ISTINYE UNIVERSITY

AO5007 - DATA SCIENCE

DR . WADHAH ZEYAD TAREQ TAREQ

House Price Prediction Using ANN Model

Authored by :

MOHAMED SOULEIMAN CHEIKH AHMED - 2333265040

SAAD ELBOUKHALI - 2333285014

House price prediction using ANN Model : “The case of California districts, derived from the 1990 census, median house prices data”

SAAD ELBOUKHALI and MOHAMED SOULEIMAN CHEIKH AHMED

DR . WADHAH ZEYAD TAREQ TAREQ

DATA SCIENCE COURSE

[1] Istinye University, Istanbul, TURKEY

Abstract. The research methodology employed in this work is machine learning algorithms, which are used to create a housing price forecast model. This study examines information on the homes in a particular California district and provides a summary of those homes' characteristics based on data from the 1990 census in an effort to increase the accuracy of housing price predictions. Next, in order to help a home seller or real estate agent make more educated decisions based on home price estimation, we provide an enhanced housing price prediction model. Based on accuracy, the studies show that the linear regression method regularly surpasses the other models in the performance of predicting home prices.

Keywords: Housing price prediction , Machine learning algorithms

1 Introduction

Our project uses machine learning techniques on the California home Prices dataset to create home price prediction models.for a variety of stakeholders, including real estate investors, purchasers, financial institutions, and urban planners, housing price prediction is essential. Precise forecasting facilitates well-informed decision-making, risk mitigation, and resource distribution.
This study's main goal is to develop and assess machine learning models that, when applied to pertinent dataset features, can reliably forecast home prices..

2 Literature Review

2.1 Historical Overview of Housing Price Prediction

Given that it reflects both the goals of individuals and families as well as the general economic health of a region, the real estate business is essential to economic development and societal advancement. According to A. H. Maslow's : Undoubtedly these physiological needs are the most pre-potent of all needs [1] .

Previous studies in housing price prediction have utilized various methodologies, ranging from traditional regression models to advanced machine learning algorithms. Key findings emphasize the importance of features such as location, size, and socioeconomic factors in determining housing prices.Housing price prediction using machine learning algorithms has gained significant attention due to its potential impact on real estate markets and decision-making processes.

By leveraging historical property data—such as location, size, amenities, and market trends—ML models can learn complex patterns and relationships to make informed predictions about future property prices.

Various machine learning algorithms, including linear regression, decision trees, random forests, and gradient boosting, have been widely used for their flexibility and ability to capture complex relationships in the data.

2.2 Comparative Analysis of Machine Learning Algorithms

The selection of machine learning algorithms plays a crucial role in achieving accurate predictions in housing price prediction. Wu's project report on showcased the practical application of machine learning algorithms in predicting housing prices [2], demonstrating the effectiveness of Support Vector Regression. Similarly, Lu et al. explored the use of algorithms such as C4.5, RIPPER, Naïve Bayesian, and AdaBoost, emphasizing the significance of leveraging advanced algorithms to enhance prediction accuracy [3],

2.3 Data Preprocessing Techniques for Housing Price Prediction

Although it is often acknowledged that data preprocessing influences machine learning tasks' results, little research has been done to quantify this effect [5],. Researchers have been more aware of and focused on the small dataset problem in recent years. and [4], and the data preprocessing is a critical step in developing accurate predictive models. Studies have shown that dealing with missing values, cleaning outliers, and normalizing data can significantly impact the performance of machine learning models. For instance, some researchers highlighted the limitations of small datasets in housing price prediction models, emphasizing the importance of data size in enhancing prediction performance.

Loading and preprocessing data are essential steps in real estate price prediction. This involves handling missing values, encoding categorical variables, and scaling numerical features. Exploratory Data Analysis (EDA) is crucial for visualizing, understanding, and summarizing the dataset's main characteristics. EDA may include examining data distribution, identifying missing values, checking for outliers, and exploring relationships between variables. Popular Python libraries for EDA include Pandas, Matplotlib, and Seaborn.

