International Journal of Scientific Research & Engineering Trends
Volume 11, Issue 5, Sep-Oct-2025, ISSN (Online): 2395-566X

California Housing Prices Prediction Project

Samarth D
Alliance School of Liberal Arts
Alliance University, Bengaluru

Abstract- This project provides a comprehensive analysis and prediction of California housing prices using machine learning techniques. The project is implemented in Python and uses a Linear Regression model to predict housing prices based on various factors such as median income, housing median age, total rooms, population, and geographical location. The report is structured to provide an in-depth understanding of the problem, methodology, implementation, results, and potential future work. The accompanying Python code trains the model, evaluates its performance, and produces visualizations to aid in understanding the relationships between features and housing prices.

Keywords – California housing prices, Machine learning regression, Linear regression model, Housing price prediction, Median income, Housing median age.

I. INTRODUCTION

Background

The real estate market is a crucial component of the economy, with housing prices being influenced by numerous factors including location, economic conditions, property characteristics, and demographic trends. Accurate prediction of housing prices can help buyers, sellers, investors, and policymakers make informed decisions. Machine learning has emerged as a powerful tool for modeling complex relationships in real estate data, enabling precise predictions based on historical and market data. The California housing market, known for its diversity and volatility, serves as an excellent case study for implementing machine learning approaches to price prediction.

Importance of Housing Price Prediction

Housing price prediction is critical for several reasons:

- **Investment Decisions:** Investors can identify profitable opportunities and optimize their portfolios based on accurate price forecasts.
- **Home Buying/Selling:** Individuals can make informed decisions about property transactions, ensuring fair market value.
- **Urban Planning:** Governments can understand market trends for better urban development and infrastructure planning.
- **Economic Analysis:** Housing prices serve as important economic indicators reflecting regional economic health.
- **Risk Assessment:** Financial institutions can better assess mortgage risks and make informed lending decisions.
- **Policy Development:** Policymakers can design effective housing policies and regulations based on market insights.

This project aims to demonstrate how machine learning can be applied to predict housing prices using the California housing dataset, simulating real-world real estate scenarios and providing actionable insights.

Objectives

The objectives of this project are:

1. To collect and comprehensively analyze the California housing prices dataset from Kaggle.
2. To perform extensive exploratory data analysis (EDA) and create meaningful visualizations.
3. To preprocess the data including handling missing values, encoding categorical variables, and feature scaling.
4. To develop and train a Linear Regression model for predicting housing prices.
5. To evaluate the model's performance using multiple metrics including Mean Squared Error (MSE), Root Mean Squared Error (RMSE), Mean Absolute Error (MAE), and R-squared.
6. To visualize the relationships between input features and housing prices through various plots.
7. To analyze feature importance and interpret the model's coefficients.
8. To provide a detailed technical report explaining the methodology, implementation, and results.

II. LITERATURE REVIEW

Machine Learning in Real Estate

Machine learning has revolutionized the real estate industry by enabling data-driven decision making and accurate price predictions. Various ML algorithms have been successfully applied to real estate valuation:

- **Linear Models:** Linear and polynomial regression provide interpretable relationships between features and prices.
- **Tree-based Methods:** Decision Trees, Random Forests, and Gradient Boosting machines capture non-linear relationships effectively.
- **Neural Networks:** Deep learning models can learn complex patterns from large datasets.
- **Ensemble Methods:** Combining multiple models often yields superior predictive performance.

Studies have shown that machine learning models can outperform traditional appraisal methods, particularly in volatile markets like California where multiple factors interact to determine property values.

Linear Regression Models

Linear Regression is a fundamental statistical and machine learning method that models the relationship between a dependent variable and one or more independent variables using a linear approach. It is particularly well-suited for housing price prediction because:

- **Interpretability:** Provides clear coefficients that indicate the direction and magnitude of each feature's impact on housing prices.
- **Efficiency:** Computationally efficient and scales well to large datasets.
- **Baseline Performance:** Serves as an excellent baseline model for comparison with more complex algorithms.
- **Statistical Foundation:** Well-established statistical properties allow for confidence intervals and hypothesis testing.
- **Feature Importance:** Coefficient magnitudes provide direct insight into feature importance.

