HOUSE PRICE PREDICTION SURVEY OF DATA MINING METHODS AND REGRESSION TECHNIQUES

Radhikareddy Chintareddy and Kusala Nagamani Reddy
Dept of Data Science
Clarkson University

Abstract— **The Real Estate industry is dynamic in terms of the prices fluctuating regularly. As accurate house prices allow better-informing parties in the real estate market, improving housing policies and real estate appraisal, a comprehensive overview of house price prediction strategies is valuable for both research and society. Although conventional methods and traditional input data remain predominant, house price prediction research is slowly adopting more advanced techniques and innovative data sources. The project mainly focuses on predicting the real-valued prices for the places and the houses by applying the appropriate ML algorithms. Algorithms like Linear regression and sklearn are used to effectively increase the accuracy. During model structure, nearly all data similarities and cleaning, outlier removal and feature engineering, dimensionality reduction, gridsearchcv for hyperparameter tuning, k-fold cross-validation, etc. are covered. Model performance was measured using evaluation matrices such as Mean Squared Error (MSE), Root Mean Squared Error (RMSE), and also Mean Absolute Error (MAE). Adjusted R-squared of 0.701 (70% of variations explained by our model). Our model identified a strong correlation between waterfront houses, Sight (if the house has already been viewed), and price. Also, there is a noticeable disparity among the zip codes. The median house price ranges from $235,000 in 98002 up to $1,260,000 in 98039. Overall, this model can aid housing developers and the common population alike. Keywords: Linear regression model, Python, Machine Learning, House Price, Decision Tree.**

I INTRODUCTION

The need for houses is not only used as a place to live, they are also often used as long-term investment instruments by investors to get additional income from these types of investments, especially for property entrepreneurs who certainly produce investments that can be considered quite promising. House is one of the main human needs besides clothing and food. In the Hierarchy of Needs, the house is one part of the Basic Needs which can be categorized into 2 parts, namely Physiological Needs and Safety Needs. If the seller of the house makes a mistake, the result will be that the house will be less accepted on the market, this can make the house sellers have a smaller possibility or even lose the opportunity to get the maximum profit from the sale. To minimize errors in setting the selling price of the house, the seller must be careful in determining the price, that's because the selling price of the house is mostly rising and rarely goes down either in the short or long term. The main drive behind the project is a prediction of real estate prices to build the best house price prediction systems using machine learning algorithms with maximum accuracy and without any loss. Many factors have to be taken into consideration for predicting house prices and try to predict efficient house pricing for customers concerning their budget as well as also according to their priorities. So, we are creating a housing cost prediction model. The price of a house depends on many factors like Area, location, population, size and number of bedrooms, bathrooms given, condition, quality, square footage, etc.

The valuation of residential properties is a pivotal aspect of the real estate market, influencing a wide range of economic and personal decisions. Traditionally, the focus has been on a limited set of parameters—primarily location, size, and the number of bedrooms—to estimate property values. However, this conventional approach often falls short of capturing the true market value of properties, as it overlooks many nuanced and significant features that can affect a house's appeal and worth. As the real estate market evolves, there is a growing recognition of the need for more sophisticated valuation models that reflect the complexities of modern properties and the diverse p of today's buyers.

Despite the critical role of accurate property valuation in real estate transactions, current methodologies remain limited in their ability to account for the full spectrum of factors that contribute to a house's market value. Many models fail to incorporate elements such as architectural style, age, renovations, technological enhancements, and environmental sustainability features. This oversight can lead to discrepancies between estimated values and market realities, posing challenges for buyers, sellers, and investors in making informed decisions. The limitation of existing models underscores the need for an innovative approach to house price prediction that leverages the capabilities of data mining.

This study aims to address the gaps in traditional house price prediction models by introducing a comprehensive data mining approach. By integrating a wide range of house features into our analysis, including those often neglected in conventional models, we seek to develop a predictive framework that more accurately reflects the factors influencing property values today. This research intends to not only enhance the accuracy of price predictions but also offer insights into the relative importance of various property features, thereby providing a richer understanding of what drives market values in the real estate sector.

Our approach employs advanced machine learning algorithms to analyze a dataset that encompasses an extensive array of property characteristics. This methodology allows us to systematically evaluate the impact of each feature on house prices, facilitating the development of a predictive model that is both nuanced and comprehensive. The study meticulously selects features based on their relevance and availability, ensuring a robust analysis that captures the complexities of the real estate market. Through this process, we aim to uncover patterns and relationships that traditional valuation methods may overlook, thereby offering a more sophisticated tool for property valuation.

