location and size. A house's pricing is influenced by a variety of complex aspects. When homeowners want to sell their homes, they frequently struggle with the challenge of choosing an asking price that is both competitive and not too expensive to turn away potential purchasers

A house's value is simply more than location and square footage. Like the features that make up a person, an educated party would want to know all aspects that give a house its value. For example, you want to sell a house and you don't know the price which you can take—it can't be too low or too high. In the past, individuals have determined the worth of their home by contrasting it with other homes in the area, although this method may not always be accurate.

Objective of study:

- Create an effective price prediction model
- Validate the model's prediction accuracy
- Identify the important home price attributes which feed the model's predictive power

Take advantage of all of the feature variables available below, and use them to analyze and predict house prices.

1. cid: a notation for a house

2. day hours: Date house was sold

3. price: Price is prediction target

4. room_bed: Number of Bedrooms/House

5. room_bath: Number of bathrooms/bedrooms

6. living_measure: square footage of the home

7. lot_measure: square footage of the lot

8. ceil: Total floors (levels) in the house

9. coast: House which has a view of a waterfront

10. sight: Has been viewed

11. condition: How good the condition is (Overall)

12. quality: grade given to the housing unit, based on the grading system

13. ceil_measure: square footage of house apart from the basement

14. basement_measure: square footage of the basement

15. yr_built: Built Year

16. yr_renovated: Year when the house was renovated

17. zip code: zip

18. lat: Latitude coordinate

19. Long: Longitude coordinate

20. living_measure15: Living room area in 2015(implies-- some renovations) This might or might not have affected the lot size area

21. lot_measure15: lot Size area in 2015(implies-- some renovations)

22. furnished: Based on the quality of the room

23. total_area: Measure of both living and lot

Company and Industry Overview

The real estate market is one of the most competitive in terms of pricing and the same tends to vary significantly based on lots of factors, forecasting property price is an important module in decision-making for both the buyers and investors in supporting budget allocation, finding property finding stratagems and determining suitable policies hence it becomes one of the prime fields to apply the concepts of machine learning to optimize and predict the prices with high accuracy.

The literature review gives a clear idea and it will serve as support for future projects. most of the authors have concluded that artificial neural networks have more influence in predicting but in the real world, other algorithms should be taken into consideration. Investors' decisions are based on the market trends to reap maximum returns.

Developers are interested in knowing the future trends for their decision making, this helps to know about the pros and cons and also helps to build the project. To accurately estimate property prices and future trends, a large amount of data that influences land prices is required for analysis, modeling, and forecasting.

The factors that affect the land price have to be studied and their impact on price has also to be modeled. It is inferred that establishing a simple Regression linear mathematical relationship for these time-series data is found not viable for prediction.

Hence it became imperative to establish a non-linear model that can well fit the data characteristics to analyze and predict future trends. As real estate is fast developing sector, the analysis and prediction of land prices using mathematical modeling and other techniques is an immediate and urgent need for decision making by all those concerned

Overview of theoretical concepts:

Data Collection and Preprocessing:

Data Gathering: Collect relevant data from various sources, which may include databases, APIs, web scraping, or manual data entry.

Data Cleaning: Handle missing values, outliers, and inconsistencies in the data.

Data Transformation: Convert data into a suitable format, scale variables, and encode categorical features.

Exploratory Data Analysis (EDA):

Data Visualization: Create plots, charts, and graphs to understand data distributions, correlations, and patterns.

Descriptive Statistics: Compute summary statistics to gain insights into the data.

Feature Engineering:

Feature Selection: Choose the most relevant variables for prediction.

Feature Extraction: Create new features from existing data to improve model performance.

Machine Learning Models:

Regression: Utilize linear regression, decision trees, random forests, or other regression algorithms for predictive tasks.

Classification: Apply classification algorithms like logistic regression, support vector machines, or neural networks for categorical outcomes.

Time Series Analysis: If your data is time-dependent, consider models like ARIMA or LSTM for time series forecasting.

Model Evaluation and Selection:

Cross-Validation: Assess model performance through techniques like k-fold cross-validation.

Metrics: Use appropriate evaluation metrics such as RMSE, MAE, accuracy, F1-score, or AUC-ROC depending on the nature of the problem.

Hyperparameter Tuning: Optimize model hyperparameters to enhance predictive accuracy.

Ensemble Methods: Combine multiple models using techniques like bagging, boosting, or stacking for improved predictions.

Deployment: Deploy the predictive model in a production environment, which may involve creating APIs or integrating it into a web application.

Ethical Considerations: Address potential bias, fairness, and privacy concerns in your predictive model.

Interpretability**:** Ensure that your model's predictions are explainable, especially if used in critical decision-making.

Data Visualization for Reporting: Create interactive dashboards or reports to communicate the results effectively.

Project Management: Plan and execute the project effectively, including setting milestones, managing data, and collaborating with stakeholders.

Documentation and Presentation: Document your work comprehensively and present your findings and methodology to a non-technical audience.

Domain Knowledge: Gain domain-specific knowledge to better understand the problem and refine your predictive model.

Tools and Libraries: Utilize programming languages (e.g., Python, R), libraries (e.g., Scikit-Learn, TensorFlow), and tools (e.g., Jupyter, Tableau) that are relevant to your project.

CHAPTER 2

Research Methodology

RESEARCH METHODOLOGY

2.1 Scope of the Study

2.2 Methodology

2.2.1 Research Design

2.2.2 Data Collection

2.2.3 Sampling Method (if applicable)

2.2.4 Data Analysis Tools

2.3 Period of Study

2.4 Utility of Research

Scope of the study:

The scope of this study is to present a thorough examination of the real estate market, complete with price predictions, data-driven insights, and suggestions. By utilizing the available feature variables and predictive modeling approaches, it seeks to help different stakeholders make educated decisions about purchasing, selling, or listing properties.

Methodology:

Software Requirements

Software requirements deal with defining resource requirements and prerequisites that need to be installed on a computer to provide the functioning of an application. These requirements need to be installed separately before the software is installed. The minimal software requirements are as follows,

1. FRONT END: PYTHON

2. IDE: JUPITER

3. OPERATING SYSTEM: WINDOWS 10

Importing the libraries

In this project, I used Python's powerful libraries to make the machine-learning models efficient. Three essential libraries NumPy, Pandas, and Sci-kit learn had been used in all the machine learning models. NumPy is a powerful library for implementing scientific computing with Python. The most important object of NumPy is the homogeneous multidimensional array.

NumPy saves us from writing inefficient and tiresome huge calculations. NumPy provides a way more elegant solution for mathematical calculations in Python. It provides an alternative to the regular Python lists. A NumPy array is similar to a regular Python list with one additional feature. You can perform calculations over all entire arrays easily, and super-fast as well.

Pandas is a flexible open-source Python library with high-performance, flexible, and expressive data structures. Pandas work better with relational and labeled data. Though Python is great for data mining and preparation, python lags greatly in practical, real-world data analysis and modeling. Pandas help greatly in filling these gaps. It is called the most powerful tool for data analysis and data manipulation.

Scikit-learn is a great open-source package providing a good chain of supervised and unsupervised algorithms. Scikit-learn is built upon scientific Python (SciPy). This library is primarily focused on modeling data. A few popular models of Scikit-learn are clustering, cross-validation, ensemble methods, feature extraction, and feature selection.

Data Collection:

We collected good amounts of data(21,634) on households from Kaggle. It contains 23 different parameters for processing the price prediction (Expected CTC). We have observed it contains both numerical and categorical data.

We do have Missing values in the room bed, room bath, sight, condition, yr built, living measure15, lot measure15, furnished, and total area.

IDX	INDEX
Cid	A notation for a house
Day_hours	The date the house was sold
Price	Price is the prediction target
Room_bed	Number of Bedrooms/House
Room_bath	Number of bathrooms/bedrooms
Living_measure	Square footage of the home
Lot_measure	Square footage of the lot
Ceil	Total floors (levels) in house
Coast	House which has a view to a waterfront
Sight	Has been viewed
Condition	How good the condition is (Overall)
Quality	The grade given to the housing unit, based on the grading system
Ceil_measure	Square footage of the house apart from the basement

Basement	Square footage of the basement
yr_built	Built Year
yr_renovated	The year when the house was renovated
Zipcode	Zip
Lat	Coordinate
Long	Coordinate
living_measure15	Living room area in 2015(implies-- some renovations) This might or might not have affected the lot size area
lot_measure15	Lot Size area in 2015(implies-- some renovations)
Furnished	Based on the quality of the room
Total_area	A measure of both living and lot

Table 1

Handling Missing data:

The important part and problem of data preprocessing is handling missing values in the dataset. Data scientists must manage missing values because it can adversely affect the operation of machine learning models. Data can be imputed in such a procedure, missing values can be filled based on the other observations. Techniques involved in imputing unknown or missing observations include:

1. Deleting the whole rows or columns with unknown or missing observations.

2. Missing values can be inferred by averaging techniques like mean, median, and mode.

3. Imputing missing observations with the most frequent values.

4. Imputing missing observations by exploring correlations.

5. Imputing missing observations by exploring similarities between cases.

Missing values are usually represented with _nan‘, ‘NA‘, or _null‘.

The 5-factor analysis of the features:

1. **CID:** House ID/Property ID.Not used for analysis
2. **Day hours:** 5-factor analysis is reflected in this column
3. **price:** Our target column value is in the 75k - 7700k range. As Mean > Median, it's **Right-Skewed**.
4. **room_bed:** The number of bedrooms ranges from 0 - 33. As the Mean slightly > Median, it's **slightly Right-Skewed.**
5. **room_bath:** The number of bathrooms ranges from 0 - 8. As the Mean is slightly < Median, it's **slightly Left-Skewed**.
6. **living_measure:** The square footage of the house range from 290 - 13,540. As Mean > Median, it's **Right-Skewed**.