2.5 Python for Housing Price Prediction

Python has emerged as a preferred tool for housing price prediction due to its versatile and extensive libraries. Libraries like Pandas provide efficient data manipulation, while NumPy offers numerical computation capabilities. Scikit-learn simplifies machine learning tasks with its user-friendly API, allowing easy implementation of regression algorithms like Linear Regression, Decision Trees, and Random Forests. Visualization libraries such as Matplotlib and Seaborn aid in data exploration and model evaluation. Python's rich ecosystem, supported by a robust community, provides numerous tutorials, examples, and open-source projects, making it easier for newcomers to get started. Its scalability and language compatibility further solidify its status as an ideal choice for home price prediction, enabling seamless integration with online applications and data pipelines.

2.4 Challenges and Future Directions

Despite advancements in machine learning for housing price prediction, several challenges remain, including the need for large and comprehensive datasets, the selection of appropriate algorithms, and the consideration of historical transaction data. Future research should focus on integrating insights from existing studies to enhance the predictive power of models and contribute to the advancement of housing economics through data-driven approaches.

III. Methodology

This section describes how we established the experiment to test the performance of machine learning algorithms for linear regression. We began by carefully selecting and preprocessing the dataset, ensuring its suitability for training and evaluation. Subsequently, we implemented various machine learning models, ranging from simple linear regression to more complex neural networks, to capture the intricate relationships within the data. Each step in the methodology was meticulously designed to ensure robustness and accuracy in our predictive models.

3.1. Data source and selection

The dataset utilized in this study originates from the second chapter of Aurélien Géron's renowned book, 'Hands-On Machine Learning with Scikit-Learn and TensorFlow'. This dataset is particularly valuable for its role as an introductory tool for implementing machine learning algorithms, as it necessitates fundamental data cleaning processes while offering a comprehensible list of variables. Moreover, its size strikes an optimal balance, being neither overly simplistic nor excessively complex.

Additionaly, the dataset comprises information gleaned from the 1990 California census, focusing on the houses situated within a specified California district and providing summary statistics based on the census data : these data were collected from "https://www.kaggle.com/datasets/camnugent/california-housing-prices"

This dataset it serves as a valuable resource for teaching the rudiments of machine learning. By engaging with this dataset, we can develop a foundational understanding of machine learning algorithms…

NO.	Variable name
1	A measure of how far west a house is; a higher value is farther west
2	A measure of how far north a house is; a higher value is farther north
3	Median age of a house within a block; a lower number is a newer building
4	Total number of rooms within a block
5	Total number of bedrooms within a block
6	Total number of people residing within a block
7	Total number of households within a block
8	Median income for households within a block (measured in tens of thousands of US Dollars)
9	Median house value for households within a block (measured in US Dollars)
10	Location of the house with respect to the ocean/sea

Table 1 : List of Variables in the Dataset

3.2. Data acquisition

In this section, we describe the process of acquiring and exploring the dataset used in our study...

We began with data exploration ,by loading the dataset and examining the first few records to understand its structure. The dataset consists of features such as longitude, latitude, housing median age, total rooms, total bedrooms, population, households, median income, median house value, and ocean proximity. This initial inspection helped us identify that the features have different scales. (Look at Table 2)

Table 2

	longitude	latitude	housing_median_age	total_rooms	total_bedrooms	population	households	median_income	median_house_value	ocean_proximity
0	-122.23	37.88	41.0	880.0	129.0	322.0	126.0	8.3252	452600.0	NEAR BAY
1	-122.22	37.86	21.0	7099.0	1106.0	2401.0	1138.0	8.3014	358500.0	NEAR BAY
2	-122.24	37.85	52.0	1467.0	190.0	496.0	177.0	7.2574	352100.0	NEAR BAY

Next, we utilized the info() function to get a summary of the dataset, revealing that the "total_bedrooms" column contains missing values. Additionally, the "ocean_proximity" column is categorical, which may require special handling since it is not directly suitable for regression models. Look at Fig. 1

```
 #   Column              Non-Null Count  Dtype
---  ------              --------------  -----
 0   longitude           20640 non-null  float64
 1   latitude            20640 non-null  float64
 2   housing_median_age  20640 non-null  float64
 3   total_rooms         20640 non-null  float64
 4   total_bedrooms      20433 non-null  float64
 5   population          20640 non-null  float64
 6   households          20640 non-null  float64
 7   median_income       20640 non-null  float64
 8   median_house_value  20640 non-null  float64
 9   ocean_proximity     20640 non-null  object
dtypes: float64(9), object(1)
memory usage: 1.6+ MB
```