The implications of this research extend far beyond academic interest, offering practical benefits for a range of stakeholders in the real estate market, including homeowners, buyers, investors, and policy-makers. By providing a more accurate and comprehensive tool for house price prediction, this study aids in the facilitation of more transparent and informed real estate transactions. Furthermore, the insights gained from this research contribute to the broader understanding of the real estate market, potentially influencing future developments in property valuation practices. Ultimately, this study not only challenges existing paradigms but also sets a new standard for the integration of data mining techniques in real estate analytics.

The proposed model is meticulously designed to yield highly accurate results by considering a wide array of factors that influence house prices. It utilizes a variety of sophisticated machine learning models, such as Multiple Linear Regression (MLR), Decision Tree Regression, Logistic Regression, and Artificial Neural Networks. This comprehensive approach makes the house pricing model immensely beneficial for a diverse group of users including buyers, property investors, and house builders. By leveraging this model, entities involved in the real estate sector along with other stakeholders can effectively evaluate current market trends and identify properties that are both high in quality and budget-friendly.

Initially, the focus of the model was primarily on analyzing the attributes that most significantly influence house prices using only Multiple Linear Regression. However, to enhance the accuracy and robustness of the results, additional methodologies such as Decision Trees and Random Forest Regressors were later integrated. These improvements were aimed at refining the predictive capabilities of the model, thereby increasing its reliability and utility.

The model begins with the compilation and utilization of a dataset sourced from a reliable provider, ensuring simplicity and ease of use. For our house price prediction, a particularly extensive dataset was selected, comprising 21,613 records and 23 distinct features, which provides a solid foundation for training the model. The selection and application of various machine learning procedures allow for the projection of future values, enhancing the model's ability to forecast property values with greater accuracy and minimal error (refer to Figure 1). This predictive model stands out as a critical tool for forecasting future property valuations, serving as an essential resource for making informed investment decisions in the real estate market.

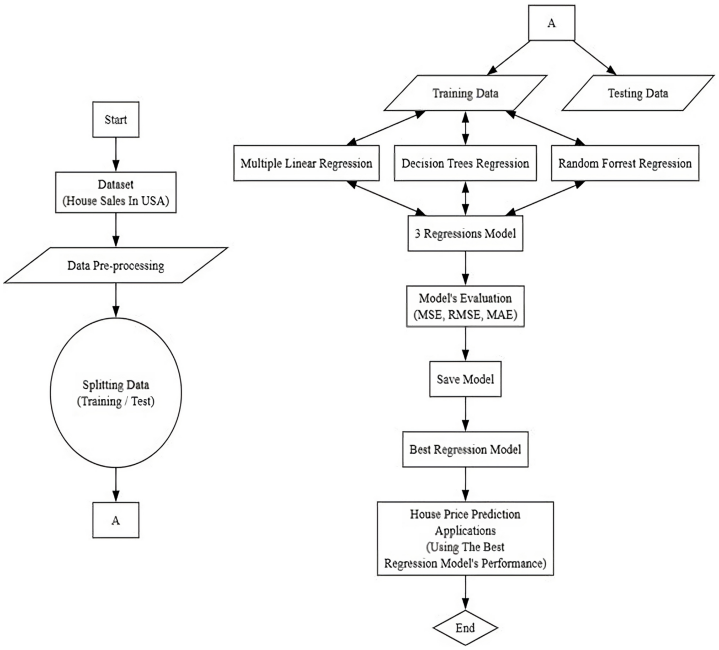

Fig. 1. Project workflow diagram

II EXPLORATION OF DATASET

In our dataset, each variable plays a crucial role in constructing a predictive model for house prices. The variables are carefully selected to capture a comprehensive picture of what drives property value, ensuring our model is both accurate and reliable.

- **cid**: This unique identifier for each house is essential for tracking individual property transactions without revealing private information.
- **day hours**: The sale date of a house could influence its price due to seasonal market fluctuations, with certain times of the year possibly fetching higher prices.
- **price**: Serving as our prediction target, this variable is the outcome we aim to predict using the other features in the dataset.
- **room_bed**: The number of bedrooms is a fundamental characteristic that affects a house's utility and appeal, making it a significant predictor of price.
- **room_bath**: The number of bathrooms is also a key feature, often reflecting the house's size and level of comfort, which in turn affects its market value.
- **living_measure**: The total living area in square footage is a direct measure of a house's size and is often correlated with price, with larger homes commanding higher prices.
- **lot_measure**: The size of the property lot can greatly

impact the value, as it pertains to the potential for expansion, landscaping, and privacy.

