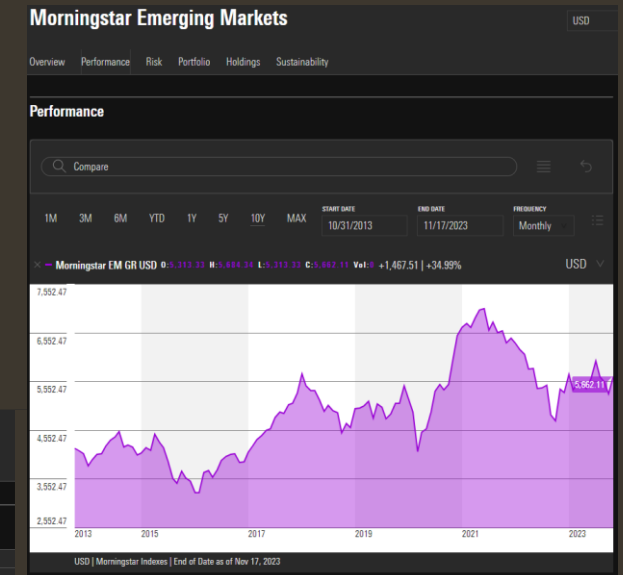
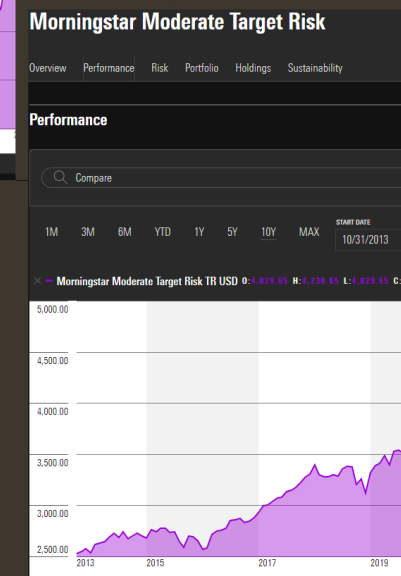
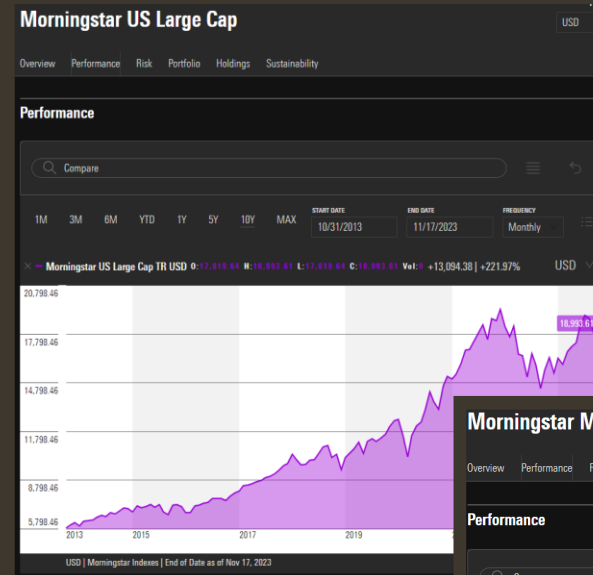


Forecasting U.S. Home Prices with Machine Learning: Insights from Morningstar Indexes and SOMA Securities



