SP/Case-Shiller U.S. National Home Price Index Prediction

Authors:

Xu, Yucheng

Date: December 22, 2023

1 Summary

This project aims to forecast the monthly change of SP/Case-Shiller U.S. National Home Price Index (HPI_{chg}). Based on data analysis, we formulate our prediction as the sum of HPI_{chg} today and a prediction on the first order difference of HPI_{chg}, in order to achieve lower bias. We apply rolling windows to train and validate Ridge model and LightGBM model respectively. Results show that the ensemble model of Ridge and LightGBM could produce a good forecast with RMSE 0.002273, 72% lower than that of the benchmark model. Besides, we found that seasonality features, time series features, money supply, interest rate, market instability, and economy play an important role in home price prediction.

2 Data Analysis

We define three variables:

$$\text{HPI: stands for SP/Case-Shiller U.S. National Home Price Index}$$

$$HPI_{chg,t} = HPI_t/HPI_{t-1} - 1 \tag{1}$$

$$\delta\text{HPI}_{chg,t} = HPI_{chg,t} - HPI_{chg,t-1}$$

Figure 1 shows that HPI has a strong trend. The result of ADF test on HPI in Figure 2 fails to reject the null hypothesis that it has a unit root. So HPI is not a stationary value. Therefore, we won't predict HPI. Instead, we forecast HPI_{chg}, which is a stationary value shown by ADF test.

Figure 3 shows that HPI_{chg} has high lag 1 autocorrelation, which is around 90%. Ideally, as long as we put lag 1 HPI_{chg} as a feature into the prediction model, there exist a solution which identically maps lag 1 HPI_{chg} to HPI_{chg} and ignores all other features. But in practice, it would be hard for the solver to find solutions that are comparably good like this. Leveraging the spirit of ResNet, we directly use lag 1 HPI_{chg} as our first forecast, and continuously build

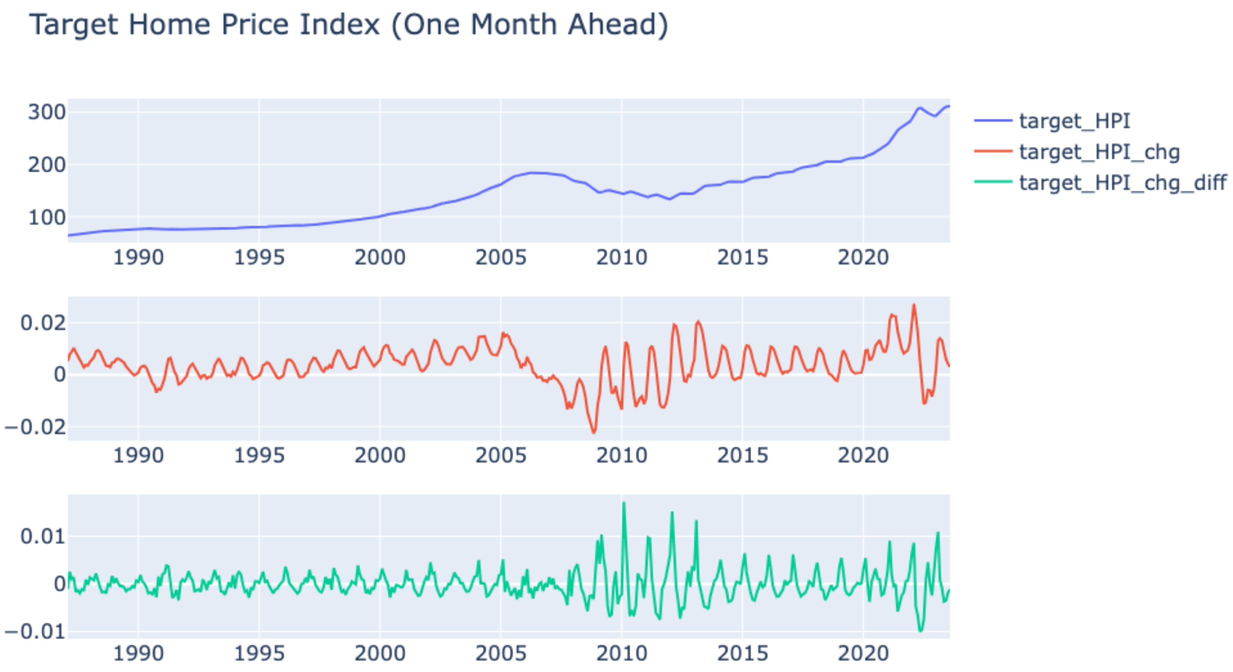
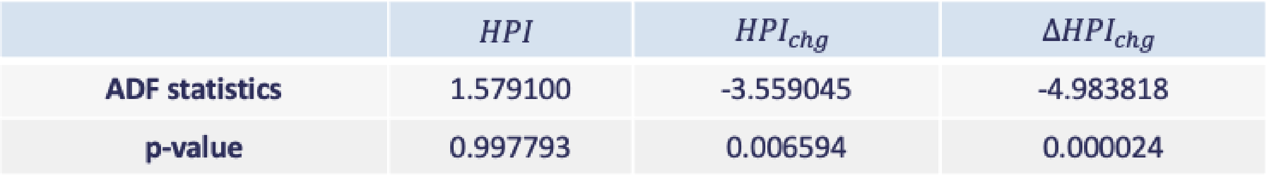


