Introduction/Background: Zillow Home Value Prediction

Datasets: [7]

- **Zillow Home Value Index (ZHVI):** A measure of the typical home value and market changes across a given region and housing type.
- **Zillow Home Value Forecast (ZHVF):** A month-ahead, quarter-ahead and year-ahead forecast of the Zillow Home Value Index (ZHVI).

Literature Review:

Paper 1 The research paper [3] "Predicting Zillow Estimation Error Using Linear Regression and Gradient Boosting" explores linear regression and gradient boosting models to improve the accuracy of the current home value estimator on Zillow. Both underperform when dealing with outliers, and due to skewed data, normalization techniques often worsen accuracy. We aim to overcome these limitations by incorporating Random Forest Regressor and SVMs. Through cross-model comparison, we hope to build a more generalizable model.

Paper 2 This research paper â€œHouse price prediction using neural networks" [6] compares the predictive performance of the ANNs - multilayer perceptron, and the autoregressive integrated moving average model (ARIMA). ANNs fail to optimize performance on small datasets used in the paper. We plan to use SVMs, as they handle small datasets better than neural networks, also making them less prone to overfitting.

Problem Definition: Zillow Home Value Prediction

Problem

Knowing the potential that a house can offer is a huge asset for all stakeholders involved, and it calls for the price of homes in a specific region, market trends, and various other economic indicators.

Motivation

The Volatility of the Real Estate Market and Its Impacts

The instability [4] of the real estate market is due to many factors such as interest rates, demographic trends, economic trends, etc. There is recent data showing significant shifts in market activity, which have critical implications.

Overvaluation

The 25% overvaluation statistic relative to its long-term fundamental value in Q2 2022 adds a layer of risk of significant wealth loss for homeowners and investors. This can also impact the Affordability Crisis [5] which is a challenge for first time buyers entering the market.

The Role of Machine Learning in Addressing Market Challenges

Applying machine learning [1] [2] to predict home values can offer essential insights for both industry professionals and government agencies.

Methods

Data Preprocessing Methods

1. **Data Cleaning** - Missing values in the growth columns of the Zillow dataset, which represent recent market trends, were handled by median imputation. Median imputation was chosen for its robustness to outliers and its ability to maintain the dataset's central tendency without being skewed by extreme values.
2. **Standardization** - To ensure consistent scaling across features, the growth columns were standardized to have a mean of 0 and a standard deviation of 1. Standardization makes it easier to compare market trends without disproportionate influence from features with larger scales, ultimately improving model performance.
3. **Regional Bias Adjustment** - Recognizing that regional differences affect housing trends, we grouped data by RegionType to calculate mean and standard deviation for each region's growth columns. This regional adjustment highlights variability across different areas, helping identify and account for regional biases that could otherwise skew predictions.

Machine Learning Algorithms/Models to be implemented

1. **Linear Regression** - Modeling the relationship between home values and their features. Implemented through scikit-learn's `LinearRegression` class
2. **Random Forest Regressor** - Improving the predictive accuracy by averaging multiple decision trees, reducing variance. Implemented using scikit-learn's `RandomForestRegressor`class
3. **Gradient Boosting Regressor** - Effective in capturing complex relationshipsAvailable in scikit-learn as `GradientBoostingRegressor`.
4. **Support Vector Machine (SVM)** - Effective in high-dimensional spaces, accommodating many features. Use scikit-learn's `svm.SVR` for regression tasks.

Midpoint Implementation Overview

For the midpoint assessment, we began by leveraging a **Linear Regression Model** to capture potential relationships between home values and their features, focusing on supervised learning techniques to predict home value growth rates. Given the nature of the Zillow dataset as time-series data, it was reasonable to hypothesize that it may exhibit a linear trend over time across features. Based on this assumption, we chose linear regression as an initial approach to model the trend.

Here, the features consist of growth rates from two previous time points (X), while the target is the growth rate at a third time point (y). This setup provides a framework for analyzing sequential trends, offering valuable insights into the housing market's temporal dynamics.

Model Setup and Evaluation

The dataset was divided into training and testing sets with an 80/20 split, allowing the model to train on a subset of the data and then evaluate its accuracy on previously unseen data. Using scikit-learn's LinearRegression model, we trained the model on the training set and generated predictions for the test set. Model performance was assessed using **Mean Absolute Error (MAE)**, **Mean Squared Error (MSE)**, and **Root Mean Squared Error (RMSE)**. RMSE, in particular, is a highly interpretable metric as it reflects error in the same units as the target variable, offering insight into how closely predictions match actual values.

