California Housing Market Analysis

Background

- California's real estate market has experienced significant fluctuations over the years, with periods of rapid growth followed by downturns and periods of stability.
- Understanding the dynamics of the real estate market is crucial due to its impact on the overall economy, housing affordability, and individuals' financial well-being.
- By conducting regression analysis on California's real estate market data, we can identify key variables driving housing prices, predict future trends, and inform policy decisions aimed at promoting housing affordability and stability.

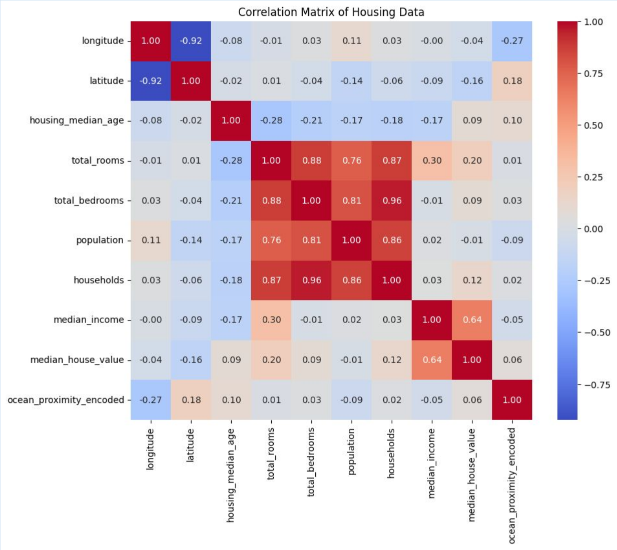
Dataset Description

The data contains information from the 1990 California census. So although it may not help us with predicting current housing prices like Zillow Zestimate, it does provide an accessible dataset to learn how predictive analysis works

Variables:

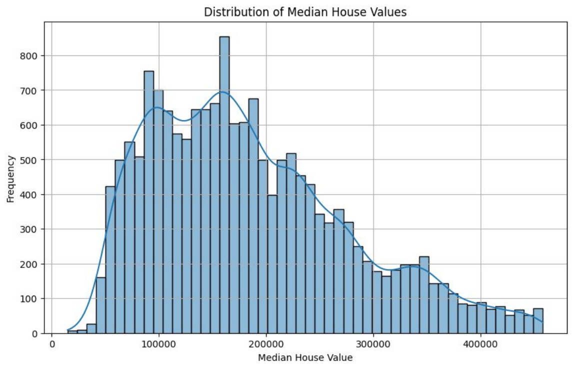
Longitude, latitude, housing_median_age, total_rooms, total_bedrooms, population, households, median_income, median_house_value, ocean_proximity

EDA

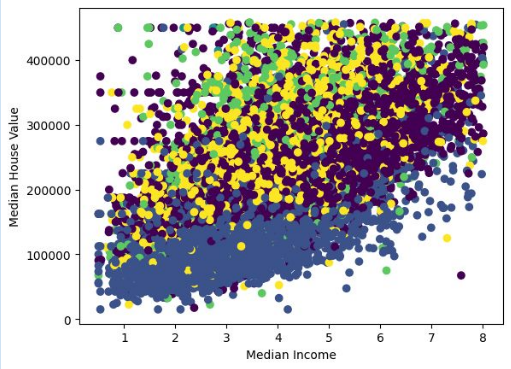


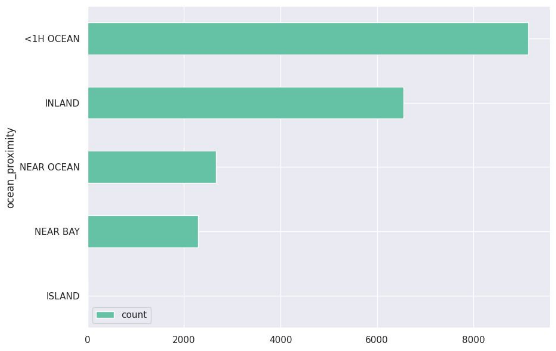
Correlation Matrix of Housing Data

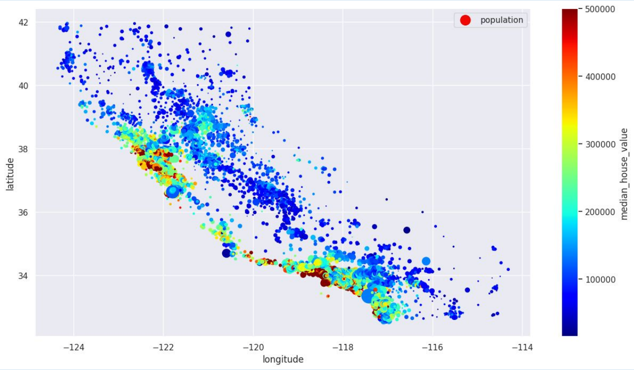
	longitude	latitude	housing_median_age	total_rooms	total_bedrooms	population	households	median_income	median_house_value	ocean_proximity_encoded
longitude	1.00	-0.92	-0.08	-0.01	0.03	0.11	0.03	-0.00	-0.04	-0.27
latitude	-0.92	1.00	-0.02	0.01	-0.04	-0.14	-0.06	-0.09	-0.16	0.18
housing_median_age	-0.08	-0.02	1.00	-0.28	-0.21	-0.17	-0.18	-0.17	0.09	0.10
total_rooms	-0.01	0.01	-0.28	1.00	0.88	0.76	0.87	0.30	0.20	0.01
total_bedrooms	0.03	-0.04	-0.21	0.88	1.00	0.81	0.96	-0.01	0.09	0.03
population	0.11	-0.14	-0.17	0.76	0.81	1.00	0.86	0.02	-0.01	-0.09
households	0.03	-0.06	-0.18	0.87	0.96	0.86	1.00	0.03	0.12	0.02
median_income	-0.00	-0.09	-0.17	0.30	-0.01	0.02	0.03	1.00	0.64	-0.05
median_house_value	-0.04	-0.16	0.09	0.20	0.09	-0.01	0.12	0.64	1.00	0.06
ocean_proximity_encoded	-0.27	0.18	0.10	0.01	0.03	-0.09	0.02	-0.05	0.06	1.00