7. **lot_measure:** The square footage of the lot range from 520 - 16,51,359. As the Mean almost doubles of Median, it's **highly right-skewed**.
8. **ceil:** The number of floors ranges from 1 - 3.5 As Mean ~ Median, it's **almost Normal Distributed**.
9. **coast:** This value represents whether the house has a waterfront view or not. It's a **categorical column**. From the above analysis, we got to know, that very few houses have waterfront views.
10. **sight:** Value ranges from 0 - 4. As Mean > Median, it's **Right-Skewed**
11. **condition:** Represents the rating of a house which ranges from 1 - 5. As Mean > Median, it's **Right-Skewed**
12. **quality:** Representing grades given to house which range from 1 - 13. As Mean > Median, it's **Right-Skewed**.
13. **ceil_measure:** The square footage of the house apart from the basement ranges from 290 - 9,410. As Mean > Median, it's **Right-Skewed**.
14. **basement:** The square footage house's basement ranges from 0 - 4,820. As Mean highly > Median, it's **Highly Right-Skewed**.
15. **yr_built:** House built year ranges from 1900 - 2015. As Mean < Median, it's **Left-Skewed**.
16. **yr_renovated:** House renovation year only 2015. So this column can be used as a **Categorical Variable** for knowing whether the house is renovated or not.
17. **zipcode:** House ZipCode ranges from 98001 - 98199. As Mean > Median, it's **Right-Skewed**.
18. **lat:** Lattitude ranges from 47.1559 - 47.7776 As Mean < Median, it's **Left-Skewed**.
19. **long:** Longitude ranges from -122.5190 to -121.315 As Mean > Median, it's **Right-Skewed**.
20. **living_measure15:** Value ranges from 399 to 6,210. As Mean > Median, it's **Right-Skewed**.
21. **lot_measure15:** Value ranges from 651 to 8,71,200. As Mean highly > Median, it's **Highly Right-Skewed**.
22. **furnished:** Representing whether the house is furnished or not. It's a **Categorical Variable**
23. **total_area** Total area of the house ranges from 1,423 to 16,52,659. As the Mean is almost double of Median, it's **Highly Right-Skewed**

From the above analysis, we got to know,

Most column's distribution is Right-Skewed and only a few features are Left-Skewed (like room_bath, yr_built, lat).

We have columns that are Categorical in nature -> coast, yr_renovated, furnished.

Exploratory Data Analysis:

EDA refers to deep analysis of data to discover different patterns and spot anomalies. Before making inferences from data it is essential to examine all your variables.

visual data analysis of the features:

Uni-Variate, Bi-Variate, Multi-Variate:

Uni-Variate: Uni-Variate in House Price Prediction, chosen attribute like price because by price is independent of each other.

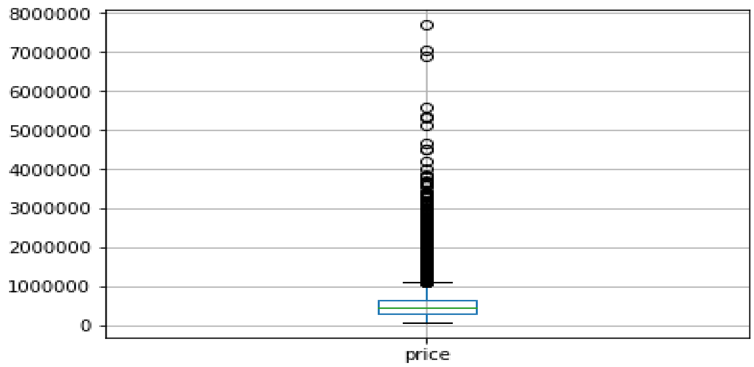


Fig 1

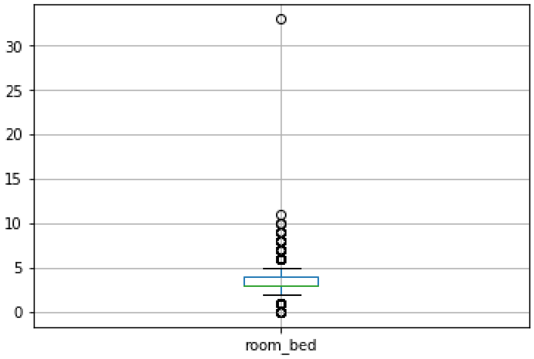


Fig 2

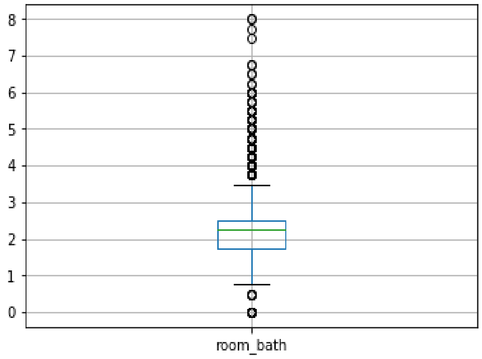
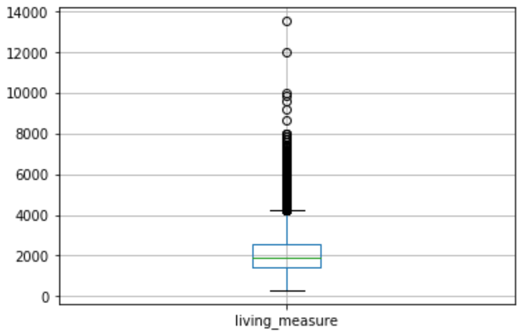


Fig 3



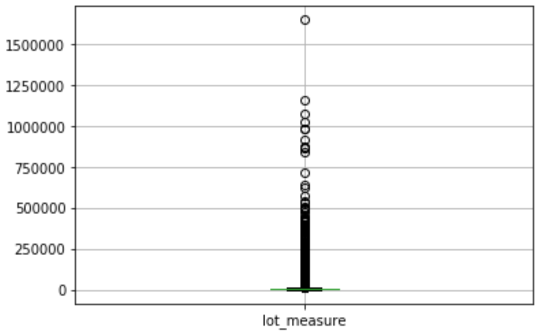
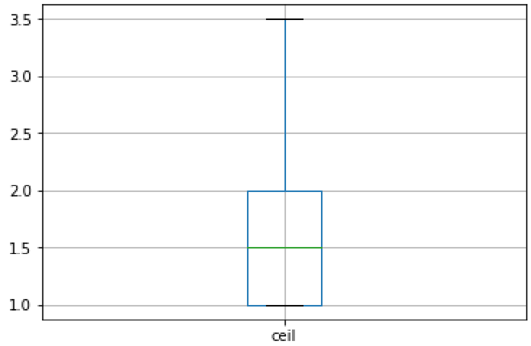


Fig 4

Fig 5



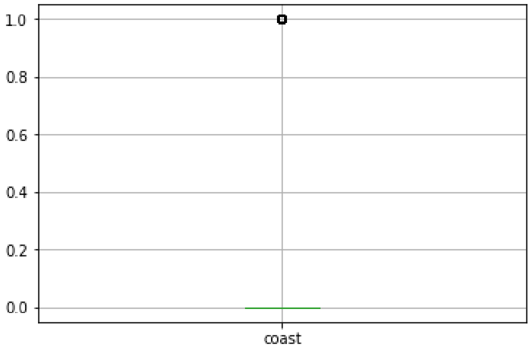
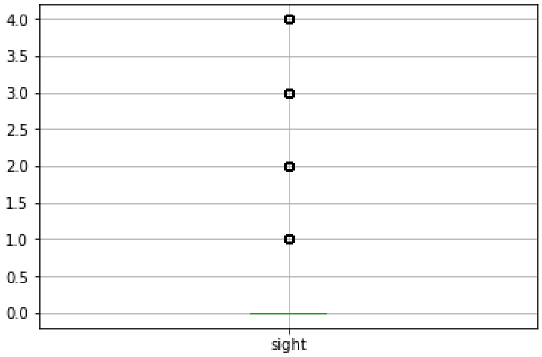


Fig 6

Fig 7



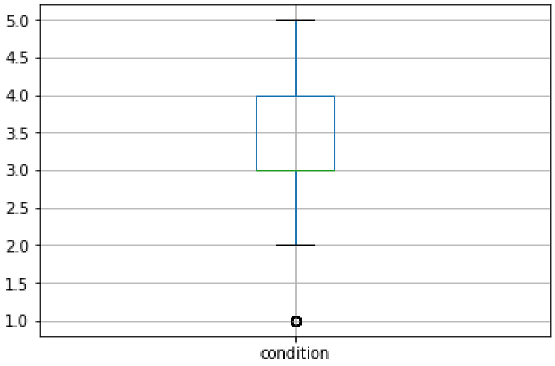
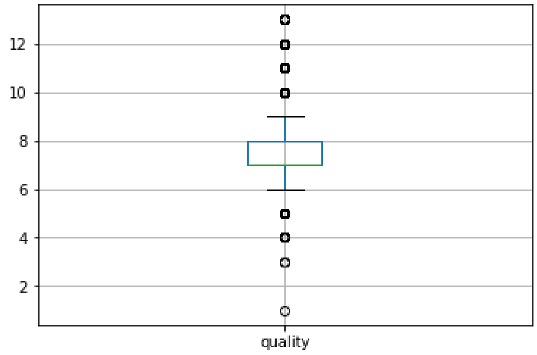


