Analyzing Housing Dynamics and Property Valuations

Introduction

The business problem that the study aims to address revolves around using data analytics to gain deeper insights into the dynamics of the real estate market. Specifically, the study seeks to understand the factors influencing housing prices and identify patterns associated with expensive properties. By doing so, the study provides valuable intelligence to stakeholders in the real estate industry, enabling them to make informed decisions regarding pricing strategies, investment opportunities, and urban development initiatives. Ultimately, the goal is to contribute to the optimization of resource allocation, promote equitable access to housing, and enhance overall market efficiency within the real estate sector.

The real estate industry plays a pivotal role in both the economy and society, with housing being a fundamental need and a significant investment for individuals and businesses alike (Zhao & Liu, 2023). The study aims to uncover hidden patterns, correlations, and insights that can inform strategic decision-making for various stakeholders, including real estate developers, investors, policymakers, and prospective homeowners, by using data analytics techniques on the provided dataset. Understanding the drivers of median house values and identifying expensive properties can guide urban planning efforts, inform marketing strategies, optimize investment portfolios, and contribute to fostering inclusive and sustainable communities(Vaidynathan et al., 2023).

Methodology

Data Exploration

The dataset under analysis, named 'California Housing Prices Data Set', provides a comprehensive collection of housing-related attributes across various regions. With a sample size of N = 20640, this dataset encompasses a diverse range of features, including longitude, latitude,

housing median age, total rooms, total bedrooms, population, households, median income, ocean proximity, median house value, and an indicator for expensive properties.

The numerical response variable, Y1, in this context, refers to the median house value in a given locality. It serves as a crucial metric for assessing the affordability and market dynamics of housing in different areas. Understanding the factors that influence median house values can aid in effective pricing strategies, informed investment decisions, and informed policy formulation within the real estate sector. The categorical response variable, Y2, indicates whether a property is considered expensive or not, based on specific criteria. This binary classification offers insights into the high-end segment of the housing market, helping to identify patterns and trends associated with luxury properties.

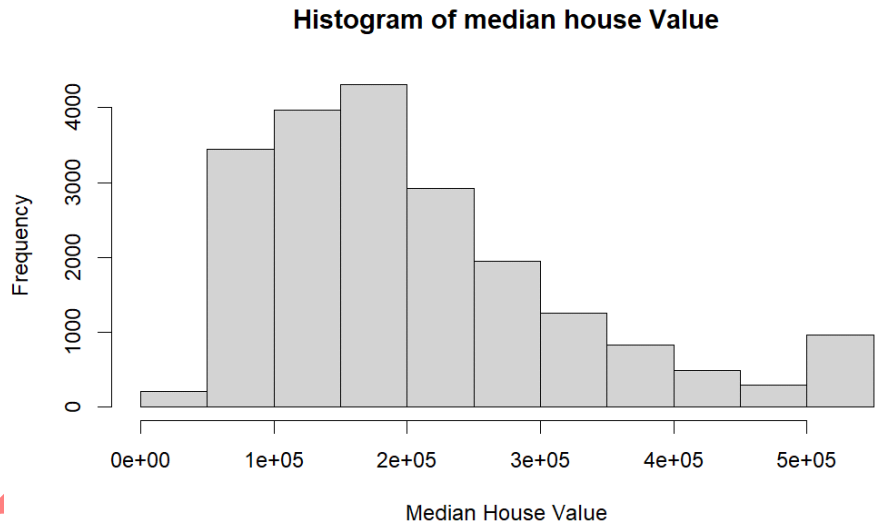

Figure 1: Histogram for Median House Value

The histogram displays the distribution of median house values across different value ranges. The x-axis represents the median house value bins, while the y-axis shows the frequency or count for each bin. The tallest bar suggests that the most frequent median house values fall between $ 200,000 and $ 299,999. The distribution is right-skewed, with fewer houses in the

higher value ranges, particularly those exceeding 300,000. The histogram provides a visual representation of how the median house values are concentrated in the lower to mid-range values, with a gradual decrease in frequency as the values increase.