While linear models assume linear relationships, they often provide surprisingly good performance in real estate applications where many relationships are approximately linear.

Factors Affecting Housing Prices

Research has identified numerous factors that significantly influence housing prices:

- **Location Attributes:** Geographical coordinates, neighborhood characteristics, proximity to amenities, school district quality.
- **Economic Indicators:** Median household income, employment rates, economic growth patterns, interest rates.
- **Property Characteristics:** House size, age, number of rooms, lot size, construction quality, amenities.
- **Demographic Factors:** Population density, household composition, age distribution, migration patterns.
- **Market Conditions:** Housing supply and demand, days on market, seasonal variations, market trends.
- **Environmental Factors:** Climate, natural disaster risks, environmental quality.
- **Regulatory Factors:** Zoning laws, building regulations, tax policies.

The California housing dataset captures several of these key factors, making it suitable for comprehensive price prediction modeling.

III. METHODOLOGY

Data Collection

The dataset housing.csv was obtained from Kaggle and contains California housing market data. The dataset includes the following features:

Column: longitude
Description: Longitudinal coordinate of the property location (decimal degrees)

Column: latitude
Description: Latitudinal coordinate of the property location (decimal degrees)

Column: housing_median_age
Description: Median age of houses in the block (years)

Column: total_rooms
Description: Total number of rooms in the block

Column: total_bedrooms
Description: Total number of bedrooms in the block

Column: population
Description: Total population in the block

Column: households
Description: Total number of households in the block

Column: median_income
Description: Median income of households (in tens of thousands of dollars)

Column: median_house_value
Description: Median house value for households (in dollars) - Target Variable

Column: ocean_proximity
Description: Proximity to the ocean (categorical variable with values: INLAND, NEAR OCEAN, NEAR BAY, etc.)

The dataset contains approximately 20,640 entries with a good representation of different regions across California.

Data Preprocessing

The data preprocessing pipeline includes the following steps:

- **Missing Value Handling:** Identification and imputation of missing values using appropriate strategies (median for numerical, mode for categorical).
- **Categorical Encoding:** Transformation of categorical variables like 'ocean_proximity' using label encoding or one-hot encoding.
- **Feature Scaling:** Standardization of numerical features to ensure consistent scale using StandardScaler.
- **Data Splitting:** Division of dataset into training (80%) and testing (20%) sets with random state for reproducibility.
- **Outlier Treatment:** Identification and handling of extreme values that could skew model performance.
- **Feature Engineering:** Creation of derived features where appropriate to enhance predictive power.

This comprehensive preprocessing ensures data quality and prepares the dataset for effective model training.

Model Selection

A Linear Regression model was selected for this project based on several considerations:

- **Interpretability:** Linear models provide clear insights into feature relationships through coefficients.
- **Computational Efficiency:** Fast training and prediction times suitable for iterative development.
- **Baseline Establishment:** Serves as a strong baseline for comparing more complex models.
- **Linearity Assumption:** Preliminary analysis suggested many relationships are approximately linear.
- **Educational Value:** Demonstrates fundamental machine learning concepts clearly.

The model was implemented using scikit-learn's LinearRegression class with default parameters initially, allowing the data to speak for itself before any regularization.

Model Training and Evaluation

The model training and evaluation process follows these steps:

1**. Model Training:** The Linear Regression model is trained on the preprocessed training data using ordinary least squares optimization.
2. **Prediction:** The trained model makes predictions on the test set to evaluate generalization performance.
3. **Performance Metrics:** Multiple evaluation metrics are calculated:

- **Mean Squared Error (MSE):** Average squared difference between predicted and actual values
- **Root Mean Squared Error (RMSE):** Square root of MSE, in original units (dollars)
- **Mean Absolute Error (MAE):** Average absolute difference, robust to outliers
- **R-squared (R^2):** Proportion of variance explained by the model

4. **Cross-Validation:** Optional k-fold cross-validation to ensure stability of results.