- **ceil**: The number of floors in a house often indicates its capacity to accommodate larger families or provide expansive views, which can be particularly valuable in certain markets.
- **coast**: Properties with a waterfront view are highly sought after and can significantly increase a house's desirability and price.
- **sight**: If a property has been viewed often, it may be indicative of high interest, which could affect its selling price due to competitive bidding.
- **condition**: The overall condition of a property is critical as it affects both the immediate usability of the house and the potential costs for new owners, influencing their willingness to pay.
- **quality**: The grade assigned to a housing unit is a summary of its construction and design quality, which can dramatically influence buyer perception and pricing.
- **ceil_measure**: The living area excluding the basement often correlates with the main living space quality, affecting the house's functionality and value.
- **basement_measure**: The size of the basement can add significant value, particularly if it's finished and usable as living space.
- **yr_built**: The age of a building can affect its value, with newer homes often fetching higher prices due to modern design and fewer anticipated repair costs.
- **yr_renovated**: Recently renovated homes may see increased value due to updated features and reduced maintenance needs.
- **zip code**: Often, the zip code can indicate the property's location within a region, capturing socio-economic factors, school districts, and proximity to amenities, which can significantly influence house prices.
- **latitude**: This geographic coordinate offers precise information about the property's north-south position, which can impact factors like climate and desirability of the location.
- **longitude**: The east-west positioning provided by the longitude coordinate can affect the property's accessibility to key locations and its overall value.
- **living_measure15**: This reflects any changes or renovations in the living area that occurred in 2015. Renovations can affect property value, and noting the year allows for the appreciation of such enhancements over time.
- **lot_measure15**: Similar to living_measure15, this variable denotes any changes to the lot size area in 2015, which could imply renovations or modifications impacting the property's outdoor space and value.
- **furnished**: The level of furnishing can directly influence a buyer's first impression and the immediate habitability of the property, affecting its market value.
- **total area**: By combining the living and lot areas, this measure provides an overall size indicator of the property, which is a fundamental aspect affecting price due to the combined utility of indoor and outdoor space.

We differentiate between two types of variables: numerical and categorical. Each type offers unique insights and is essential for the construction of a predictive model.

Numerical variables are those that exist along a continuum and represent quantities. They can be measured and expressed with numerical values, allowing for mathematical operations to be performed on them. In our dataset, some of numerical variables include:

- **price**: The value of the property that we are aiming to predict.
- **living_measure**: The square footage of the living area within the house.
- **lot_measure**: The square footage of the property's lot.
- **ceil_measure**: The square footage of the living area excluding the basement.
- **basement**: The square footage of the basement area.
- **living_measure15**: The square footage of the living area of the 15 nearest neighbors in 2015.
- **lot_measure15**: The square footage of the lot areas of the 15 nearest neighbors in 2015.
- **total_area**: The combined square footage of the living area and lot, providing an overall measure of space.

These variables can often be directly correlated with the property price, as they reflect size, space, and utilitarian aspects of the properties, which are typically quantifiable.

On the other hand, categorical variables represent characteristics or qualities that can be categorized and typically take on a limited number of possible values. They are not inherently numerical but can be coded or transformed into numerical values for analysis. Some of the categorical variables in our dataset include:

- **room_bed**: The number of bedrooms, which categorizes the house by the sleeping capacity.
- **room_bath**: The number of bathrooms, categorizing the house by its facilities.
- **ceil**: The number of floors or levels in the house, categorizing its vertical size.
- **coast**: An indicator of whether the house has a view of the waterfront, a feature with two categories: yes or no.
- **sight**: A measure of how many times the property has been viewed, which can be an indicator of interest or desirability.
- **condition**: The overall condition of the property, often rated on a standard scale.
- **quality**: The construction and design grade given to the housing unit, usually based on a defined grading system.
- **furnished**: The level of furnishing in the property, which can influence buyer perception and thus affect price.

Understanding these variable types is vital for appropriately processing the data and ensuring that the model reflects the underlying patterns within the housing market. Numerical variables can directly enter most models, while categorical

variables may need encoding or transformation to be properly utilized.