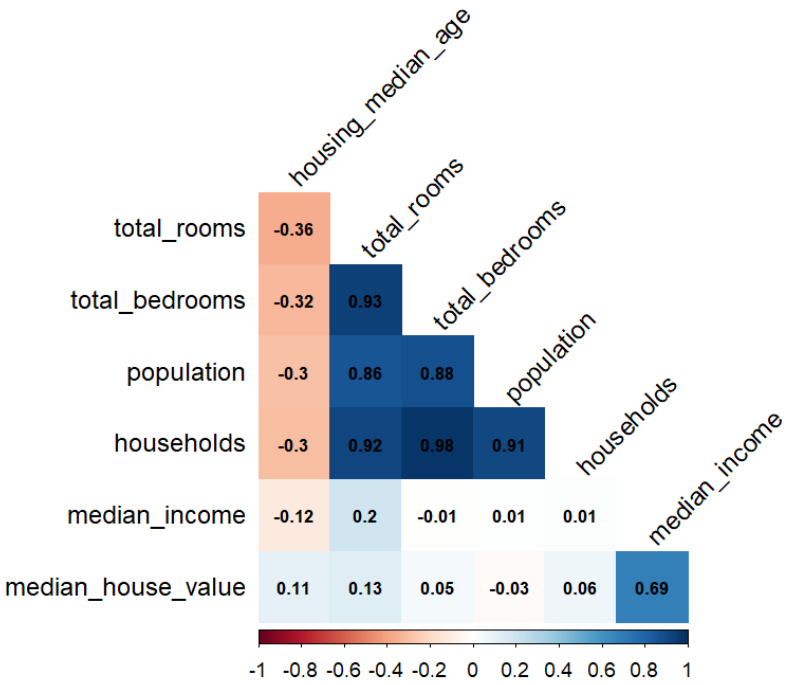

Figure 2: Correlation Matrix

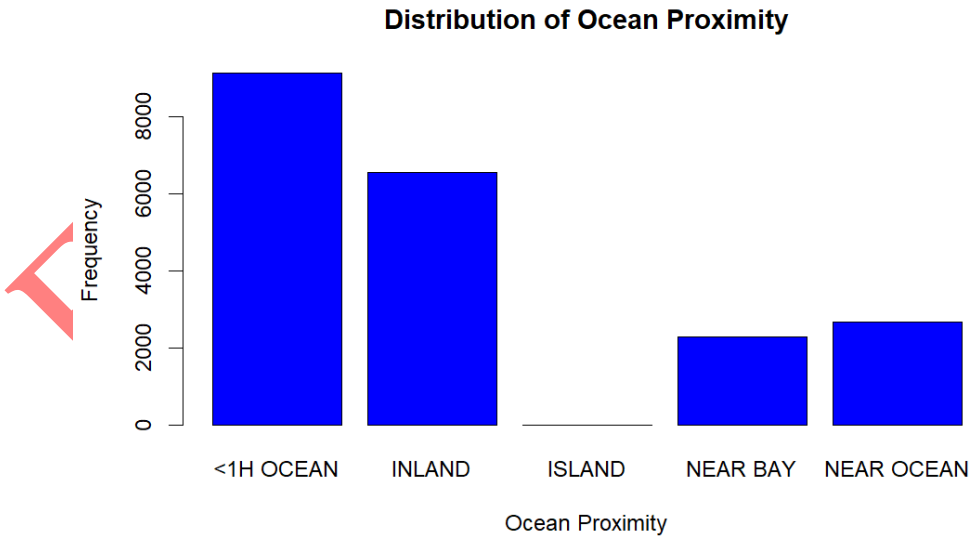

Figure 3: Distribution of Ocean Proximity

Data preprocessing

The data underwent thorough examination to identify and address missing values. Utilizing R's functionalities, each variable within the dataset was examined to identify any instances of missing data. The presence of missing values was assessed across the dataset. Fortunately, the scrutiny revealed a lack of missing values, affirming the dataset's completeness. This absence of missing data underscores the dataset's reliability and completeness, providing a solid foundation for subsequent analyses and ensuring that the conclusions drawn from the dataset are robust and representative of the underlying population.

The dataset was randomly partitioned, with approximately 60% of the observations allocated to the training set and the remaining 40% assigned to the validation set. This random allocation helped ensure that both sets were representative of the overall dataset, capturing a diverse range of observations and preserving the underlying distribution of key variables. By incorporating a random element into the partitioning process, bias was minimized, and the resulting models were less likely to overfit to specific patterns present in the training data (Hassanat et al., 2022).

Modeling

To address the problem at hand, which involves analyzing the relationship between various predictors and the target variables (median house value and flag_expensive), several data mining techniques and algorithms were employed. Firstly, for the regression modeling aimed at predicting the median house value (Y1), a stepwise regression approach was utilized. This method iteratively selects the most significant predictors based on their contribution to minimizing the model's error, thereby creating a parsimonious model that includes only the most relevant variables (Morozova et al., 2015). Additionally, a regression tree model, specifically

CART (Classification and Regression Trees), was employed. CART is a non-parametric decision tree algorithm that recursively partitions the data into subsets based on the predictors' values, optimizing splits to maximize the homogeneity of the resulting groups in terms of the target variable (Kern et al., 2019). CART produces a tree-like structure that provides interpretable insights into the relationships between predictors and the target variable by recursively splitting the data.