Figure 1: Plot of HPI, $\mathrm{HPI}_{\mathrm{chg}}$, $\delta\mathrm{HPI}_{\mathrm{chg}}$



	HPI	HPI_{chg}	ΔHPI_{chg}
ADF statistics	1.579100	-3.559045	-4.983818
p-value	0.997793	0.006594	0.000024

Figure 2: ADF test statistics of HPI, $\mathrm{HPI}_{\mathrm{chg}}$, $\delta\mathrm{HPI}_{\mathrm{chg}}$

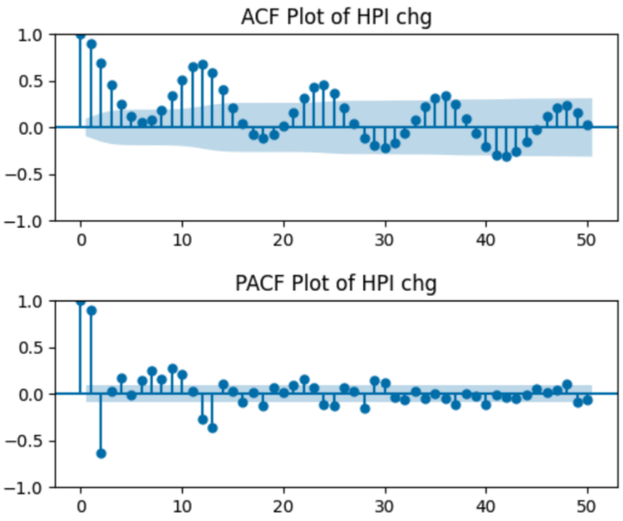


Figure 3: ACF and PACF plot of $\mathrm{HPI}_{\mathrm{chg}}$

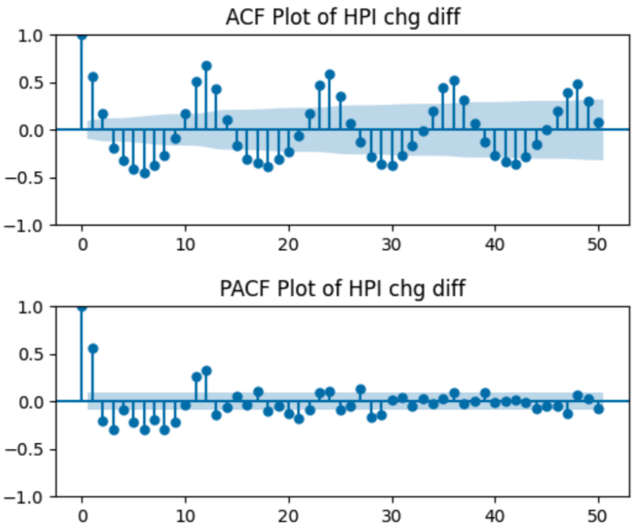


Figure 4: ACF and PACF plot of δ $\mathrm{HPI}_{\mathrm{chg}}$

model to predict the residual of it.

$$\text{predicted HPI}_{\text{chg,t}} = HPI_{chg,t-1} + residual_{t-1} \tag{2}$$

The residual is actually $\delta\text{HPI}_{\text{chg}}$, the first order difference of HPI_{chg}. From its ACF and PACF plots and ADF test statistics, we learn that this is a stationary time series with seasonality. The following part of the project will focus on predicting it.

3 Features

3.1 Features Included

In total, we include 127 features. A complete list of features is given in Appendix. In general, we try to find features that provide seasonality information, time series information, features that have an impact house demand or supply, features that serve as proxies of house demand and supply. They can be summarized into following groups.

- Seasonality Features
- Time Series Features
- Population
- Economy and Production
- Employment, Income, Asset
- Interest Rate and Loan
- Delinquency Rate and Default Risk
- Equity Market and Commodity Market
- Real Estate Related Investment
- House Supply

- Rent
- Interest Rate and Loan

And we also apply some statistical transformations on these features.

- **change**: the growth rate of a feature from last period to this period. Many features, like GDP, have strong trend. The change of these feature is closer to normal distribution. Besides, since we predict HPI_{chg}, the growth rate is more informative.
- **diff**: the first order difference of a feature.
- **ewma**: exponentially weight moving average of a features. Some features are daily. We calculate their ewma to incorporate more information while considering the information decay in time.
- **std** and **ewma std**: standard deviation of a features. Since volatility is persistent in financial world, std of a features would help predict the magnitude of true label.