Fig. 1 : Data columns

Additionaly , we handled missing values and we counted the number of missing values in each column to assess whether it is feasible to drop rows with missing values or if we need to use imputation techniques. This step is crucial for ensuring that our model receives complete data, as missing values can significantly impact model performance (Look at fig.2)

```
longitude             0
latitude              0
housing_median_age    0
total_rooms           0
total_bedrooms      207
population            0
households            0
median_income         0
median_house_value    0
ocean_proximity       0
dtype: int64
```

Fig.2 Counting numbers of missing values

In Table 3 , we generated a statistical summary of the dataset using the describe() function. This summary provided insights into the distribution of each feature, highlighting the large differences between the mean and maximum values for certain features. Identifying these variations helps in recognizing potential outliers that might affect model performance.

Table 3:

	longitude	latitude	housing_median_age	total_rooms	total_bedrooms	population	households	median_income	median_house_value
count	20640.000000	20640.000000	20640.000000	20640.000000	20433.000000	20640.000000	20640.000000	20640.000000	20640.000000
mean	-119.569704	35.631861	28.639486	2635.763081	537.870553	1425.476744	499.539680	3.870671	206855.816909
std	2.003532	2.135952	12.585558	2181.615252	421.385070	1132.462122	382.329753	1.899822	115395.615874
min	-124.350000	32.540000	1.000000	2.000000	1.000000	3.000000	1.000000	0.499900	14999.000000
25%	-121.800000	33.930000	18.000000	1447.750000	296.000000	787.000000	280.000000	2.563400	119600.000000
50%	-118.490000	34.260000	29.000000	2127.000000	435.000000	1166.000000	409.000000	3.534800	179700.000000
75%	-118.010000	37.710000	37.000000	3148.000000	647.000000	1725.000000	605.000000	4.743250	264725.000000
max	-114.310000	41.950000	52.000000	39320.000000	6445.000000	35682.000000	6082.000000	15.000100	500001.000000

To visualize the distribution of missing values, we created a heatmap, which clearly indicated the presence of missing data in the "total_bedrooms" column. So, we generated a heatmap of the feature correlations to understand the relationships between different variables , (look at Fig 3) .