Secondly, for the classification modeling targeting the flag_expensive variable (Y2), logistic regression and decision tree algorithms were applied. Logistic regression is a classic statistical method used for binary classification problems, wherein it models the probability of the target variable (flag_expensive) being in a certain category based on the predictor variables. Logistic regression quantifies the relationship between the predictors and the log-odds of the target variable, allowing for probabilistic predictions by estimating the coefficients of the predictors (Shipe et al., 2019). Additionally, a classification CART model was employed. Similar to the regression tree model, CART for classification partitions the data based on predictor variables, but in this case, it predicts the class label (i.e., whether an item is expensive or not) rather than a continuous outcome. These techniques collectively provided a comprehensive framework for understanding and predicting both continuous and categorical outcomes within the real estate domain, thereby facilitating informed decision-making processes.

The chosen data mining techniques align closely with the problem statement and dataset characteristics, primarily due to their flexibility in handling both continuous and categorical target variables, as well as their ability to capture nonlinear relationships between predictors and outcomes. Regression modeling techniques, such as stepwise regression and regression trees, were well-suited for predicting the continuous variable of median house value,

enabling the exploration of complex interactions between housing features and market dynamics. Similarly, classification algorithms, including logistic regression and decision trees, were apt for modeling the binary outcome of property expensiveness, offering interpretable insights into the factors influencing housing affordability. By leveraging these techniques, the analysis could uncover nuanced patterns within the real estate dataset, facilitating informed decision-making processes tailored to the needs of stakeholders.

Results

Table 1

Descriptive Statistics

		housing_median_age	total_rooms	total bedrooms	population	households	median income
N	Valid	20640	20640	20433	20640	20640	20640
Mean		28.6395	2635.7631	537.8706	1425.4767	499.5397	3.8707
Median		29.0000	2127.0000	435.0000	1166.0000	409.0000	3.5348
Minimum		1.00	2.00	1.00	3.00	1.00	.50
Maximum		52.00	39320.00	6445.00	35682.00	6082.00	15.00
Percentiles	25	18.0000	1447.2500	296.0000	787.0000	280.0000	2.5628
	50	29.0000	2127.0000	435.0000	1166.0000	409.0000	3.5348
	75	37.0000	3148.0000	647.0000	1725.0000	605.0000	4.7436

The descriptive statistics table provides an overview of the central tendency, dispersion, and distribution of the variables housing_median_age, total_rooms, total bedrooms, population, households, and median income. The mean values indicate the averages for each variable, with housing_median_age (M = 28.64), total_rooms (M = 2635.76), total bedrooms (M = 537.87), population (M = 1425.48), households (M = 499.54), and median income (M = 3.87). The median represents the middle values for housing_median_age (29.00), total_rooms (2127.00),

total bedrooms (435.00), population (1166.00), households (409.00), and median income (3.53). The minimum and maximum values highlight the ranges of the data. Additionally, the 25th, 50th (median), and 75th percentiles provide insights into the distribution of the data. This information helps understand the characteristics of the dataset and identify potential outliers or skewness in the data distribution.

Regression Modeling

1. **Linear Regression**

Table 2: Coefficients of Regression

Variable	Estimate	Std. Error	t-value	p-value
(Intercept)	−2,206,000	111,600	-19.775	<.001***
Longitude	−26,010	1,293	-20.113	<.001***
Latitude	−24,590	1,277	-19.253	<.001***
Housing Median Age	1,061	56.50	18.771	<.001***
Total Rooms	-6.424	1.020	-6.298	<.001***
Total Bedrooms	88.33	8.696	10.158	<.001***
Population	-42.00	1.472	-28.521	<.001***
Households	75.63	9.463	7.993	<.001***
Median Income	39,410	437.20	90.152	<.001***
Ocean Proximity (Inland)	−39,240	2,237	-17.541	<.001***
Ocean Proximity (Island)	171,400	34,140	5.019	<.001***
Ocean Proximity (Near Bay)	−4,624	2,466	-1.875	.061
Ocean Proximity (Near Ocean)	5,559	2,019	2.753	.006**

*p<0.05, **p<0.01, *** p<0.001

The linear regression model with median_house_value as the dependent variable and longitude, latitude, housing_median_age, total_rooms, total bedrooms, population, households, median income, and ocean proximity as predictors showed significant effects for most variables.

Compared to houses near the ocean (<1H OCEAN), houses inland had significantly lower median values by $39,240 (p < .001), while houses on islands had significantly higher median values by $171,400 (p < .001). Houses near bays did not significantly differ from those near the ocean (p = .061). Increases in housing_median_age (B = 1,061, p < .001), total bedrooms (B = 88.33, p < .001), households (B = 75.63, p < .001), and median income (B = 39,410, p < .001) were associated with higher median house values. In contrast, increases in longitude (B = -26,010, p < .001), latitude (B = -24,590, p < .001), total_rooms (B = -6.424, p < .001), and population (B = -42.00, p < .001) were associated with lower median house values.