III METHODS

III-A DESCRIPTION OF DATA AND PRE-PROCESSING

The first phase of our investigation focuses on the pivotal role of regression techniques in predicting house prices. The research employs a comprehensive dataset detailing house sales in King County, USA, renowned for its extensive coverage and robustness. This dataset provides a solid foundation for the predictive modeling, encompassing sales listings recorded between May 2014 and May 2015. It is composed of 21,613 individual data entries, each characterized by 23 distinct features that encapsulate a wide array of attributes relevant to the housing market. The dataset's integrity and accessibility are ensured by its availability on the official Kaggle website, a platform esteemed for its reliability and the quality of datasets it hosts. By utilizing this dataset, the study aims to distill insights from the data through the application of various regression techniques, each contributing uniquely to the accuracy of the house price predictions.

During the critical phase of Data Pre-processing, our methodology adopts a multi-faceted approach to ensure data quality and relevance. The dataset undergoes a thorough examination, utilizing graphical representations and data checking protocols, to reveal underlying patterns and anomalies. Several data visualization techniques, along with Exploratory Data Analysis (EDA), are employed to scrutinize the dataset's structure and content. This initial analysis is instrumental in identifying any duplicates, missing values, and outliers that could skew the results of the predictive modeling. By addressing these potential issues in the early stages, we aim to maintain the data's integrity and ensure that subsequent findings are both valid and reliable. The meticulous pre-processing phase is not solely about cleansing the data but is also about understanding the myriad factors that influence house prices, thereby providing a clear and unobstructed view of the real estate market dynamics.

Once the data has been rigorously examined and refined, the next pivotal step in our process involves the partitioning of the dataset. This critical division creates two subsets of data: the training set and the testing set. The first partition, which constitutes the training data, is leveraged to build and fine-tune the predictive model. This subset is where the model learns the intricate relationship between the various features and the house prices, thus becoming equipped to make predictions. The second partition, the testing data, serves as a new and unseen dataset for the model, used exclusively to evaluate the model's predictive prowess. This bifurcation is a standard practice in machine learning and is crucial for validating the model's performance. It ensures that the model is not only accurate with the data it was trained on but also retains its precision when confronted with new, unseen data, thereby confirming the model's generalizability and robustness in real-world scenarios. Through this rigorous process, we anticipate developing a model that stands as a testament to the efficacy of machine learning in real estate valuation.

III-B DATA VISUALIZATION AND EXPLORATORY DATA ANALYSIS(EDA)

The process of visualization and Exploratory Data Analysis (EDA) is a cornerstone of our methodological framework, serving to unravel the complex relationships within the housing dataset. Excluding identifiers and locational markers such as ID, Date, Zipcode, Latitude, and Longitude, a meticulous examination of the dataset's features against the target variable — the price of a house — is conducted. This phase delves into the intricate details of data distribution, which is partitioned into two distinct categories: the continuous numerical features and the discrete numerical features. For the former, attributes like the size of the living space, land area, number of floors, and size of the living area above ground level are scrutinized. This nuanced analysis helps in identifying patterns and trends that continuous variables exhibit in relation to house prices, thereby setting the foundation for sophisticated predictive modeling.

Discrete numerical features, on the other hand, encompass attributes such as waterfront presence, view quality, overall condition, and grade of the house. These categorical variables are quantified and analyzed for their individual and collective impact on the house pricing. The discrete nature of these features requires a different analytical approach to discern how each categorical level corresponds to variations in price. This bifurcation in the analysis allows for a more nuanced understanding of how both continuous and discrete factors contribute to the property's valuation, leading to insights that are critical for stakeholders looking to understand the nuances of the real estate market.

Subsequent to the initial stages of analysis comes the crucial assessment of data distribution normality within the features. Several variables, including the price, exhibit distributions that are marred by the presence of outliers, indicating deviations from normality. Features like the size of living spaces, both on the whole lot and above ground level, as well as the average living and land space sizes of the 15 closest neighbors, demonstrate abnormal data distributions. These anomalies are pivotal in understanding the outliers that could potentially skew the model's predictions if left unaddressed. Normality checks are integral to preparing the dataset for the application of regression techniques, as they often assume normal distribution of variables.