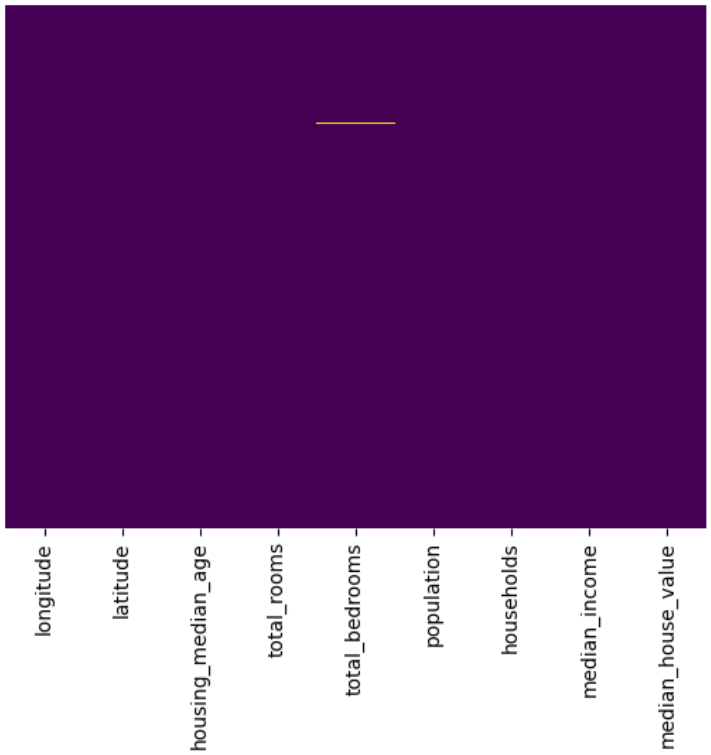


Fig.3 : Visualization of the distribution of missing values

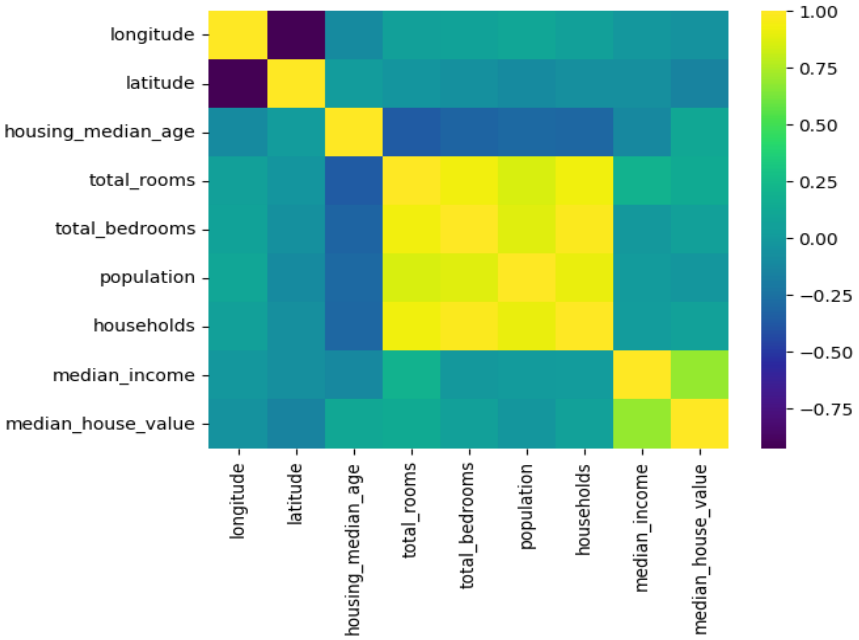


Fig. 4. Visualization of distrubutions of features

By looking at Fig.4 , we used a pair plot to examine the distributions and relationships between features. We used a pair plot to examine the distributions and relationships between features. This visualization provided a comprehensive overview of the dataset, revealing the presence of outliers in the "median_house_value" column, which made it non-normally distributed. This step is essential for understanding the data's overall structure and identifying potential issues that could impact the model's performance**…**Looking at the bottom right corner we can see that the median house value column contains outliers that make i non normally distributed (look at Fig 5)

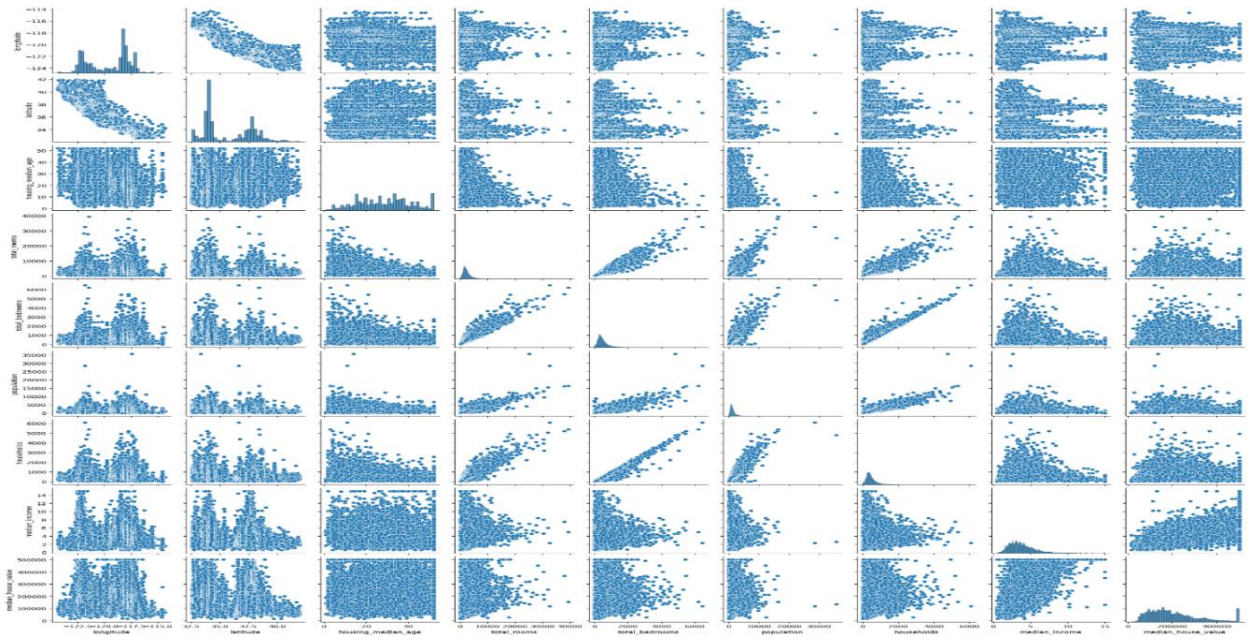


Fig. 5. Pair Plot of Housing Dataset Features

3.3 Data Preprocessing

- Handling missing values :