The Root Mean Square Error (RMSE) is a measure of the differences between values predicted by a model and the observed values. In the context of linear regression, a lower RMSE indicates that the model's predictions are closer to the actual observed values, suggesting better performance. In this case, the RMSE of 68124.94 suggests that, on average, the predicted median house values from the linear regression model are approximately $68124.94 away from the actual observed values. This value provides an indication of the overall accuracy of the model in predicting median house values based on the given predictor variables.

2. CART Regression

Here are 4 end nodes with their corresponding paths:

1. Node 460±3 (4%): This node represents instances where the median income is < 5.6, ocean proximity is INLAND, median income is < 3.4, and housing_median_age is >= 28.

2. Node 376±3 (4%): This node represents instances where the median income is < 5.6, ocean proximity is INLAND, median income is >= 3.4, median income is < 3.2, longitude is >= -118, and latitude is < 34.

3. Node 217±3 (14%): This node represents instances where the median income is < 5.6, ocean proximity is INLAND, median income is >= 3.2, and longitude is < -118.

4. Node 154±3 (10%): This node represents instances where the median income is < 5.6, ocean proximity is INLAND, median income is < 3.4, and median income is >= 3.4.

The variable importance measure in decision trees is typically based on the decrease in impurity or the decrease in node impurity that a variable provides when splitting the data. From the tree structure, it appears that median income is the most important variable, as it is used as the initial split and appears multiple times in the subsequent splits. The ocean proximity variable also seems to be highly important, as it is used as the second split after median income. Other important variables include housing_median_age, longitude, and latitude, as they are used to further split the data down the tree.

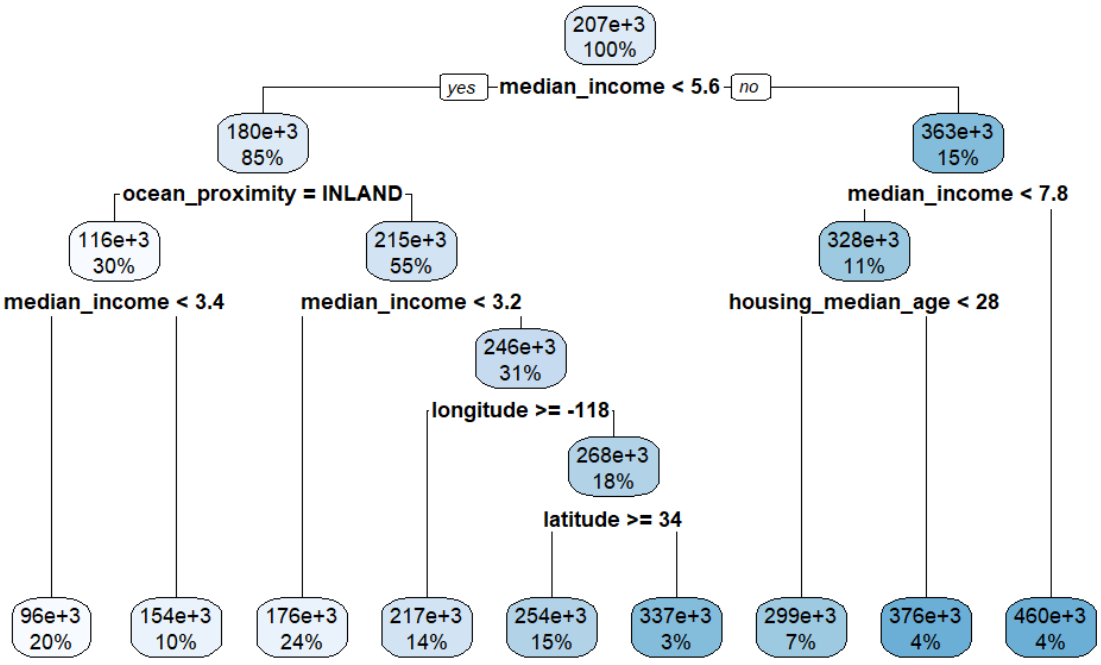

Figure 4: CART Regression

The results suggest that median income and ocean proximity are the primary drivers for predicting or classifying the target variable, followed by housing_median_age, longitude, and

latitude. However, it's important to note that variable importance can be influenced by the specific dataset, the target variable, and the algorithm's hyperparameters, among other factors. The RMSE value you provided for the CART model on the validation set is 74200.63. This means that, on average, the predicted median house values from the model differ from the actual values by approximately $74,200.63. It's essential to interpret this value in the context of the median house values in your dataset to gauge the performance of the model accurately.

