# House Price Prediction

Submitted in Partial Fulfilment of requirements for the Award of certificate of

Post Graduate Program in Data Science and Business Analytics

**Capstone Project Report**

Submitted to


Submitted by:

Deepti Joshi

Vaishnav U

Under the guidance of

Mr. Abhay Poddar

Batch: PGPDSBA July 2022

Year of Completion: 2023

# CERTIFICATE OF COMPLETION SIGNED BY MENTOR

This is to certify that the participants Deepti Joshi, Vaishnav U who are the students of Great Learning, have successfully completed their project on House Price Prediction.

This project is the record of authentic work carried out by them during the academic year July 2022- July 2023.

Mentor : Mr Abhay Poddar

Date: 2 July 2023.

Place:

7/2/23, 5:59 PM Gmail - Request for Project Certificate Sign-off

Gmail Deepti Joshi <deeptij552@gmail.com>

**Request for Project Certificate Sign-off**

**Abhay Poddar** <abhaypoddar76@gmail.com> 2 July 2023 at 13:44
To: Deepti Joshi <deeptij552@gmail.com>

Hi Deepti,

Please find the completion certificate.

This is to state and certify that this project was undertaken and completed under my guidance.

Thanks & Regards,

Abhay Poddar

[Quoted text hidden]

# ACKNOWLEDGEMENTS

I would like to extend my gratitude to the individuals who have supported and contributed to the successful completion of this academic final project.
I am deeply thankful to my project mentor, Mr Abhay Poddar, for their guidance, expertise, and unwavering support. Their invaluable insights, feedback, and encouragement have played a pivotal role in shaping this final project.
I would also like to acknowledge the faculty members at Great Learning who have provided valuable instruction and knowledge during my academic program. Their dedication to teaching and commitment to excellence have been instrumental in my growth as a student.
I am grateful for the assistance and collaboration provided by my classmates and peers who have offered their insights, ideas, and feedback during the development of this project. Their contributions have enriched the overall quality of the final outcome.
Lastly, I would like to express my heartfelt thanks to my family and loved ones for their unwavering support, understanding, and encouragement throughout this academic endeavour. Their belief in my abilities and constant motivation have been instrumental in my academic success.
Completing this academic final project has been a challenging yet rewarding experience, and I am grateful for the knowledge, skills, and personal growth it has provided. I am indebted to all those who have played a part in this journey and have contributed to the successful completion of this project.

# Contents



# List of tables



# List of figures

# GLOSSARY OF TERMS / ABBREVIATIONS

| *Term* | *Definition* |
|---|---|
| *RMSE* | *Root mean square error* |
| *R2* | *The coefficient of determination* |
| *MAPE* | *Mean absolute percentage error* |
| *GB* | *Gradient Boosting* |
| *XGB* | *Extreme Gradient Boosting* |
| *EDA* | *Exploratory data analysis* |
| *SVR* | *Support Vector Regression* |

*Table 1: Glossary of terms/ abbreviations*

# EXECUTIVE SUMMARY

## 1. Problem Statement

*A house 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 may expect — it can't be too low or too high. To find house price you usually try to find similar properties in your neighborhood and based on gathered data you will try to assess your house price.*

## 2. Data Description

*The dataset used in this project comprises 20,932 housing records with 17 columns, including both numerical and categorical variables. The data has undergone extensive preprocessing, which involved addressing missing values and removing any duplicates. The dataset has been refined to ensure its integrity and reliability for analysis.*

*The dataset includes various features such as the number of floors, square footage of living space, square footage of the basement, location details, renovation status, quality indicators, and more. These attributes provide valuable insights into the factors influencing house prices. For instance, the number of floors positively affects property prices, with properties having more floors generally commanding higher prices. The square footage of the living space and basement also play a significant role, as larger areas correlate with higher prices.*

*Additionally, the dataset reveals the impact of location on housing prices. Properties in Medina tend to have higher prices compared to those in Auburn, indicating the importance of location in determining property values. Other factors, such as furnished status, proximity to the coast, and renovation status, also contribute to price variations. Furnished properties, those near the coast, and recently renovated homes tend to command higher prices.*

*Furthermore, the data exploration highlights correlations between different attributes and house prices. For example, living area measures exhibit the strongest positive correlation with price, while the number of bedrooms and bathrooms also show moderate positive correlations. On the other hand, the age of the house has negative correlations with certain attributes like ceiling quality and the number of bathrooms.*

*This comprehensive analysis of the dataset provides valuable insights into the housing market, enabling us to better understand the factors driving house prices. These insights can inform decision-making processes and facilitate more accurate price predictions in the real estate industry.*

## 3. Main Results

*After evaluating various machine learning models, the XGBoost Regressor emerged as the top-performing model for predicting property prices. The model was fine-tuned using Grid Search, optimizing hyperparameters such as learning rate, maximum depth, and regularization terms. The model achieved impressive performance, with low root mean squared error (RMSE), high coefficient of determination (R2), and low mean absolute percentage error (MAPE) on both training and test data. The analysis of important features highlighted the significant impact of factors like furnishing, quality rating, location, and property age on property prices. These findings provide valuable insights for businesses in the real estate industry, helping them make informed decisions regarding pricing strategies, property improvements, and location considerations. The XGBoost Regressor can be confidently recommended as the final model for accurate property price predictions.*

