Daniel Rodriguez, Kzzy Centeno, Mir Khan, Sriharsha Aitharaju

Hanqin Cai

STA 4724

December 6th, 2023

Forecasting U.S. Median Home Sale Prices with Machine Learning: Insights from Morningstar Indexes and SOMA Securities

INTRODUCTION

In the U.S. housing market, understanding and forecasting home prices is a complex yet critical challenge. Our project aims to predict median home sale prices in the United States by processing and analyzing data from diverse financial sources, including Treasury Securities Holdings, Morningstar Index Funds, and Zillow Median Home Sale Prices. The Treasury Securities Holdings data provides a foundational perspective on financial rates influencing various sectors, including the housing market. The Morningstar Index Funds data offers insights into the broader economic landscape, serving as a proxy for understanding disposable income and economic growth trends. The Zillow dataset, our target variable, reflects direct insights into the U.S. housing market, encapsulating the culmination of various economic influences.

Our methodology includes data preprocessing, data collection, collation, and feature selection, to ensure the accuracy of our analysis. We address challenges such as dataset misalignments and refine our dataset to a focused collection of 188 data points with 19 highly relevant features. We employ and compare several machine learning models, including KNN Regression, Gradient Boosted Trees, Random Forest, and Support Vector Regression. Each

model is tuned and evaluated based on its performance metrics, such as Root Mean Square Error (RMSE) and computational intensity.

This paper aims to not only present a comparative analysis of these models but also to shed light on the broader implications of our findings. We explore the relationship between key economic indicators and the housing market, offering insights that are valuable for investors, policymakers, and industry practitioners.

PREPROCESSING

The foundation of our project involved the collection of three distinct datasets: Treasury Securities Holdings, Morningstar Index Funds, and Zillow Median Home Sale Prices. The Treasury Securities Holdings are related to Treasury Issuance and Treasury Rates, which provide insights into foundational financial rates affecting various sectors, including real estate. The Morningstar Index Funds data offers a comprehensive view of the stock market's asset class performances, reflecting broader economic trends. The Zillow dataset, our target variable, offers direct insights into the U.S. housing market, capturing the impact of economic and financial factors on home sale prices.

Our analysis required the integration of these datasets, each with different inception dates, posing a significant challenge in aligning the data. The SOMA Securities data began in 2003, Morningstar records dated back to 1991, and the Zillow data started in 2008. To address this, we initially sorted the data into monthly increments, reducing the data points from 753 to 384. Further refinement was necessary to align with the Zillow dataset. We decided to omit all

records before 2008, resulting in a dataset of 188 rows, enhancing the consistency and accuracy of our analysis see Figure 1 and Figure 2 for the consolidation.