However, the relatively high error values from these metrics suggest that this dataset may exhibit a nonlinear relationship. This indicates a need for more complex modeling techniques that can capture nonlinear trends, potentially involving higher-dimensional features like regional and economic factors.

CS 7641 and CS 4641 Methods Identified

- **Supervised Learning**: The focus here is on supervised learning methods, as we have labeled data (historical home values).
    - Algorithms: Linear Regression, Random Forest Regressor, Gradient Boosting Regressor, Support Vector Machine (SVMs)

Results and Discussion

Quantitative Metrics

1. **Mean Absolute Error (MAE)**:

$$\text{MAE} = \frac{1}{n}\sum_{i=1}^{n}|y_i - \hat{y}_i|$$

- **Root Mean Squared Error (RMSE)**:

$$\text{RMSE} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}$$

- **R-squared (Coefficient of Determination)**:

$$R^2 = 1 - \frac{\sum_{i=1}^{n} (y_i - \hat{y}_i)^2}{\sum_{i=1}^{n} (y_i - \bar{y})^2}$$

These metrics provide insight into the model's accuracy and its performance in capturing home value growth rates. The MAE reflects the average absolute difference between predicted and actual values, giving a straightforward indication of prediction error. The MSE, which squares the error differences, further penalizes larger errors, highlighting any significant mispredictions in the model. Finally, RMSE presents an interpretable measure of error in the same units as the target variable, making it a practical indicator for evaluating prediction accuracy in the context of home value growth rates.

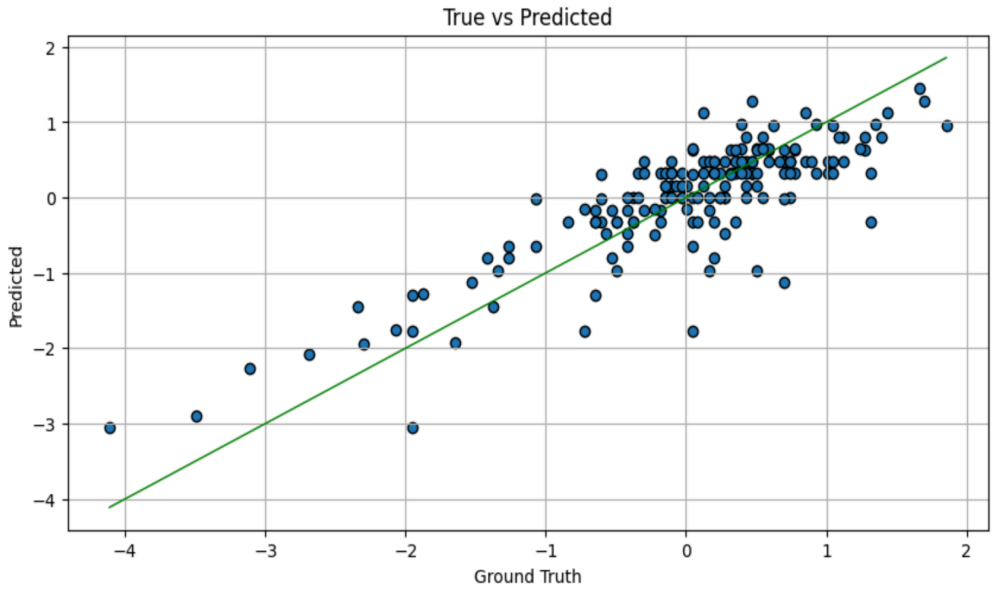
The evaluation metrics for the **Linear Regression model** are as follows:

- Mean Absolute Error (MAE): 0.3899
- Mean Squared Error (MSE): 0.2590
- Root Mean Squared Error (RMSE): 0.5090