3.2 Features' Correlation with ylabel

Out of 127 features in total, 24 of them have a correlation with y label higher than 10%. Time series features, seasonality features, income and asset, interest rate, and money supply are those highly correlated with true label.

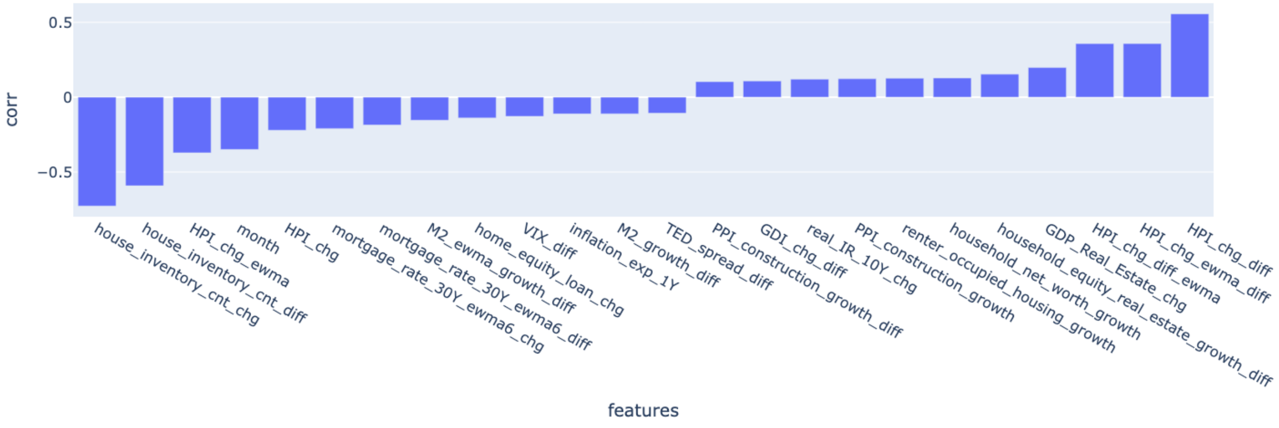


Figure 5: Correlation between features and y label (δHPI$_{chg}$)

3.3 Correlation between Features

Figure 6 illustrates the correlation between features, where cold color implies low correlation and warm color implies high correlation. While most of the features are weakly correlated, there are indeed a few highly correlation features, namely HPI chg and HPI chg ewma. So we should be careful with the multicolinearity problem when modeling.