the dataset contained missing values in the "total_bedrooms" column. We addressed this issue by filling the missing values with the mean of the respective column. This approach ensures that we maintain the integrity of the data while avoiding the introduction of bias.

- Cleaning outliers : The outliers can significantly impact the performance of machine learning models. To address this, we used the Interquartile Range (IQR) method to identify and remove outliers. The IQR method involves calculating the first quartile (Q1) and the third quartile (Q3) and then determining the upper and lower bounds. Data points outside these bounds are considered outliers and were subsequently removed. This process helps in reducing the potential for overfitting and improves the robustness of the model.So we can said that the preprocessing steps have been completed , (look at Fig 6).

#	Column	Non-Null Count	Dtype
0	longitude	20640 non-null	float64
1	latitude	20640 non-null	float64
2	housing_median_age	20640 non-null	float64
3	total_rooms	20640 non-null	float64
4	total_bedrooms	20640 non-null	float64
5	population	20640 non-null	float64
6	households	20640 non-null	float64
7	median_income	20640 non-null	float64
8	median_house_value	20640 non-null	float64
9	ocean_proximity	20640 non-null	object

Fig.6 : Dataset Summary and Structure After Preprocessing

- Splitting the Dataset : To evaluate the performance of our models, we split the dataset into training and testing sets. The training set consists of 70% of the data, while the testing set comprises the remaining 30%. This split allows us to train our models on a substantial portion of the data while reserving a separate set for evaluation to ensure that our model's performance generalizes well to unseen data.

- Feature Scaling: Feature scaling is a crucial preprocessing step, especially for algorithms that are sensitive to the scale of data. We used the MinMaxScaler to normalize the features in the dataset. The scaler was trained only on the training data to prevent data leakage, ensuring that information from the testing set did not influence the training process. The scaled features were then used for both training and testing.

By carefully handling missing values, cleaning outliers, splitting the dataset, and scaling the features, we prepared the data for subsequent modeling and analysis.

3.4 Analysis Procedure

We go over the machine learning techniques we used on our dataset in this part. In this specific study…We will perform different experiments to see the optimal choices for hyperparameters and architecture and we will perform different experiments to see the optimal choices for hyperparameters and architecture…

- Model 1: is the first model we will start simple and increase the complexity in the other models to see if their is going to be any improvment .
- Model 2 : In this model we increased the complexity of the model to see if it was going to do better or worse and the changing the parameters in the next experiments .
- Model 3 :Is the third model for us .

- Model 4 : It’s the last model .

IV. Results

4.1 Model Performance

We evaluated the performance of four different models using key metrics: Mean Absolute Error (MAE), Mean Squared Error (MSE), and R^2 score. The results are summarized in Table 4 below.

Table 4 :

Model	MAE	MSE	R^2
Model 1	201,720.71	53,504,714,352.27	-3.076
Model 2	45,020.74	4,587,054,005.69	0.651
Model 3	38,601.44	3,451,255,368.55	0.737
Model 4	37,889.84	3,373,955,145.33	0.743

4.2 Model Comparison

Model 1

Model 1 shows the poorest performance with the highest MAE and MSE, and a negative R^2 score, indicating that the model does not fit the data well.

- Mean Absolute Error (MAE): 201,720.71
- Mean Squared Error (MSE): 53,504,714,352.27
- R^2 Score: -3.076

Model 2

Model 2 improves significantly, reducing both MAE and MSE, and achieves a positive R^2 score of 0.651, indicating a better fit.

- Mean Absolute Error (MAE): 45,020.74
- Mean Squared Error (MSE): 4,587,054,005.69
- R^2 Score: 0.651

Model 3

Model 3 further improves on Model 2, with lower MAE and MSE, and a higher R^2 score of 0.737.