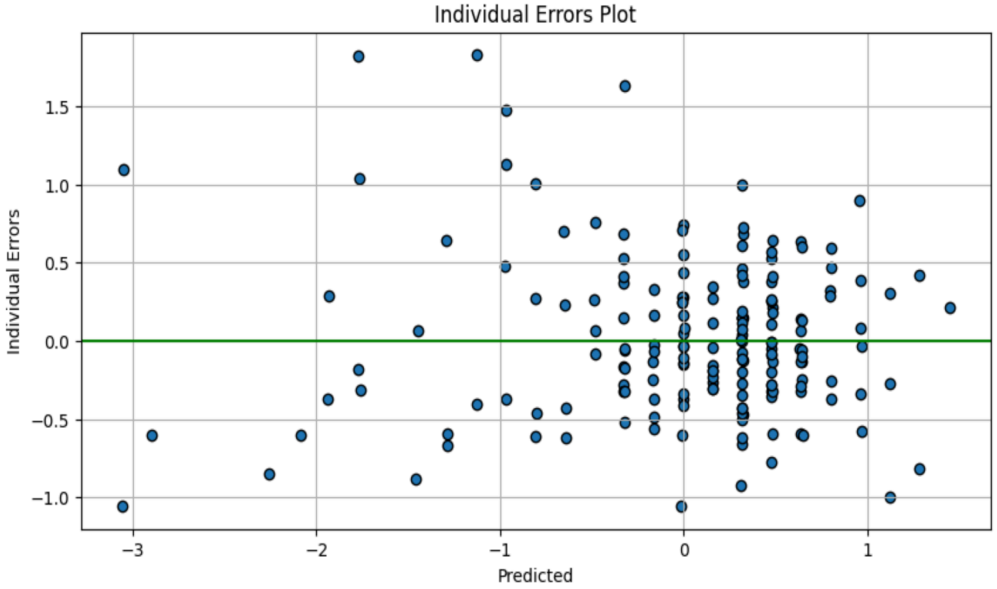
These metrics indicate that while the model performs reasonably well in capturing general trends, the relatively high RMSE suggests a moderate level of prediction variance. This implies that the dataset may contain nonlinear relationships or complex patterns that the linear model struggles to capture accurately, particularly across different regions and economic factors. To address this, exploring more sophisticated algorithms, such as ensemble or nonlinear regression models, may be necessary to better represent the dynamics of the housing market.

Visualizations and Model Analysis (3 Implemented for the Final Report)

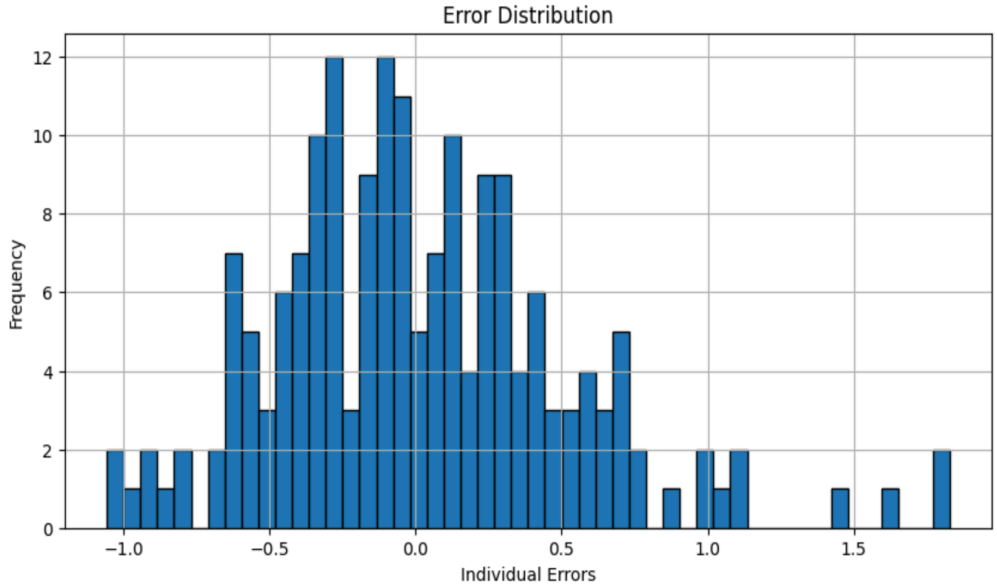
- **Model 1: Linear Regression**



The following plot illustrates the correlation between actual values and the predictions generated by the model. The closer the points are to the green line, the more accurate the predictions are. However, the dispersion of points around this line indicates that the model struggles to capture precise values, especially at the extremes of the target range. This suggests that the linear model may not fully capture the complexity of the underlying data, as there seems to be substantial error for both high and low predictions, implying potential nonlinear relationships that are not accounted for by a simple linear approach.