Fig 8

Fig 9



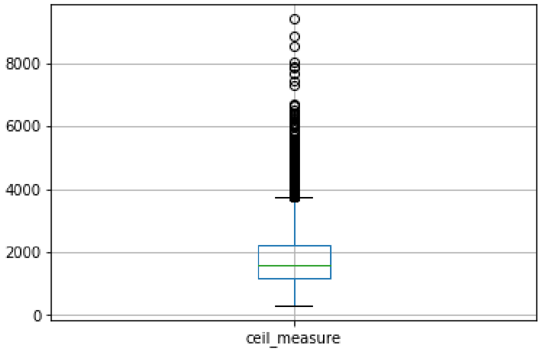


Fig 10

Fig 11

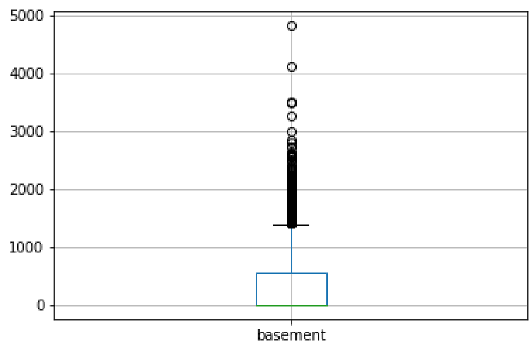


Fig 12

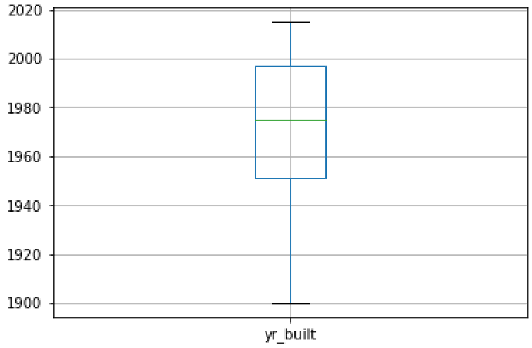


Fig 13

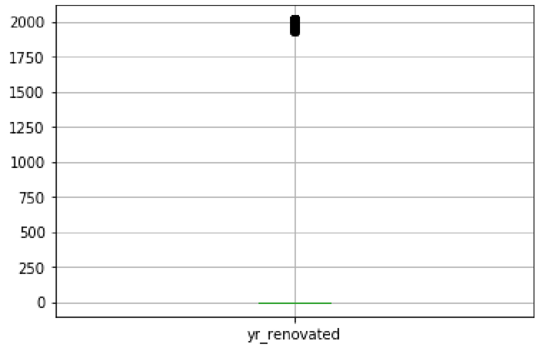


Fig 14

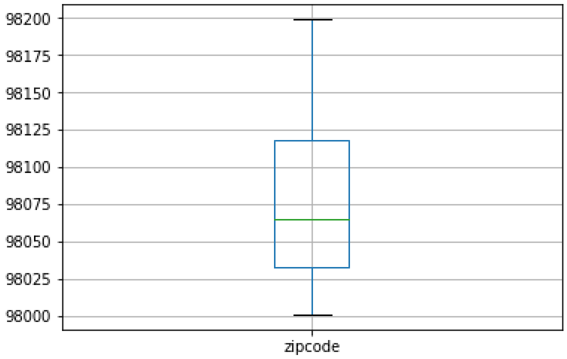


Fig 15

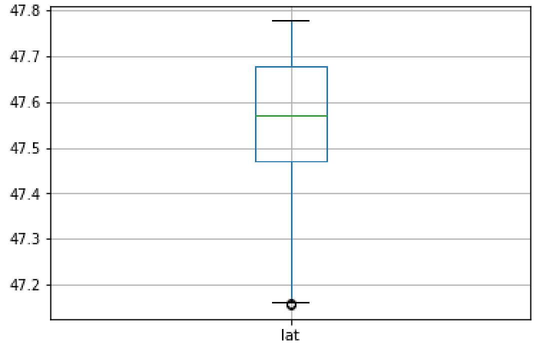


Fig 16

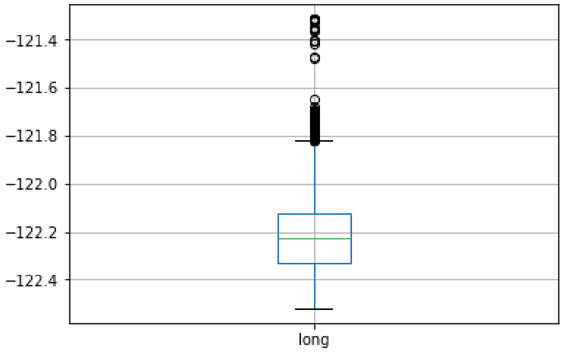


Fig 17

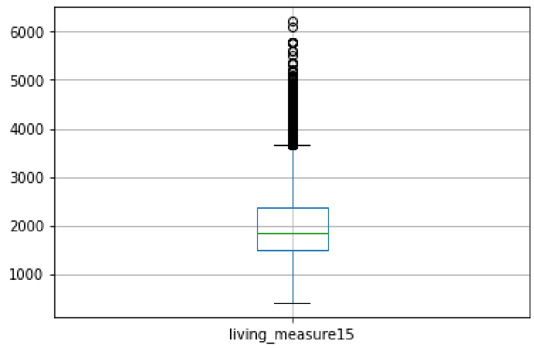


Fig 18

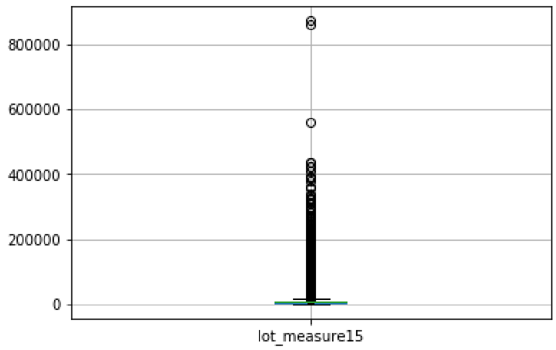


Fig 19

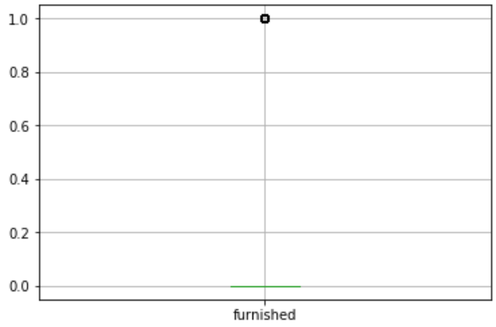


Fig 20

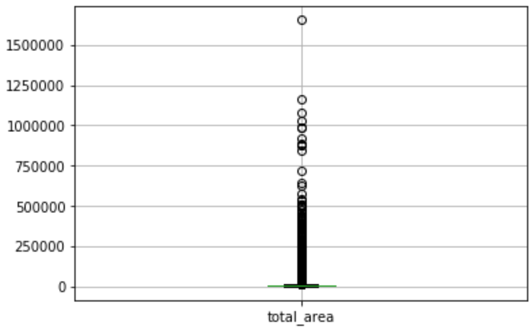


Fig 21

We can see, that there are a lot of features that have outliers. So we might need to treat those before building a model.

Bi-Variate:

From the below pair plot, we observed/deduced below

1. **price:** price distribution is Right-Skewed as we deduced earlier from our 5-factor analysis
2. **room_bed:** our target variable (price) and room_bed plot is not linear. Its distribution has a lot of Gaussians
3. **room_bath:** Its plot with price has a **somewhat linear relationship**. Distribution has several Gaussians.
4. **living_measure:** The plot against price has a **strong linear relationship**. It also has a linear relationship with the room bath variable. So **might remove one of these 2**. Distribution is Right-Skewed.
5. **lot_measure: No clear relationship** with price.

6. **ceil: No clear relationship** with price. We can see, it's **have 6 unique values** only. Therefore, we can **convert this column into a categorical column** for values.
7. **coast: No clear relationship** with price. It's a **categorical variable with 2 unique values**.
8. **sight: No clear relationship** with price. This has **5 unique values**. This can be **converted to a Categorical variable**.
9. **condition: No clear relationship** with price. This has **5 unique values**. It can be **converted to a Categorical variable**.
10. **quality: Somewhat linear relationship with price**. Has **discrete values from 1 - 13. It can be converted to a Categorical variable**.
11. **ceil_measure: Strong linear relationship with price**. Also with room_bath and living_measure features. Distribution is **Right-Skewed**.
12. **basement: No clear relationship** with price.
13. **yr_built: No clear relationship** with price.
14. **yr_renovated: No clear relationship** with price. Have **2 unique values. Can be converted to Categorical Variable** which tells whether the house is renovated or not.
15. **zipcode, lat, long: No clear relationship** with price or any other feature.
16. **living_measure15: Somewhat linear relationship with target feature**. It's the same as living_measure. Therefore we can drop this variable.
17. **lot_measure15: No clear relationship** with price or any other feature.
18. **furnished: No clear relationship** with price or any other feature. **2 unique values so can be converted to Categorical Variable**
19. **total_area: No clear relationship with price**. But it has a **Very Strong linear relationship with lot_measure**. So one of it can be dropped.