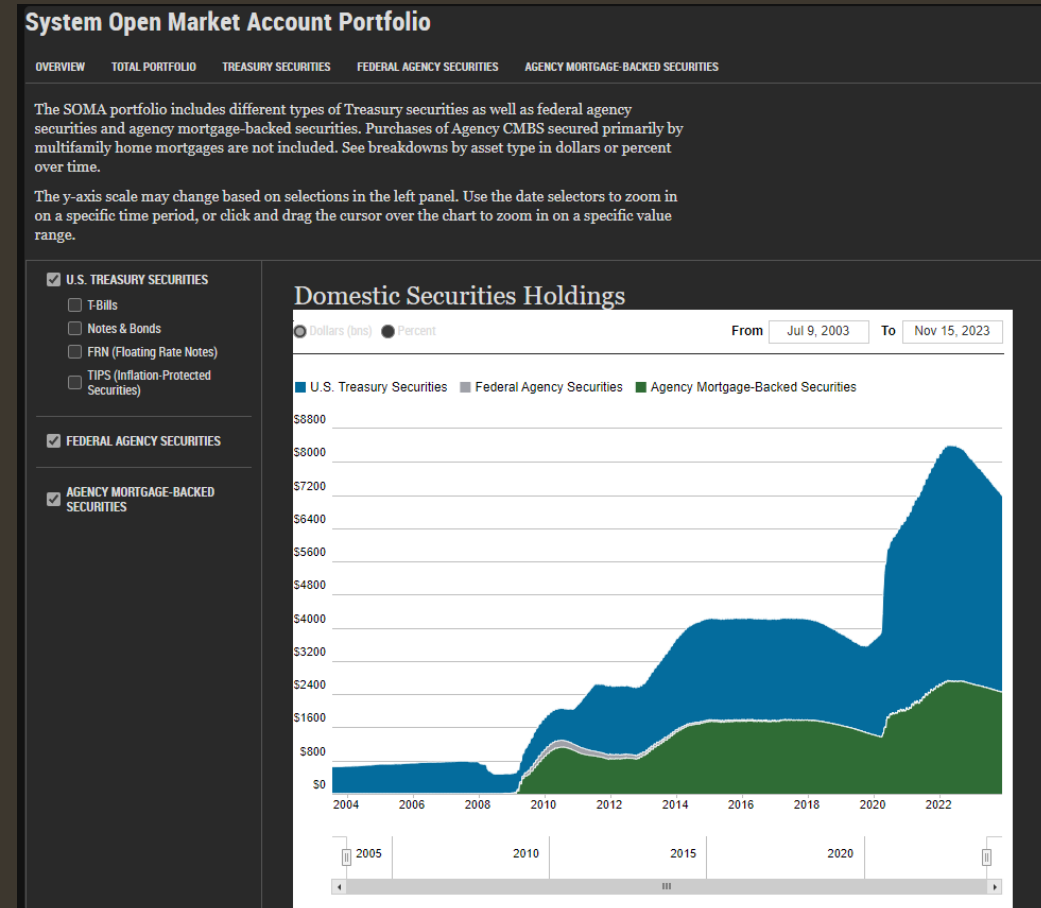
Datasets – Morningstar Index Funds

- Morningstar Index Funds
 - Moderate Target Risk
 - Developed Markets
 - Emerging Markets
 - US Large Cap
 - US Small Cap