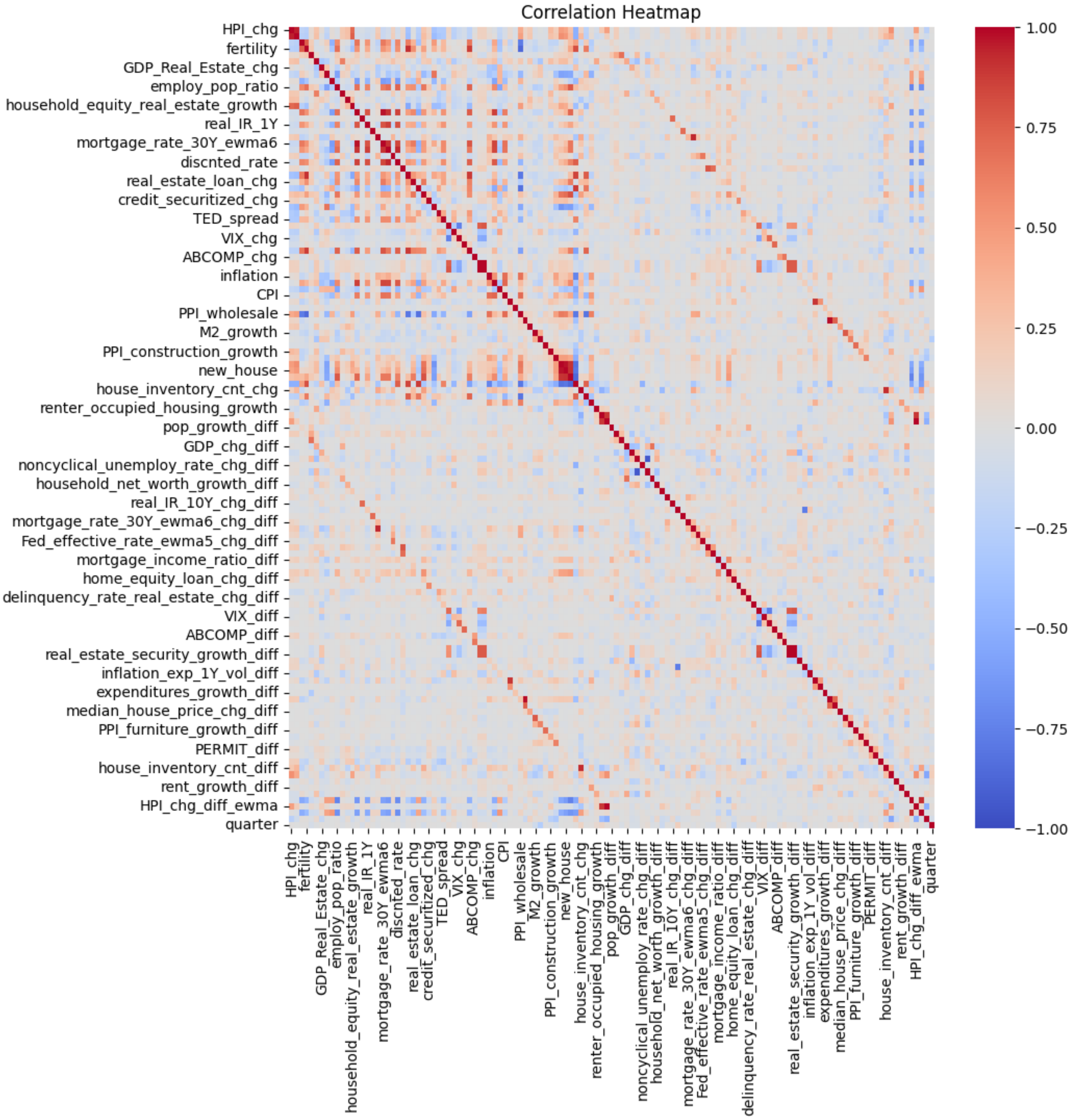


Figure 6: Correlation between features

4 Modeling

In this section, we will first discuss the models we use, including Ridge and LightGBM. Following that, we explain our model training and validation methods, as well as the model evaluation metric.

4.1 Linear Regression: Ridge

4.1.1 Reason to choose the model

- Linear model has few parameters. It is a robust model on our small dataset with only 440 data points.
- Ridge regression add a l2 penalty term to shrink the coefficient. This reduces overfitting and increase the model generalizability.
- Ridge makes sure that solution exists when there are more regressors than data points.

4.1.2 Data Processing

- **Outliers**: Ridge regression is sensitive to outliers. Ideally, the minimum zscore and maximum zscore of each features should be around -3 and 3. But many of the features has fatter tails. Therefore, we calculate the 1% and 99% percentile of each feature based on only training data to avoid forward looking. Then, we clip the data points outsides this range in both training and testing sets.
- **Missing Value**: Missing value of training set and testing set are filled by the average value of this feature in the training set.
- **Categorical Features**: We apply one-hot-encoding to categorical features, including month and quarter. One category is dropped for each feature to avoid perfect collinearity.

4.1.3 Feature Selection

- **Dimension Reduction**: Many of our features are correlated with each other. Multicolinearity will make the inverse of second moment of input data unstable, and thus the standard error will of coefficient estimators will be larger and the forecast will be less accurate. However, if we simply drop some of correlated features, we will lose some useful information. So we apply PCA to reduce the dimensions and make sure the regressors are orthogonal.

  Specifically, We first standardize the data, because PCA is affected by scale. Then we take the first n components that could explain 99% of the variance. Both the StandardScaler and PCA model are fitted on the training dataset, and then applied on both training and testing sets, in order to avoid forward looking.

- **t-statistics**: Including irrelevant regressors will increase the standard error of estimators and lead to inaccurate prediction. So before training the Ridge model, we first fit a OLS. We only keep the regressors with a p-value smaller than 1%. With a 99% of chance, these regressors are statistically significant. It should be noticed that we use White standard error, considering that the residual of time series are heteroskedastic.

4.2 Tree-based Model: LightGBM

4.2.1 Reason to choose the model

- **Nonparametric**: LightGBM is nonparametric. It does not rely strong assumptions regarding the shape of the relationship between $\Delta \text{HPI}_{\text{chg}}$ and features. They are also able to detect non-linear relationships, which Ridge fails to do.

- **Good Interpretability**: The weak learner of LightGBM is decision tree. Its hierarchical decision process has good interpretability.

- **Low Bias**: LightGBM is a boosting model. Each weak learner fits the residual of it previous model. So LightGBM has lower bias than decision tree. Compared with Gradient

Boosting Decision Tree, it has regularization terms in objective function to reduce overfitting. It uses higher order Taylor expansion to more accurately estimate the objective function. Compared with XGBoost, it focuses more on samples with larger residual in previous model to further reduce bias.