## 4. Recommendations

1. *Develop an AI-powered house valuation tool that incorporates all relevant feature variables to provide homeowners with accurate estimates of their property's value.*
2. *Market the accuracy and comprehensive nature of the tool, emphasizing that it goes beyond location and square footage to consider multiple factors influencing house prices.*
3. *Provide comparative market analysis reports to homeowners, offering insights into recent sales data, market trends, and property comparisons in their neighborhood or desired location.*
4. *Partner with real estate agents and agencies to integrate the house valuation tool into their services, enabling them to provide accurate and data-driven pricing recommendations to homeowners.*
5. *Implement a user-friendly interface for the tool, ensuring it is intuitive, visually appealing, and easy to navigate.*
6. *Continuously update and improve the predictive model by incorporating new data and market trends, and gather feedback from users to identify areas for enhancement.*
7. *Consider offering premium services such as in-depth property reports, personalized consultations, and access to additional data and insights.*
8. *Build trust and credibility in the real estate industry through success stories, testimonials, and content marketing initiatives.*

*By following these recommendations, you can create a successful business that provides accurate house price predictions and valuation services, while establishing yourself as a trusted resource in the real estate market.*

# Section 1. : Introduction

*The business model proposed in this project aims to revolutionize the assessment of house value by going beyond traditional methods that focus solely on location and square footage. By considering a comprehensive range of factors that contribute to a house's worth, the model seeks to provide homeowners, sellers, and real estate professionals with a more accurate estimation of house prices. This will enable them to make informed decisions, set competitive asking prices, and facilitate efficient transactions in the real estate market.*

## *1.1. Problem Statement, Need of the Study and Objective*

*A house 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 may expect — it can't be too low or too high. To find house price you usually try to find similar properties in your neighborhood and based on gathered data you will try to assess your house price*

### *Need of the Study:*

*The need for this study or project arises from the desire to accurately assess the value of a house beyond just considering its location and square footage. By recognizing that a house's value is influenced by various factors, it becomes essential to develop a method or system that takes into account these aspects. Homeowners or sellers who wish to sell their house need to have an idea of the price they can expect. By conducting this study or project, individuals can gain a better understanding of the factors that contribute to a house's value and make informed decisions when determining the asking price*

### *Objectives*

*The objective of the House Price Prediction Data Science Project is to develop a reliable and accurate predictive model for estimating house prices. This model aims to incorporate various influential factors, such as location, size, amenities, and historical sales data, to provide actionable insights for homebuyers, sellers, real estate agents, and financial institutions. By leveraging machine learning algorithms and comprehensive datasets, the project seeks to enhance decision-making in the real estate industry and contribute to the field of data-driven real estate analysis.*

# Section 2. : EDA and Insights

## *2.1 Data Report*

### *Sample of Dataset:*

| dayhours | price | room_bed | room_bath | living_measure | lot_measure | ceil | coast | sight | ... | basement | yr_built | yr_renovated | zipcode | lat | long | living |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 2015-04-27 | 600000 | 4.0 | 1.75 | 3050.0 | 9440.0 | 1 | 0 | 0.0 | ... | 1250.0 | 1966 | 0 | 98034 | 47.7228 | -122.183 | |
| 2015-03-17 | 190000 | 2.0 | 1.00 | 670.0 | 3101.0 | 1 | 0 | 0.0 | ... | 0.0 | 1948 | 0 | 98118 | 47.5546 | -122.274 | |
| 2014-08-20 | 735000 | 4.0 | 2.75 | 3040.0 | 2415.0 | 2 | 1 | 4.0 | ... | 0.0 | 1966 | 0 | 98118 | 47.5188 | -122.256 | |
| 2014-10-10 | 257000 | 3.0 | 2.50 | 1740.0 | 3721.0 | 2 | 0 | 0.0 | ... | 0.0 | 2009 | 0 | 98002 | 47.3363 | -122.213 | |
| 2015-02-18 | 450000 | 2.0 | 1.00 | 1120.0 | 4590.0 | 1 | 0 | 0.0 | ... | 0.0 | 1924 | 0 | 98118 | 47.5663 | -122.285 | |

lumns

Table 2: Sample of Dataset

### *Summary of the data:*

- *Shape of dataset : number of rows: 21613 , number of columns : 23*
- *There are few missing values. And special character like "$" is present in the data.*
- *Duplicate values: There are no duplicate values in the data.*
- *Data types-We can see there are 12 columns of float data type, 4 columns of integer data* type *and 6 columns of object data type, and 1 column of datetime data type.*
- *Some column should be float instead of object. It will be taken care of in data pre-processing.*