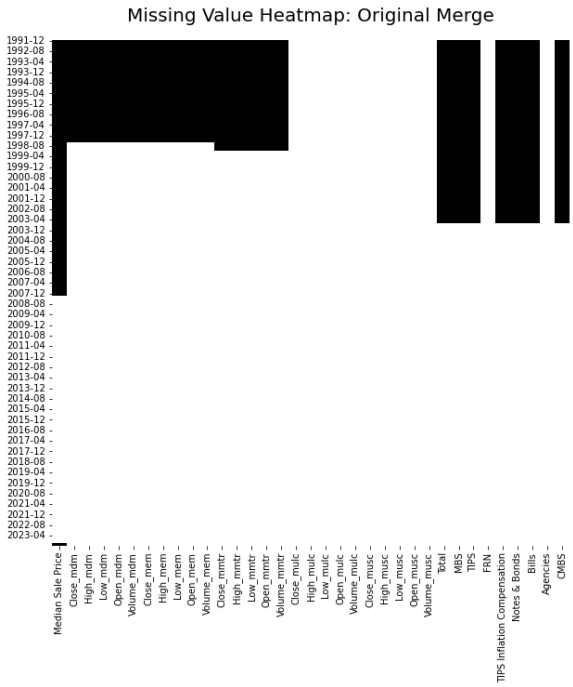

Figure 1

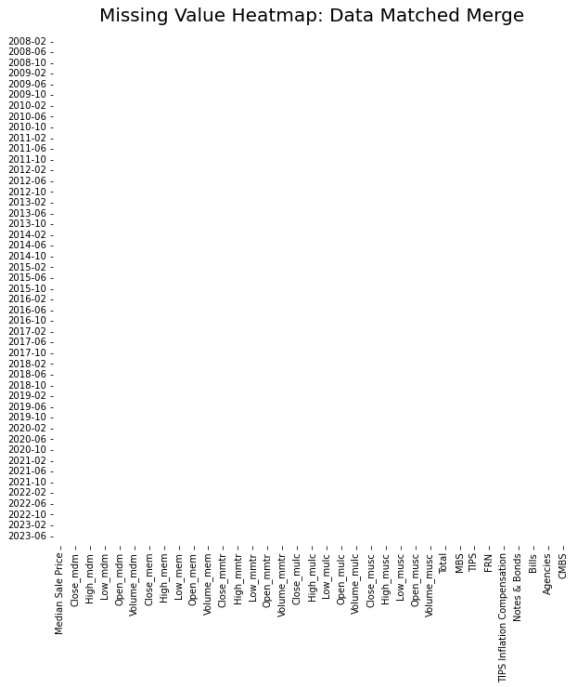

Figure 2

The next step was feature selection, where we reduced the number of features from 35, see Figure 3, to 19, see Figure 4. This was achieved using a correlation matrix to identify the most impactful features, aiming for one feature per ten data points. This process resulted in a refined dataset of 188 data points with 19 highly relevant features, allowing us to utilize our models to their full potential.