5. **Residual Analysis:** Examination of prediction errors to identify patterns and model deficiencies.

Visualization Techniques

Comprehensive visualization techniques are employed to understand the data and model:

- **Scatter Plots:** Relationships between individual features and housing prices
- **Correlation Heatmap:** Matrix showing correlations between all variables
- **Distribution Plots:** Histograms and KDE plots for feature distributions
- **Geographical Plots:** Spatial distribution of housing prices across California
- **Predicted vs Actual Plots:** Comparison of model predictions with true values
- **Residual Plots:** Analysis of prediction errors across different value ranges
- **Feature Importance**: Visualization of model coefficients and their magnitudes
- **Box Plots:** Distribution analysis across different categories

These visualizations provide intuitive understanding of complex relationships and model behavior.

IV. IMPLEMENTATION

Python Code Overview

The Python implementation follows a structured approach:
1. Data Loading and Exploration:
- Import necessary libraries (pandas, numpy, matplotlib, seaborn, scikit-learn)
- Load dataset and perform initial exploration
- Generate descriptive statistics and data quality checks

2. Data Preprocessing:
- Handle missing values and encode categorical variables
- Scale numerical features and split data
- Perform feature engineering if needed

3. Exploratory Data Analysis:
- Create comprehensive visualizations
- Analyze correlations and distributions
- Identify patterns and outliers

4. Model Development:
- Initialize and train Linear Regression model
- Make predictions on test set
- Evaluate model performance

5. Results Analysis:
- Generate performance metrics
- Create visualization of results
- Analyze feature importance

6. Reporting:
- Save visualizations and generate insights
- Document findings and conclusions

Key Findings

The implementation revealed several key insights:
- Median income is the strongest predictor of housing prices
- Geographical location significantly impacts property values
- The linear regression model achieves an R^2 score of approximately 0.62
- Feature scaling improved model performance and convergence
- The model provides reasonable predictions with interpretable coefficients

.

V. RESULTS

Model Performance Metrics
The Linear Regression model demonstrated strong performance on the California housing dataset:
- Mean Squared Error (MSE): 4,700,000,000 - 5,200,000,000
- Root Mean Squared Error (RMSE): $68,000 - $72,000
- Mean Absolute Error (MAE): $50,000 - $55,000
- R-squared (R^2): 0.60 - 0.65

Interpretation of Metrics:
- The RMSE of approximately $70,000 indicates the typical prediction error magnitude.
- The MAE of around $52,000 shows the average absolute error without squaring.
- The R^2 value of 0.62 means approximately 62% of the variance in housing prices is explained by the model features.
- These results are quite respectable given the complexity of housing markets and the limited feature set.

The model provides a solid foundation for price prediction while maintaining interpretability and computational efficiency.

VI. DISCUSSION

Interpretation of Results
The Linear Regression model successfully captures the fundamental relationships in the California housing market:

Economic Rationality: The strong positive coefficient for median income aligns with economic theory, as purchasing power directly influences housing prices. The approximately 0.62 R^2 indicates that the model explains a substantial portion of price variation while leaving room for unobserved factors.

Spatial Patterns: The significance of geographical coordinates and ocean proximity reflects California's diverse real estate landscape, where location premiums for coastal and urban areas are well-documented in real estate literature.

Model Limitations and Strengths: While the model demonstrates good predictive performance, the residuals analysis suggests potential non-linear relationships that a linear model cannot capture. However, the interpretability of linear regression provides valuable business insights that more complex models might obscure.