### *Data Description:*

| | count | mean | std | min | 25% | 50% | 75% | max |
|---|---|---|---|---|---|---|---|---|
| cid | 21613.0 | 4.580302e+09 | 2.876566e+09 | 1.000102e+06 | 2.123049e+09 | 3.904930e+09 | 7.308900e+09 | 9.900000e+09 |
| price | 21613.0 | 5.401822e+05 | 3.673622e+05 | 7.500000e+04 | 3.219500e+05 | 4.500000e+05 | 6.450000e+05 | 7.700000e+06 |
| room_bed | 21505.0 | 3.371355e+00 | 9.302886e-01 | 0.000000e+00 | 3.000000e+00 | 3.000000e+00 | 4.000000e+00 | 3.300000e+01 |
| room_bath | 21505.0 | 2.115171e+00 | 7.702481e-01 | 0.000000e+00 | 1.750000e+00 | 2.250000e+00 | 2.500000e+00 | 8.000000e+00 |
| living_measure | 21596.0 | 2.079861e+03 | 9.184961e+02 | 2.900000e+02 | 1.429250e+03 | 1.910000e+03 | 2.550000e+03 | 1.354000e+04 |
| lot_measure | 21571.0 | 1.510458e+04 | 4.142362e+04 | 5.200000e+02 | 5.040000e+03 | 7.618000e+03 | 1.068450e+04 | 1.651359e+06 |
| sight | 21556.0 | 2.343663e-01 | 7.664376e-01 | 0.000000e+00 | 0.000000e+00 | 0.000000e+00 | 0.000000e+00 | 4.000000e+00 |
| quality | 21612.0 | 7.656857e+00 | 1.175484e+00 | 1.000000e+00 | 7.000000e+00 | 7.000000e+00 | 8.000000e+00 | 1.300000e+01 |
| ceil_measure | 21612.0 | 1.788367e+03 | 8.281025e+02 | 2.900000e+02 | 1.190000e+03 | 1.560000e+03 | 2.210000e+03 | 9.410000e+03 |
| basement | 21612.0 | 2.915225e+02 | 4.425808e+02 | 0.000000e+00 | 0.000000e+00 | 0.000000e+00 | 5.600000e+02 | 4.820000e+03 |
| yr_renovated | 21613.0 | 8.440226e+01 | 4.016792e+02 | 0.000000e+00 | 0.000000e+00 | 0.000000e+00 | 0.000000e+00 | 2.015000e+03 |
| zipcode | 21613.0 | 9.807794e+04 | 5.350503e+01 | 9.800100e+04 | 9.803300e+04 | 9.806500e+04 | 9.811800e+04 | 9.819900e+04 |
| lat | 21613.0 | 4.756005e+01 | 1.385637e-01 | 4.715590e+01 | 4.747100e+01 | 4.757180e+01 | 4.767800e+01 | 4.777760e+01 |
| living_measure15 | 21447.0 | 1.987066e+03 | 6.855196e+02 | 3.990000e+02 | 1.490000e+03 | 1.840000e+03 | 2.360000e+03 | 6.210000e+03 |
| lot_measure15 | 21584.0 | 1.276654e+04 | 2.728699e+04 | 6.510000e+02 | 5.100000e+03 | 7.620000e+03 | 1.008700e+04 | 8.712000e+05 |
| furnished | 21584.0 | 1.967198e-01 | 3.975279e-01 | 0.000000e+00 | 0.000000e+00 | 0.000000e+00 | 0.000000e+00 | 1.000000e+00 |

Table 3: Data Description

- *The describe method will assist in determining how data has been distributed for numerical and categorical values. We can plainly see the minimum, mean, and maximum values, as well as the different percentile values.*
- *We can also find some values 0 as minimum value for furnished, basement, room_bed, room_bath which is meaningless and to be taken care while data pre-processing.*
- *Also, the mean and median for the attributes are same which shows that the data is normally distributed.*
- *Also, by seeing the difference in the values we can say there are outliers present in the data.*

### *2.2 Data Pre-processing:*

***Data type validation:***

- *Columns 'ceil', 'coast', 'condition', 'yr_built', 'long', 'total_area' contains unique character like "$". We are converting these into null values.*
- *We are converting 'ceil', 'yr_built', 'long', 'total_area', into float data type which were originally in object data type.*
- *And columns 'coast', 'condition', 'quality', 'furnished' 'sight' into object data types.*
- *We have removed rows or columns with significant missing data.*
- *There are no duplicate values in data*

***Feature Selection/Extraction:***

- *Creating new column 'is_renovated' to get the measure of how long the house has not been renovated.*
- *Creating a new column 'location' to store the geocoded location information.*
- *Creating new column 'years_old' to get the measure of how long the house has not been renovated.*
- *We will drop all rows where room bed and bath are zero.*
- *As living_measure and lot_measure show high collinearity we will remove it from column.*
- *After outlier treatment we will remove 'Sold_date', 'yr_built', 'yr_renovated', 'zipcode' from dataset.*

*Outlier Treatment:*

| | % OUTLIERS |
|---|---|
| total_area | 11.1555 |
| lot_measure15 | 10.1354 |
| price | 5.3686 |
| yr_renovated | 4.1980 |
| is_renovated | 4.1980 |
| ceil_measure | 2.8347 |
| room_bath | 2.5856 |
| living_measure15 | 2.5197 |
| room_bed | 2.4445 |
| basement | 2.2659 |
| ceil | 0.0000 |
| yr_built | 0.0000 |
| zipcode | 0.0000 |
| years_old | 0.0000 |

*Table 4: Outliers*

- *We will remove all the values higher than 99.7th percentile in price, lot_measure15 & total_area.*

*The data has undergone thorough pre-processing and cleaning, resulting in a refined and reliable dataset. Missing values have been addressed, outliers have been detected and handled, and duplicates have been removed. Inconsistent and erroneous entries have been corrected or eliminated, ensuring data quality. Through these pre-processing steps, the data is now prepared to reveal valuable patterns and insights, laying the foundation for accurate and reliable outcomes.*

## *Info of the cleaned data*

*First 5 rows:*

| | price | room_bed | room_bath | ceil | coast | sight | condition | quality | ceil_measure | basement | living_measure15 | lot_measure15 | furnished | total_area | is_re |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 0 | 600000 | 4.0 | 1.75 | 1.0 | 0 | 0.0 | 3 | 8.0 | 1800.0 | 1250.0 | 2020.0 | 8660.0 | 0.0 | 12490.0 | |
| 1 | 190000 | 2.0 | 1.00 | 1.0 | 0 | 0.0 | 4 | 6.0 | 670.0 | 0.0 | 1660.0 | 4100.0 | 0.0 | 3771.0 | |
| 2 | 735000 | 4.0 | 2.75 | 2.0 | 1 | 4.0 | 3 | 8.0 | 3040.0 | 0.0 | 2620.0 | 2433.0 | 0.0 | 5455.0 | |
| 3 | 257000 | 3.0 | 2.50 | 2.0 | 0 | 0.0 | 3 | 8.0 | 1740.0 | 0.0 | 2030.0 | 3794.0 | 0.0 | 5461.0 | |
| 4 | 450000 | 2.0 | 1.00 | 1.0 | 0 | 0.0 | 3 | 7.0 | 1120.0 | 0.0 | 1120.0 | 5100.0 | 0.0 | 5710.0 | |