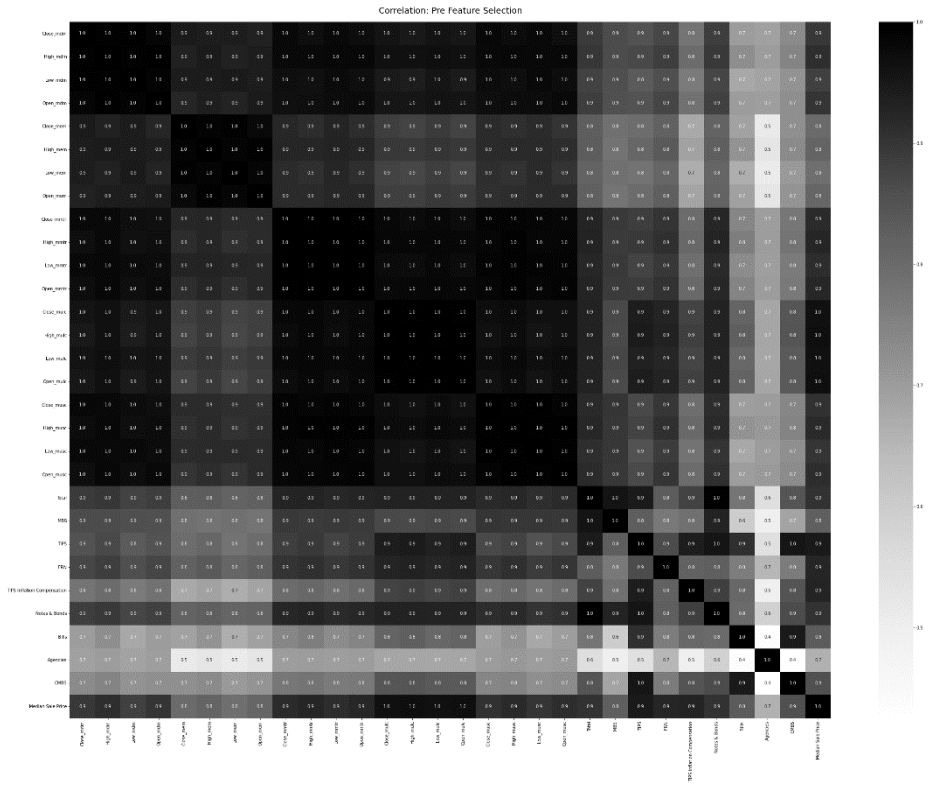

Figure 4

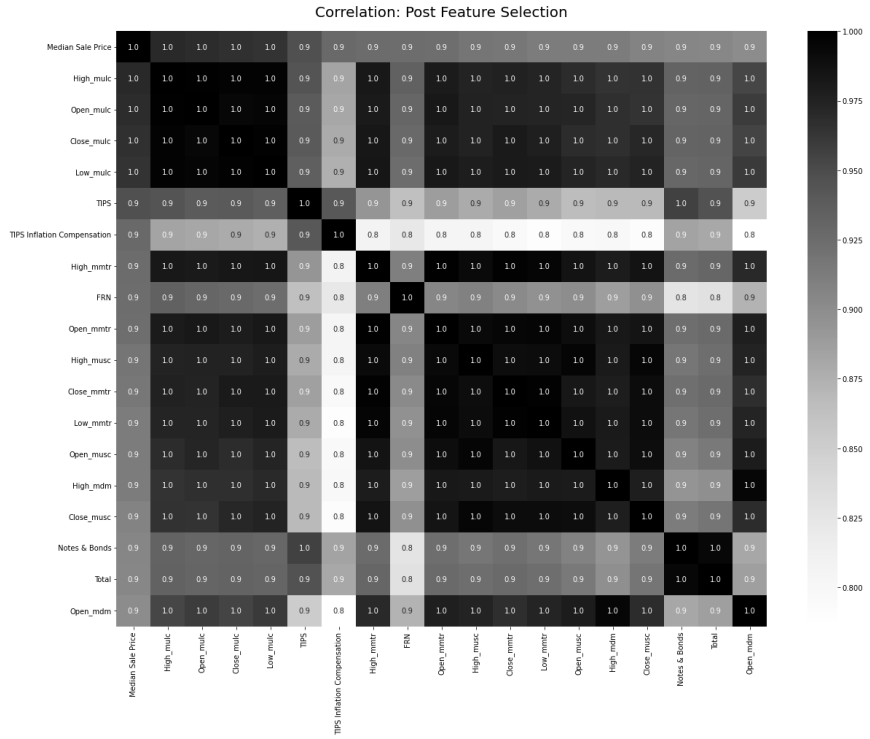

Figure 5

Normalization is a key preprocessing step in our data analysis, essential for ensuring that each feature contributes equally to the machine learning models, regardless of their original scale or unit. This process is crucial because it prevents features with larger scales from disproportionately influencing the model's predictions. We achieved normalization by employing Scikit-learn's StandardScaler, which standardizes features by removing the mean and scaling to unit variance. This approach ensures that our diverse financial indicators are on a comparable scale, facilitating more effective and unbiased machine learning analysis.

MODELS

In our project, we employed four distinct machine learning models to forecast U.S. median home prices. We utilized GridSearchCV for hyperparameter tuning, working through multiple combinations of parameter tunes, cross-validating as it goes to determine which hyperparameter combination gives the best performance. Below is an overview of these models, their hyperparameters, and the process of tuning them for optimal performance.

K-Nearest Neighbors (KNN) Regression predicts outcomes based on the attributes of the 'k' closest training examples in the feature space. It is a type of instance-based learning where the function is only approximated locally, and all computation is deferred until function evaluation.
Key Hyperparameters:

- Number of Neighbors (k): Determines how many neighbors influence the prediction. A smaller k makes the model sensitive to noise, while a larger k makes it computationally expensive.

Gradient Boosted Trees (GBT) is an ensemble learning method that builds trees in a sequential manner, where each tree attempts to correct the errors of its predecessor. This approach combines multiple weak predictive models to create a strong overall model. Key Hyperparameters:

- Number of Trees: Controls the number of sequential trees built.
- Max Depth of Trees: Determines the maximum depth of each tree.
- Minimum Samples Split & Leaf: Regulates the number of samples required to split an internal node and to be at a leaf node.