Distribution of Median House Values



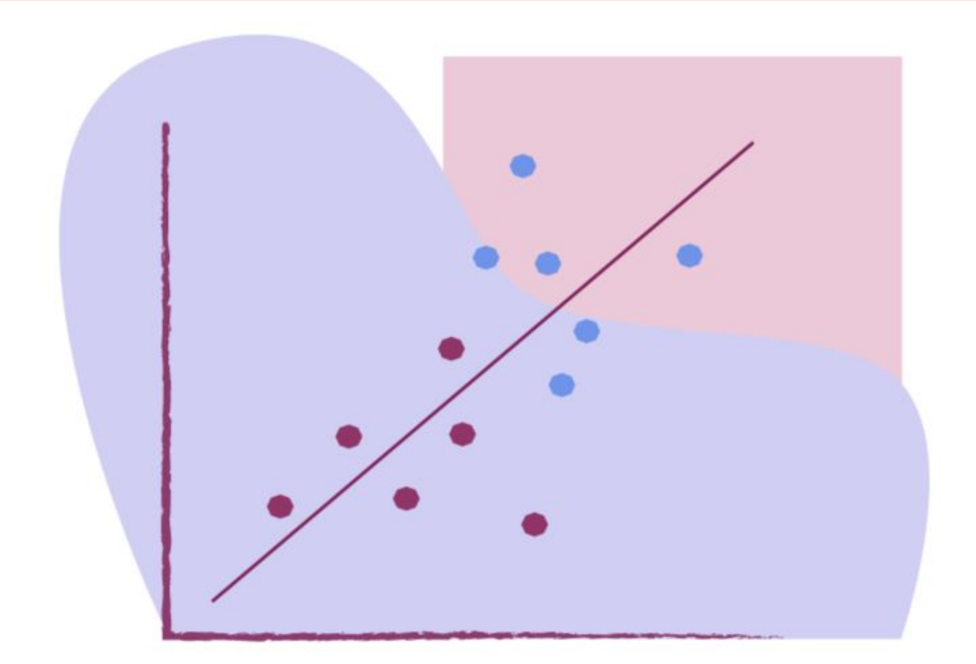




Linear Regression

Craig Cultice

Kaushik Pendiyala

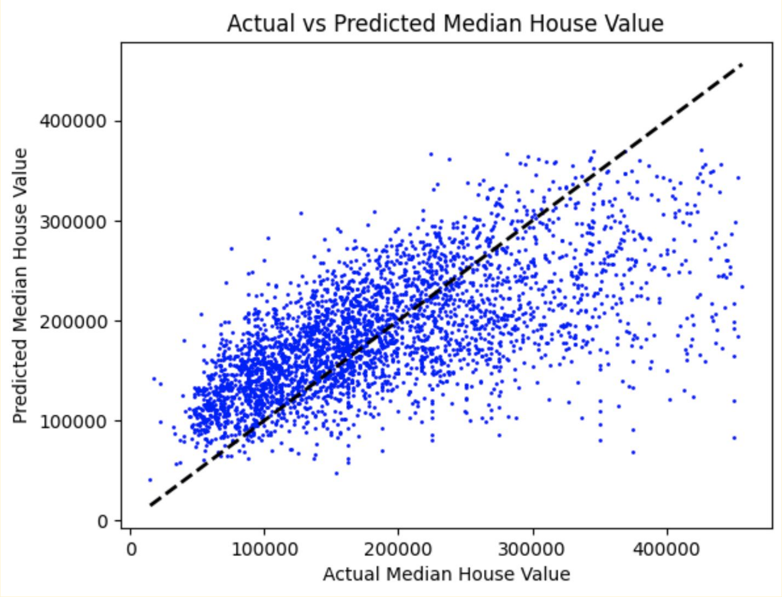


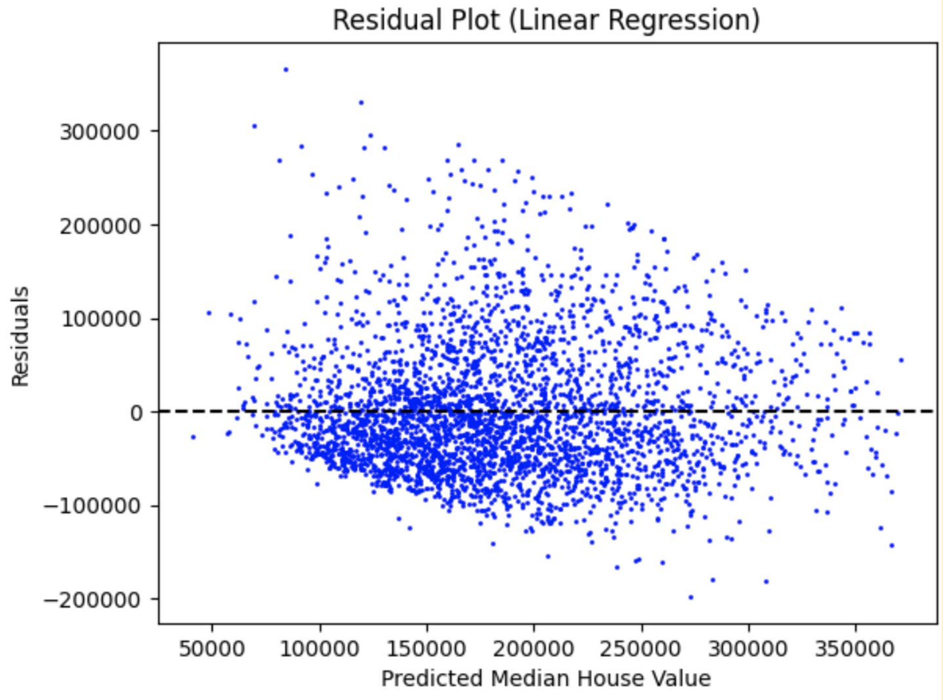
Goal Of Linear Regression

- The goal is to estimate the linear relationship between an independent variable and one or more explanatory variables
- Through the splitting of the data into testing and training data sets, we created this model in order to find the optimal regression line that fit our observed housing data.