Variable Importance

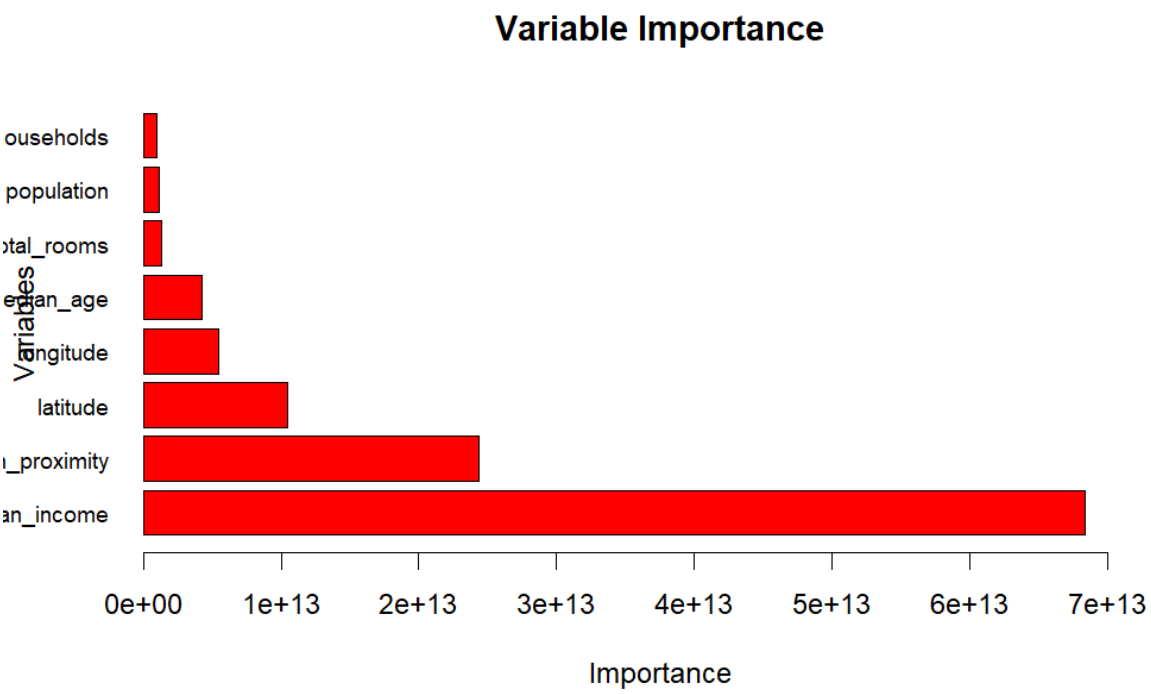

Figure 5: Variable importance for CART Regression

The variable importance values indicate the relative significance of each predictor variable in the regression model. A higher value suggests that the variable has a more substantial impact on predicting the outcome variable, while a lower value indicates less influence. In this case, the variable importance values reveal that "median income" and "ocean proximity" are the most critical predictors, with extremely high values compared to the other variables. These variables

contribute significantly to the model's ability to predict median house values. "Latitude" and "longitude" also have relatively high importance values, suggesting that location plays a significant role in determining house prices. On the other hand, variables such as "housing_median_age," "total_rooms," "population," and "households" have lower importance values, indicating comparatively lesser influence on the model's predictions. Overall, these variable importance values help prioritize variables for further analysis and informed decision-making in predictive modeling tasks.

3. Logistic Regression

Table 3: Coefficients of Regression

Variable	Coefficient	Std. Error	z value	p-value
(Intercept)	-136.1	7.79	-17.473	< 0.001 ***
Longitude	-1.544	0.094	-16.375	< 0.001 ***
Latitude	-1.572	0.102	-15.398	< 0.001 ***
Housing Median Age	0.03863	0.002789	13.854	< 0.001 ***
Total Rooms	0.0002191	0.000047	4.661	0.000***
Population	-0.002429	0.0001076	-22.568	< 0.001***
Households	0.00579	0.000283	20.462	< 0.001***
Median Income	1.075	0.02834	37.919	< 0.001 **
Ocean Proximity (Inland)	-0.3637	0.1409	-2.582	0.010*
Ocean Proximity (Island)	13.57	155	0.088	0.930
Ocean Proximity (Near Bay)	-0.007394	0.1079	-0.069	0.945
Ocean Proximity (Near Ocean)	-0.002345	0.08845	-0.027	0.979

Significance codes: $^{***}p < 0.001$, $^{**}p < 0.01$, $^{*}p < 0.05$

The logistic regression model examined the effects of longitude, latitude, housing_median_age, total_rooms, population, households, median income, and ocean proximity

on the binary outcome of flag_expensive. Compared to houses near the ocean (<1H OCEAN), houses inland had lower odds of being expensive (B = -0.3637, p = .01), while there was no significant difference for houses on islands or near bays. Increases in housing_median_age (B = 0.03863, p < .001), total_rooms (B = 0.0002191, p < .001), households (B = 0.00579, p < .001), and median income (B = 1.075, p < .001) were associated with higher odds of being expensive. In contrast, increases in longitude (B = -1.544, p < .001), latitude (B = -1.572, p < .001), and population (B = -0.002429, p < .001) were associated with lower odds of being expensive.

