Daniel Rodriguez & Kemoye Williams

Teng Zhang

MAP 4191

December 2023

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

Introduction

In the landscape of financial analytics, the application of machine learning to predict market trends has become increasingly pivotal. Our research, "Forecasting U.S. Home Prices with Machine Learning: Insights from Morningstar Indexes and SOMA Securities," intends to leverage the predictive power of machine learning in the realm of real estate, particularly focusing on the U.S. housing market.

This study integrates diverse datasets, including the U.S. Treasury Securities Holdings [Fig. 1.], Morningstar Index Fund [Fig. 2.]s, and Zillow Median Home Sale Prices, to construct a comprehensive model that forecasts U.S. home prices. The Treasury Securities Holdings offer a foundational perspective on financial rates influencing various sectors, including real estate. In contrast, the Morningstar Index Funds provide a macroeconomic view of the stock market's performance, reflecting broader economic trends that indirectly impact the housing market. The Zillow dataset, our target variable, directly correlates with the U.S. housing market, offering insights into home sale prices.

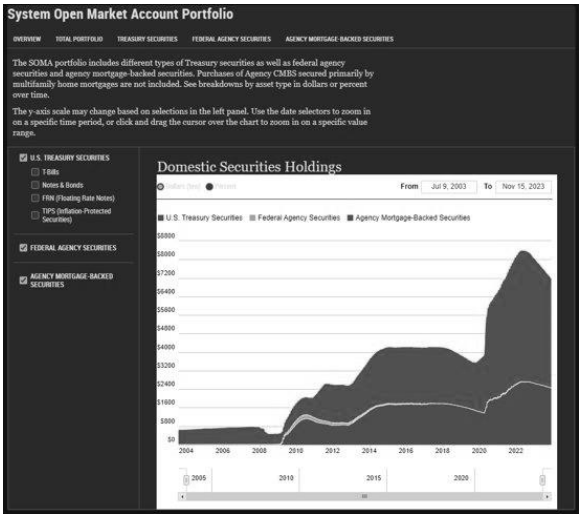


Fig. 1. Morningstar Index Funds Data Sourcing Example

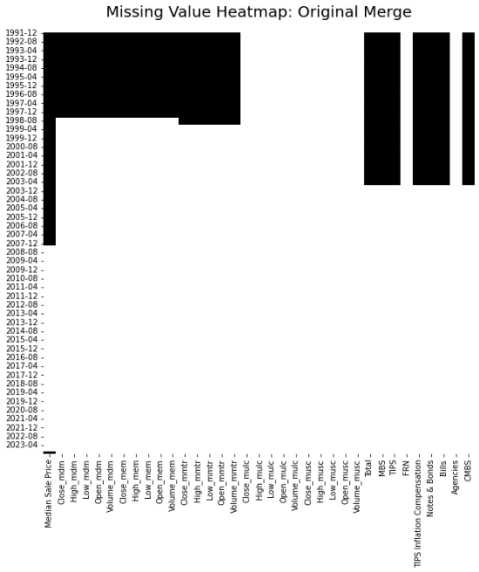
Fig. 2. System Open Market Account (SOMA) Holding of Domestic Securities

Our approach to data preprocessing involved alignment and refinement of these datasets, ensuring consistency and accuracy in our analysis. We navigated through challenges such as varying inception dates and data sparsity, ultimately consolidating our data into a coherent and focused dataset for model application.

The core of our research lies in the application of advanced machine learning models, including KNN Regression, Gradient Boosted Trees, and Decision Trees. These models were rigorously tested and tuned using techniques like GridSearchCV to optimize their performance. Our analysis delves into the computational intensity and efficiency of these models, providing a comparative perspective on their effectiveness in predicting U.S. home prices.

Implementation/Preprocessing

Our data acquisition process leveraged a selenium script that harvested Morningstar Index Fund historical data, US Federal Reserve APIs, and a Zillow data export. The Treasury Securities Holdings data, starting from 2003, provided insights into foundational financial rates influencing various sectors. The Morningstar Index Funds data, dating back to 1991, offered a broad perspective on stock market performances and economic trends. Lastly, the Zillow dataset, commencing in 2008, served as our target variable, directly reflecting the U.S. housing market dynamics. The initial challenge in our data collation process was the alignment of these datasets, each with different inception dates. To address this, we first consolidated the data into monthly increments, reducing the data points from 753 [Fig. 3.] to 384 [Fig. 4.]. Further refinement led us to focus on data post-2008, aligning with the Zillow dataset. This strategic decision resulted in a more focused dataset comprising 188 rows.