The investigation of data distribution extends beyond the search for normality. It is also focused on how each feature interacts with the target variable: the price. Visual evidence presented in Figure 3 illustrates the significant impact of certain features on house pricing. For example, the attribute of being a waterfront property shows a notable correlation with house price, suggesting that waterfront properties are often valued higher. Continuous numerical features like the

size of the living space are instrumental in forecasting price, with larger living spaces commonly driving higher prices. Similarly, the age of the house, deduced from the year of construction, is found to have a significant influence on price, with newer or more recently renovated properties often fetching higher market values.

For numerically discrete features, the analysis uncovers that variables such as View, Waterfront, and Grade have a pronounced effect on escalating the sale price. The impact of these discrete features is not linear but rather shows an exponential relationship with price, reflecting the premium placed on exceptional views, waterfront access, and higher construction grades. These features, with their categorical nature, contribute a layered complexity to the dataset, necessitating discrete analysis to capture the subtleties in how they modify the house's value.

The condition of a house also plays a vital role, although its effect on pricing follows a more stable trend. Unlike the aforementioned features where the effect on price can dramatically increase, the condition of the house influences the price in a more predictable and steady manner. This observation is elucidated through bar plot visualizations in Figure 2, where the relationship between house condition and price can be interpreted with greater clarity. Such visualizations are instrumental in conveying the influence of discrete numerical features, offering an immediate graphical representation of the data that can be intuitively understood by researchers and stakeholders alike.

Overall, these stages of analysis embody the thorough and multifaceted approach required to handle the complexity of real estate data. They underscore the necessity of both broad and detailed examinations of the factors at play, paving the way for a predictive model that not only accounts for the diverse array of features within a dataset but also appreciates the subtleties of their interrelationships. The nuanced insights garnered from this process are invaluable, laying a robust groundwork for the predictive tasks that follow and ensuring that the resulting model is well-calibrated to the realities of the housing market.

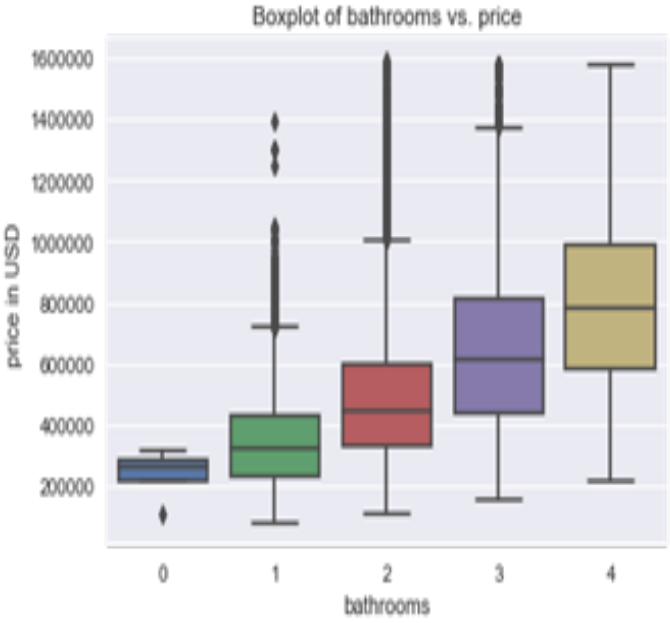

Fig. 2. Price Vs Discrete variable

Fig. 3. Categorical feature plot

III-C DATA CLEANSING

Prior to delving into the detailed analysis of continuous and discrete numerical features within the dataset, a crucial preliminary step is undertaken: the meticulous scanning for and handling of NULL values. This data cleaning process is not only about exclusion but also preparation, as it involves dropping certain features that are deemed non-essential for the predictive modeling task. Specifically, features such as ID, Date, Zipcode, Latitude, and Longitude are removed from the dataset. This deliberate elimination is guided by the understanding that these attributes, while useful for identification and location purposes, do not contribute meaningfully to the model's ability to predict house prices.

The data is then rigorously prepared for model training; this phase is essential as clean and well-prepared data forms the backbone of any reliable machine learning model. During this phase, tasks such as handling missing values, excising outliers, and encoding categorical variables are executed. These actions fall under the umbrella of feature engineering, a process critical for transforming raw data into a format that is suitable for modeling.

Feature engineering also includes the application of feature selection techniques which are instrumental in pinpointing the most influential attributes for the prediction of house prices. The objective is to distill the dataset down to the

variables that hold the most predictive power, thereby streamlining the modeling process and potentially enhancing the performance of the predictive algorithms. Once the most relevant features are selected, the data must be processed to optimize the models' ability to discern and learn from underlying patterns.