4.2.2 Feature Selection

Considering that the number of features is not large for LightGBM, we don't select features anymore. But we leverage SHAP (SHapley Additive exPlanations) to evaluate the performance of each feature. SHAP values measure the contribution of each feature in our LightGBM model. It has an addictive nature: SHAP values of all features sum up to the difference baseline prediction and current model prediction. Therefore, we define **RMSE gain** to measure the contribution of a feature to our RMSE decrease:

$$\begin{gathered}\text{prediction without feature k} = \text{prediction} - \text{SHAP of feature k}\\ \text{RMSE without feature k} = RMSE(\text{prediction without feature k}, \text{true label})\\ \text{RMSE gain} = \text{RMSE without feature k} - \text{RMSE}\end{gathered} \tag{3}$$

RMSE gain is calculated on a valid set. Higher RMSE gain means larger contribution to a good model prediction.

4.3 Model Training and Testing Methods

We applied a monthly rolling window in modeling training and testing, in order to alleviate the regime change problem. In each month, the training dataset is the most recent available data with a size no larger than maximum training data size, and the testing data is the one month ahead data point. We set the minimum training data size as 60. The maximum training data size for Ridge is set to be 180, while that of LightGBM is set to be 440. This is because LightGBM is able to handle larger dataset.

In the case of LightGBM model, a validation set is needed for calculating SHAP and best number of estimators. In each rolling window, we train two models:

- **Model 1**: trained on first 80% of the training set and validated on the rest 20% of the training set. Only used for estimating best number of estimators and feature performance.

- **Model 2**: trained on the whole training set. Used for prediction.

4.4 Model Evaluation Metric

RMSE is the primary metric for model evaluation in our project. This is for two reasons:

- RMSE is convex, easy for algorithm to learn.

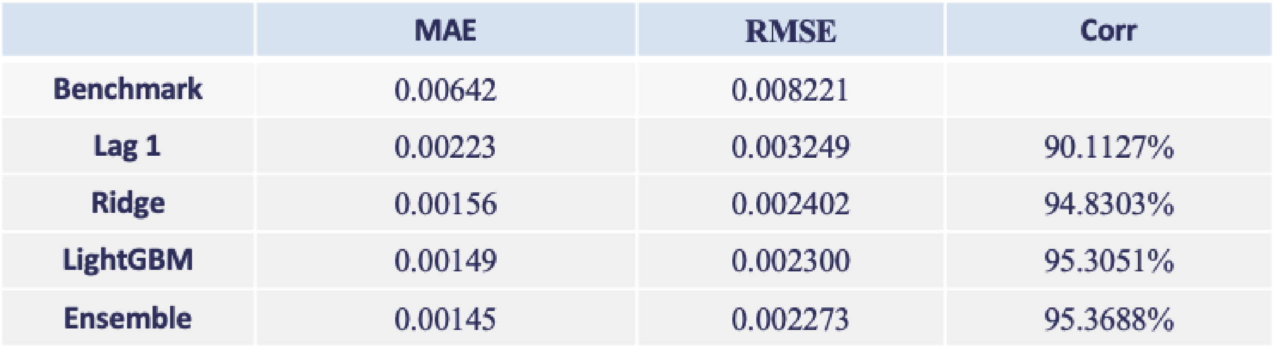
- Intuitively, we want to penalize larger error more.

MAE and correlation between true label and prediction are also provided.

5 Result and Findings

5.1 Model Performance

This section present the model performance in outsample prediction.



	MAE	RMSE	Corr
Benchmark	0.00642	0.008221	
Lag 1	0.00223	0.003249	90.1127%
Ridge	0.00156	0.002402	94.8303%
LightGBM	0.00149	0.002300	95.3051%
Ensemble	0.00145	0.002273	95.3688%

Figure 7: Model Performance

To clarify, true label is the HPI change. Benchmark model always predicts 0. Lag 1 model use HPI change in this month as prediction. Ensemble model use the simple average of Ridge prediction and LightGBM prediction as prediction.

Both Ridge and LightGBM capture extra preditive information besides lag 1 HPI change. LightGBM performs better than Ridge, as it is able to discover the non-linear relationship between features and true label. The ensemble model has the most robust performance, which implies that Ridge and LightGBM have slightly difference focus when exploiting the features.