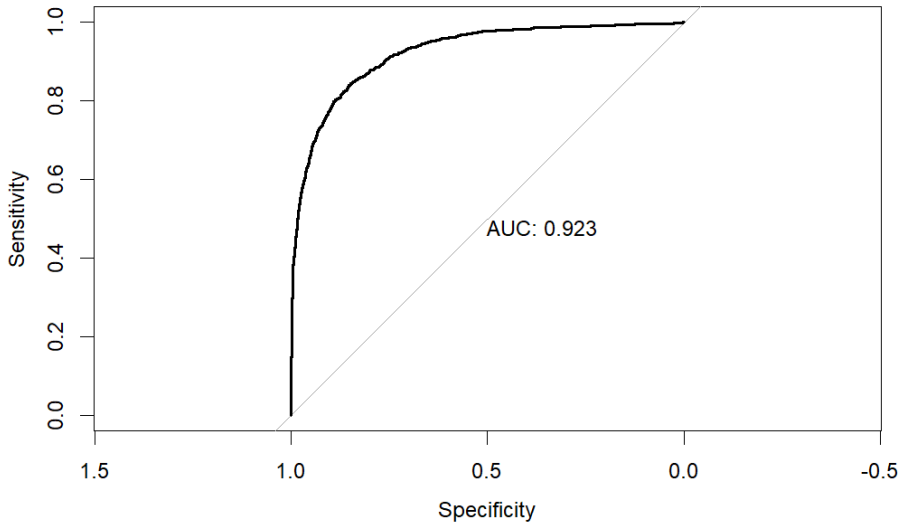

Figure 6: ROC

The confusion matrix provides a comprehensive overview of the classification model's performance. With an accuracy of 88.32%, the model correctly classified approximately 88.32% of all instances. The sensitivity, also known as the true positive rate, measures the proportion of actual positive cases correctly identified by the model, which stands at 67.60%. On the other hand, the specificity, or true negative rate, indicates the proportion of actual negative cases correctly identified, achieving a high value of 94.80%. This suggests that the model is effective in identifying both positive and negative instances. The AUROC (Area Under the Receiver

Operating Characteristic Curve) value of 0.9228 further validates the model's performance, indicating a strong ability to distinguish between positive and negative cases. Overall, the results indicate that the model strikes a good balance between sensitivity and specificity, rendering it a reliable choice for binary classification tasks.

CART Classification

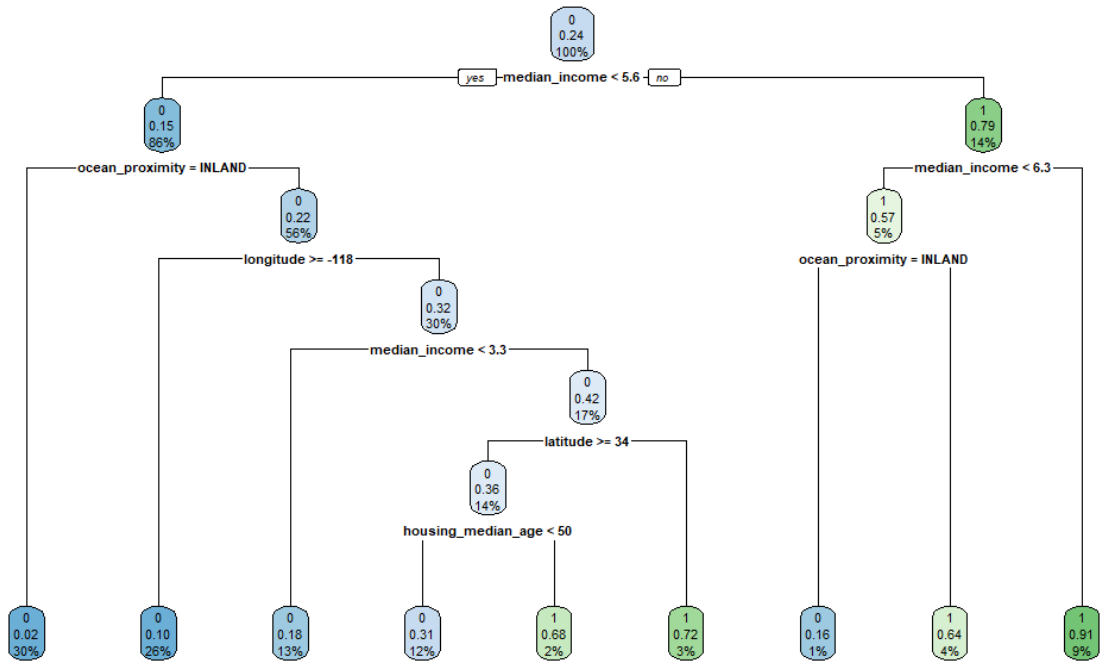

Figure 6: CART Classification

Here are 4 end nodes with their corresponding paths:

1. Node 1 (4%): This node represents instances where the median income is >= 6.3, median income is >= 8.3, ocean proximity is INLAND, and housing_median_age is >= 50.