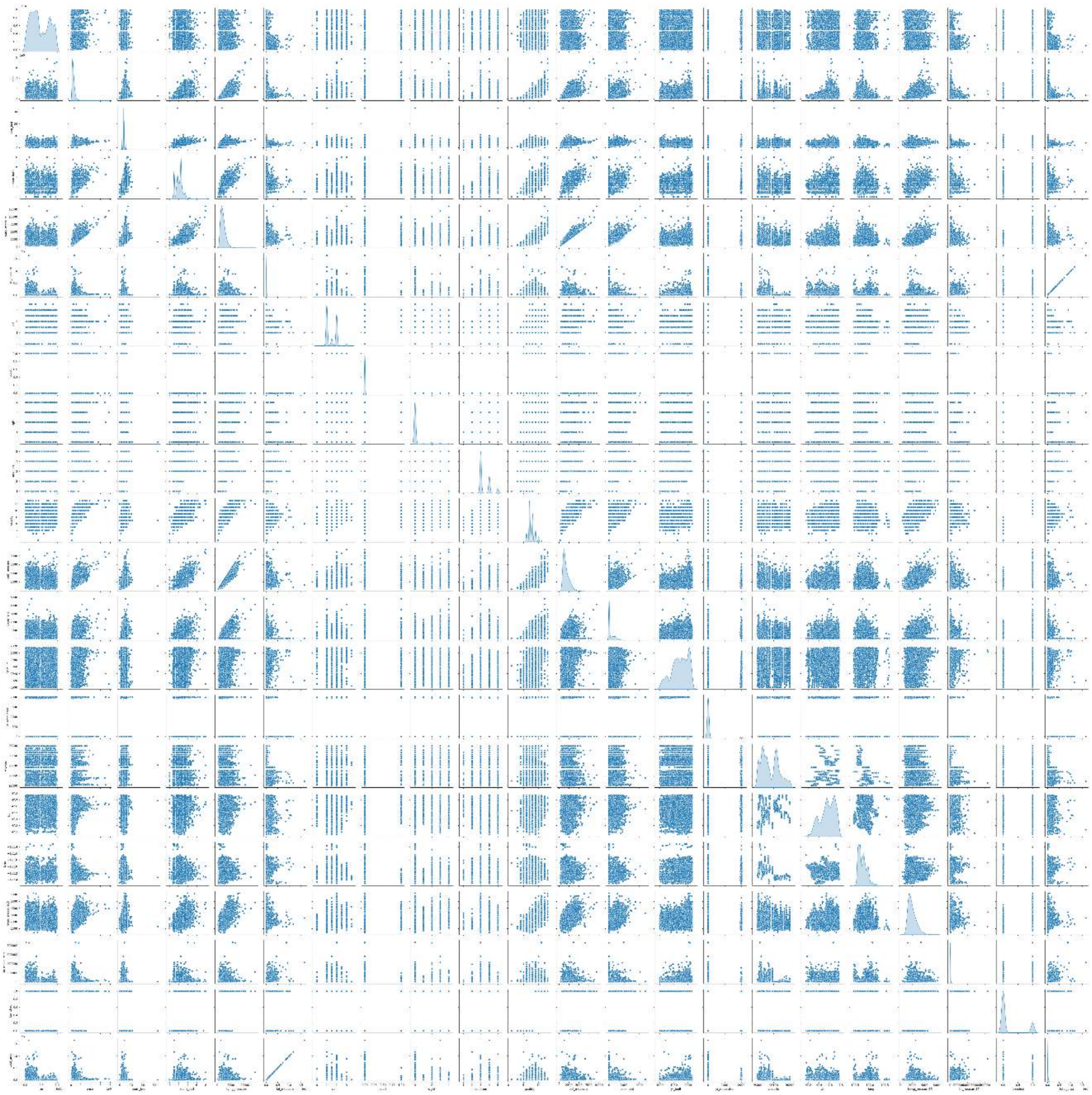


Fig 22

We have linear relationships in the below featues as we got to know from the matrix

1. **price**: room_bath, living_measure, quality, living_measure15, furnished
2. **living_measure**: price, room_bath. So we can consider dropping 'room_bath' variable.
3. **quality**: price, room_bath, living_measure
4. **ceil_measure**: price, room_bath, living_measure, quality
5. **living_measure15**: price, living_measure, quality. So we can consider dropping living_measure15 as well. As it's giving the same info as living_measure.
6. **lot_measure15**: lot_measure. Therefore, we can consider dropping lot_measure15, as it's giving the same info.

7. **furnished**: quality
8. **total_area**: lot_measure, lot_measure15. Therefore, we can consider dropping total_area feature as well. As it's giving the same info as lot_measure.

We can plot heatmap and can easily confirm our above findings

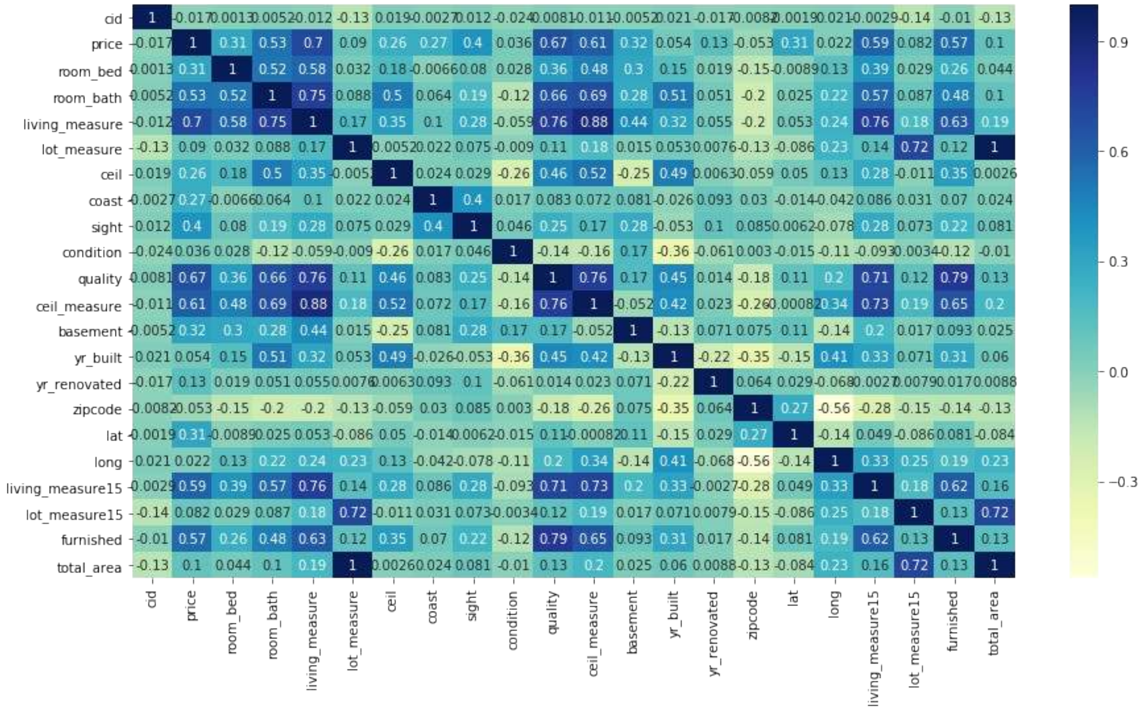


Fig 23

CHAPTER 3

DATA ANALYSIS AND INTERPRETATION

DATA ANALYSIS AND INTERPRETATION :

Linear regression model

The Linear regression model shows a linear relationship between a dependent and one or more independent variables, hence called linear regression. Since linear regression shows the linear relationship, which means it finds how the value of the dependent variable changes according to the value of the independent variable.

Model evaluation against Linear Regression Mean Absolute Error (MAE):

It is the simplest error metric used in regression problems. It is the sum of the average of the absolute difference between the predicted and actual values.

Mean Square Error (MSE): MSE is like the MAE, but the only difference is that it squares the difference between actual and predicted output values before summing them all instead of using the absolute value.

Root Mean Squared Error (RSME): RSME provides information about the short-term performance of a model by allowing a term-by-term comparison of the actual difference between the estimated and the measured value.

R Squared (R2): R Squared metric is generally used for explanatory purposes and indicates the goodness or fit of a set of predicted output values to the actual output values.

Linear Regression (with Ridge and Lasso)

Ridge regression is a technique used to analyze multi-linear regression (multi-collinear), also known as L2 regularization.