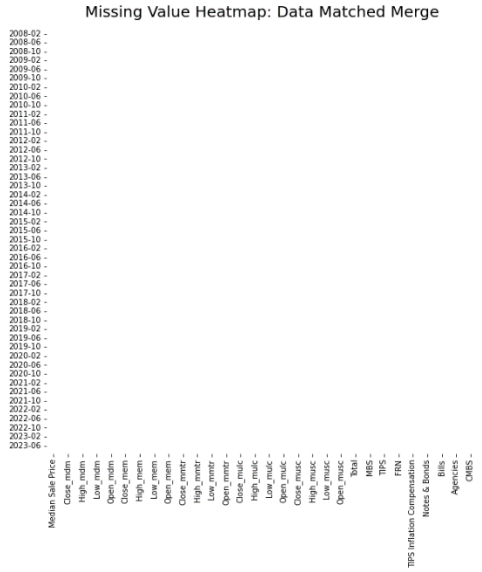


Fig. 3. A heatmap illustrating the missing values in the dataset.

Fig. 4. A heatmap illustrating the reduced and aligned dataset, based on DateTime index.

In the feature selection phase, we employed a correlation matrix to identify the most impactful features. Our objective was to maintain a ratio of one feature per ten data points, which led us to narrow down from 35 [Fig. 5.] features to 19 [Fig. 6.]. This refinement enhanced the efficacy of our models, ensuring that we utilized only the most relevant features in our analysis.

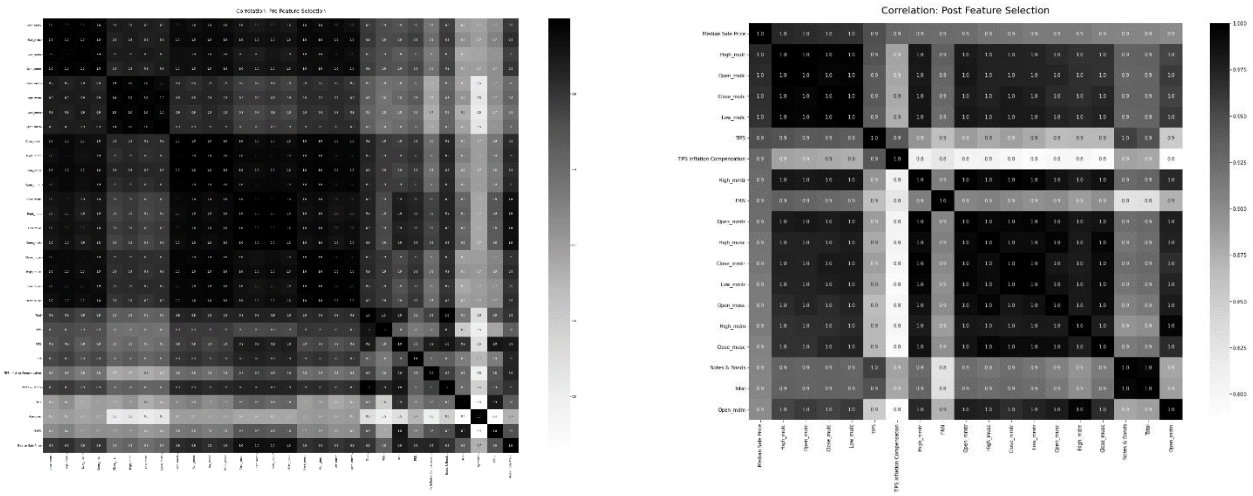


Fig. 5. A correlation matrix showing the original 38 features in the dataset, with color grading based on their correlation, darker indicating greater correlation between variables.

Fig. 6. A correlation matrix showing the reduced 19 features in the dataset, with color grading the same color grading.

The scaling of data was an essential step in our preprocessing. We utilized the StandardScaler from Scikit-Learn to standardize our features. This process involved subtracting the mean and scaling to unit variance, which is crucial in models that are sensitive to the scale of input data, such as KNN Regression. Scaling ensured that our models could interpret and process the data more effectively, leading to more accurate predictions.

Models

In our study, we employed three distinct machine learning models to forecast U.S. home prices: KNN Regression, Gradient Boosted Trees (GBT), and Decision Trees.

KNN Regression

KNN Regression operates on the principle of identifying the majority class among its k-nearest neighbors 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 and potentially over-smooth.
- Weights: Dictates whether each neighbor contributes equally ('uniform') or inversely to their distance ('distance').
- Algorithm: The method used to compute nearest neighbors ('ball_tree', 'kd_tree', 'brute', 'auto'), affecting computation speed and memory usage.
- Leaf Size: Impacts the speed of constructing and querying the tree, relevant for 'ball_tree' and 'kd_tree' algorithms.