*Table 5: Clean Data*

*Summary of the cleaned data:*

- *Shape of dataset: number of rows: 20932, number of columns: 17.*
- *There are no missing values.*
- *Duplicate values: There are no duplicate values in the data.*
- *Data types-We can see there are 8 columns of float data type, 2 columns of integer data type and 7 columns of object data type.*

## *2.3 EDA, insights:*

*During the EDA phase of the house price prediction project, several interesting insights were uncovered regarding the relationship between house prices and various features within the dataset.*

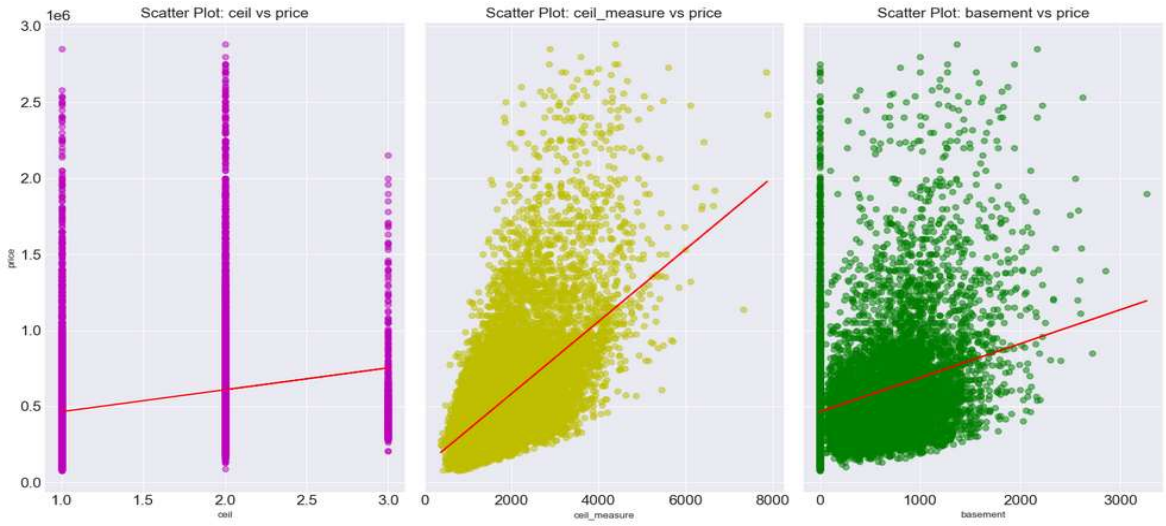


*Figure 1: Price vs. ceil, basement, ceil_measure*

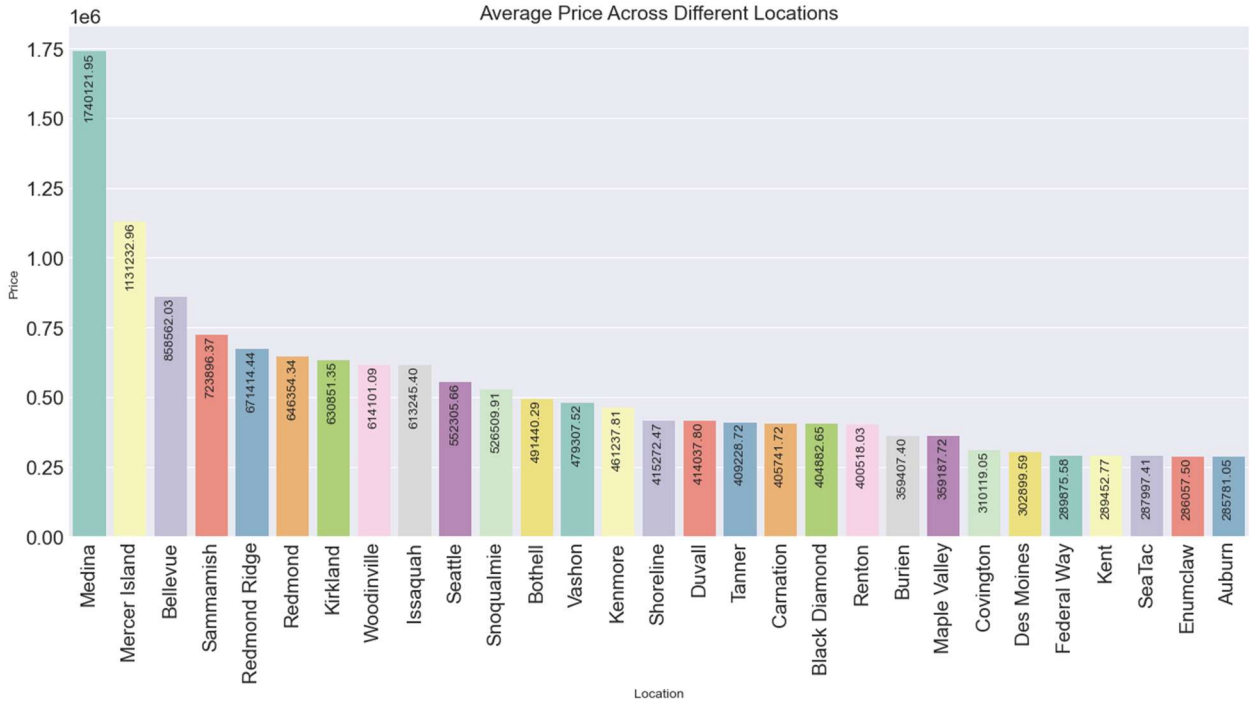


*Figure 2: Price vs. Location*

*In the exploratory data analysis, it was observed that the number of floors in a property positively influenced its price. Properties with more floors tended to have higher prices. Additionally, a strong*