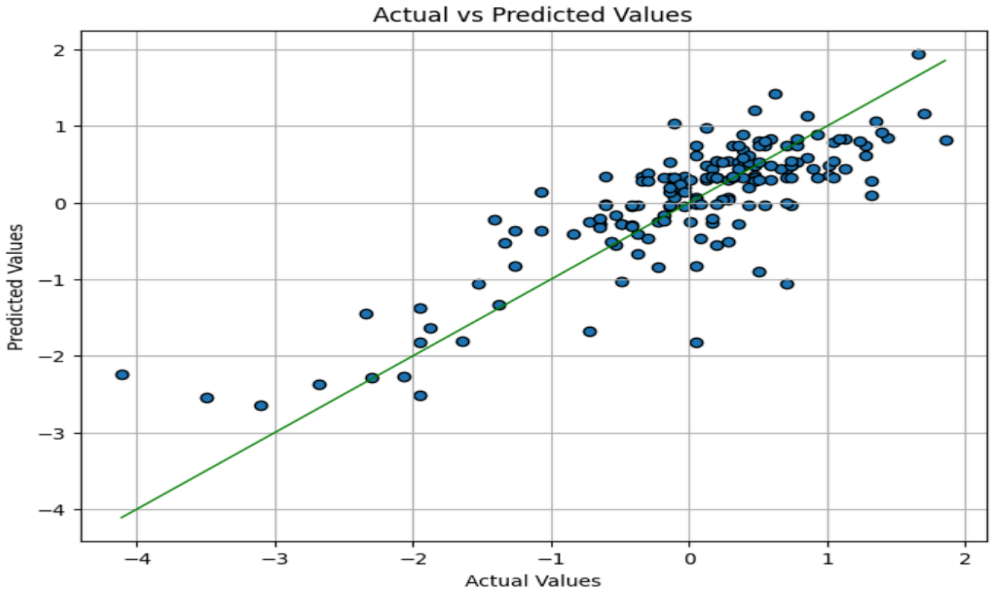


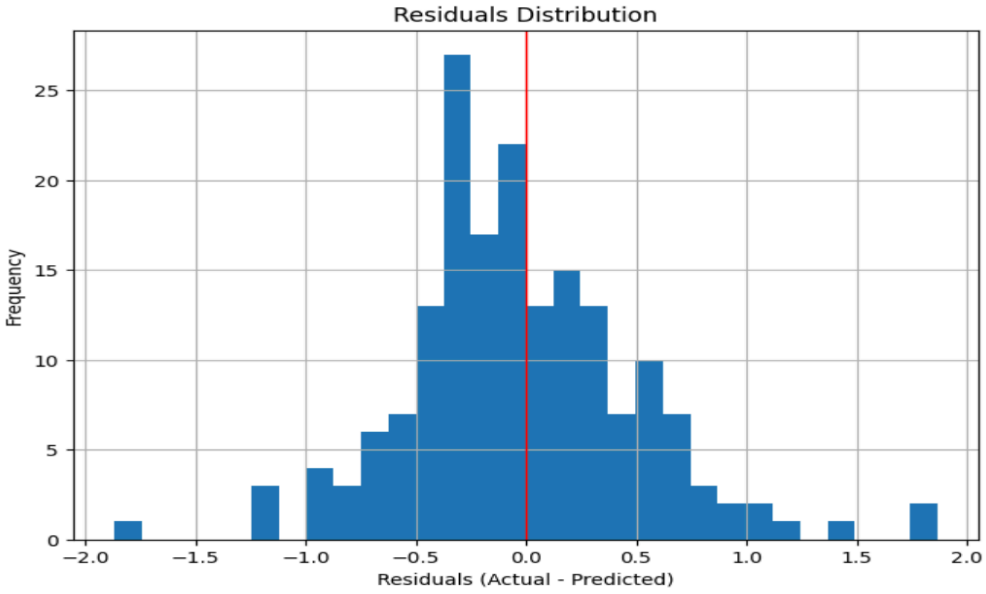
In this plot, we see the distribution of individual prediction errors. Ideally, we would expect errors to be randomly distributed around zero, but here, the pattern of clustering around certain values suggests systematic biases in the modelâ€™s predictions. The green line at zero helps to indicate that while many predictions are close to accurate (falling near the line), a significant number of errors deviate in both positive and negative directions. This variance in error values further supports the need to consider alternative models that might handle complex patterns more effectively.