This entails standardizing numerical values to a common scale, and converting categorical values into a format that can be understood by machine learning algorithms through one-hot encoding. These preparatory steps are crucial for eliminating potential biases that could distort the models' learning phase and for ensuring that each feature contributes appropriately to the predictive task at hand.

An additional step in the data preparation process involves examining the dataset for multicollinearity using the Pearson correlation coefficient. The resulting analysis is often visualized through a heatmap, as indicated in Figure 4, which provides a vivid representation of the correlation between different variables. Through this visual exploration, a notable correlation was detected between 'Living Measure' and 'Living Measure 15', which denotes the living space in relation to the neighboring 15 properties.

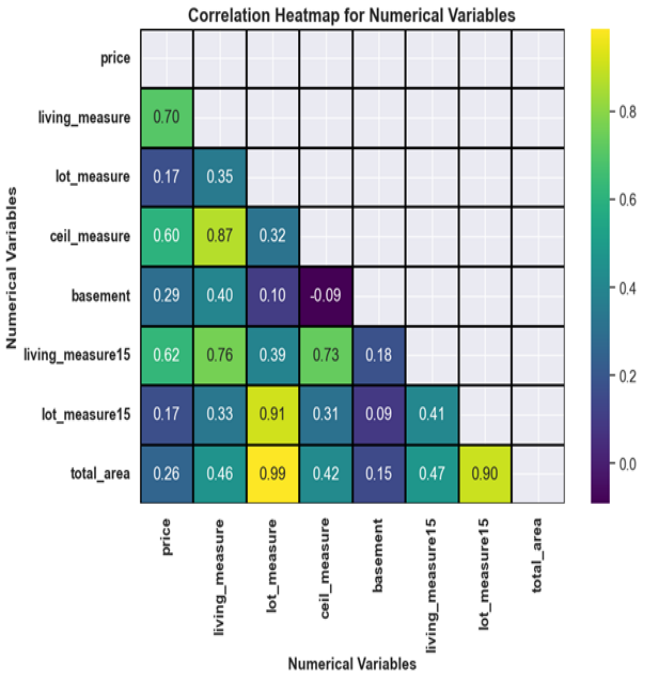

Fig. 4. Pearson correlation heat map.

- There is a strong positive correlation between `living_measure` and `living_measure15` (0.76), which indicates that the living space of a house is highly similar to the average living space of the 15 closest neighbors. This could suggest that houses within the same neighborhood tend to have similar living space sizes.
- `lot_measure` and `lot_measure15` also share a strong positive correlation (0.91), much like the living measures. This again suggests a consistency in lot sizes within local neighborhoods.
- A very strong correlation is observed between `total_area` and both `living_measure` and `lot_measure` (0.99 and 0.90, respectively), which is expected since total area is likely a sum or combination of living and lot area.
- `ceil_measure` has a relatively strong correlation with `living_measure` (0.87), implying that houses with larger living spaces also tend to have larger ceiling measurements, possibly reflecting overall larger and possibly taller houses.
- `price` has a moderately strong correlation with `living_measure` (0.70), suggesting that as the living space of a house increases, so does the price, which is a common trend in real estate markets.
- The `basement` variable has low to moderate correlations with other features, indicating that it might not be as strongly connected to the size of other house features or the overall pricing as other variables are.

III-D CLUSTERING & SCALING THE DATA

Scaling is done on the data before clustering to ensure that each feature contributes equally to the clustering process, preventing biased results where certain features dominate due to their scale. Here two clusters are formed and analysis of the clusters reveals clear differences: Cluster 0 generally consists of higher-priced, larger, better-conditioned houses with superior views and furnishings, while Cluster 1 represents properties with lower values across these metrics. However, when regression analysis is applied to each cluster separately, it becomes apparent that the clusters only explain less than half of the overall variability in the dataset. Given below is (Table 1) Shows the results after performing linear regression on clusters separately. This suggests that the clusters might not fully capture the diversity present in the entire dataset, possibly due to the influence of categorical variables. Further investigation is needed to refine the clustering approach and improve its accuracy.

TABLE I
CLUSTER REGRESSION RESULTS

Cluster	Train RMSE	Test RMSE	Training Score	Test Score
High Cluster (Cluster 0)	180325.283	179476.109	0.437610	0.432128
Low Cluster (Cluster 1)	128868.297	128001.968	0.421834	0.415300

From the table,we can say:

Train RMSE (Root Mean Square Error):

- High Cluster (Cluster 0) has a Train RMSE of 180325.283, suggesting that when the model is applied to the training data of higher-priced and presumably better-conditioned houses, the average error is higher compared to the Low Cluster.
- Low Cluster (Cluster 1) shows a Train RMSE of 128868.297, indicating that the model predicts lower-

priced houses with a smaller average error.