The linear regression model performed with scores 0.73 & .72 in training data set and validation data set respectively

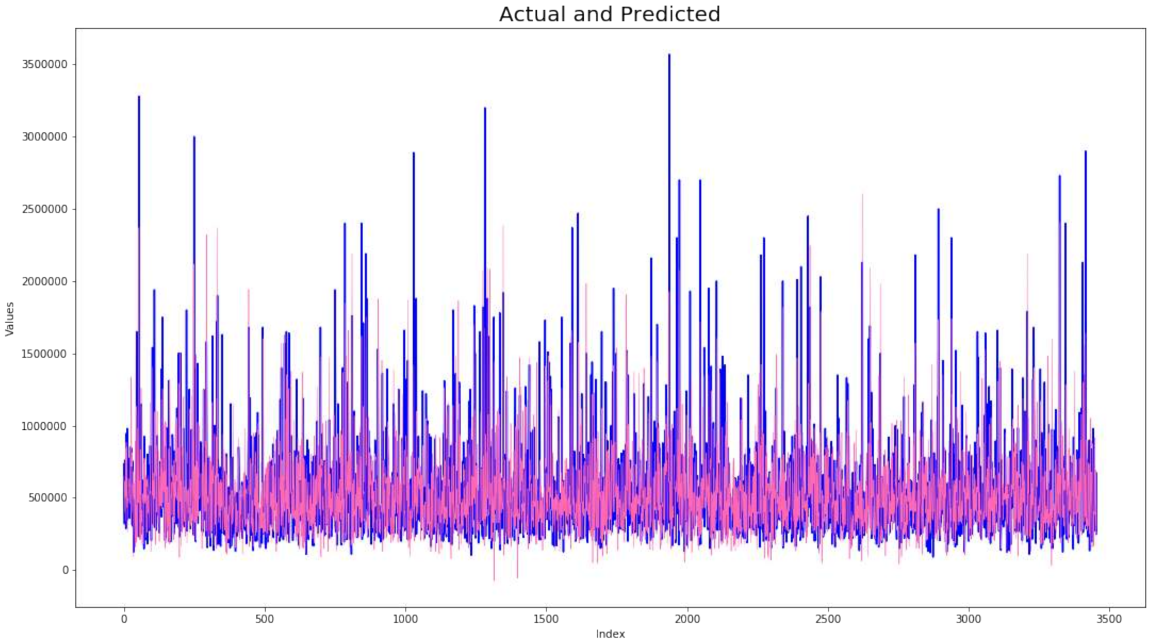


Fig 24

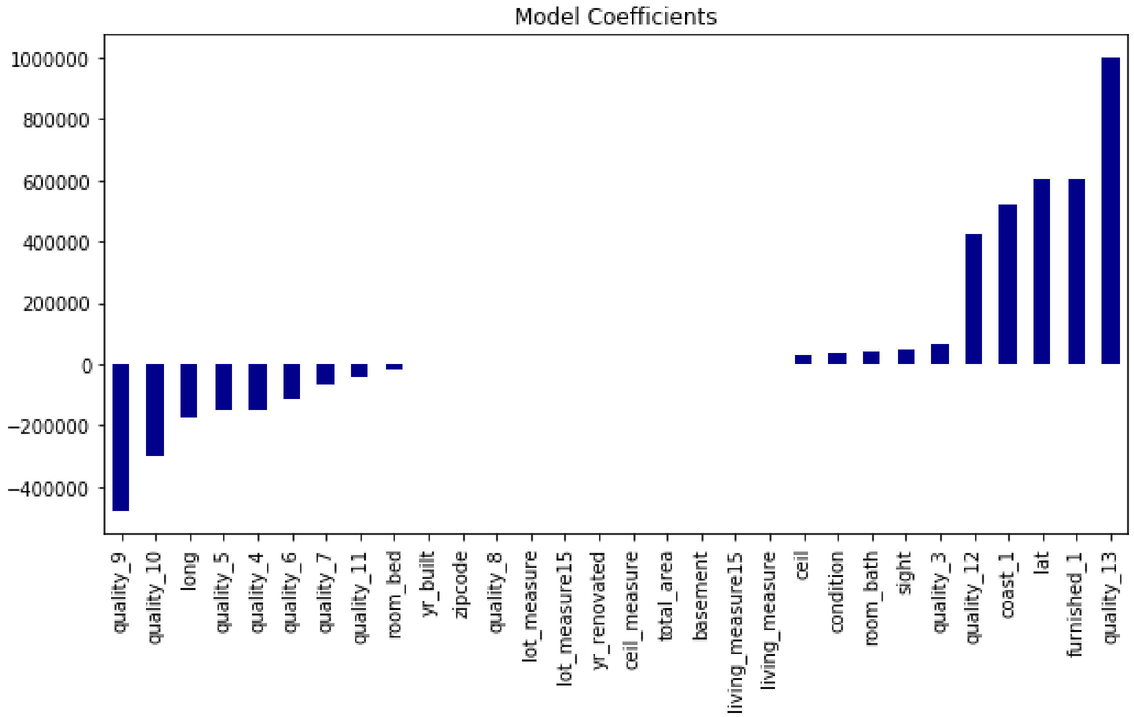


Fig 25

Lasso model

Lasso. It stands for – Least Absolute Shrinkage and the Selection Operator is a technique where data points are shrunk towards a central point, like the mean. Lasso is also known as L1 regularization.

The lasso linear regression model performed with scores of 0.73 & .72 in the training data set and validation data set respectively. The coefficients of 1 variable in lasso model is almost '0', signifying that the variable with '0' coefficient can be dropped.

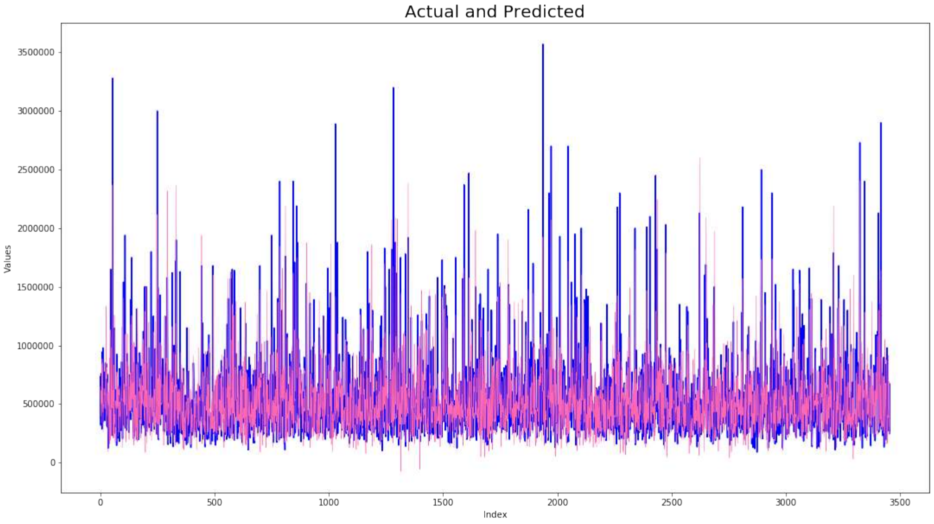


Fig 26

Ridge model

Ridge regression is a technique used to analyze multi-linear regression (multi-collinear), also known as L2 regularization.

The Ridge linear regression model performed with scores of 0.73 & .72 in the training data set and validation data set respectively. The coefficient of variables in the ridge model is all non-zero, indicating that none of the variables can be dropped.

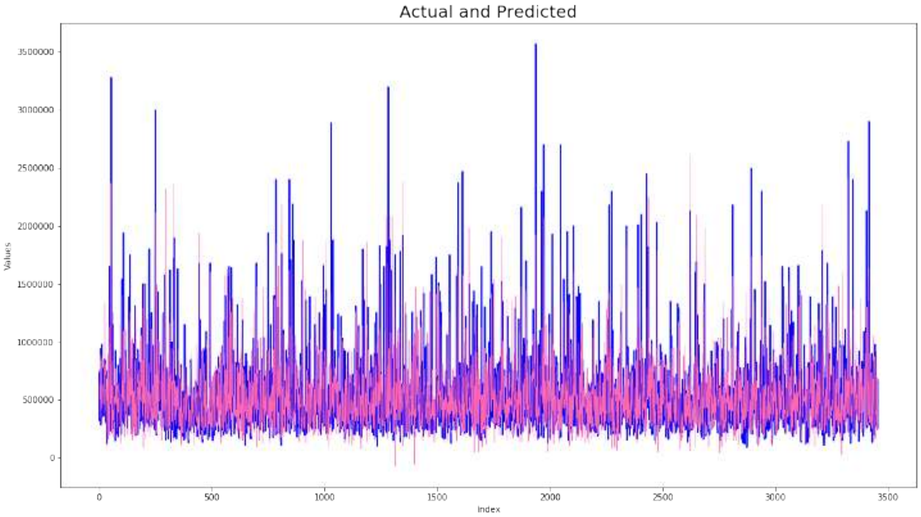


Fig 27

KNN regressor model and decision tree models

The KNN regressor model and decision tree models have not performed well in comparison with linear regression models

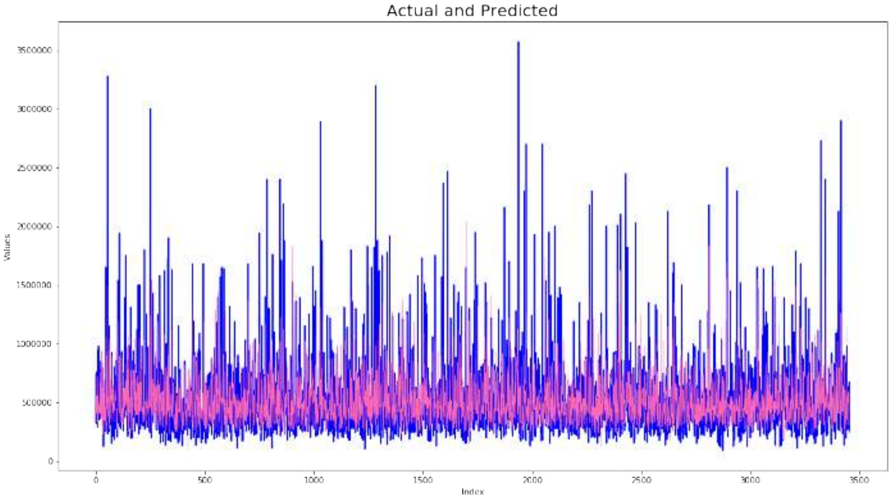


Fig 28

Random forest regression

Random Forest is an ensemble technique capable of performing both regression and classification tasks with the use of multiple decision trees and a technique called Bootstrap and Aggregation, commonly known as bagging.

The basic idea behind this is to combine multiple decision trees in determining the final output rather than relying on individual decision trees.