Datasets – Treasury Securities Holdings

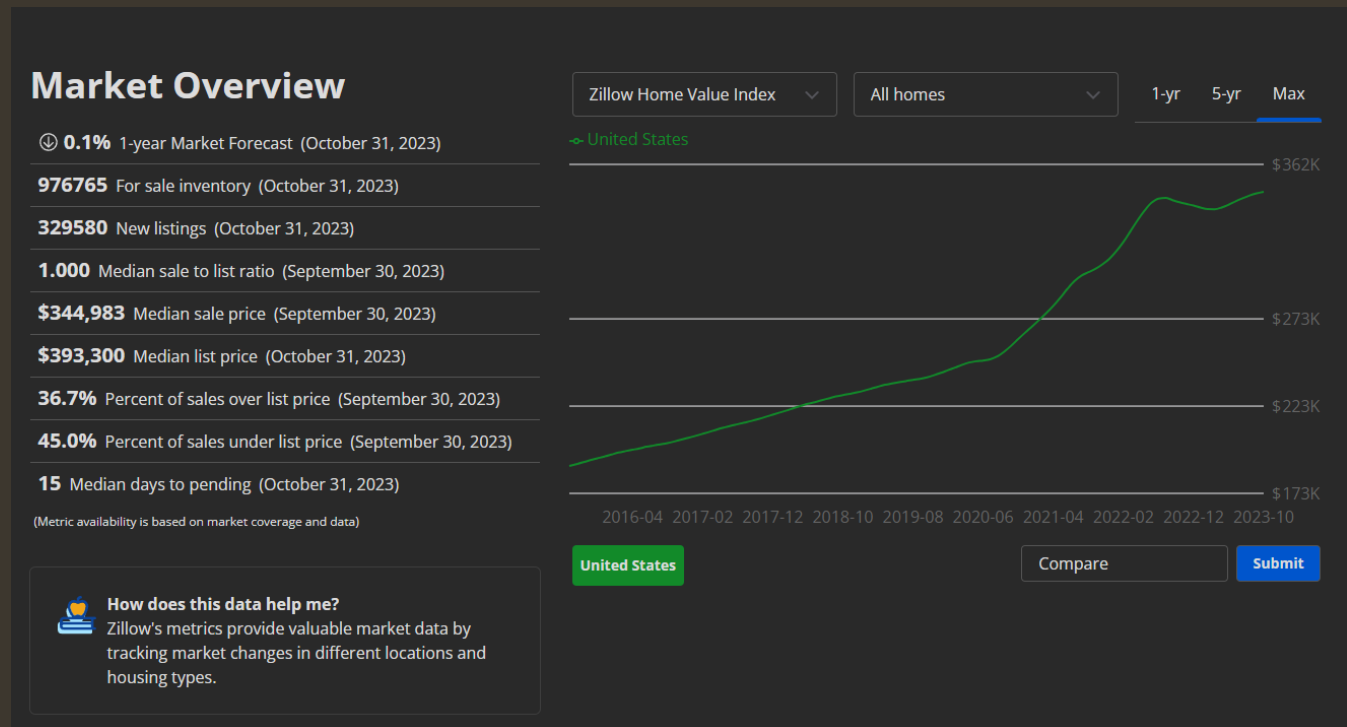
- Federal Reserve Bank of New York
 - System Open Market Account (SOMA)
Holding of Domestic Securities
 - U.S Treasury Securities
 - Federal Agency Securities
 - Agency Mortgage-Backed Securities



Datasets – Zillow Median Home Sale Price

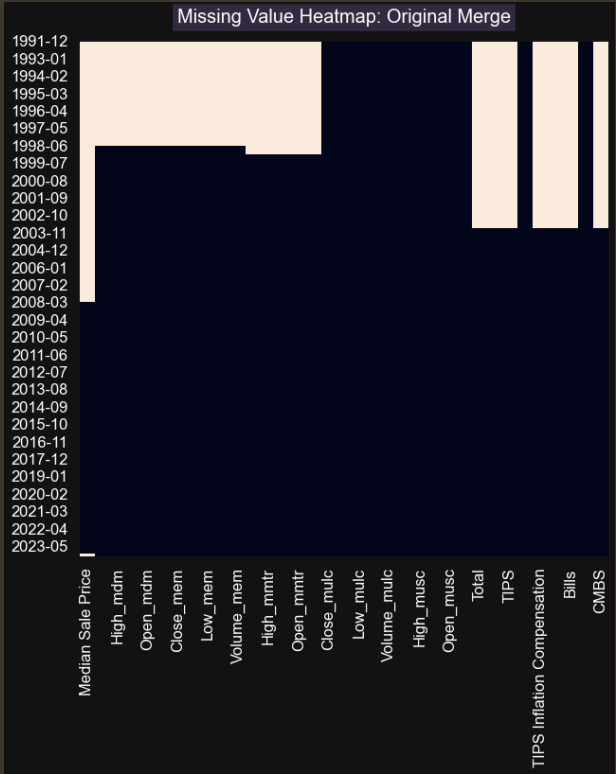
- Zillow Home Values Index
 - Median U.S. Home Sale Price

From Source: Zillow Home Value Index (ZHVI), built from the ground up by measuring monthly changes in property level Zestimates, captures both the level and home values across a wide variety of geographies and housing types.