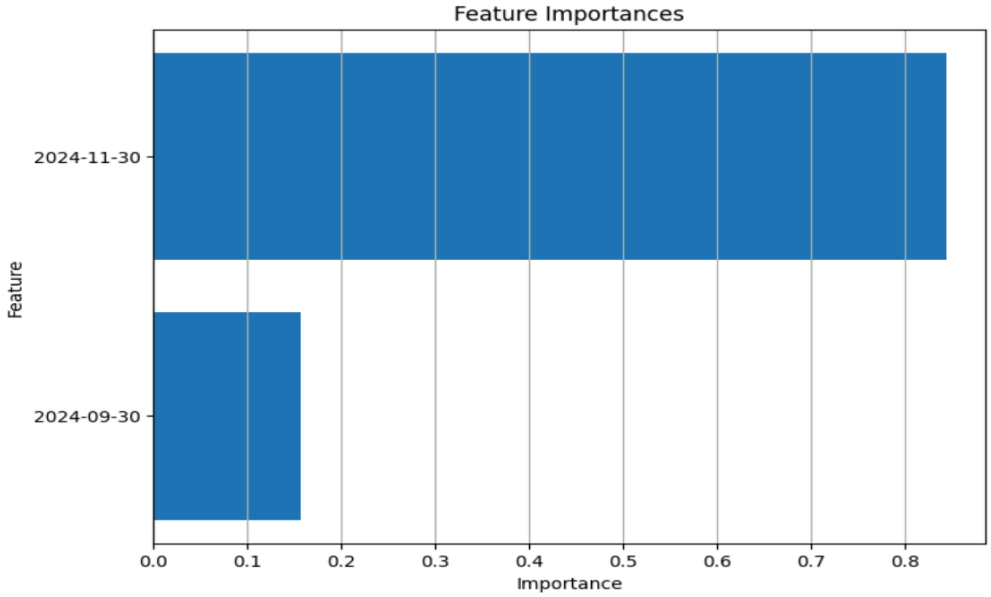


The third visualization, which displays the distribution of individual errors, reveals a roughly symmetric pattern around zero. This shape indicates that, while the model makes a fair number of predictions close to the actual values, the spread of errors is substantial, with a long tail on both sides of the distribution. This gives us further proof that the linear model does not perfectly fit the dataset, and the distribution of errors suggests that we would need additional feature engineering or more sophisticated machine learning models, such as Random Forest Regressor, Gradient Boosting Regressor, Support Vector Machine (SVMs), which could reduce this spread of error and help in improving the predictive accuracy of our model.