*positive correlation was found between the square footage of the living space. Furthermore, the square footage of the basement also played a role in determining property prices, with larger basement areas generally leading to higher prices.*

*The analysis revealed that properties located in Medina tend to have higher prices compared to properties in Auburn. This suggests that the location of a property has a significant impact on its price, with Medina being associated with higher-priced homes and Auburn with relatively lower-priced homes.*

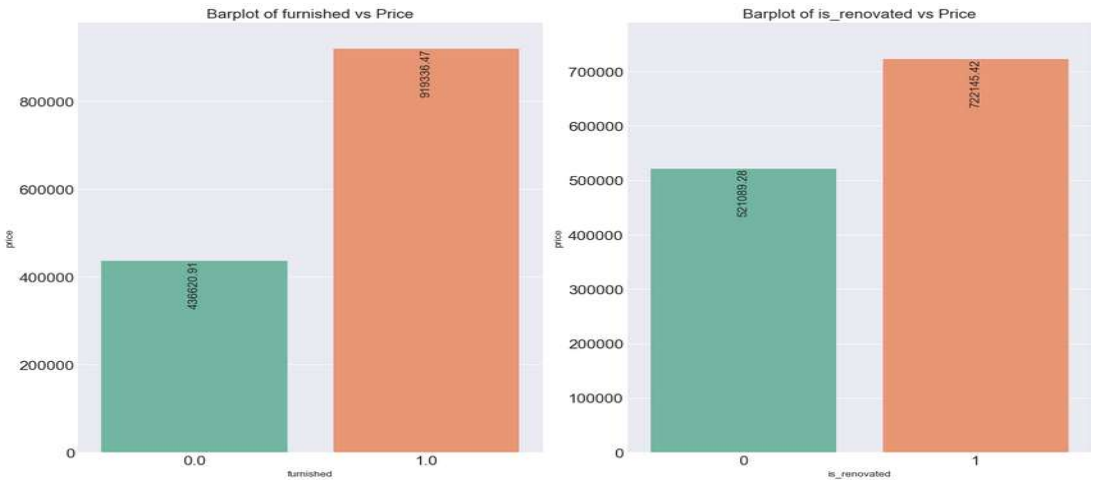


*Figure 3: Price vs. furnished and Renovation*

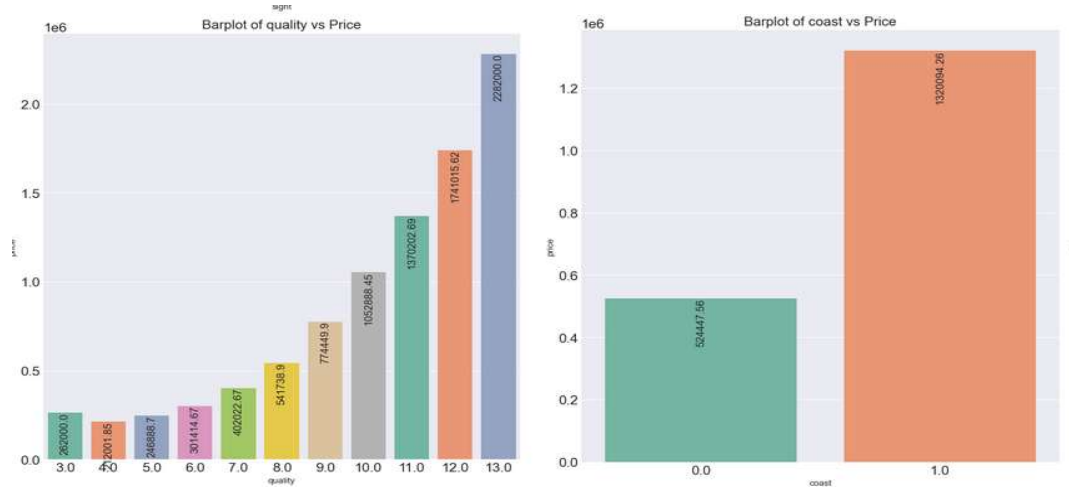


*Figure 4: Price vs. Quality, Coast.*

*The analysis showed that furnished properties have higher prices due to added value from furniture and amenities. Properties near the coast command higher prices due to desirability. Recently renovated properties also have higher prices, reflecting buyer preference for move-in ready homes. Higher-*

*quality properties, with superior materials and craftsmanship, are priced higher due to durability and aesthetic appeal.*