Linear Regression + Residual Plot

- Positive correlation (0.412 R squared)
- Points are closely clustered around line of perfect fit at lower house values
- Increasing variance with predicted house value





Interestingly, the larger the actual median house value was, the worse the linear regression model performed (underpredicted)

Evaluation Metrics

Mean Squared Error = 4898991080.538166

Root Mean Squared Error = 69992.7930576

R^2 = 0.416514286111604

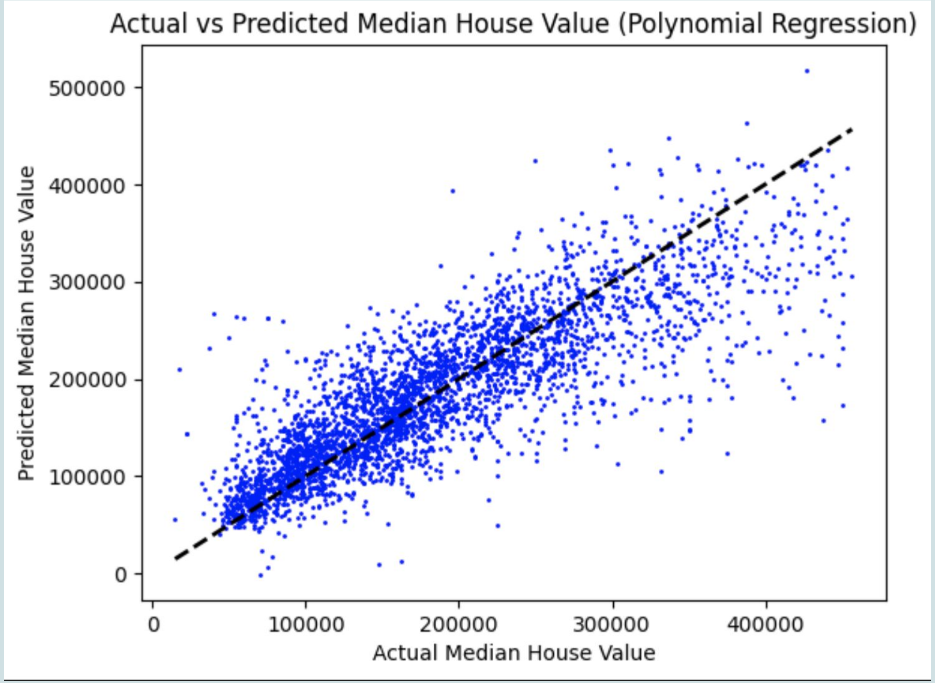
Normalized Absolute Error (NAE) = 0.12138889412287455

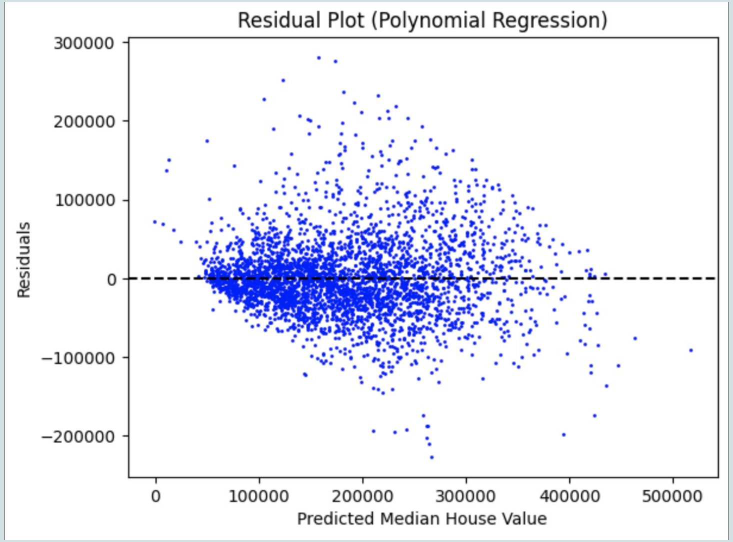
Explained Variance Score = 0.416517832799261

Median Absolute Error = 44842.981385752035

Polynomial Regression + Residual Plot

- We also experimented with polynomial regression
  - We determined degree 3 yielded the best polynomial graph





Evaluation Metrics

Mean Squared Error = 2,544,375,312.0174394

Root Mean Squared Error = 50441.8012367

R-squared (Polynomial Regression): 0.6969566547630011

Normalized Absolute Error (NAE) (Polynomial Regression): 0.08125612346000907

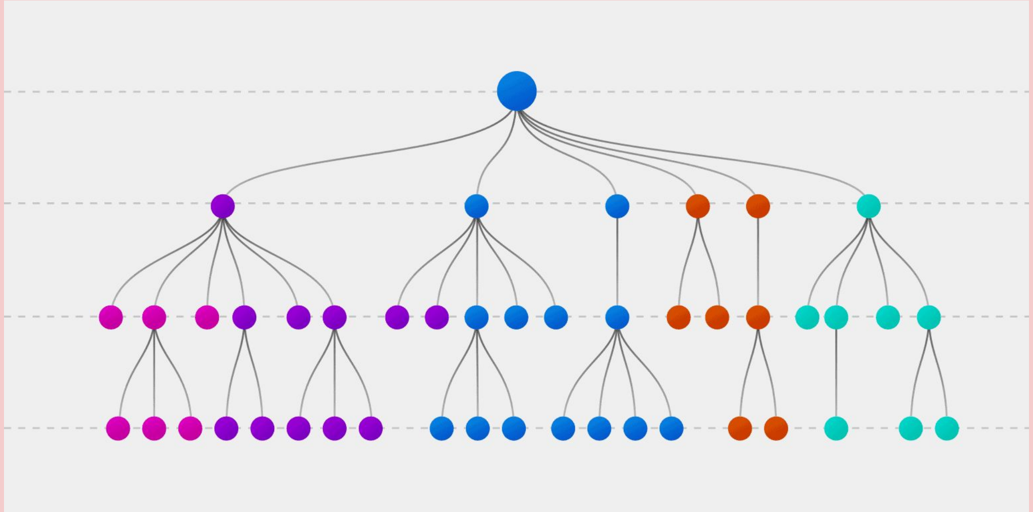
Explained Variance Score (Polynomial Regression): 0.6969940229931412