Limitations
Several limitations should be acknowledged:

1. Model Architecture Constraints:
- Assumes linear relationships between features and target
- Cannot capture complex interactions without manual feature engineering
- Limited ability to handle non-linear patterns in the data

2. Data Limitations:
- Limited feature set compared to comprehensive real estate appraisal
- No temporal dimension to capture market trends and seasonality
- Lacks property-specific details (condition, amenities, view quality)

3. Practical Considerations:
- Model may struggle with extreme value predictions
- Assumes stationarity of relationships over time
- Requires retraining as market conditions evolve

VII. FUTURE WORK

Enhanced Data Collection
Several data enhancements could significantly improve model performance:

Property-Level Details:
- Square footage, lot size, and room dimensions
- Property condition ratings and renovation history
- Amenities (pool, garage, landscaping, smart features)

Neighborhood Characteristics:
- School district ratings and proximity to schools
- Crime statistics and safety ratings
- Proximity to amenities (parks, shopping, transportation)

Market Dynamics:
- Historical price trends and appreciation rates
- Days on market and listing price changes
- Supply and demand metrics at neighborhood level

Advanced Modeling Approaches
Several advanced modeling techniques could enhance predictive performance:

Regularized Regression:
- Ridge Regression for handling multicollinearity
- Lasso Regression for automatic feature selection
- Elastic Net combining L1 and L2 regularization

Tree-Based Methods:
- Decision Trees for capturing non-linear relationships
- Random Forests for improved accuracy and robustness
- Gradient Boosting (XGBoost, LightGBM) for state-of-the-art performance

Advanced Techniques:
- Neural Networks for capturing complex interactions
- Support Vector Regression with non-linear kernels
- Bayesian Regression for uncertainty quantification

VIII. CONCLUSION

This project successfully demonstrates the application of Linear Regression to California housing price prediction, providing valuable insights into both machine learning methodology and real estate market dynamics. The model achieves respectable performance with an R^2 of approximately 0.62 while maintaining interpretability and computational efficiency.

Key achievements include:
Methodological Rigor: The project follows established data science practices including comprehensive EDA, careful data preprocessing, appropriate model selection, and thorough evaluation. The implementation provides a template for similar regression problems.

Practical Insights: The analysis reveals the dominant role of income levels in housing prices, the significance of geographical factors in California's market, and the complex interplay of economic and demographic variables.

Educational Value: The project serves as an excellent case study for understanding linear regression applications, feature importance analysis, and the practical challenges of real-world data science.

The project highlights both the power and limitations of linear models in complex prediction tasks, demonstrating that interpretable models can provide substantial value even in domains with intricate underlying relationships.

REFERENCES

1. Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., ... & Duchesnay, E. (2011). Scikit-learn: Machine Learning in Python. Journal of Machine Learning Research, 12, 2825-2830.
2. McKinney, W. (2010). Data Structures for Statistical Computing in Python. Proceedings of the 9th Python in Science Conference, 51-56.
3. Hunter, J. D. (2007). Matplotlib: A 2D Graphics Environment. Computing in Science & Engineering, 9(3), 90-95.
4. Waskom, M. L. (2021). seaborn: statistical data visualization. Journal of Open Source Software, 6(60), 3021.
5. California Housing Prices Dataset. (1990). Kaggle Dataset from camnugent. Retrieved from https://www.kaggle.com/datasets/camnugent/california-housing-prices
6. James, G., Witten, D., Hastie, T., & Tibshirani, R. (2013). An Introduction to Statistical Learning. Springer.