- Mean Absolute Error (MAE): 38,601.44
- Mean Squared Error (MSE): 3,451,255,368.55
- R^2 Score: 0.737

Model 4

Model 4 shows the best performance among the four models, with the lowest MAE and MSE, and the highest R^2 score of 0.743, indicating it has the best predictive accuracy for the house price data.

- Mean Absolute Error (MAE): 37,889.84
- Mean Squared Error (MSE): 3,373,955,145.33
- R^2 Score: 0.743

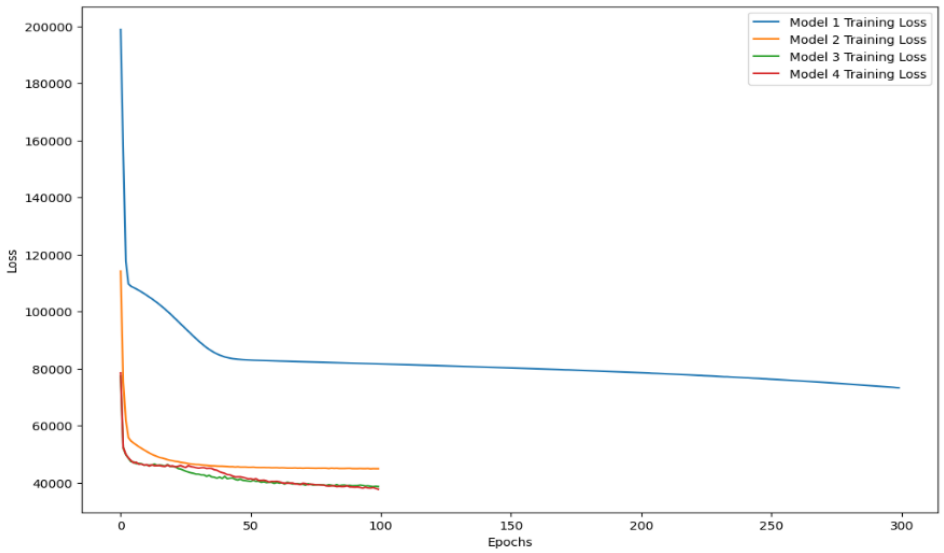
Based on these results, Model 4 is the most effective model for predicting house prices in our study, as it provides the most accurate predictions with the lowest error metrics and the highest R^2 score.

4.3 Visualization of Results

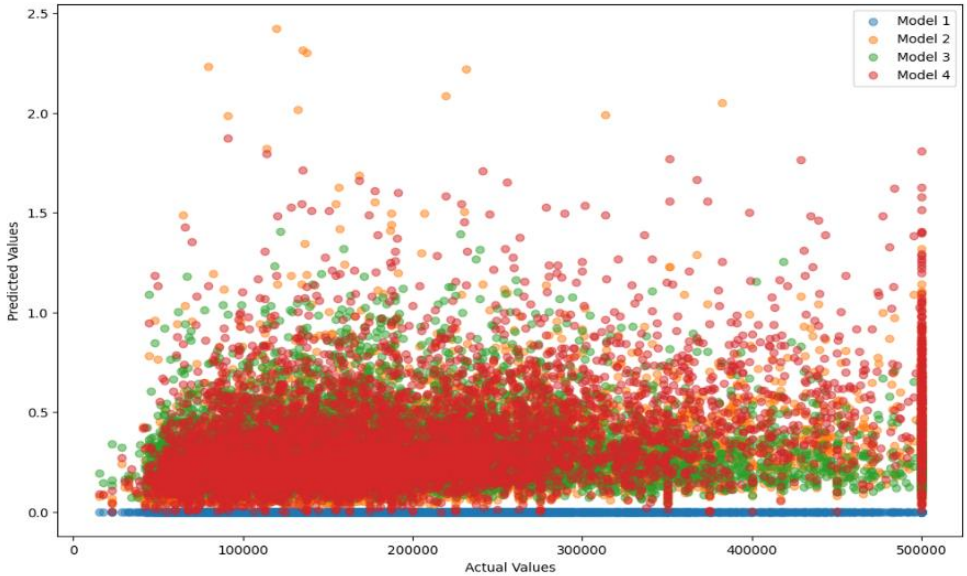
To better understand the performance of the models, we visualized the training loss curves, predicted vs. actual values, and the distribution of residuals.