Random Forest (RF) is another ensemble learning method that operates by constructing multiple decision trees during training and outputting the average prediction of the individual trees. It offers high accuracy and handles large data sets with higher dimensionality well. Key Hyperparameters:

- Number of Trees: The count of trees in the forest.
- Maximum Depth of Trees: The maximum depth of each tree.
- Subsample: Fraction of samples used for fitting each tree.

Support Vector Regression (SVR) applies the principles of Support Vector Machines for regression problems. It works by mapping input features into high-dimensional feature spaces and finding a hyperplane that best fits the data. Key Hyperparameters:

- C (Regularization): Controls the trade-off between achieving a low error on the training data and minimizing the norm of the weights.

- Gamma: Defines the influence of a single training example; low values imply 'far' and high values imply 'close.'
- Epsilon: Specifies the epsilon-tube within which no penalty is associated in the training loss function with points predicted within a distance epsilon from the actual value.

The performance of each model, both before and after tuning, is summarized in Figure 6. This comparison includes the base and tuned scores, along with the Root Mean Square Error (RMSE) for each model. The best hyperparameters were selected based on their performance in reducing RMSE and improving the model's accuracy. The tuning process involved evaluating various combinations of hyperparameters and selecting the set that yielded the most accurate predictions with the lowest error rates. For an example of this see Figure 7 for the optimal k value being selected for KNN Regression.

Model	Base Score	Base RMSE	Tuned Score	Tuned RMSE
Gradient Boosted Regressor	0.9837	7480.8441	0.99	5852.3235
Random Forest Regressor	0.9851	7152.8823	0.9876	6514.0365
Support Vector Regressor	0.9767	8952.2102	0.9817	7925.677
KNN Regressor	0.9853	7093.6745	0.9863	6851.0314

Figure 6: The table displays the results of each model pre and post tuning.