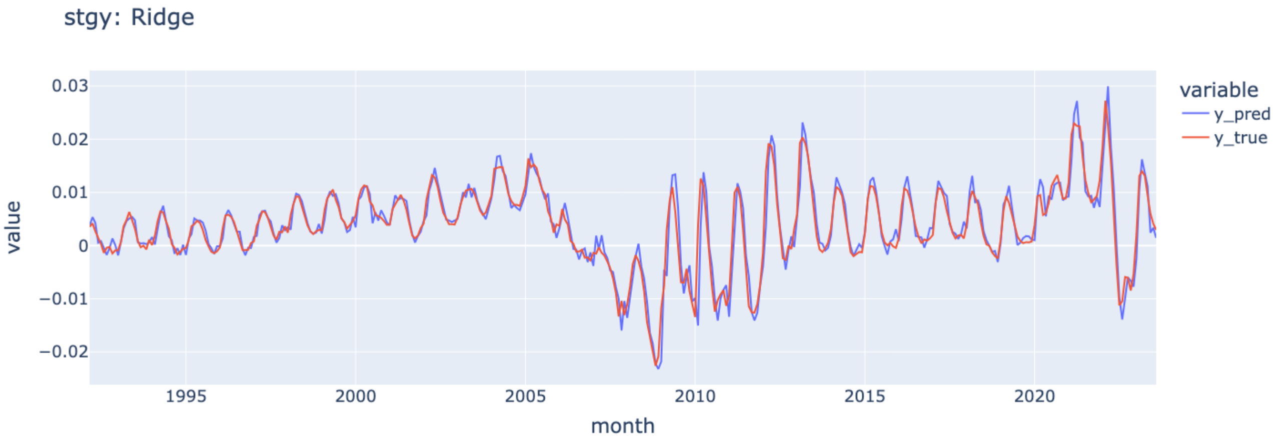


Figure 8: Prediction of Ridge

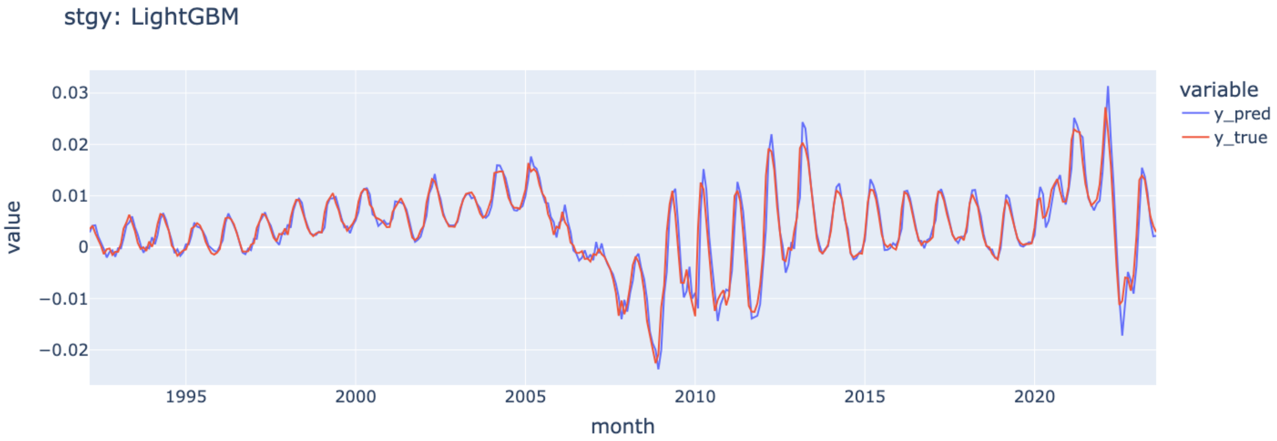


Figure 9: Prediction of LightGBM

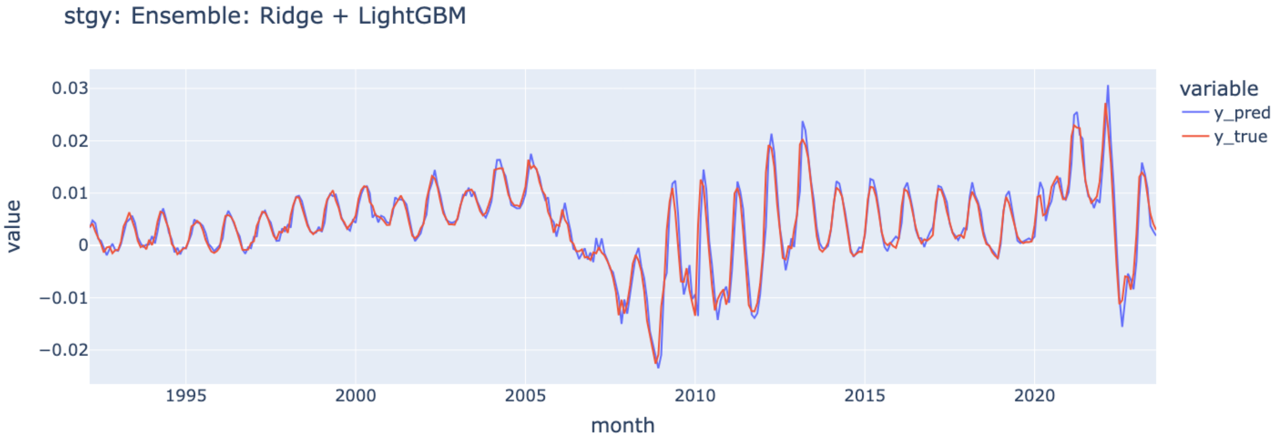


Figure 10: Prediction of Ensemble model

5.2 Features Performance

We are also interested in figuring out the features with large contribution to a good model prediction. Since LightGBM outperforms Ridge, we focus on the features contribution in LightGBM. Figure 11 visualize the RMSE Gain defined in Section 4.2.2 of the top 20 features. Since feature 'month' has a RMSE gain far larger than those of other features, we also provide Figure 12 which excludes feature 'month'.