Median Absolute Error (Polynomial Regression): 25357.59298658371

Decision Tree

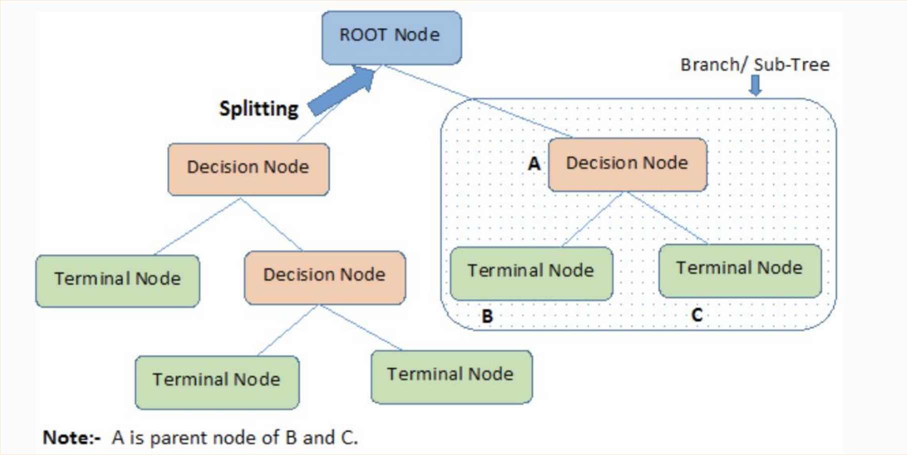
Zeeva Chaver

Xuyu Song



Goal Of Decision Tree

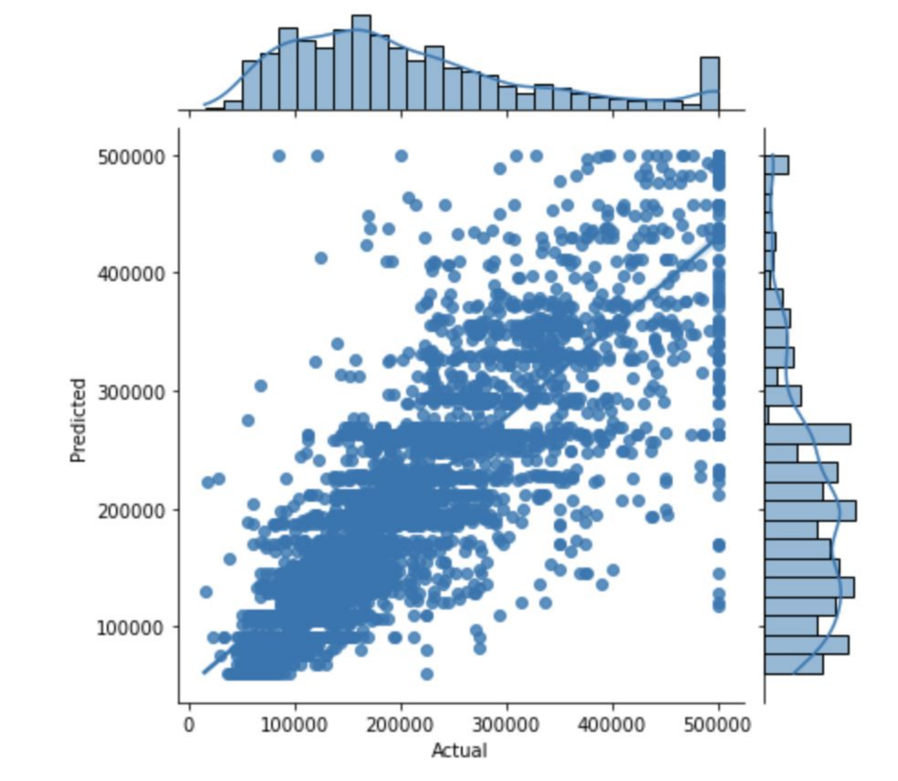
- To create a model that learns simple decision rules from training data to apply on (and predict values for) test data
- Process: starting with root node (entire dataset/test), split each non-leaf node into sub-nodes (subsets) according to decisions
- Decisions/conditions of a node define how the dataset or subsets will be further divided
- End goal is to create homogeneous groups/subsets of data (represented by leaf nodes) in terms of target variable