- **Model 2: Random Forest Regressor**

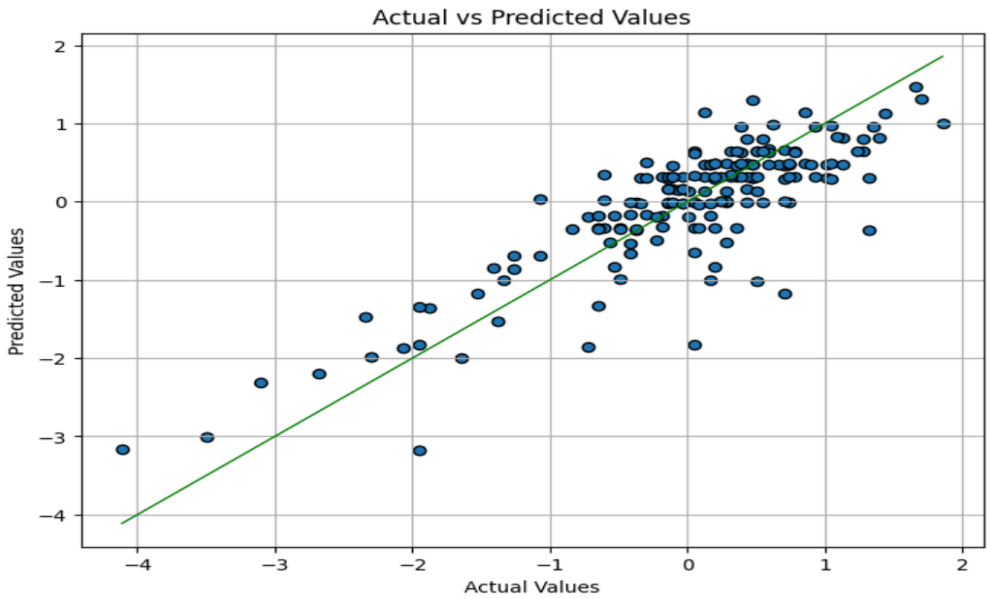


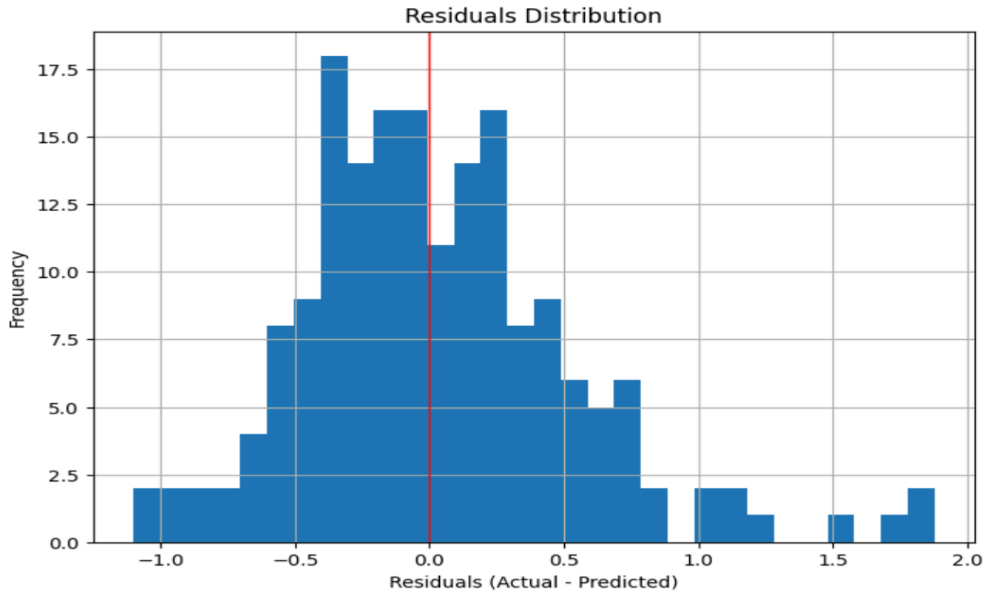




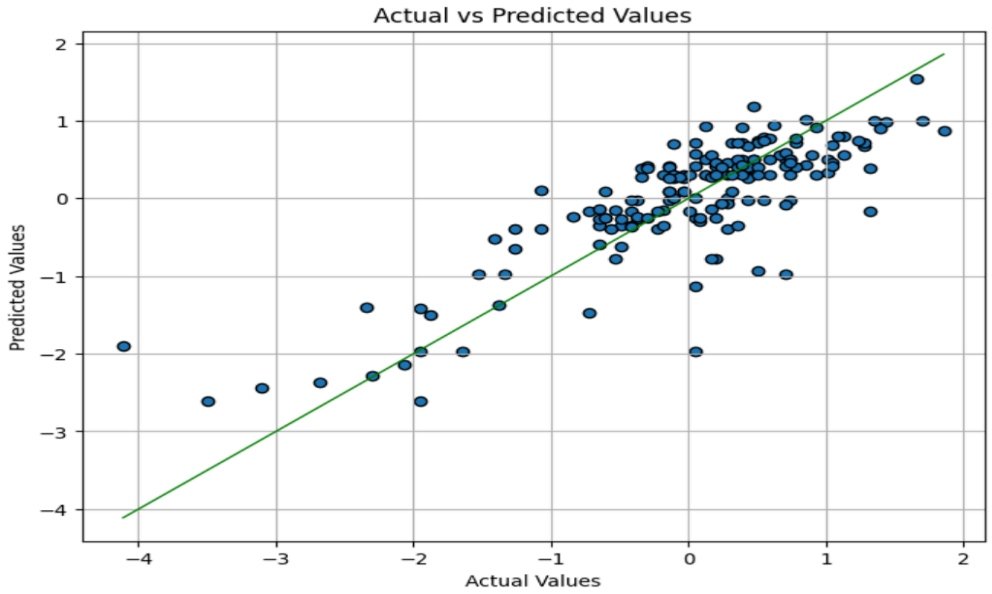
Here, we can observe that with the Random Forest Regressor, while the actual vs predicted values are clustered similar to Linear Regression, the Residual Distribution is more symmetric around 0, indicating a higher frequency of residuals (individual errors) that are closer to 0. This indicates a more optimal performance. The feature importance plot showcases how each of the 2 features impact the prediction made. Feature 1, the time series feature â€œ2024-11-30â€ contributes significantly more to the regressorâ€™s predictions than Feature 2: â€œ2024-09-30â€ .