***Correlation Analysis:***

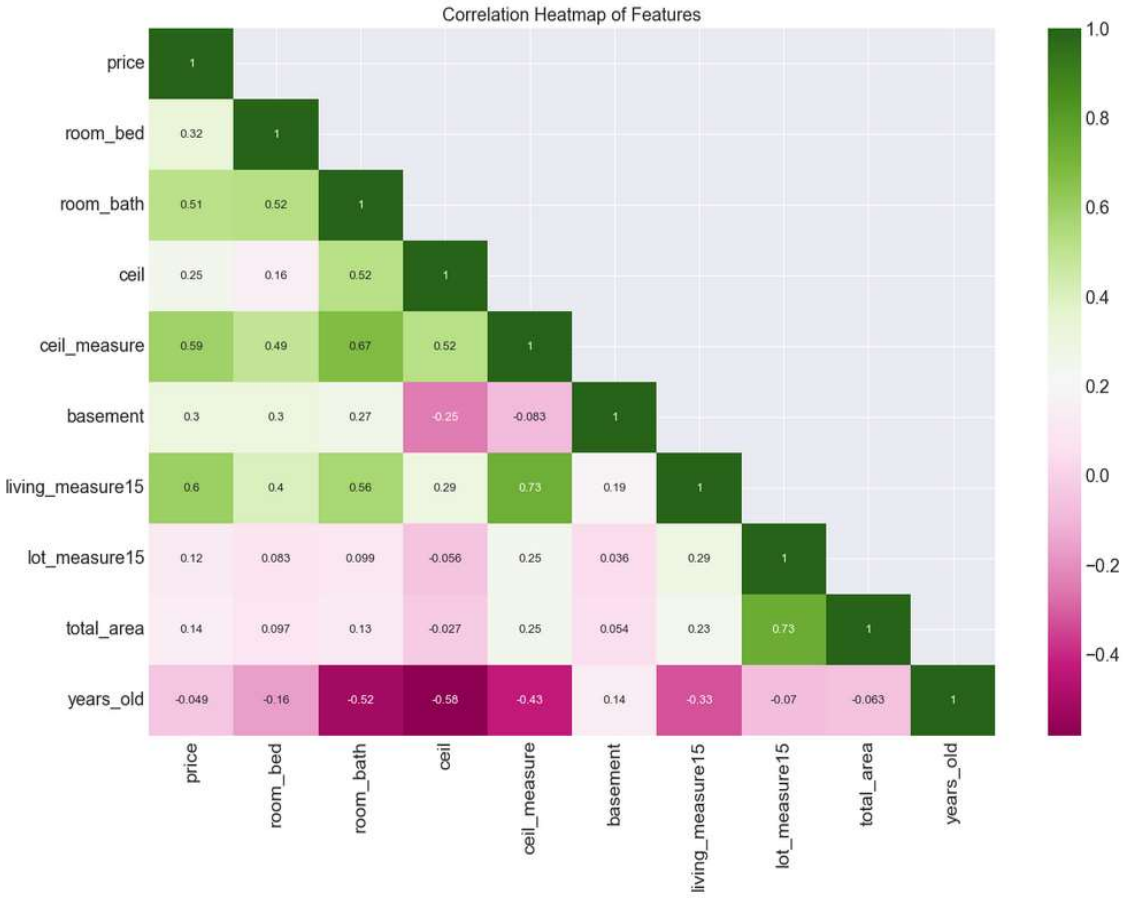


*Figure 5: Correlation Analysis.*

*The correlation table heatmap reveals several insights about the relationships between various attributes and house prices. The strongest positive correlation exists between price and living area measures. Additionally, the number of bedrooms and bathrooms show a moderate positive correlation, while the age of the house has negative correlations with certain attributes like ceiling quality and the number of bathrooms. Overall, the heatmap provides valuable information for understanding the factors influencing house prices. However, it's essential to remember that correlation does not imply causation, and further analysis is necessary to draw definitive conclusions.*

## Section 3. : Model Development

| Model Comparison Table: | | | | | | | |
|---|---|---|---|---|---|---|---|
| Model | Model Type | Train RMSE | Test RMSE | Train R2 | Test R2 | Train MAPE | Test MAPE |
| Linear Regression | Base model | 0.5181 | 0.5116 | 0.7307 | 0.7401 | 2.6894 | 2.5268 |
| Decision Tree | Base model | 0.0111 | 0.6553 | 0.9999 | 0.5734 | 0.0050 | 4.0773 |
| Random Forest | Base model | 0.1735 | 0.4485 | 0.9698 | 0.8002 | 0.9003 | 3.2399 |
| Gradient Boosting | Base model | 0.445 | 0.4619 | 0.8013 | 0.7881 | 2.3757 | 2.386 |
| Bagging Regression | Base model | 0.207 | 0.4728 | 0.957 | 0.778 | 1.0759 | 3.3163 |
| SVR | Base model | 0.4044 | 0.4502 | 0.8359 | 0.7987 | 2.0085 | 2.6143 |
| XGB | Base model | 0.2668 | 0.4294 | 0.9286 | 0.8169 | 1.7005 | 3.1007 |
| SVR | TUNED MODEL | 0.4044 | 0.4502 | 0.8359 | 0.7987 | 2.0085 | 2.6143 |
| XGB | TUNED MODEL | 0.3142 | 0.4186 | 0.901 | 0.826 | 1.9013 | 2.6536 |
| Bagging Regression | TUNED MODEL | 0.227 | 0.4418 | 0.9483 | 0.8062 | 1.1919 | 2.9907 |
| Stacking regressor | Meta mode-Linear Reg. | 0.3115 | 0.4131 | 0.9027 | 0.8305 | 1.8791 | 2.7706 |
| Stacking regressor | Meta mode-SVR | 0.3198 | 0.4121 | 0.8974 | 0.8314 | 1.6854 | 2.8109 |
| Stacking regressor | Meta mode-XGB | 0.3419 | 0.4297 | 0.8828 | 0.8166 | 1.8291 | 2.6762 |
| Stacking regressor | Meta mode-Bagging Reg. | 0.3589 | 0.4462 | 0.8708 | 0.8022 | 1.8675 | 2.8193 |

*Table 6: Model comparison table*

*We have evaluated various machine learning models. The models considered include Linear Regression, Decision Tree, Random Forest, Gradient Boosting, Bagging Regression, Support Vector Regression (SVR), and XGBoost (XGB). Additionally experimented with tuning some of these models. Here stacking regressor with linear regression as meta-model outperforms all other models. But due to interpretability issue we choose not to use stacking regression.*
*Tuned XGB Regressor is the chosen final model.*