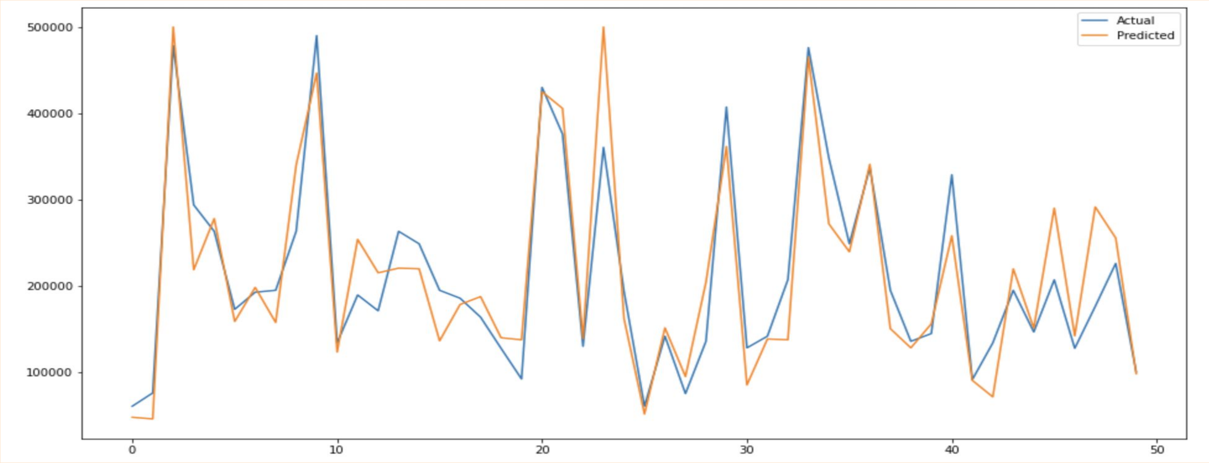
Note:- A is parent node of B and C.

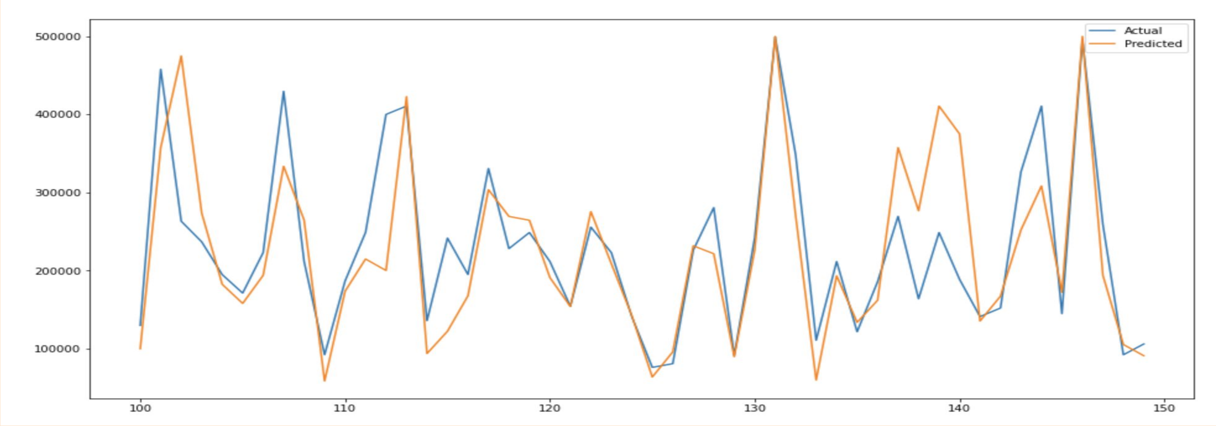
https://www.kdnuggets.com/2020/01/decision-tree-algorithm-explained.html#:~:text=The%20goal%20of%20using%20a,the%20root%20of%20the%20tree .

Predicted & Actual



- Compared to when the actual median house value is relatively small, decision tree model performs worse when the actual median house value was relatively high, and the actual median house value was more unpredictable.
- This phenomenon suggests that higher-priced houses have more influencing factors, making them more difficult to predict.