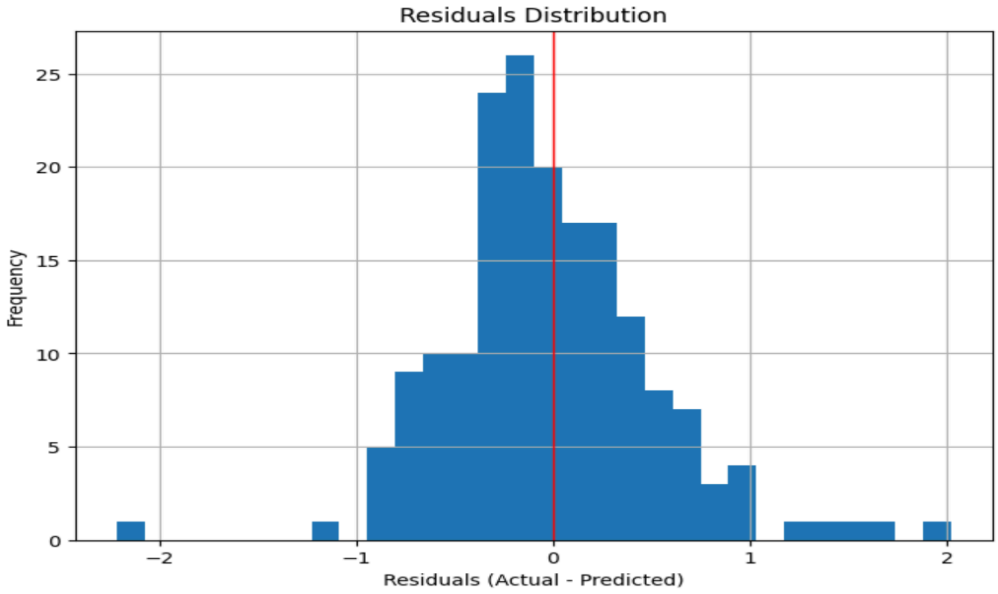
- **Model 3: Support Vector Regressor**





- **Model 4: Gradient Boosting Regressor**





Observing the Residuals Distribution for Models 3 and 4, we can observe that the residuals get further clustered around 0 through incremental error correction with the decision tree classifiers. The Gradient Boosting Regressor Model, with similar performance metrics to Linear Regression has the best clustered distribution around 0, showcasing better performance.

Comparisons

Although the Gradient Boosting Regressor and Linear Regression show comparable error metrics such as Mean Absolute Error (MAE) and Mean Squared Error (MSE), the residual distributions highlight why Gradient Boosting Regressor is a better choice.

- **Linear Regression:**
    1. MAE: 0.38996
    2. MSE: 0.25904
    3. RMSE: 0.50896
- **SVM:**
    1. MAE: 0.38796
    2. MSE: 0.26246
    3. RMSE: 0.51231
- **Random Forest Regressor:**
    1. MAE: 0.39967
    2. MSE: 0.27492
    3. RMSE: 0.52433
- **Gradient Boosting Regressor:**
    1. MAE: 0.39573
    2. MSE: 0.27834
    3. RMSE: 0.52757

Residual Distribution of Linear Regression:

1. It shows a broader spread, with more extreme residuals on both ends of the scale.
2. This indicates that Linear Regression struggles to capture certain patterns in the data, leading to larger errors for some predictions.
3. There is slight asymmetry in the distribution, suggesting the model might not be completely unbiased in its predictions.

Residual Distribution of Gradient Boosting Regressor:

1. It is more concentrated around zero, with fewer extreme values.
2. This indicates that Gradient Boosting Regressor is better at minimizing prediction errors, especially for outliers or complex patterns in the data.
3. The distribution is more symmetric, suggesting that the model captures the underlying data trends more effectively and without significant bias.

When comparing the four models- Linear Regression, SVM, Random Forest Regressor, and Gradient Boosting Regressorâ€”based on their MAE and MSE, the differences in values are relatively small, indicating a negligible difference in overall performance. Linear Regression achieves the lowest MSE (0.25904) and a competitive MAE (0.38996), making it slightly better at minimizing both average errors and large deviations. SVM has the lowest MAE (0.38796), with an MSE of 0.26246, which is only marginally higher than that of Linear Regression, showing its ability to produce similarly accurate predictions. Gradient Boosting Regressor follows closely with an MAE of 0.39573 and an MSE of 0.27834, demonstrating comparable performance and strong adaptability to complex patterns. Random Forest Regressor, while slightly higher in both MAE (0.39967) and MSE (0.27492), still remains close enough to the others to indicate no drastic drop in

performance. Overall, the differences in MAE and MSE across these models are minimal, with each performing well enough to be viable, though Linear Regression and Gradient Boosting Regressor have slight edges depending on the complexity of the dataset's relationships.

The other models, SVM and Random Forest Regressor, fail to perform as well as Gradient Boosting Regressor and Linear Regression. While SVM achieves a slightly lower MAE than Gradient Boosting and Linear Regression, its higher MSE and RMSE imply it struggles with larger errors, making it far less reliable across multiple scenarios. The Random Forest Regressor, on the other hand, has the highest MAE, MSE, and RMSE among all models, indicating poorer performance in capturing the dataset's patterns. These results showcase that Gradient Boosting Regressor and Linear Regression are better suited for this specific dataset, with Gradient Boosting having the upper hand due to its tighter residual distribution.