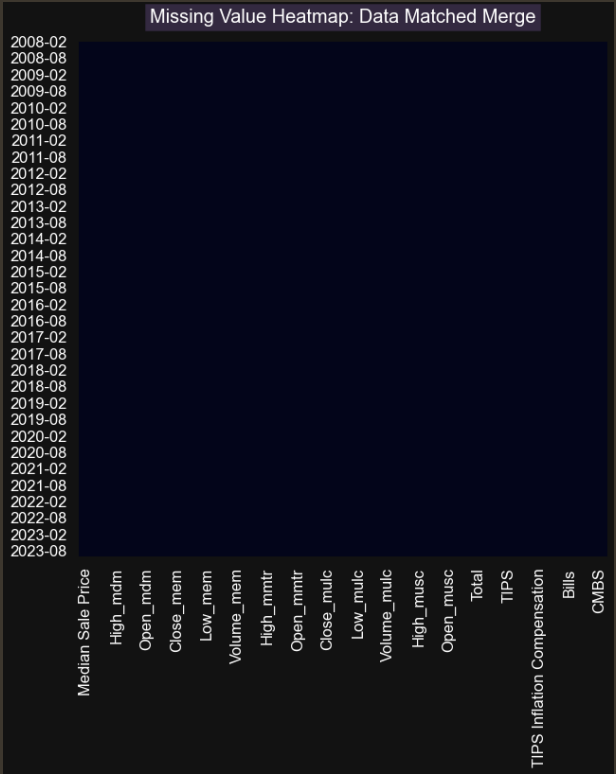


Data Preprocessing – Missing Data

Missing Values

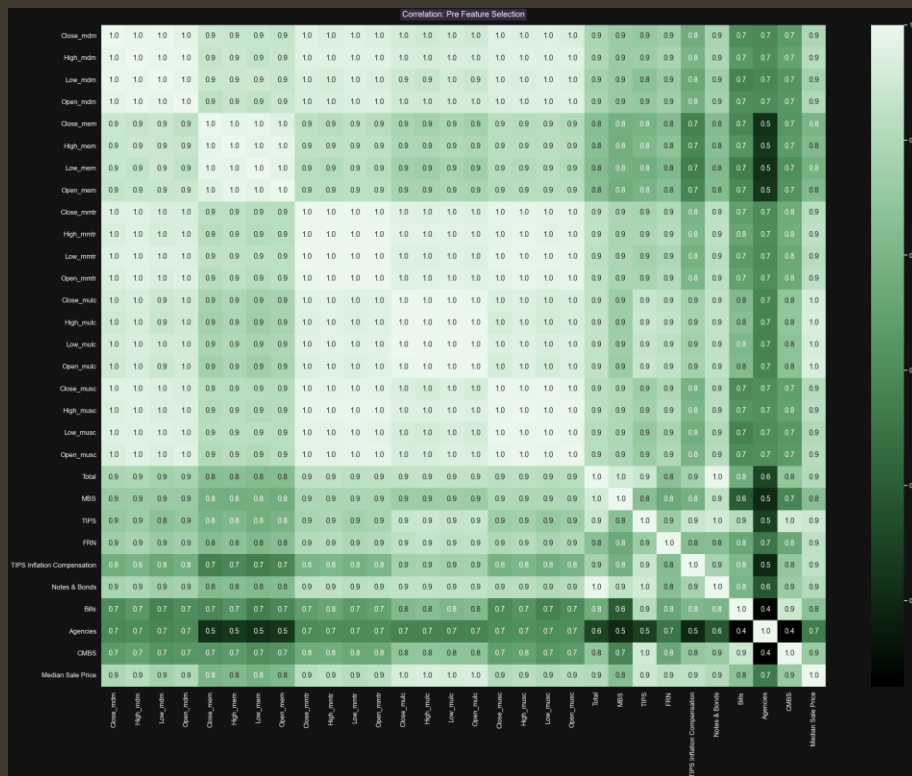


Data Alignment on Dates

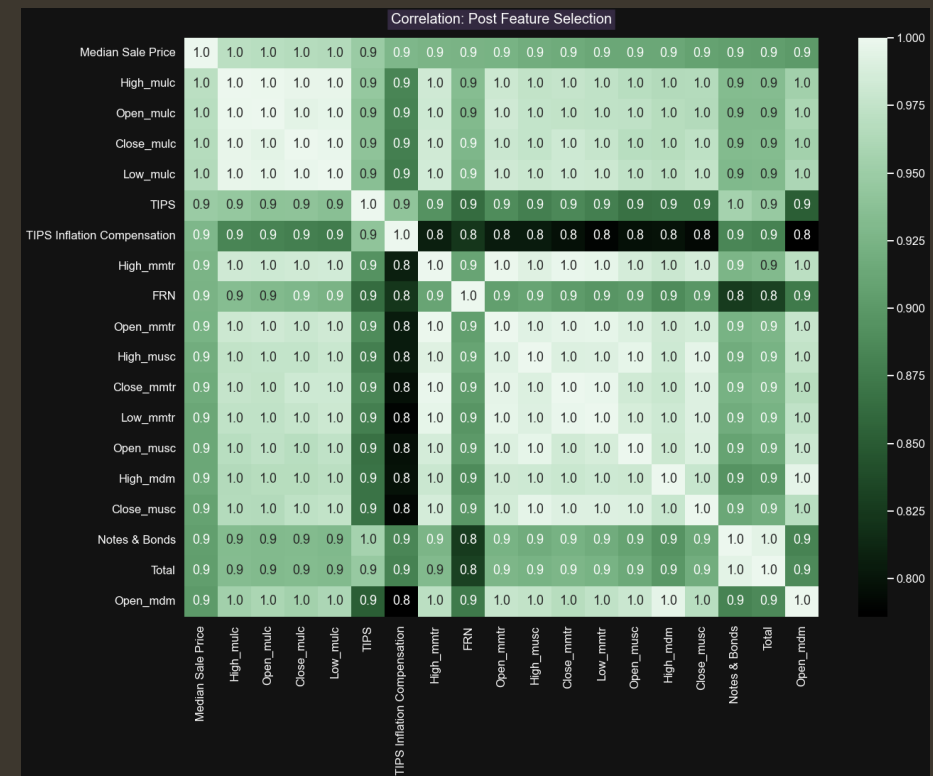


Data Preprocessing – Feature Selection

Correlation Matrix: Pre-Reduction



Correlation Matrix: Post-Reduction



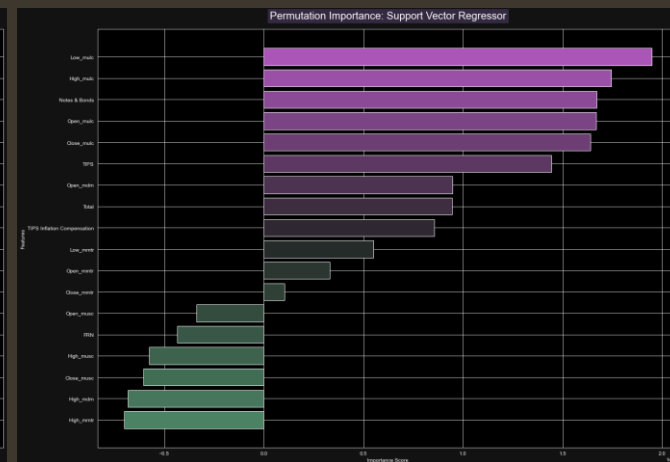
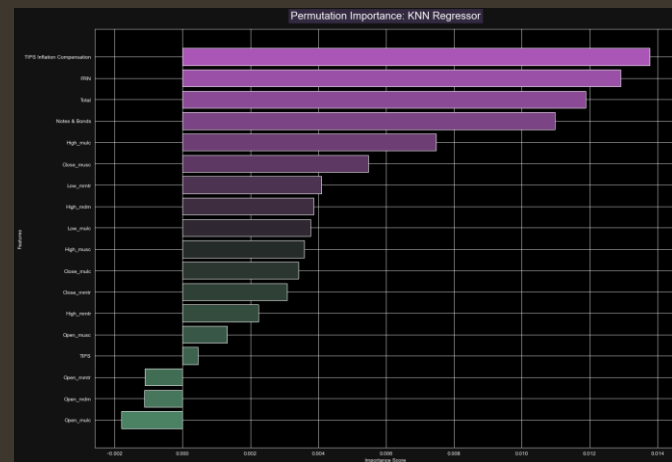