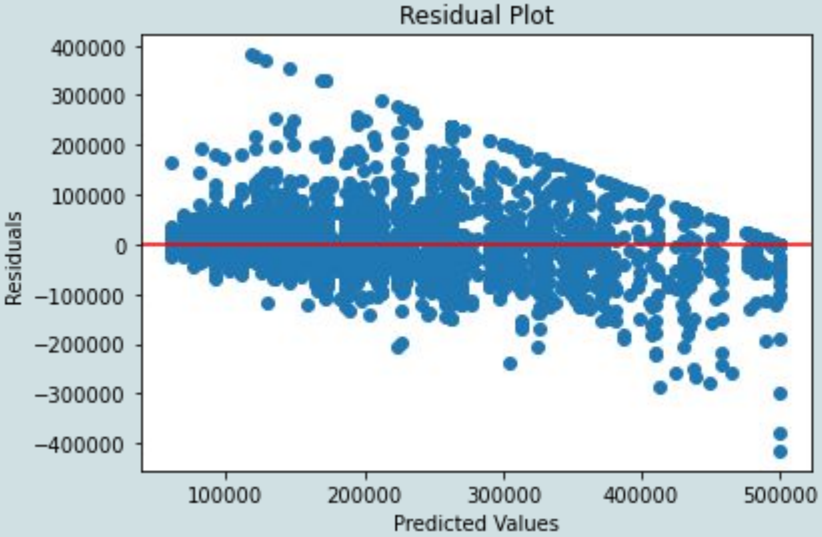


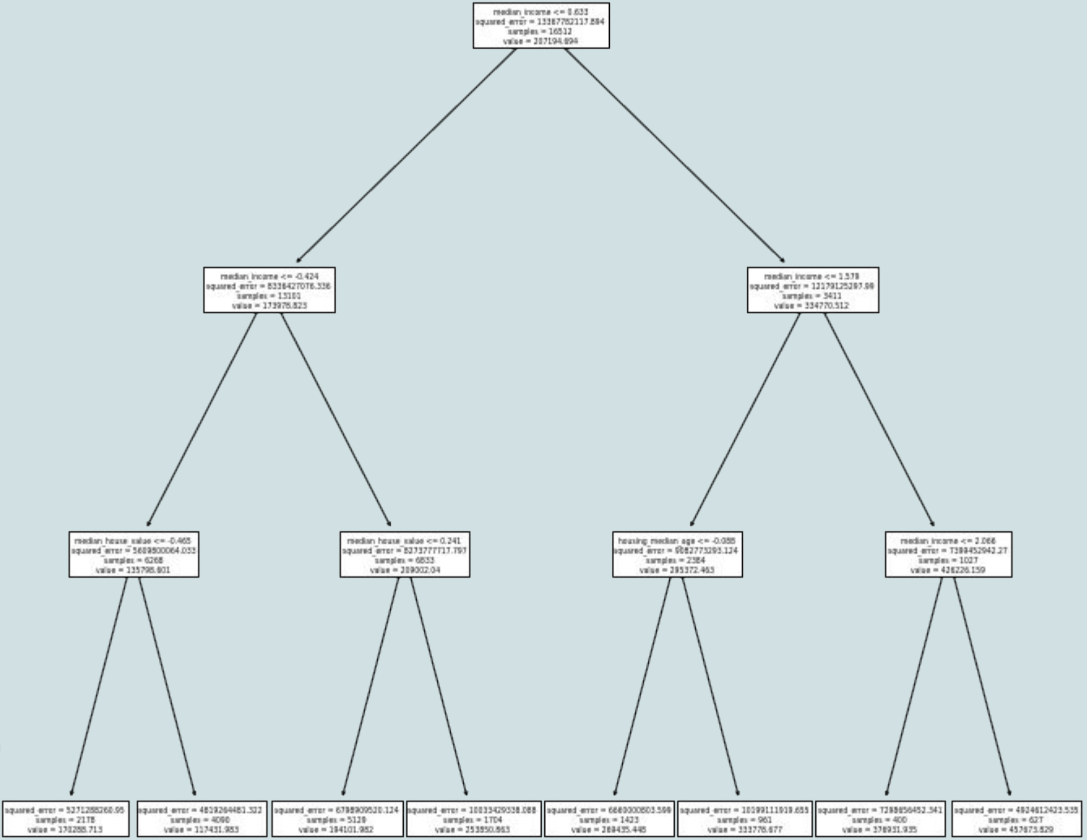


➜ Line plots of sample data of predicted & actual median house value

Regardless of whether the house prices are high or low, there are times when predictions are very accurate and times when there are significant discrepancies in predictions.

Final Visualization and Residual Plot





Evaluation Metrics

Mean Squared Error = 3713191064.8847675

Root Mean Squared Error = 60935.95871802435

R-square Score: 0.7166387649549398

Median Absolute Error : 40354.08196128783

Random Forest Model

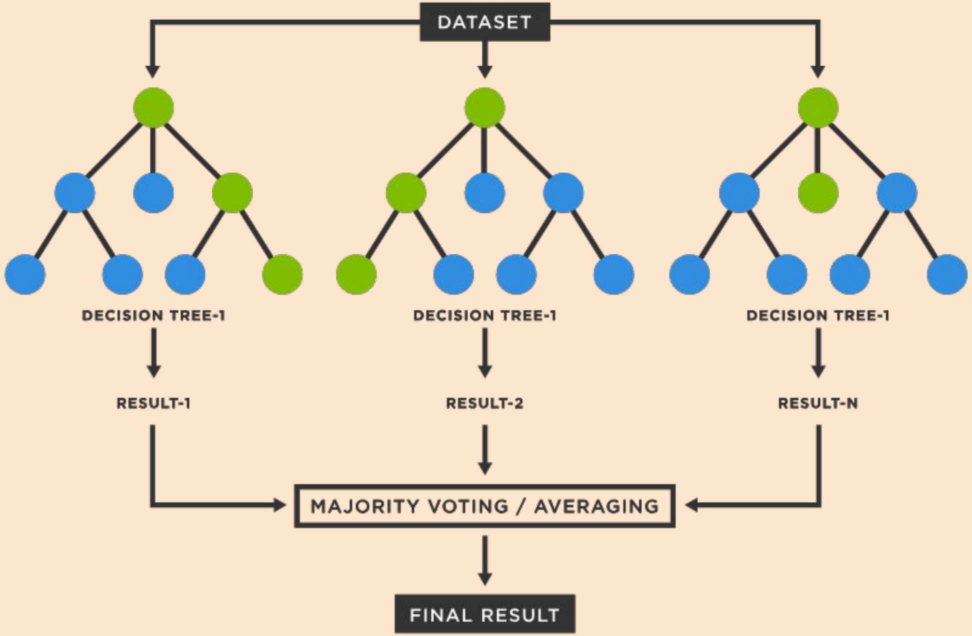
Ravinit Chand

Eric Kye

Revanth Rao

Goal Of Random Forest Model

- Random Forest Model is a machine learning algorithm that will build decision trees from split testing and training datasets. The main goal is to reduce the overfitting of individual decision trees to the training data by averaging and combining multiple trees to make more calculated and accurate predictions.