***Step by step model building***

1. *Create an XGBoost Regressor:*
2. *Define the Parameter Grid : Define a parameter grid containing different hyperparameter values. These hyperparameters are like knobs that can be adjusted to optimize the model's performance.*

*The parameters used are:*

***n_estimators** : The number of boosting rounds or trees in the ensemble. The values to try are 100, 200, and 300.*
***max_depth** : The maximum depth of each tree in the ensemble. The values to try are 3, 5, and*
***learning_rate** : The step size shrinkage used to prevent overfitting. The values to try are 0.1, 0.01, and 0.001.*
***reg_alpha** : L1 regularization term to prevent overfitting. The values to try are 0.01, 0.1, and 1.*
***reg_lambda** : L2 regularization term to prevent overfitting. The values to try are 0.01, 0.1, and 1.*

3. ***Perform Grid Search** : Grid Search is a technique that tries out all possible combinations of hyperparameters from the defined parameter grid. It evaluates each combination using cross-validation to find the best set of hyperparameters that give the most accurate predictions.*
4. ***Retrieve Best Hyperparameters and Model** : Once the Grid Search is complete, the best hyperparameters and the corresponding model that performed the best during the search is retrieve.*

***Best Hyperparameters (XGBRegressor)**:*
*'learning_rate': 0.1*
*'max_depth': 5,*
*'n_estimators': 300*
*'reg_alpha': 1*
*'reg_lambda': 1*

5. ***Train the Best Model**: With the best hyperparameters train the XGBoost model again.*
6. ***Make Predictions and Evaluate:** After training, trained model is used to make predictions. Then various evaluation metrics like MAPE, R2, RMSE, etc. are used to assess the model's performance.*

***Model Validation***

- *Train RMSE: 0.3142*
- *Test RMSE: 0.4186*
- *Train R2: 0.901*
- *Test R2: 0.826*
- *Train MAPE: 1.9013*
- *Test MAPE: 2.6536*

*RMSE (Root Mean Squared Error): With an RMSE of 0.3142 on the training data and 0.4186 on the test data, the model's predictions have an average error of approximately 0.3142 units and* 0.4186 *units*

*from the actual values, respectively. The lower RMSE values indicate the model's accurate predictions and its ability to generalize well to unseen data.*
*R2 (Coefficient of Determination): The model explains approximately 90.1% of the variance in the target variable on the training data (R2 = 0.901) and around 82.6% on the test data (R2 = 0.826). These high R2 values demonstrate the model's strong ability to capture the variability in the target variable, both on the data it was trained on and on new, unseen data.*
*MAPE (Mean Absolute Percentage Error): The MAPE of 1.9013% on the training data and 2.6536% on the test data suggest that, on average, the model's predictions deviate by approximately 1.9013% and 2.6536% from the actual values, respectively. These low MAPE values demonstrate the model's ability to make accurate predictions on both the training and test data.*

***Important Features (XGBRegressor):***

- *furnished_1.0: 0.5045*
- *quality_7: 0.0811*
- *quality_6: 0.0695*
- *location_Bellevue: 0.0311*
- *coast_1.0: 0.0244*
- *location_Federal Way: 0.0226*
- *sight: 0.0222*
- *location_Covington: 0.0220*
- *location_Seattle: 0.0204*
- *location_Kent: 0.0173*
- *quality_8: 0.0168*
- *location_Kirkland: 0.0154*
- *years_old: 0.0144*
- *quality_10: 0.0125*
- *room_bath: 0.0108*
- *ceil_measure: 0.0102*
- *condition_3.0: 0.0100*
- *quality_5: 0.0097*
- *living_measure15: 0.0097*
- *location_Redmond: 0.0095*
- *location_Maple Valley: 0.0088*
- *basement: 0.0078*
- *location_Others: 0.0065*
- *is_renovated_1: 0.0062*
- *location_Renton: 0.0049*
- *location_Sammamish: 0.0047*
- *condition_5.0: 0.0044*

- *location_Issaquah: 0.0041*
- *lot_measure15: 0.0036*
- *quality_11: 0.0029*
- *quality_9: 0.0025*
- *total_area: 0.0024*
- *ceil: 0.0020*
- *condition_2.0: 0.0015*
- *quality_4: 0.0014*
- *room_bed: 0.0011*
- *condition_4.0: 0.0008*
- *quality_3: 0.0006*
- *quality_2: 0.0000*