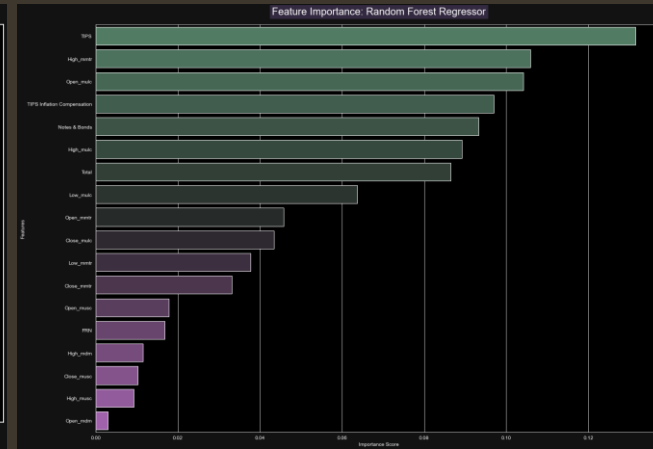
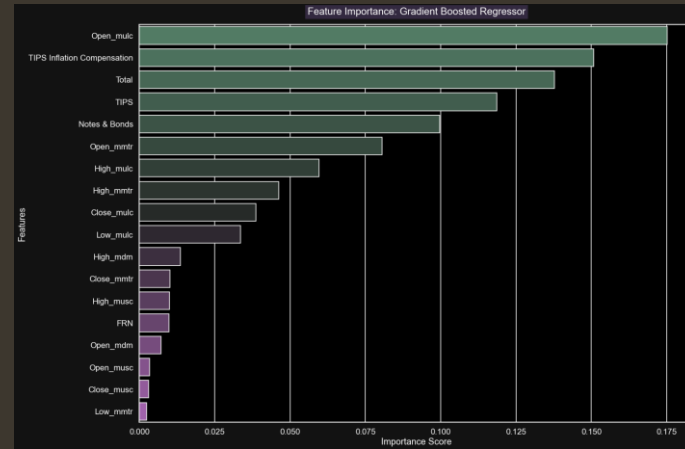
Models

- KNN Regression - It makes predictions by identifying the majority class (k) of its k-nearest neighbors in the feature space
- Gradient Boosted Trees - GBT is an ensemble learning method that builds trees sequentially. Each tree corrects the errors of the previous ones, improving the overall model.
- Support Vector Regression - SVR is a type of Support Vector Machine designed for regression tasks. It identifies a hyperplane that best represents the trend in the data. Effective for capturing complex relationships in high-dimensional spaces
- Random Forest - Random Forest is an ensemble learning method that builds multiple decision trees during training and merges them together to get a more accurate and stable prediction.