Test RMSE:

- High Cluster (Cluster 0) has a Test RMSE of 179476.109, which is quite close to its Train RMSE, suggesting a consistent model performance from training to testing in higher-priced houses.
- Low Cluster (Cluster 1) has a Test RMSE of 128001.968, also mirroring its training performance closely.

Training Score:

- High Cluster (Cluster 0) has a training score of 0.437610, and Low Cluster (Cluster 1) has a training score of 0.421834. These scores are measures of the model's accuracy, with 1 being a perfect score. The scores suggest that both models are performing moderately well, but there is considerable room for improvement.

Test Score:

- High Cluster (Cluster 0) has a test score of 0.432128 and Low Cluster (Cluster 1) has a test score of 0.415300. The test scores are lower than the training scores for both clusters, which is common in model performance due to overfitting or model complexity. However, the decline is not drastic, indicating that the model maintains some level of generalizability.

IV RESULTS

In the next step, the linear regression analysis conducted on the dataset yielded a model that explains approximately 65% of the variance in both the training and test data. While this demonstrates a moderate level of predictive ability, it suggests that the model may not fully capture the complexity of the data. Given is the regression equation to predict the price of the house based on the features. To further explore the importance of different features in predicting house prices, feature importance analysis was performed. The results of this analysis are presented in the accompanying feature importance plot, which highlights the relative significance of each feature in the regression model.

$$\begin{aligned}\text{Price} = &-14207.35 \times \text{room\_bed} + 21372.84 \times \text{room\_bath} \\ &+ 25695.58 \times \text{ceil} + 140365.88 \times \text{coast} \\ &+ 28660.38 \times \text{sight} + 20246.03 \times \text{condition} \\ &+ 78870.96 \times \text{quality} + 84.86 \times \text{ceil\_measure} \\ &+ 114.64 \times \text{basement} + 58.63 \times \text{living\_measure15} \\ &- 4.15 \times \text{lot\_measure15} + 62563.28 \times \text{furnished} \\ &- 1.28 \times \text{total\_area} + 2255.73 \times \text{house\_age} \quad (1)\end{aligned}$$

Moving forward, additional regression techniques will be explored to assess their efficacy in capturing the underlying patterns in the data. Specifically, decision trees, random forests, and artificial neural network (ANN) regressors will be employed. These algorithms offer the advantage of capturing non-linear relationships between features and target variables, which may better reflect the underlying structure of the dataset. By comparing the performance of these non-linear regression models with that of linear regression, we aim to identify the most suitable approach for predicting house prices in our dataset.

TABLE II

RESULTS BEFORE HYPERPARAMETER TUNING

Regression Model	Train RMSE	Test RMSE	Training Score	Test Score
Linear Regression	146602.06	146173.57	0.6592	0.6510
Decision Tree Regressor	8286.19	181869.34	0.9989	0.4598
Random Forest Regressor	49285.74	129987.63	0.9615	0.7240

TABLE III

RESULTS AFTER HYPERPARAMETER TUNING

Regression Model	Train RMSE	Test RMSE	Training Score	Test Score
Linear Regression	146602.06	146173.57	0.6592	0.6510
Decision Tree Regressor	127342.60	142514.23	0.7429	0.6683
Random Forest Regressor	120945.44	133225.85	0.7680	0.7101

The tables summarize the performance metrics of various regression models both before and after the process of hyperparameter tuning. The metrics used to evaluate the models include the Root Mean Square Error (RMSE) on both training and testing data, as well as scores that presumably reflect the models' accuracy on training and testing sets.

From the pre-tuning results, the Decision Tree Regressor exhibits a discrepancy between training and testing performance, suggesting overfitting. The Random Forest Regressor, while having higher RMSE values, shows more consistency between training and testing scores, indicating better model generalization.