Random Forest has multiple decision trees as base learning models. Randomly perform row sampling and feature sampling from the dataset forming sample datasets for every model.

The random forest model has performed well in the training and validation set. There is scope for further analysis of this model

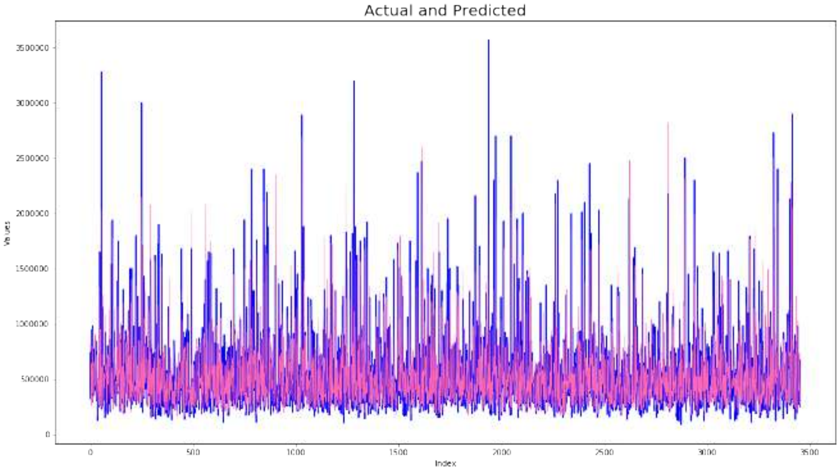


Fig 29

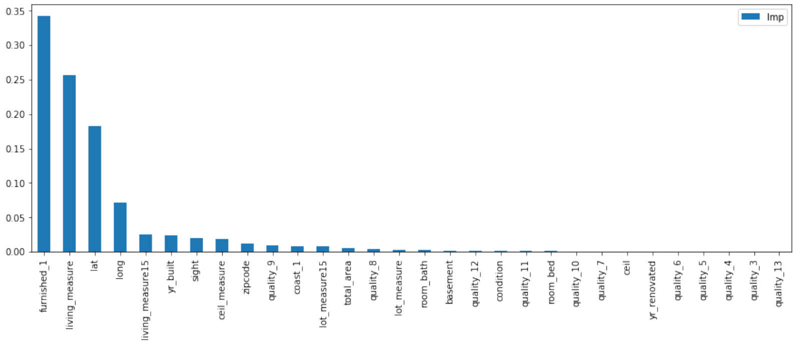


Fig 30

Gradient Boost Regressor

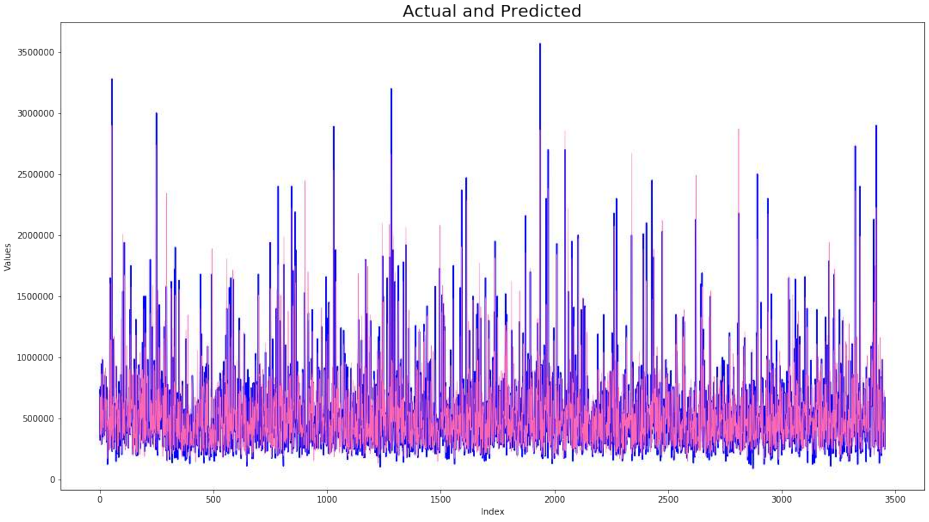


Fig 31

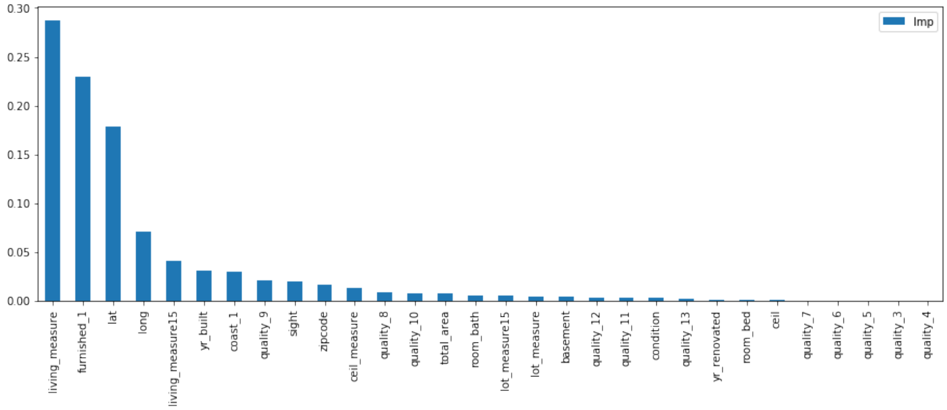


Fig 32

XGBOOST REGRESSOR

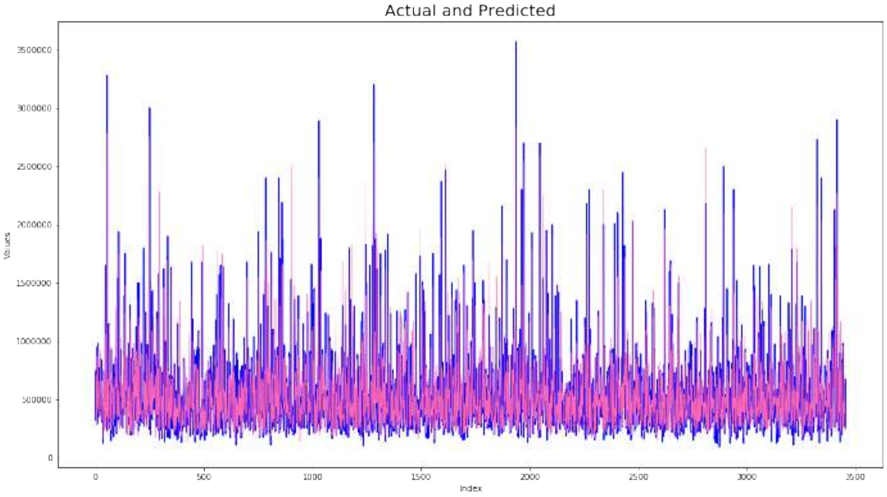


Fig 33

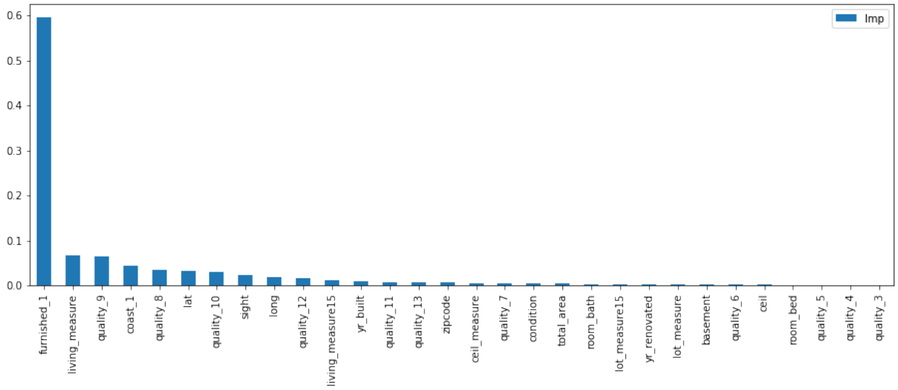


Fig 34

ADABOOST REGRESSOR

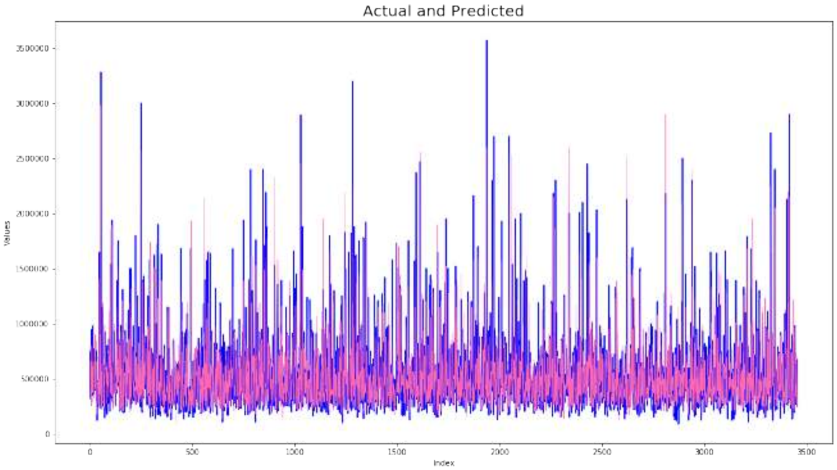


Fig 35

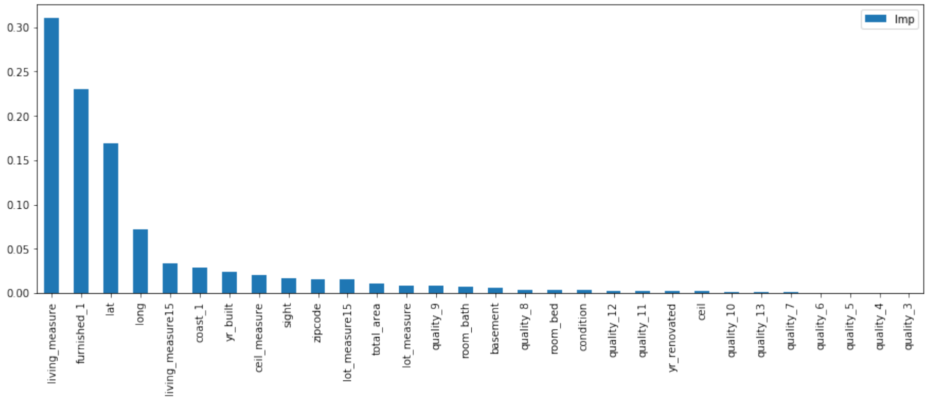


Fig 36

BAGGING REGRESSION

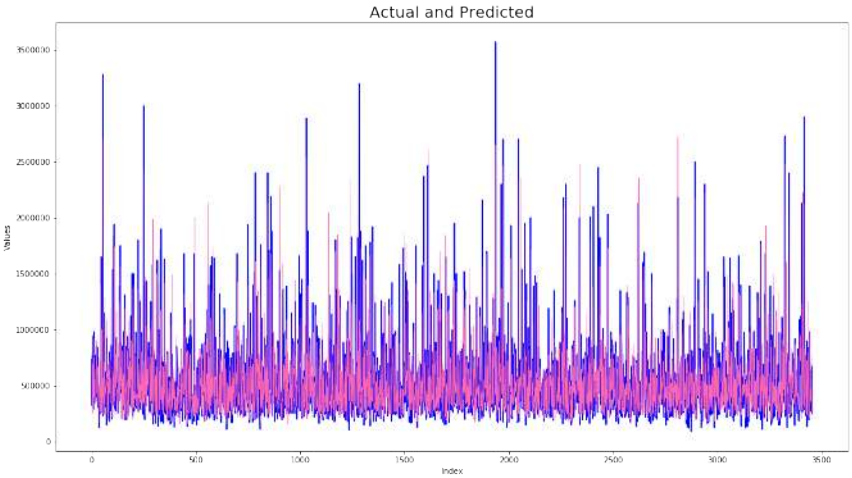


Fig 37

RANDOM FOREST HYPERTUNE

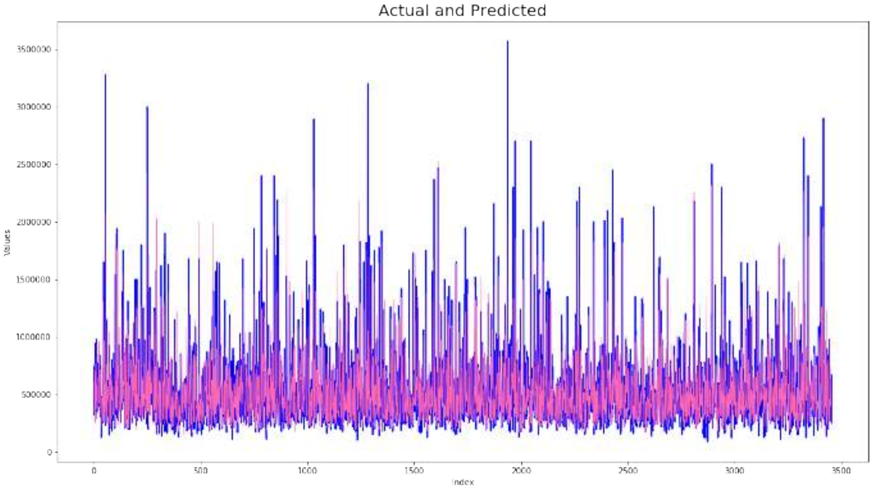


Fig 38

GRADIENT BOOST HYPERTUNE

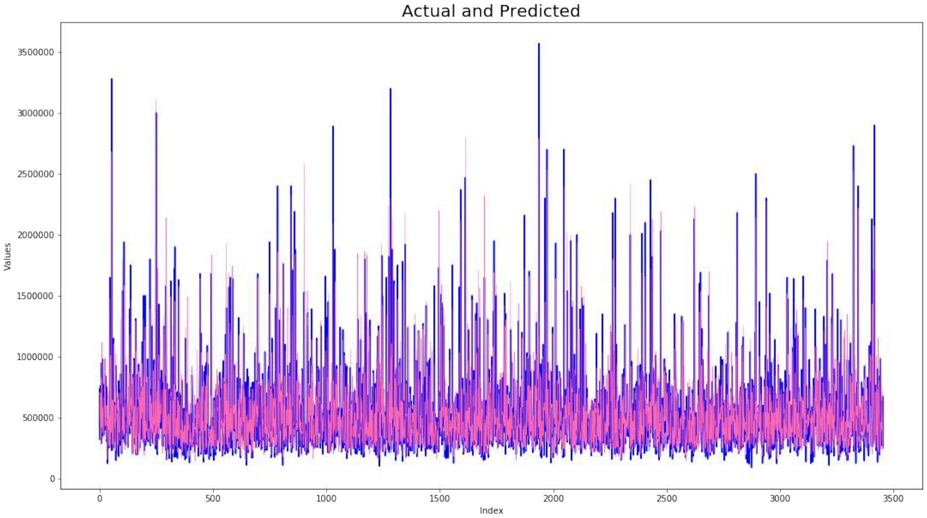


Fig 39

ADABOOST HYPERTUNE

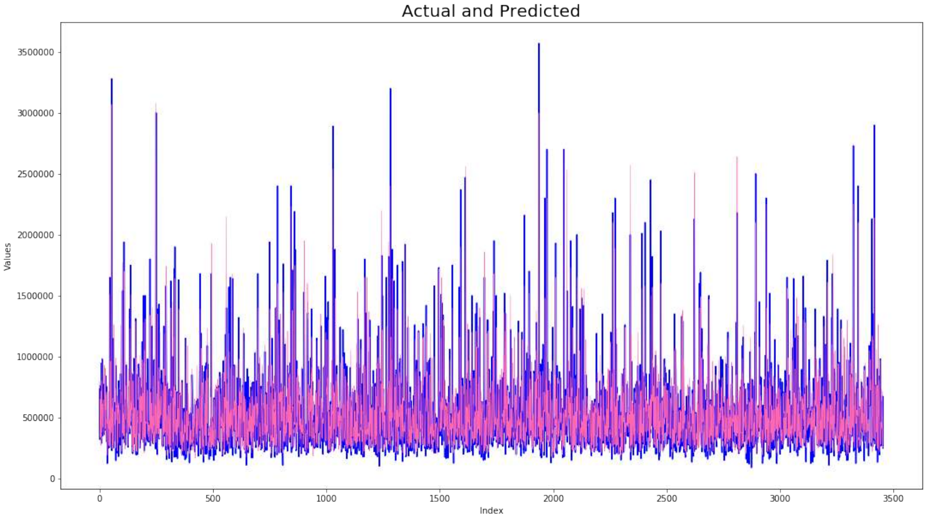


Fig 40

XGBoost Regressor

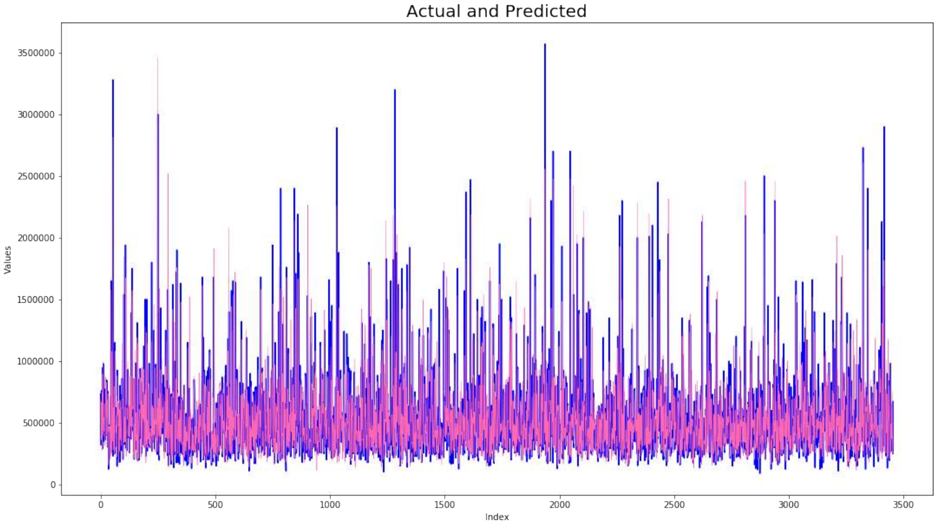


Fig 41

Executing xgb_3_ht on a test data set

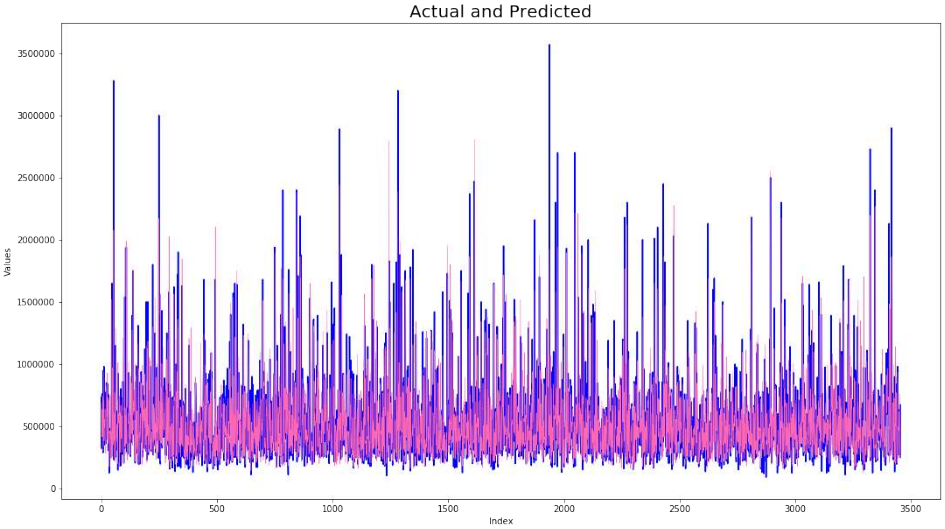


Fig 42

CALCULATING CONFIDENCE INTERVAL ON THE FINAL SELECTED MODEL at 95% ALPHA

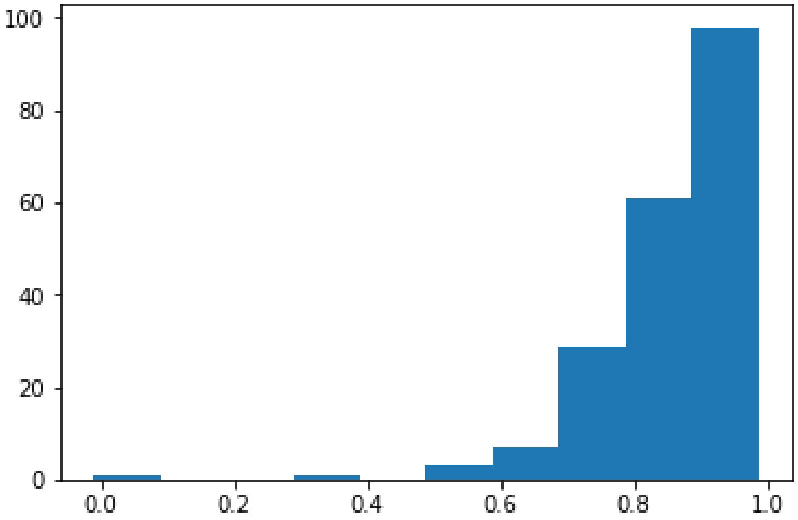


Fig 43

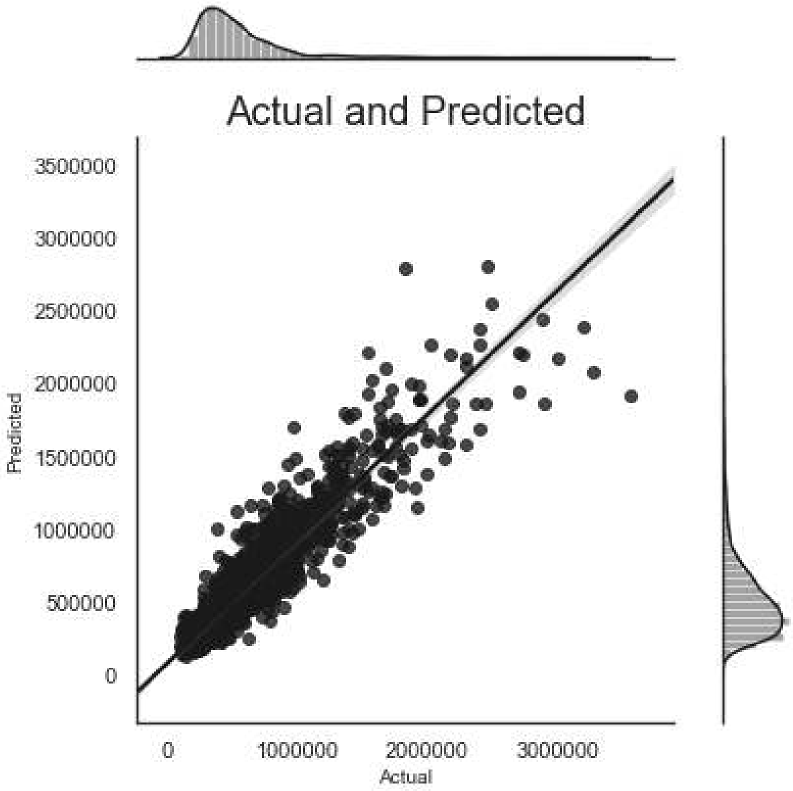

Actual and Predicted
3500000
3000000
2500000
2000000
1500000
1000000
500000
0
Predicted
0
1000000
2000000
3000000
Actual

Fig 44

CHAPTER 4

FINDINGS, RECOMMENDATIONS AND CONCLUSION

FINDINGS, RECOMMENDATIONS AND CONCLUSION

Findings Based on Observations

- The experiment is done to pre-process the data and evaluate the prediction accuracy of the models. The experiment has multiple stages that are required to get the prediction results. These stages can be defined as:
- Pre-processing: Datasets will be checked and pre-processed using the methods. These methods have various ways of handling data. Thus, the preprocessing is done on multiple iterations where each time the accuracy will be evaluated with the used combination.
- Data splitting: dividing the dataset into two parts is essential to train the model with one and use the other in the evaluation. The dataset will be split into 80% for training and 20% for testing.

- Evaluation: the accuracy of a dataset will be evaluated by measuring the R2 and RMSE rate when training the model alongside an evaluation of the actual prices on the test dataset with the prices that are being predicted by the model.