Hyperparameter Tuning

- GridSearchCV
- KNN: Key hyperparameter is the number of neighbors(k). Found the optimal k value $k = 2$. The higher in k we went the worse the model got.
- GBT: The number of trees, max depth of trees, minimum samples split and minimum samples leaf are the most important hyperparameters.
- RF: Learning rate, number of trees, maximum depth of trees, subsample (fraction of samples used for fitting) are the most important hyperparameters.
- SVR: C (regularization parameter), kernel type (linear, poly, rbf) , epsilon (tolerance for termination criterion) are the important hyperparameters

Feature/Permutation Importance



Gradient Boosted Regressor	KNN Regressor	Support Vector Regressor	Random Forest Regressor
Open_mulc	TIPS Inflation Compensation	Low_mulc	TIPS
TIPS Inflation Compensation	FRN	High_mulc	High_mmtr
Total	Total	Notes & Bonds	Open_mulc
TIPS	Notes & Bonds	Open_mulc	TIPS Inflation Compensation
Notes & Bonds	High_mulc	Close_mulc	Notes & Bonds

Feature/Permutation Importance

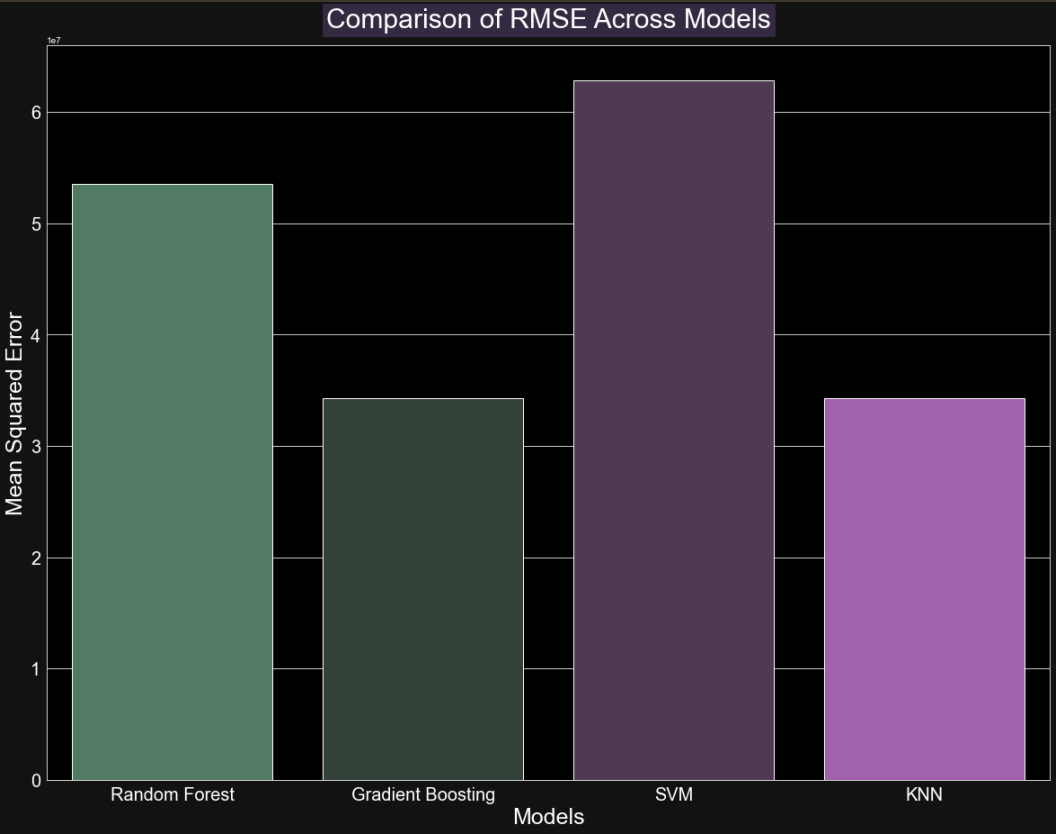
Top 5 Predictors Across Models

Computational Intensity

- Random Forest:
 - Computation Intensity: Moderate
- Gradient Boosting:
 - Computation Intensity: High
- Support Vector Machine (SVM):
 - Computation Intensity: Moderate to High
- K-Nearest Neighbors (KNN) Regression:
 - Computation Intensity: Low to Moderate

Model Efficiency Comparison

Model	R ²
Gradient Boosted Regressor	0.99
Random Forest Regressor	0.98
Support Vector Machine	0.98
KNN	0.98



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/>