To summarize, seasonality features, time series features, money supply, interest rate, market instability, and economy play an important role in home price prediction. Such a result makes sense fundamentally. More money supply means more money available for investment. It also potentially leads to high inflation, and thus further stimulates people's investment willingness. Interest rate reflects the fund cost of house purchasing. Besides, people are more likely to buy house in a stable and expanding economy.

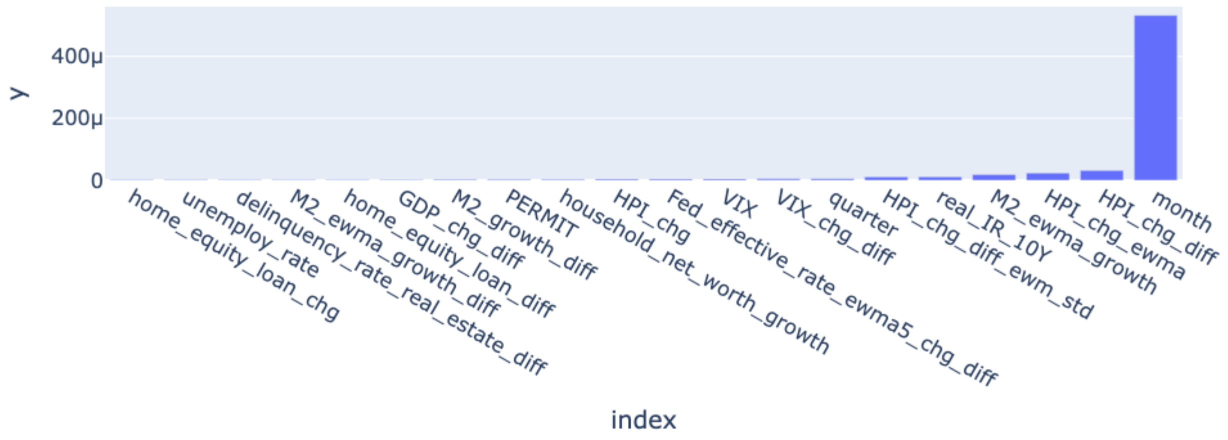


Figure 11: RMSE gain

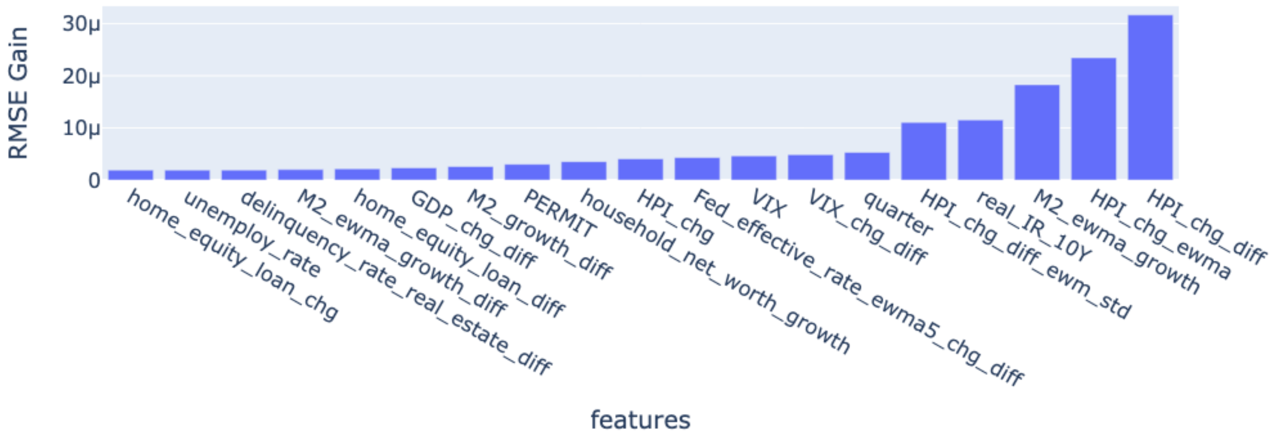


Figure 12: RMSE gain (exclude feature ’month’)

Appendices

A Features

A.0.1 Seasonality Features

- **month**: current month
- **quater**: current quarter

A.0.2 Time Series

- **HPI chg**: monthly change of HPI
- **HPI chg ewma**: exponentialy weighted average of **HPI chg**