DATASET

DECISION TREE-1

DECISION TREE-1

DECISION TREE-1

RESULT-1

RESULT-2

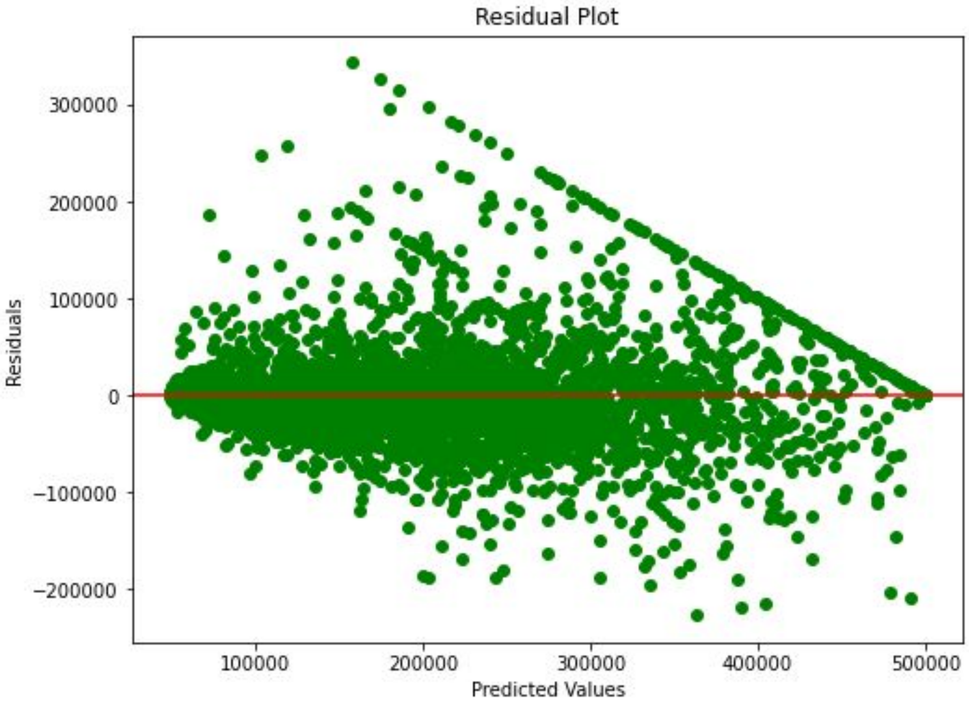
RESULT-N

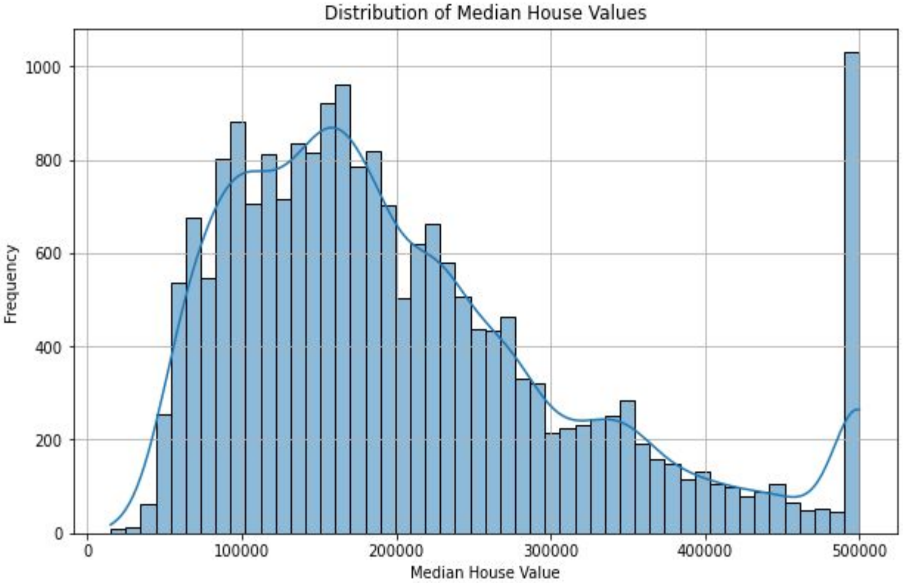
MAJORITY VOTING / AVERAGING

FINAL RESULT

Residual Plot For Random Forest Model

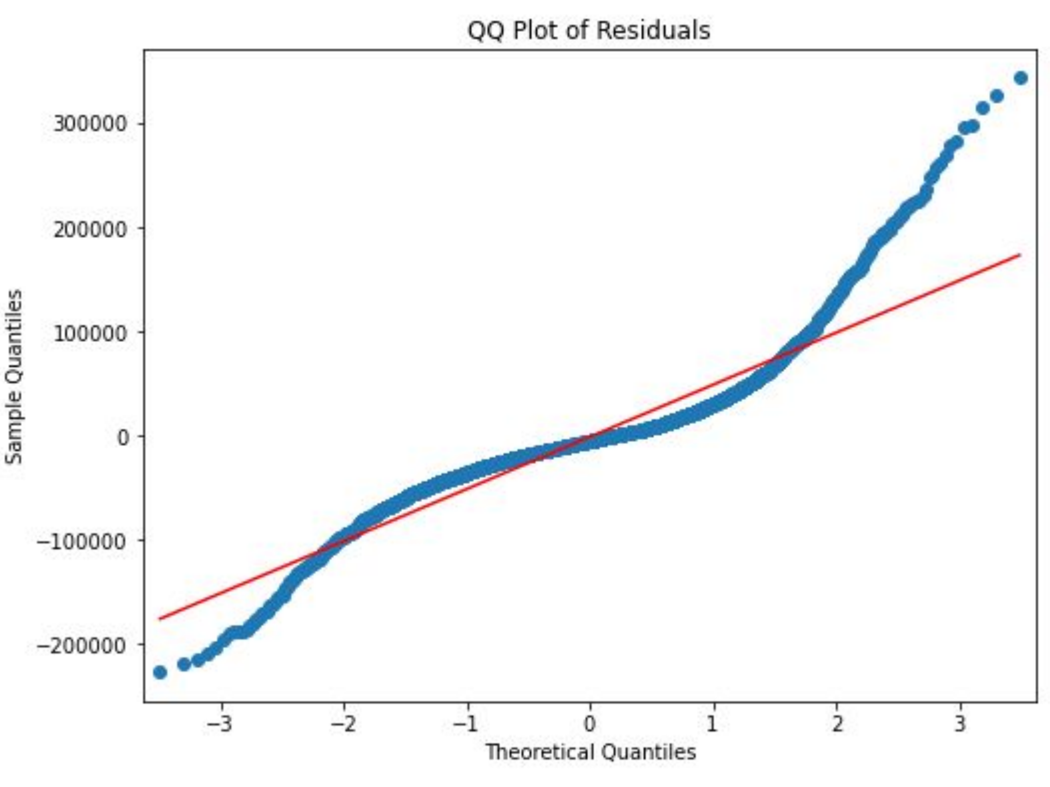
- Predicted model x axis, residuals on y
- One row of points that appears to be linear but negative
- Hard cap (arrow) on house prices, which could've caused line





QQ Plot Model

- Points seem to follow a normal distribution in the middle
- Heavy tail shown by sharper slopes on the ends, more sharper than a normal distribution
- Heteroscedasticity evidence since variability of data isn't consistent