2. Node 0.16 (1%): This node represents instances where the median income is < 5.6, ocean proximity is INLAND, longitude is >= -118, median income is < 3.3, latitude is >= 34, and housing_median_age is >= 50.

3. Node 0.12 (12%): This node represents instances where the median income is < 5.6, ocean proximity is INLAND, longitude is >= -118, median income is < 3.3, and latitude is < 34.

4. Node 0.02 (30%): This node represents instances where the median income is < 5.6, ocean proximity is INLAND, and longitude is < -118.

The variable importance measure in decision trees is typically based on the decrease in impurity or the decrease in node impurity that a variable provides when splitting the data. From the tree structure, it appears that median income is the most important variable, as it is used as the initial split and appears multiple times in the subsequent splits. The ocean proximity and longitude variables also appear to be highly important, as they are used to further split the data down the tree.

Other important variables include latitude and housing_median_age, as they are used for splitting the data at lower levels of the tree. However, it's worth noting that the tree structure suggests that total_rooms, population, and households are not being used for splitting the data in this specific model, which could indicate that they are relatively less important predictors in this context. The results suggest that median income, ocean proximity, longitude, latitude, and housing_median_age are the primary drivers for predicting or classifying the target variable in this decision tree model. However, as mentioned earlier, variable importance can be influenced by the specific dataset, the target variable, and the algorithm's hyperparameters, among other factors.

Variable Importance

The variable importance table reveals the factors contributing significantly to predicting the target outcome. Notably, median income emerges as the most influential predictor, with an importance score of 1452.17, indicating its substantial impact on the outcome. Latitude and ocean proximity follow closely, emphasizing the relevance of the geographic aspect, while longitude also plays a notable role. Housing_median_age, total_rooms, population, and households exhibit comparatively lower importance but still contribute meaningfully to the predictive model. This insight underscores the significance of socioeconomic and spatial

dimensions in understanding and forecasting the target variable, thereby informing strategic interventions and resource allocation in relevant domains.

Table 3: Variable Importance

Variable	**Importance**
median income	1452.17
latitude	267.15
ocean proximity	260.42
longitude	223.98
housing_median_age	68.70
total_rooms	16.35
population	9.70
households	8.24