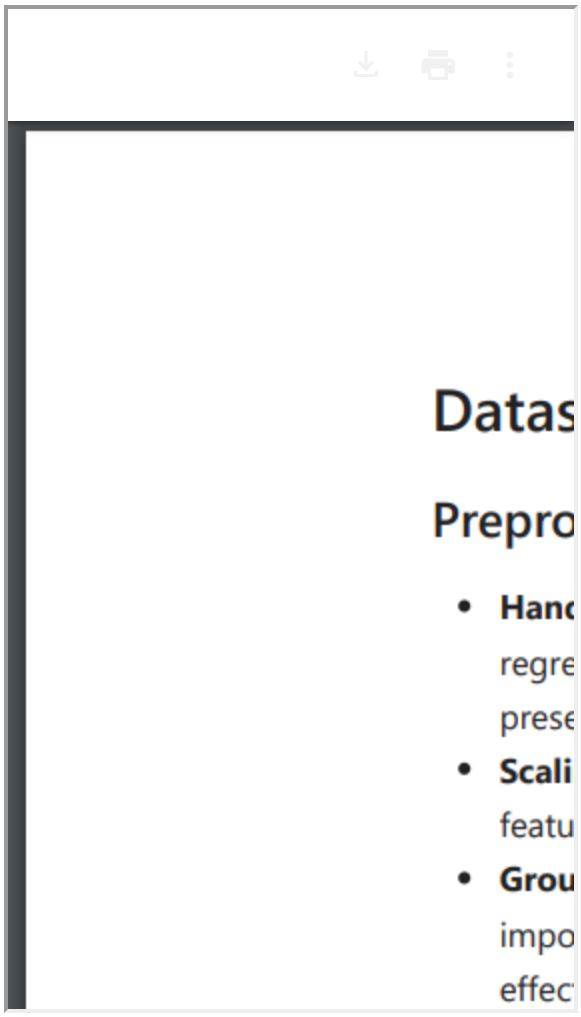
Next Steps

To improve predictive performance,we will look into further feature engineering to incorporate higher-dimensional features, such as economic factors and regional differences, which may have a substantial impact on home value trends. These adjustments will aim to capture more of the variance in the dataset and reduce the observed error, ultimately leading to a model that better reflects the complexity of housing market dynamics. These improvements will help us reach the project goals outlined below and achieve the expected results using advanced machine learning methods. With these higher dimensional features, better nonlinear techniques like Neural Networks (potentially Recurrent Neural Networks or Long Short Term Memory Networks) should fit the training data better

Project Goals

- **Accuracy**: Get MAE and RMSE below a set threshold (e.g., MAE < 0.4 (scaled by the size of the dataset: 894 entries)) and compare different methods of supervised learning based on accuracy.
- **Sustainability**: Using efficient algorithms for reducing computational costs (especially crucial when it comes to md-sized and larger datasets).
- **Ethical Considerations**: Ensuring the model doesn't incorporate biased features, leading to unfair treatment for specific demographic groups.

Notebook PDF



Finally, the Contributions Table and Gantt Chart remain the same as previously shared, with each individual member responsible for their respective action items.

References

1. Q. Truong, M. Nguyen, H. Dang, and B. Mei, â€œHousing Price Prediction via Improved Machine Learning Techniques," Procedia Computer Science, vol. 174, pp. 433-442, 2020, doi: https://doi.org/10.1016/j.procs.2020.06.111.
2. A. P. Singh, K. Rastogi, and S. Rajpoot, â€œHouse Price Prediction Using Machine Learning," IEEE Xplore, Dec. 01, 2021. https://ieeexplore.ieee.org/document/9725552
3. D. Sangani, K. Erickson, and M. A. Hasan, â€œPredicting Zillow Estimation Error Using Linear Regression and Gradient Boosting," IEEE Xplore, Oct. 01, 2017. https://ieeexplore.ieee.org/abstract/document/8108793
4. V. Calanog, K. Fagan, and T. Metcalfe, â€œVolatility in 2022, Uncertainty in 2023," CRE Real Estate Issues, vol. 47, no. 2, pp. 1-15, Feb. 2023, Accessed: Oct. 03, 2024. [Online]. Available: https://cre.org/real-estate-issues/volatility-in-2022-uncertainty-in-2023/
5. K. Reuben, S. Lei, â€œWhat the Housing Crisis Means for State and Local Governments - Lincoln Institute of Land Policy," Lincoln Institute of Land Policy, Jan. 13, 2017. https://www.lincolninst.edu/publications/articles/what-housing-crisis-means-state-local-governments/
6. W. T. Lim, L. Wang, Y. Wang, and Q. Chang, â€œHousing price prediction using neural networks," 2016 12th International Conference on Natural Computation,

Fuzzy Systems and Knowledge Discovery (ICNC-FSKD), Aug. 2016, doi: https://doi.org/10.1109/fskd.2016.7603227.
7. Zillow, â€œHousing Data - Zillow Research," *Zillow Research*, 2011. https://www.zillow.com/research/data/