Gradient Boosted Trees

GBT is an ensemble learning method that builds trees in a sequential manner. Each tree in the sequence focuses on correcting the errors made by the previous one, thereby continuously improving the model's accuracy. Key Hyperparameters:

- Learning Rate: Controls the contribution of each tree to the final outcome. A lower rate requires more trees but can yield a more generalized model.
- Number of Trees (n_estimators): The total number of sequential trees to be modeled. Too many trees can lead to overfitting.
- Max Depth of Trees: Determines the maximum depth of each tree. Deeper trees can capture more complex patterns but might overfit.
- Min Samples Split & Min Samples Leaf: These parameters control the minimum number of samples required to split an internal node and to be at a leaf node, respectively, affecting the tree's depth and overfitting.
- Max Features: The number of features to consider when looking for the best split, influencing the diversity of each tree.

Decision Trees

Decision Tree Regression in an ensemble context involves constructing multiple decision trees during training. The final prediction is derived by averaging the predictions of each individual tree. This approach generally results in a more accurate and stable model compared to using a single decision tree. Key Hyperparameters:

- Learning Rate: Similar to GBT, it controls the rate at which the model learns.
- Number of Trees (n_estimators): More trees increase accuracy but also computational cost.
- Maximum Depth of Trees: Sets the depth limit for each tree, impacting the model's complexity.

- Subsample: The fraction of samples used for fitting individual base learners, affecting the variance of the model.
- Min Samples Split & Min Samples Leaf: Define the conditions for further splitting of the trees, crucial for preventing overfitting.
- Max Features: Determines how many features each tree should consider, impacting the diversity and accuracy of the model.

Testing

The testing phase involved evaluating the baseline performance of each model, followed by hyperparameter tuning to optimize their performance. The hyperparameter tuning was conducted using GridSearchCV, which systematically worked through multiple combinations of parameter tunes, cross-validating as it went to determine which tune gives the best performance [Fig. 7.]. The results from the testing phase provided insights into the effectiveness of each model and their respective configurations. The Gradient Boosted Decision Tree emerged as the most effective model, achieving the highest R^2 value and the lowest RMSE after tuning.

Model	Base Score	Base RMSE	Tuned Score	Tuned RMSE
Decision Tree Regression	0.957	12154.0403	0.9626	11332.2089
KNN Regression	0.9853	7093.6745	0.9842	7371.8885
Gradient Boosted Decision Tree	0.9837	7480.8441	0.99	5852.3235

Fig. 7. This table illustrates the results of each algorithm. Base represents the model with no tuning and Tuned represent the model after tuning for the aforementioned hyperparameters.

Conclusion

Our study, leveraging machine learning and economic data analysis, has provided significant insights into the U.S. housing market. By meticulously preprocessing and aligning data from Treasury Securities Holdings, Morningstar Index Funds, and Zillow Median Home Sale Prices, we have successfully developed a robust dataset that captures the multifaceted nature of the housing market influenced by broader economic trends.

The implementation of machine learning models, namely KNN Regression, Gradient Boosted Trees, and Decision Tree Regression, has demonstrated the potential of predictive analytics in real estate forecasting. Our findings reveal that the Gradient Boosted Decision Tree, fine-tuned through hyperparameter optimization, emerged as the most effective model. This model's superior efficacy, evidenced by its high R^2 value and the lowest RMSE, underscores its capabilities as an accurate tool for our dataset.

However, our research also highlights areas for potential improvement. The exploration of additional models or hybrid approaches could provide further enhancements in predictive accuracy. Specifically noting that Gradient Boosted Decision Trees have a high computational power requirement which would be alleviated through further tuning of the package itself or a hybrid approach. Moreover, KNN Regression, where we found k = 2 being the optimal number, followed Gradient Boosted Decision Trees closely in results while having a lower to moderate computational intensity. If we were to continue this study a focus on performance would be key in ensuring a truly efficient model.

Sources

- 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/
- https://scikit-learn.org/stable/modules/generated/sklearn.tree.DecisionTreeRegressor.html#sklearn.tree.DecisionTreeRegressor
- https://scikit-learn.org/stable/modules/generated/sklearn.neighbors.KNeighborsRegressor.html#sklearn.neighbors.KNeighborsRegressor
- https://scikit-learn.org/stable/modules/generated/sklearn.ensemble.GradientBoostingRegressor.html