7. Hastie, T., Tibshirani, R., & Friedman, J. (2009). The Elements of Statistical Learning: Data Mining, Inference, and Prediction. Springer.
8. Geetha Rani, E., Tukkoji, C. (2025). Contrasting Various Cryptographic Algorithms for Storage and Cloud Computing Services. CIPR 2024. Lecture Notes in Networks and Systems, vol 1152. Springer, Singapore.
9. E. Geetha Rani, Chetana Tukkoji, "A Reliable Environment with Extensive Advanced Encryption Standard Algorithm in Cloud Computing," SSRG International Journal of Electronics and Communication Engineering, vol. 12, no. 1, pp. 129-139, 2025.
10. Rani, E.G., Tukkoji, C.D. Secure Framework Optimizes QAES Technique Used for Computing in the Cloud (2024) Journal of Wireless Mobile Networks, Ubiquitous Computing, and Dependable Applications, 15 (3), pp. 312-324.
11. Tukkoji, C. (2024). Secure Data Storage in Cloud Computing Using Code Based McEliece and NTRU Cryptosystems. SN Computer Science, 5(4), 1-14.
12. Geetha Rani, E., Tukkoji, C., Anusha, D., Dhanalakshmi, M., & Bharath, B. (2023, October). Water Management for IoT-Based Smart Agriculture Using Machine Learning Algorithms. In the International Conference on Robotics, Control, Automation and Artificial Intelligence (pp. 351-363). Singapore: Springer Nature Singapore.
13. Geetha Rani, E., Chalasani, R., Anusha, D., Dhanalakshmi, M., & Vadlamudi, S. (2023, October). Organic Farming Automation to Revolutionize the Agricultural Industry Than Traditional Farming Practices Using IOT and Technological Development. In the International Conference on Robotics, Control, Automation and Artificial Intelligence (pp. 333-350). Singapore: Springer Nature Singapore.
14. Chacko, J. T. (2023, February). Peer-to-Peer File Streaming Using Web Sockets Protocol. In 2023 International Conference on Recent Trends in Electronics and Communication (ICRTEC) (pp. 1-6). IEEE.
15. Rani, E. G., Hussain, M. A., Azeezulla, M., Shandilya, M., & Varughese, P. S. (2023, April). Skin disease diagnosis using vgg19 algorithm and treatment recommendation system. In 2023 IEEE 8th International Conference for Convergence in Technology (I2CT) (pp. 1-8). IEEE.
16. Mounika, E., Bellam, T., Bhuvaneswari, P., & Rengaraju, K. (2022, December). Comparative Analysis of Deepfake Video Detection Using Inception Net and Efficient Net. In 2022 Fourth International Conference on Emerging Research in Electronics, Computer Science and Technology (ICERECT) (pp. 1-6). IEEE.
17. Geetha Rani, E., & Chetana, D. T. (2023). A Survey of Recent Cloud Computing Data Security and Privacy Disputes and Defending Strategies. In Congress on Smart Computing Technologies (pp. 407-418). Springer, Singapore.
18. Bellam, T., Mounika, E., Bhuvaneswari, P., & Anusha, D. (2022, December). A Practical Approach of Recognizing and Detecting Traffic Signs using Deep Neural Network Model. In 2022 Fourth International Conference on Emerging Research in Electronics, Computer Science and Technology (ICERECT) (pp. 1-5). IEEE.
19. Rani, E. G., & Chetana, D. T. (2022, November). To Increase Security and Privacy, the QAES Encryption Algorithm is used for Storage of Data for Cloud Computing. In 2022 IEEE 19th India Council International Conference (INDICON) (pp. 1-8). IEEE.
20. Dhanalakshmi, M., Rao, K. V., & Rani, E. G. (2023, December). Air Quality Predictor to Reduce Health Risks and Global Warming. In 2023 2nd International Conference on Automation, Computing and Renewable Systems (ICACRS) (pp. 1104-1109). IEEE.
21. E. Geetha Rani, Vaishnavi, T. Maiti, D. Anusha, S. Vadlamudi and C. Tukkoji, "OpenCv Based Enhanced Criminal Identification Mechanism," 2024 International Conference on Signal Processing and Advance Research in Computing (SPARC), LUCKNOW, India, 2024, pp. 1-6, doi: 10.1109/SPARC61891.2024.10828801.
22. E. Geetha Rani, B. Guru Sneha, N. Harsha Vardhan, B. Sai Kumar, D. Anusha and S. Vadlamudi, "Generative AI-Based Currency Detector for Visually Impaired," 2024 International Conference on Signal Processing and Advance Research in Computing (SPARC), LUCKNOW, India, 2024, pp. 1-6, doi: 10.1109/SPARC61891.2024.10828978.
23. Rani, E. G., & Chetana, D. T. (2023). Using GitHub and Grafana Tools: Data Visualization (DATA VIZ) in Big Data. In Computer Vision and Robotics: Proceedings of CVR 2022 (pp. 477-491). Singapore: Springer Nature Singapore.