- Performance: alongside the evaluation metrics, the required time to train the model will be measured to show the algorithm varies in terms of time.
- Correlation: The correlation between the available features and house price will be evaluated using the Pearson Coefficient Correlation to identify whether the features have a negative, positive, or zero correlation with the house price.

Findings Based on Analysis of Data

- Pre-processing methods played a significant role in providing the final prediction accuracy, as shown in the experiment sequence in both public and local data.
- outlier, as suggested by gave a worse outcome than Isolation Forest where it has improved the prediction accuracy.
- The performance of trained models has been measured by evaluating the RMSE, R2 metrics, MAE, and MSE.
- The accuracy has been evaluated by plotting the actual prices on the predicted values, as shown below

General findings

Recommendation based on findings

- Many machine learning algorithms are used to predict. However, previous research has shown a comparison between all algorithms.
- Therefore, using these algorithms is beneficial so that the result can be as near to the claimed results.
- However, the prediction accuracy of these algorithms depends heavily on the given data when training the model.
- If the data is in bad shape, the model will be overfitted and inefficient, which means that data pre-processing is an important part of this experiment and will affect the final results.
- Thus, multiple combinations of pre-processing methods need to be tested before getting the data ready to be used in the train

Suggestions for areas of improvement

Scope for Future Research

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

- The used pre-processing methods do help in the prediction accuracy. However, experimenting with different combinations of pre-processing methods to achieve better prediction accuracy.

- Making use of the available features and if they could be combined as binning features has shown that the data improved.

- Training the datasets with different regression methods such as Elastic net regression that combines both L1 and L2 norms. To expand the comparison and check the performance.

- The correlation has shown the association in the local data. Thus, attempting to enhance the local data is required to make it rich with features that vary and can provide a strong correlation relationship.