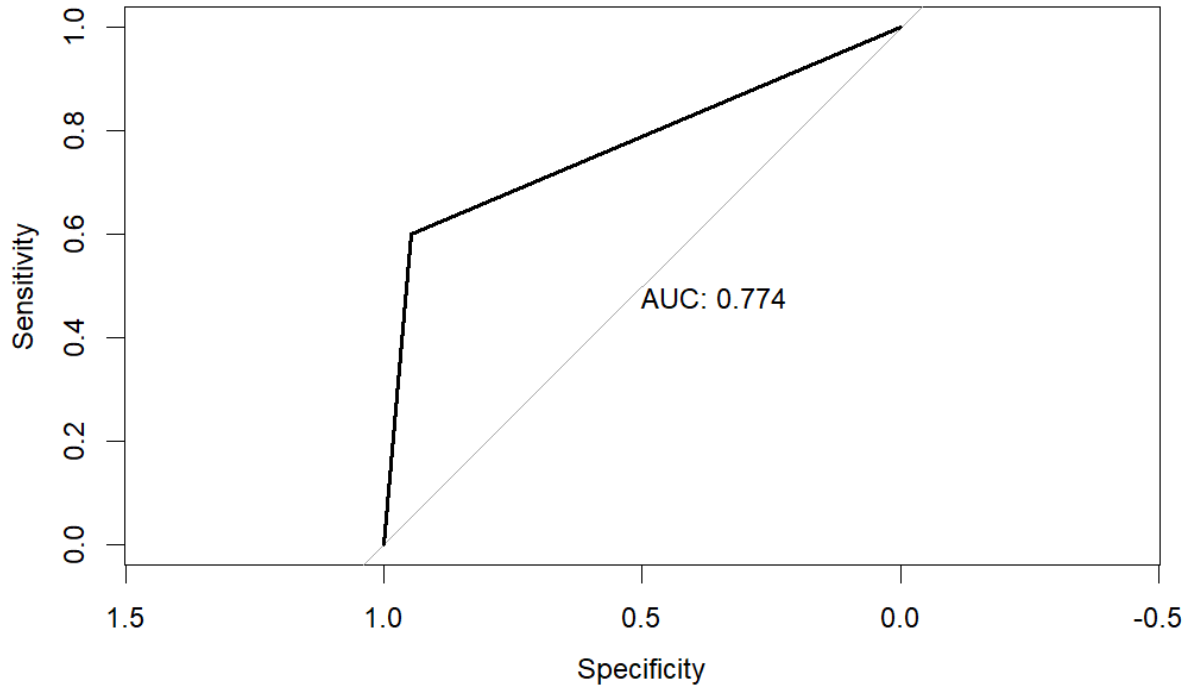

Figure 7: ROC for CART

The confusion matrix provides insights into the performance of the CART classification model. With an accuracy of 86.47%, the model demonstrates a commendable overall predictive ability. However, its sensitivity, measuring the proportion of correctly identified positive cases, is relatively modest at 60.07%. This suggests that the model may struggle to accurately classify instances belonging to the positive class. Conversely, the specificity, indicating the proportion of accurately identified negative cases, is high at 94.72%, reflecting the model's proficiency in correctly identifying instances of the negative class. The AUROC value of 0.77 indicates the model's overall discriminatory power, although it may benefit from further refinement, particularly in enhancing sensitivity to improve its ability to effectively detect positive cases.

Discussion

The linear regression analysis highlighted the pivotal role of several predictors in determining house prices, with median income emerging as a prominent factor, alongside housing median age and geographical variables such as longitude and latitude. These findings suggest that economic factors and spatial attributes have a significant impact on property values, aligning with conventional wisdom in real estate. Furthermore, the CART regression and classification models reinforced the importance of median income and geographic features in predicting both house values and the likelihood of properties being classified as expensive. This convergence of results across different modeling approaches corroborates the importance of these key variables in understanding the dynamics of the housing market. The emphasis on median income underscores its status as a fundamental determinant of housing affordability and demand, while the geographical factors highlight the enduring influence of location on property values.

The findings of the analysis have significant implications for various stakeholders in the real estate industry and beyond. Real estate agents and property investors can utilize predictive models to make informed decisions about property pricing, investment strategies, and market positioning. Understanding the factors driving house prices can also aid policymakers and urban planners in implementing effective housing policies, urban development plans, and initiatives for affordable housing (van Doorn et al., 2019). Moreover, financial institutions and mortgage lenders can leverage these models to assess property valuations, risk management, and lending practices, enhancing overall financial stability and market efficiency. Despite the valuable insights provided by our models, several limitations and challenges need to be acknowledged. One limitation is the reliance on historical data, which may not fully capture dynamic market trends and economic fluctuations. Furthermore, the models' predictive accuracy may be influenced by unobserved factors and external variables not included in the analysis, such as neighborhood characteristics, property amenities, and market sentiment.

To address these limitations and enhance the robustness of predictive models, future research could explore the incorporation of alternative data sources, such as geospatial data, social media sentiment analysis, and real-time market indicators. Furthermore, conducting longitudinal studies and incorporating time-series analysis techniques can capture temporal trends and seasonality in housing markets. Collaborations with industry partners and stakeholders can also facilitate access to proprietary data and domain expertise, fostering interdisciplinary research and knowledge exchange. Additionally, applying advanced machine learning algorithms, ensemble methods, and model ensembles can further improve predictive accuracy and generalization performance, paving the way for more effective decision-making and strategic planning in the real estate sector.

References

Hassanat, A. B., Tarawneh, A. S., Abed, S. S., Altarawneh, G. A., Alrashidi, M., & Alghamdi, M. (2022). RDPVR: Random data partitioning with voting rule for machine learning from class-imbalanced datasets. *Electronics*, *11*(2), 228. https://doi.org/10.3390/electronics11020228

Kern, C., Klausch, T., & Kreuter, F. (2019, April 11). *Tree-based machine learning methods for survey research*. Survey research methods. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC7425836/

Morozova, O., Levina, O., Uusküla, A., & Heimer, R. (2015). Comparison of subset selection methods in linear regression in the context of health-related quality of life and substance abuse in Russia. *BMC Medical Research Methodology*, *15*(1). https://doi.org/10.1186/s12874-015-0066-2

Shipe, M. E., Deppen, S. A., Farjah, F., & Grogan, E. L. (2019). Developing prediction models for clinical use using logistic regression: An overview. *Journal of Thoracic Disease*, *11*(S4). https://doi.org/10.21037/jtd.2019.01.25

Vaidynathan, D., Kayal, P., & Maiti, M. (2023). Effects of economic factors on median list and selling prices in the U.S. housing market. *Data Science and Management*, *6*(4), 199–207. https://doi.org/10.1016/j.dsm.2023.08.001

van Doorn, L., Arnold, A., & Rapoport, E. (2019). In the age of cities: The impact of urbanization on house prices and affordability. *Hot Property*, 3–13. https://doi.org/10.1007/978-3-030-11674-3_1

Zhao, C., & Liu, F. (2023). Impact of housing policies on the real estate market - systematic literature review. *Heliyon*, *9*(10). https://doi.org/10.1016/j.heliyon.2023.e20704