The graph 1 shows the training loss over epochs for each model, and helps in understanding how well the models learn over time and if they suffer from overfitting or underfitting.

This scatter plot (Graph 2) compares the predicted house prices against the actual values for the test set, and helps in assessing how closely the predicted values align with the actual values.

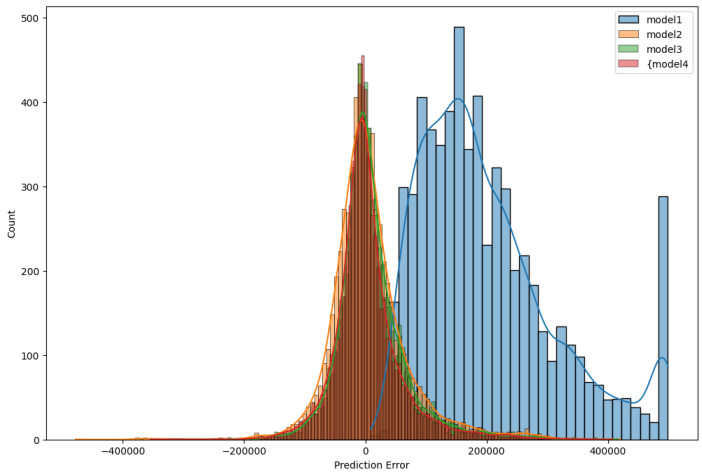


Graph 1 : Loss Curves

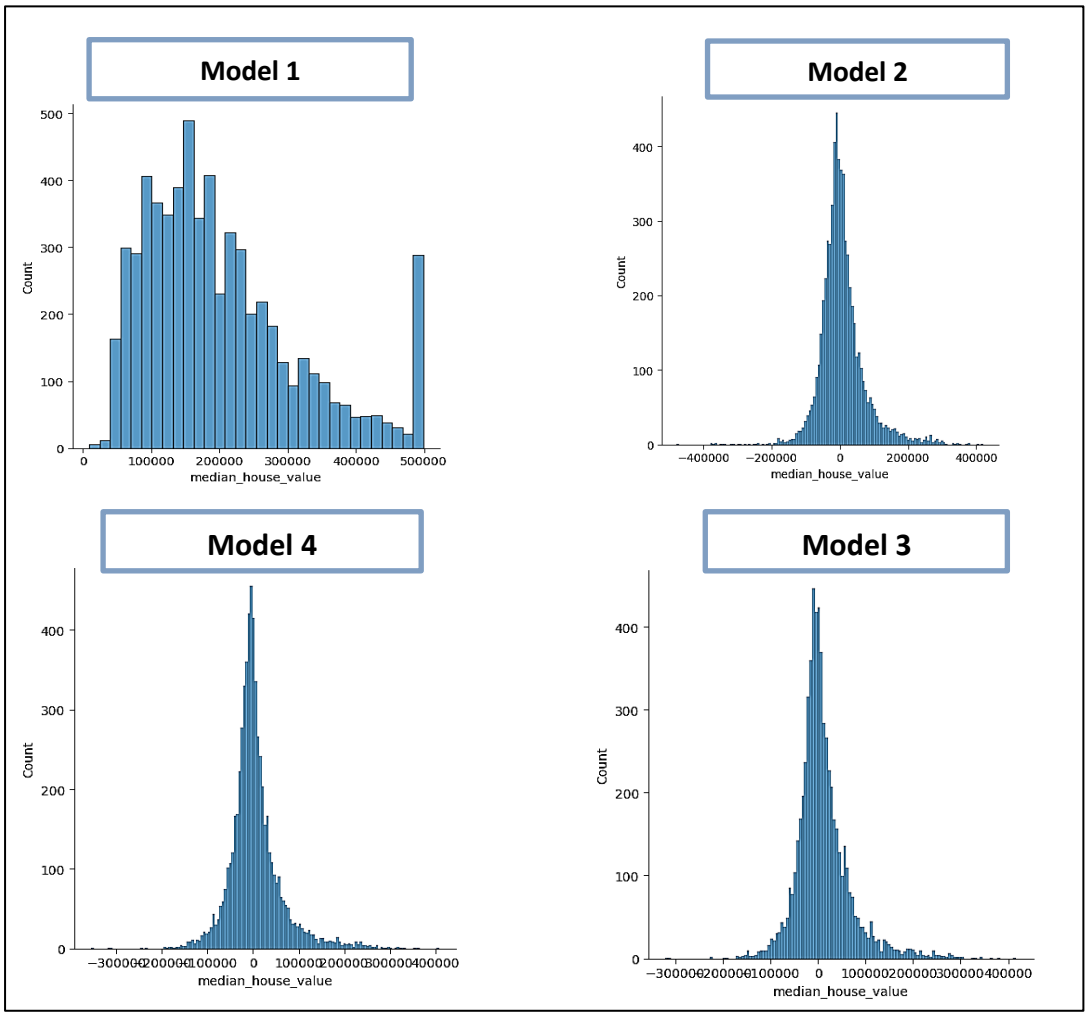


Graph 2 : Predicted vs Actual Values

This histogram (Graph 3) displays the distribution of prediction errors across all predictions , also examining the distribution of prediction errors provides insights into the model's overall performance and any patterns or outliers in prediction accuracy.



Graph 3: Prediction Error Distribution



Graph 4: The models Reduals

V . Discussion

We can see in the (graph 4) , that model 4 residuals are centered around zero more than the other model To evaluate the performance of different machine learning models for house price prediction, we implemented and tested four models: Model 1, Model 2, Model 3, and Model 4. The models were evaluated using key performance metrics such as Mean Absolute Error (MAE), Mean Squared Error (MSE), and R^2 score. The results are summarized in Table 1…

Additionally,we explored several models, including simple linear regression and more complex algorithms, and assessed their performance using key metrics such as Mean Absolute Error (MAE), Mean Squared Error (MSE), and R^2 score. Among the four models we tested, Model 4 demonstrated the highest predictive accuracy, with the lowest MAE and MSE, and the highest R^2 score. This indicates that Model 4 is the most effective in capturing the complex relationships within the data and providing reliable predictions.