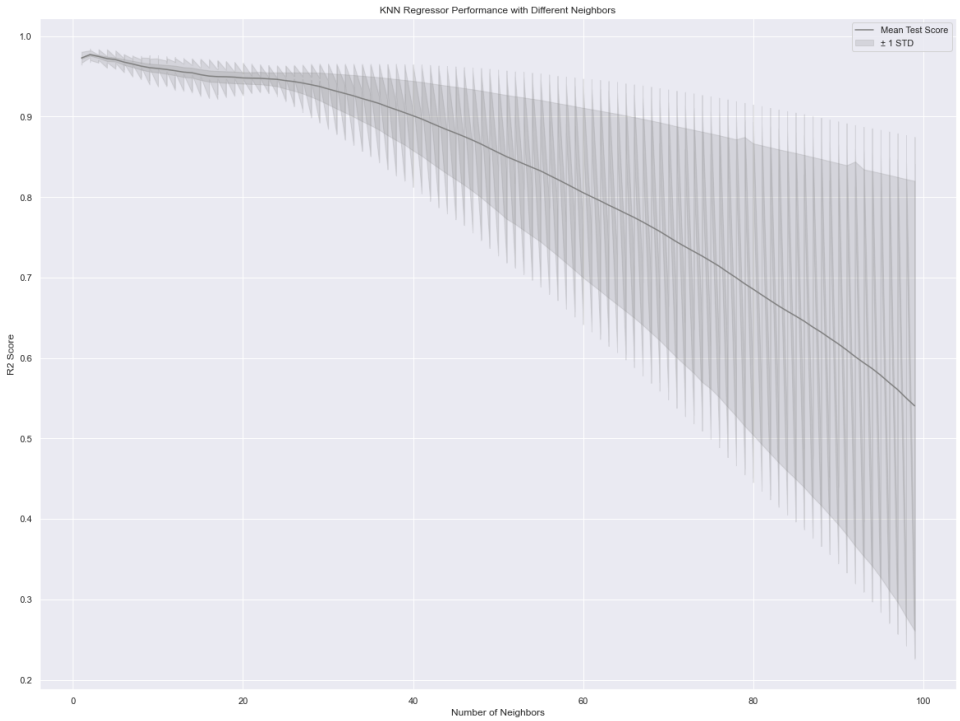

Figure 7: KNN Regressor Performance with Different Neighbors

RESULTS

In our comprehensive analysis, the Gradient Boosted Regressor emerged as the most effective model, boasting an impressive R^2 of 0.99 and the lowest Root Mean Square Error (RMSE). This model's high accuracy underscores its effectiveness in capturing the complexities of the U.S. housing market and its sensitivity to various economic indicators.

A critical aspect of our analysis was understanding the feature importance across all models see Figure 8 for results. This analysis provided valuable insights into which economic indicators most significantly impact U.S. median home sale prices. Commonly important features across models included 'Open_mulc', 'TIPS Inflation Compensation', 'Total', 'TIPS', and 'Notes & Bonds'. These features represent key economic factors such as market openness,

inflation expectations, and government securities, which are instrumental in shaping the housing market. Notably Morningstar US Large Capital Stocks (MULC) posed a significant interest to our team as the index fund is comprised primarily of major blue-chip stocks, particularly those in the technology sector. For instance, significant components of the U.S. Large Capital Stocks, including prominent FAANG members, have shown a strong linkage with housing market trends. This is exemplified by the substantial portfolio weights of leading companies like Microsoft Corp, Apple Inc, and Amazon.com Inc in the technology and consumer cyclical sectors. The influence of these corporations extends beyond their respective industries, reflecting a broader economic impact that resonates in the real estate domain.