Evaluation Metrics

Mean Squared Error = 2,502,287,622.42

Root Mean Squared Error = 50,022.87

R^2 = 0.81

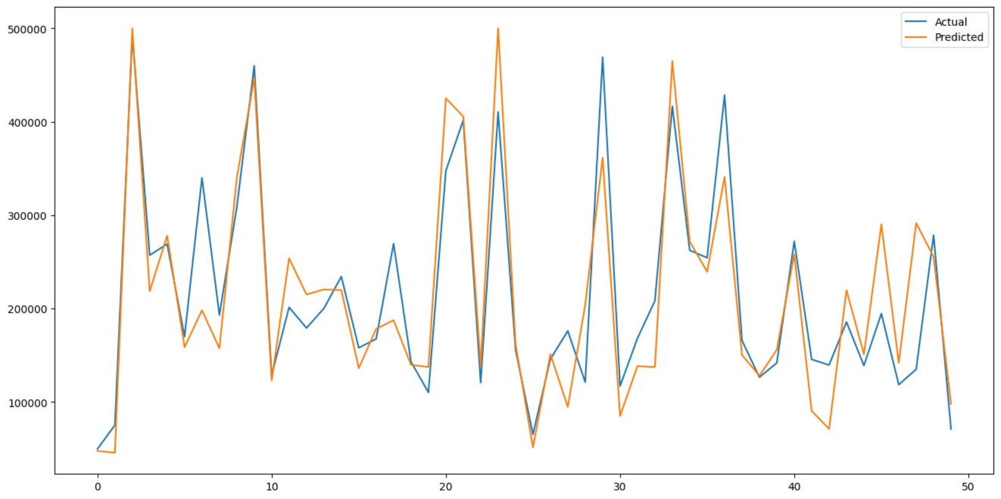
Normalized Absolute Error (NAE) = 0.18

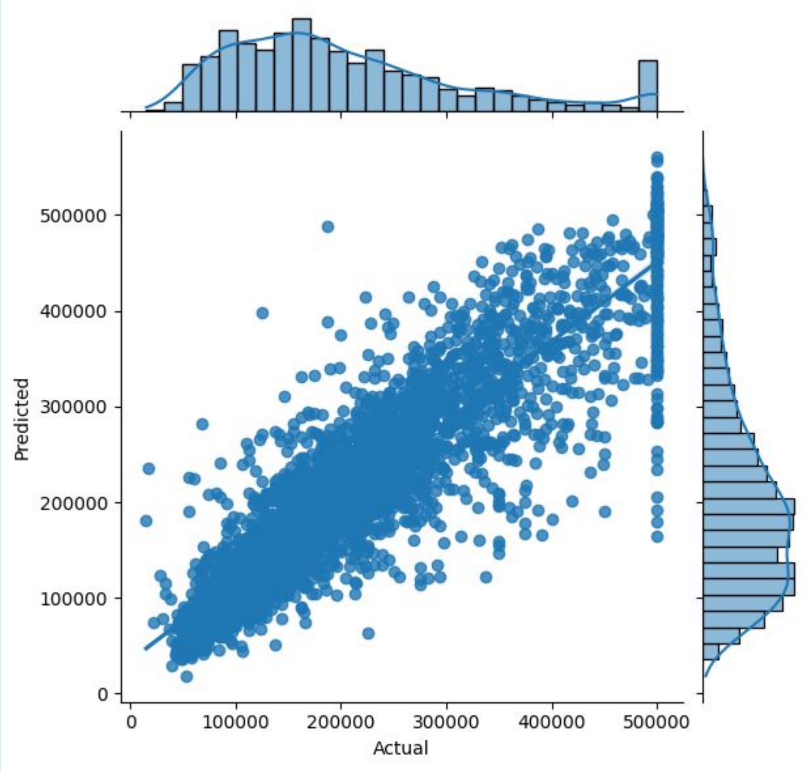
Explained Variance Score = 0.81

Median Absolute Error = 19834.01

XGBoost

- XGBoost implements gradient boosting, iteratively training decision trees to minimize the objective function by correcting errors from previous models, while incorporating regularization, pruning, and feature importance analysis for improved predictive performance.





Evaluation Metrics

Mean Squared Error (MSE): 2338148776.6440325

Root Mean Squared Error: 48354.40803736545

Mean Absolute Error (MAE): 32137.68109414744

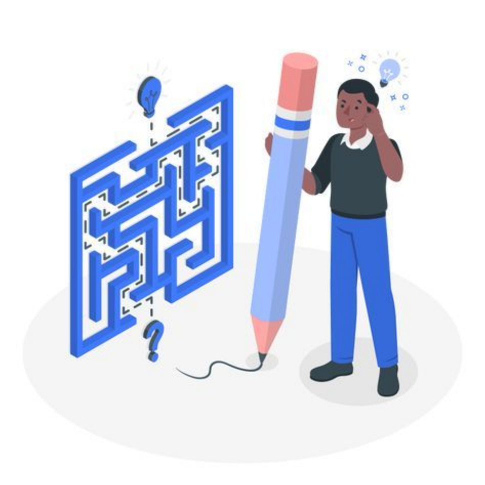
R-squared (R2) Score: 0.8215710655628462

Discussion

- XGBoost performed the best with an R^2 value of 0.82
- Linear Regression had the worst performance with an R^2 value of 0.41
- In general, none of the models achieved a value higher than 0.82
- Could work on improving the model performance (e.g. Feature Engineering, Feature Selection, Hyperparameter tuning, Regularization etc.)
- Variables in the dataset are limited. A better dataset would include information about inflation, demographics, interest rates, etc.

Some Challenges We Encountered

- Dataset from 1990. A newer dataset would’ve been more relevant and useful
- Determining the best way to fill NA values (median vs 0)
- Understanding MSE size in relation to median house price (leading to very high values)
- “Lines” of data points in plot graphs (residuals, actual-predicted), due to hard cap of median house price
- Scaling data before vs. after splitting the dataset into train and test sets
    - We ended up scaling after splitting (StandardScaler)



THANK YOU