Our findings emphasize the importance of using advanced machine learning techniques and comprehensive datasets in developing robust predictive models for housing prices. By improving the accuracy of housing price predictions, stakeholders such as real estate investors, purchasers, financial institutions, and urban planners can make more informed decisions, mitigate risks, and allocate resources more efficiently.

VI. Conclusion

Overall, this study contributes to the growing field of real estate analytics by demonstrating the potential of machine learning in improving housing price predictions and supporting data-driven decision-making processes. So,future research should focus on integrating more recent and diverse datasets, experimenting with additional machine learning algorithms, and incorporating other relevant features to further enhance the accuracy and applicability of housing price prediction models. Additionally, exploring the impact of external factors such as economic changes and policy developments could provide deeper insights into housing market dynamics.

VII. References

1. Maslow, A.H.: A theory of human motivation. Psychol. Rev. 50(4), 370-396 (1943). https://doi.org/10.1037/h0054346

2. Wu, J.Y.: Housing Price Prediction Using Supported Vector Regression. Master's Project, San Jose State University, Spring 2017. SJSU ScholarWorks. Available at: https://scholarworks.sjsu.edu/etd_projects/588/

3. Fahriah, K., Kamarudin, T., Triyono, T., Rizaldy, R.: The Adaboost Integration to the C4.5 Algorithm in Improving Study Interest Classification Accuracy. Indones. J. Inf. Syst. (IJIS) 6(2), 130 (2024). https://doi.org/10.24167/ijis.v6i2.313

4. Lateh, M.A., Hussain, M.A., Hashim, A.H.A., Khalid, M.: Handling a Small Dataset Problem in Prediction Model by Employing Artificial Data Generation Approach: A Review. J. Phys. Conf. Ser. 892, 012016 (2017). https://doi.org/10.1088/1742-6596/892/1/012016

5. García, S., Luengo, J., Herrera, F.: Towards Explaining the Effects of Data Preprocessing on Machine Learning. In: 2019 IEEE 35th International Conference on Data Engineering (ICDE), pp. 2126-2133. IEEE, New York (2019).https://doi.org/10.1109/ICDE.2019.00213