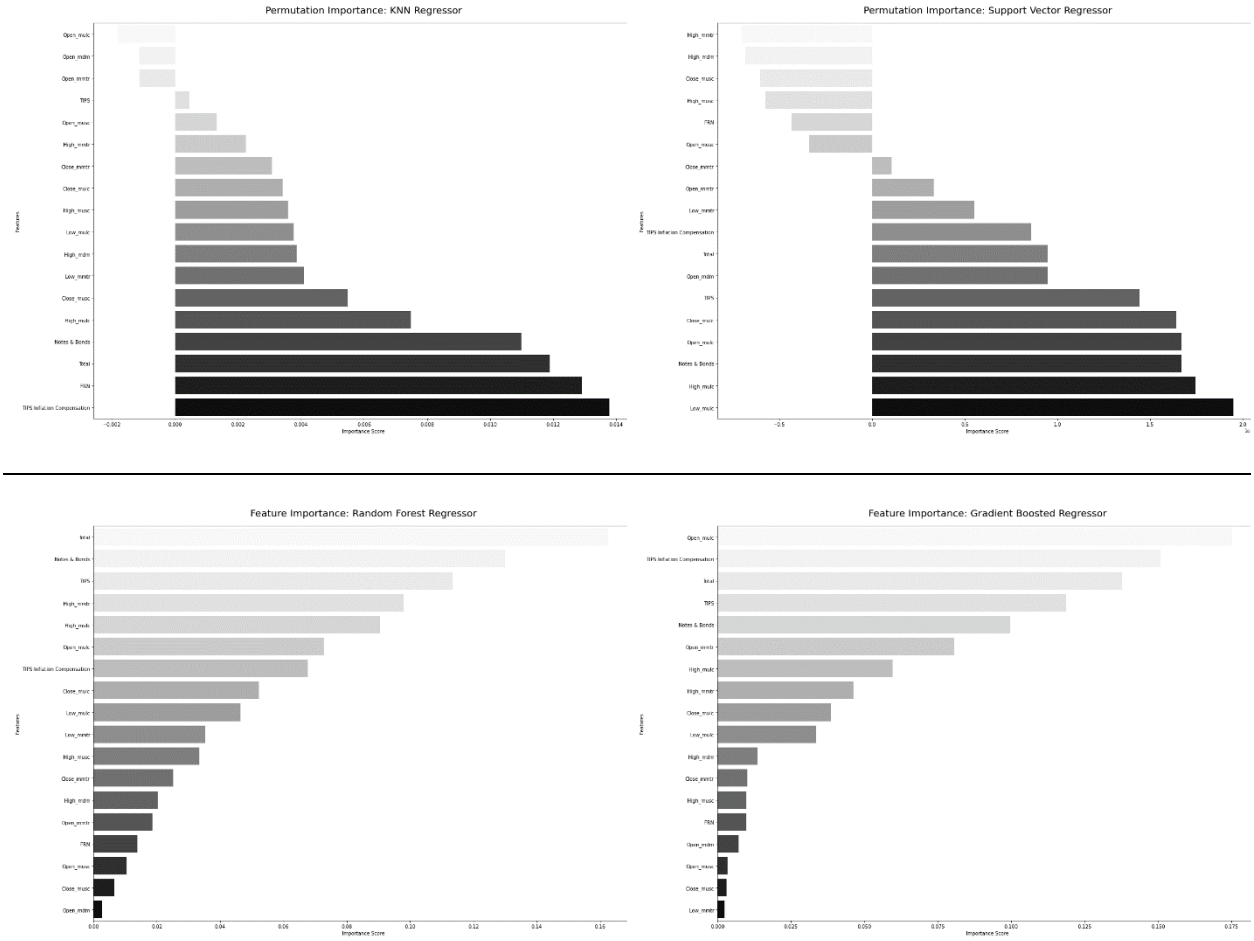

Figure 8: Feature Importance for RF and GB, Permutation Importance for KNN and SVR.

CONCLUSION

Our analysis concluded with the Gradient Boosted Regressor achieving the best Root Mean Square Error (RMSE). However, it's crucial to note that this model, while highly accurate, also exhibited high computational intensity. This aspect is an important consideration, especially in scenarios where computational resources or time are constrained. In such instances, models like the Random Forest or KNN Regressor, which offer a balance between accuracy and computational efficiency, might be more suitable. While our models have provided valuable insights, there are several avenues for further enhancement:

- Enhanced Datasets: Testing additional economic indicators, incorporating regional housing market trends, and analyzing the impact of national economic policies could enhance the models' predictive power.
- Efficiency Tuning: Further optimization of the models, particularly in terms of computational efficiency, could make them more practical for real-time analysis and forecasting.
- Deep Learning Approaches: Exploring deep learning techniques could uncover complex patterns in the data that traditional machine learning models might miss, especially when considering expanding upon the dataset.

With all that said, our project underscores the potential of machine learning in real estate market analysis. The ability to accurately forecast housing prices can significantly benefit investors, policymakers, and industry stakeholders. Moreover, the insights gained from feature importance analysis, particularly the impact of economic indicators like market openness and inflation expectations, can guide strategic decision-making in the real estate sector.

REFERENCES

- https://indexes.morningstar.com/indexes/details/morningstar-us-large-cap-FSUSA00KH5?currency=USD&variant=TR&tab=performance
- https://indexes.morningstar.com/indexes/details/morningstar-us-small-cap-FSUSA00KGS?currency=USD&variant=TR&tab=performance
- https://indexes.morningstar.com/indexes/details/morningstar-developed-markets-ex-us-FS00009P5R?currency=USD&variant=TR&tab=performance
- https://indexes.morningstar.com/indexes/details/morningstar-emerging-markets-FS00009P5Q?currency=USD&variant=TR&tab=performance
- https://indexes.morningstar.com/indexes/details/morningstar-moderate-target-risk-FSUSA09PYI?tab=performance
- https://www.newyorkfed.org/data-and-statistics/data-visualization/system-open-market-account-portfolio
- https://www.zillow.com/research/data/