Post-tuning, there is a notable improvement across all models, suggesting that hyperparameter tuning has effectively enhanced model performance and possibly reduced overfitting

V DISCUSSIONS

Hyperparameter tuning is an essential step in building robust machine learning models, as it significantly influences their performance. The meticulous process of hyperparameter optimization for regression models like the Decision Tree and Random Forest Regressors has been demonstrated to refine predictive accuracy. For the Decision Tree Regressor, tuning involved identifying the optimal maximum depth, minimum samples per leaf, and minimum samples required to split a node. The ideal parameters discovered through techniques such as grid search—maximum depth of 20, minimum samples per leaf of 30, and minimum samples split of 15—were instrumental in improving the model's performance.

In a similar vein, the Random Forest Regressor underwent a thorough hyperparameter search. The parameters that yielded the best results were a maximum depth of 10, a maximum of 6 features to consider when looking for the best split, a minimum of 3 samples required to be at a leaf node, and 30 samples required to split an internal node, with the forest consisting of 101 estimators. These hyperparameters are crucial because they control the complexity of the model and, therefore, its ability to generalize from the training data to unseen test data.

The purpose of hyperparameter tuning goes beyond mere adjustment of model parameters; it targets the reduction of error while ensuring the model does not overfit the training data. Overfitting occurs when a model learns the details and noise in the training data to an extent that it negatively impacts the model's performance on new data. The optimal hyperparameters are those that strike a delicate balance between the model's ability to learn from the training data and its capacity to generalize to new, unseen data.

Post-tuning, the adjusted models were set against the baseline Linear Regression model to compare performance. The results indicated a marked improvement, showcasing the effectiveness of the tuning process. Both the Decision Tree and Random Forest regressors not only improved their own performance metrics but also outstripped the Linear Regression model in terms of training and test scores. This comparison illustrates the considerable impact that careful hyperparameter optimization can have on the performance of complex models.

The standout performance was noted in the Random Forest Regressor, which, after tuning, achieved the most favorable test RMSE and test scores among the models. This indicates a significant enhancement in the model's ability to predict accurately on data it hasn't encountered before, a critical measure of success for real-world applications. Such advancements in generalization performance emphasize the Random Forest's robustness, making it a potentially more reliable choice for predicting house prices. Overall, the improvements post-tuning elucidate the pivotal role that hyperparameter tuning holds in the development of machine learning models, particularly in fields like real estate where accurate predictions can have substantial economic implications. The value of this process is clear: it not only refines the models for greater accuracy but also ensures their applicability to a wide range of real-world scenarios.

VI CONCLUSIONS

In conclusion, our comprehensive analysis of housing data within King County has yielded valuable insights for real estate market strategies. We observed that waterfront properties and upscale neighborhoods like Medina, Clyde Hill, and Mercer Island command significantly higher prices. Additionally, attributes such as house grade and bedroom count emerged as strong predictors of price. Timing-wise, launching campaigns in March/April aligns with peak sales periods in Q2.

Our modeling efforts, particularly the hyperparameter tuning of regression algorithms, significantly enhanced predictive performance. Notably, the Random Forest Regressor exhibited superior performance without signs of underfitting, underscoring its effectiveness in capturing complex relationships within the data. Furthermore, future enhancements could explore alternative regression algorithms like Ridge, Lasso, Bagging, and Boosting, potentially yielding further improvements in model accuracy.

In future endeavors, incorporating additional datasets on commuting times, income distributions, longer-term trends, and school rankings could provide richer insights and refine our predictive models. Ultimately, our findings lay the groundwork for developing more tailored and effective real estate market strategies, empowering stakeholders to make informed decisions and capitalize on emerging opportunities in the housing market.

References

[1] Frontiers — Differential Impact of COVID-19 Risk Factors on Ethnicities in the United States (frontiersin.org)

[2] A Comprehensive Guide on Hyperparameter Tuning and its Techniques (analyticsvidhya.com)

[3] Regression Analysis: Step by Step Articles, Videos, Simple Definitions (statisticshowto.com)

[4] Regression analysis—ArcGIS Insights — Documentation

[5] Subhani Shaik, Uppu Ravibabu. Classification of EMG Signal Analysis based on Curvelet Transform and Random Forest tree Method. Paper selected for Journal of Theoretical and Applied Information Technology (JATIT). 95. 15

[6] Shiva Keertan J, Subhani Shaik. Machine Learning Algorithms Prediction, for International Oil Journal Price of Innovative Technology and Exploring Engineering. 8(8). 16.

[7] KP Surya Teja, Vigneswar Reddy and Subhani Shaik, Flight Delay Prediction Using Machine Learning Algorithm XGBoost, Jour of Adv Research in Dynamical & Control Systems. 11(5).