- The factors that have been studied in this study have a weak correlation with the sale price. Hence, by adding more factors to the local dataset that affect the house price, such as GDP, average income, and the population. To increase the number of factors that have an impact on house prices. This could also lead to a better finding for questions 1 and 2.

The results answer the research questions as follows:

Question 1 – Which machine learning algorithm performs better and has the most accurate result in house price prediction? And why?

Lasso made the best performance overall when both R2 and RMSE scores were taken into consideration. It has achieved the best performance due to its L1 norm regularization for assigning zero weights to insignificant features.

Question 2 – What are the factors that have affected house prices in Malmö over the years?   The number of crimes, repo, lending, and deposit rates have a weak correlation with house prices. This means there are lower likelihood relationships between these factors and sale price. However, when these factors increase the house price decreases. Besides, inflation and year have changed house prices positively, which means when these factors increase, the house prices increase

Conclusion

We have built different models on 2 datasets. The performance (score and 95% confidence interval scores) of the model built on dataset-1 is better than dataset-2 as the 95% confidence interval of dataset-1 is very narrow compared to that of dataset-2. Even though the score of the dataset-2 model is higher, the model has a very vast range of performance scores.

The top key features to consider for pricing a property are:'furnished_1', 'yr_built', 'living_measure','quality_8', 'lot_measure15', 'quality_9', 'ceil_measure', 'total_area'. These are almost similar in both models

So, one needs to thoroughly introspect its property on the parameters suggested and list its price accordingly, similarly if one wants to buy a house - one needs to check the features suggested above in-house and calculate the predicted price. The same can then be compared to the listed price.

For further improvisation, the datasets can be made by treating outliers in different ways and hypertuning the ensemble models. Making polynomial features and improvising the model performance can also be explored further.

References:

1. https://www.scribd.com/document/549175078/Capstone-Interim-Report-HR-CTC-prediction
2. https://thesai.org/Downloads/Volume8No10/Paper_42-Modeling_House_Price_Prediction_using_Linear_Regression.pdf
3. https://www.diva-portal.org/smash/get/diva2:1456610/FULLTEXT01.pdf
4. https://medium.com/@blogsupport/decisive-housing-price-prediction-88c8f90a1733