## Section 4. Final Recommendation

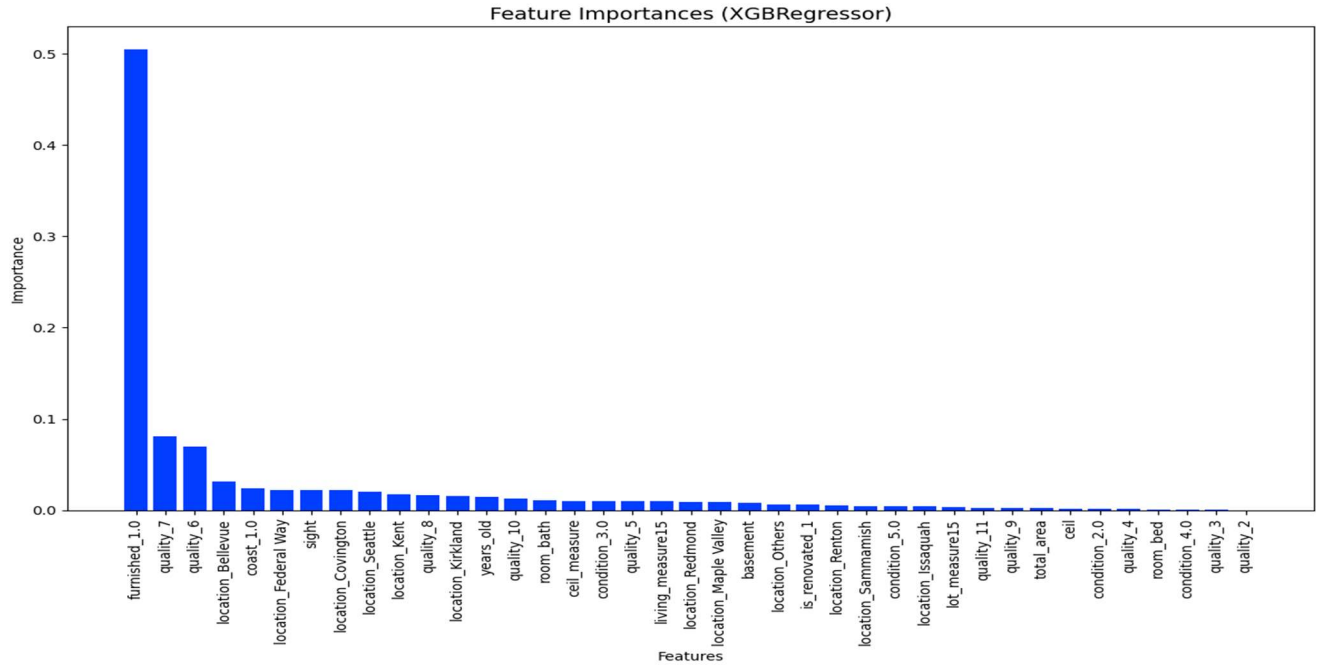


*Figure 6: Feature Importances*

1. ***Develop an AI-Powered House Valuation Tool:*** *Create an advanced online platform or mobile application that utilizes artificial intelligence algorithms to analyse and predict house prices. Incorporate all the relevant feature variables identified in the problem statement, such as the number of floors, square footage of living space and basement, location, property condition, and amenities. This tool will provide homeowners with an accurate estimate of their house's value, eliminating the guesswork involved in pricing their property.*

2. ***Market the Accuracy and Comprehensive Nature of the Tool:*** *Highlight the fact that your house valuation tool takes into account a wide range of features and variables, providing a holistic assessment of a property's value. Educate potential users about the limitations of relying solely on location and square footage, emphasizing that your tool provides a more comprehensive analysis based on multiple factors. Position your tool as a reliable and trusted source for determining house prices.*
3. ***Provide Comparative Market Analysis Reports:*** *Alongside the house valuation tool, offer detailed comparative market analysis (CMA) reports to homeowners. These reports will include insights into recent sales data, market trends, and property comparisons in their neighbourhood or desired location. By providing this information, you empower homeowners with a deeper understanding of their local real estate market, allowing them to make more informed decisions when pricing their property.*
4. ***Partner with Real Estate Agents and Agencies:*** *Collaborate with local real estate agents and agencies to integrate your house valuation tool into their services. Offer them access to the platform and provide training to ensure they can effectively use the tool to assist their clients. This partnership will enable agents to provide accurate and data-driven pricing recommendations to homeowners, strengthening their credibility and enhancing their service offering.*
5. ***Implement a User-Friendly Interface: Ensure*** *that the user interface of your house valuation tool is intuitive, easy to navigate, and visually appealing. Provide clear instructions and guidance throughout the valuation process. Consider incorporating interactive features, such as sliders or visual representations, to enhance the user experience. A user-friendly interface will encourage homeowners to use the tool repeatedly and recommend it to others.*
6. ***Continuously Update and Improve the Model:*** *Regularly update your predictive model and algorithms to incorporate new data and market trends. Monitor the performance of the tool and gather feedback from users to identify areas for improvement. By staying up-to-date with the latest information and continuously refining your model, you will ensure the accuracy and relevance of your house valuation tool.*
7. ***Offer Premium Services:*** *Consider offering premium services, such as in-depth property reports, personalized consultations, or access to additional data and insights. These premium offerings can generate additional revenue streams while providing added value to users who require more comprehensive and detailed information about their property's value.*
8. ***Establish Trust and Credibility:*** *Build trust and credibility in the real estate industry by sharing success stories, testimonials, and case studies from homeowners who have used your house valuation tool. Highlight the accuracy of the tool and its ability to provide fair market value estimates. Engage in content marketing initiatives, such as blog posts and educational resources, to position your brand as an authority in the field of house price prediction.*

*Remember to comply with local regulations and privacy policies when collecting and utilizing user data. Regularly update your platform with the latest features and functionalities to stay ahead of the*

*competition. By implementing these recommendations, you can create a successful and innovative business focused on accurate house price prediction and valuation.*

## Bibliography

https://www.enjoyalgorithms.com/
https://www.kaggle.com/
https://www.analyticsvidhya.com/

## Appendix

***Data dictionary***

*1. cid: a notation for a house*
*2. dayhours: 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: quare footage of the lot*
*8. ceil: Total floors (levels) in house*
*9. coast: House which has a view to 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 grading system*
*13. ceil_measure: square footage of house apart from basement*
*14. basement_measure: square footage of the basement*
*15. yr_built: Built Year*
*16. yr_renovated: Year when house was renovated*
*17. zipcode: zip*
*18. lat: Latitude coordinate*
*19. long: Longitude coordinate*
*20. living_measure15: Living room area in 2015*
*21. lot_measure15: lotSize area in 2015(implies-- some renovations)*
*22. furnished: Based on the quality of room*
*23. total_area: Measure of both living and lot*