A.0.3 Population

- **pop growth**: Annual growth rate of population
- **working pop growth**: Monthly growth rate of working age population
- **fertility**: Annual fertility rate

A.0.4 Economy and Production

- **GDP chg**: Seasonal GDP growth rate
- **GDP Real Estate chg**: Growth rate of value added of real estate, rental, and leasing in GDP

A.0.5 Employment, Income, Asset

- **unemploy rate**: Unemployment rate
- **noncyclical unemploy rate chg**: Quarterly change of unemployment rate which excludes the unemployment arising from fluctuation in aggregate demand

- **GDI chg**: Quarterly growth rate of gross domestic income
- **household net worth growth**: Quarterly growth rate of households' net worth
- **household equity real estate growth**: Quarterly growth rate of households' equity in real estate

A.0.6 Interest Rate and Loan

- **real IR 10Y**: 10-Year real interest rate
- **real IR 10Y chg**: monthly change of 10-Year real interest rate
- **real IR 1Y**: 1-Year real interest rate
- **real IR 1Y chg**: monthly change of 1-Year real interest rate
- **mortgage rate 30Y ewma6**: exponentially moving average of 30-Year fixed rate mortgage average. The half life decay window is 6 days.
- **mortgage rate 30Y change**: monthly change of **mortgage rate 30Y ewma6**
- **Fed effective rate ewma5**: exponentially moving average of Federal Funds effective rate, which is the average interest rate at which commercial banks borrow and lend their reserve to each other. The half life decay window is 5 days.
- **Fed effective rate ewma5 chg**: monthly change of **Fed effective rate ewma5**
- **discnted rate**: discounted rate, the rate hich Fed charges for short-term loan
- **discnted rate chg**: monthly change of **discnted rate**
- **mortgage income ratio**: mortgage payments as a percent of personal income
- **real estate loan ewma**: exponentially moving average of real estate loan in all commercial banks
- **home equity loan**: home equity loan

- **home equity loan chg**: quarterly growth of home equity loan
- **credit securitized chg**: monthly change of total consumer credit owned and securitized

A.0.7 Delinquency Rate and Default Risk

- **delinquency rate real estate**: delinquency rate on commercial real estate loans
- **delinquency rate real estate chg**: quarterly change of **delinquency rate real estate**
- **TED spread**: the difference between the interest rates on interbank loans and on short-term U.S. government debt (T-bills), which measure default risk

A.0.8 Equity Market and Commodity Market

- **Wilshire 5000 chg**: change of Wilshire 5000 index
- **VIX**: VIX
- **VIX chg**: change of VIX
- **gold chg**: change of PPI in gold ore and silver ore mining industry

A.0.9 Real Estate Investment

- **ABCOMP**: asset-backed commercial paper outstanding
- **ABCOMP chg**: change of **ABCOMP**
- **real estate trust growth**: growth of Wilshire US Real Estate Investment Trust Total Market Index
- **real estate security growth**: growth of Wilshire US Real Estate Securities Price Index

A.0.10 Inflation, Price Index, Consumption

- **inflation**

- **inflation exp 1Y**: 1-Year expected inflation
- **inflation exp 1Y vol**: Exponentially moving standard deviation of **inflation exp 1Y** with a half life decay window of 3 months
- **CPI**
- **CPI growth**: monthly change of CPI
- **expenditures growth**: annual change of household operation expenditures
- **PPI wholesale**: PPI in wholesale trade industry
- **PPI wholesale growth**: monthly growth of **PPI wholesale**
- **median house price chg**: quarterly change of median sales price of houses sold

A.0.11 Money Supply

- **M2 growth**: monthly change of M2
- **M2 ewma**: exponentially weighted moving average of M2, with a half life decay window of 1 day
- **M2 ewma growth**: monthly change of **M2 ewma**

A.0.12 House Supply

- **PPI furniture growth**: monthly change of PPI of household furniture manufacturing industry
- **PPI construction growth**: monthly change of PPI of construction materials
- **PPI concrete growth**: monthly change of PPI of cement and concrete product manufacturing industry
- **PERMIT**: total unit of new privately-owned housing units authorized in permit-issuing places

- **new house**: total unit of new privately-owned housing
- **house inventory chg**: total unit of housing inventory
- **house inventory cnt**: active listing count of housing inventory
- **house inventory cnt chg**: monthly change of **house inventory cnt**

A.0.13 Rent

- **rental vacancy**: rental vacancy rate
- **rent growth**: monthly growth of CPI of rent
- **renter occupied housing growth**: quarterly growth of renter occupied housing units

A.0.14 Others: First Order Difference

- **feature diff**: Here **feature** refer to all features mentioned above, except categorical features. Since we predict $\delta\ HPI_{chg}$, we calculate the first order difference of these features.
- **HPI chg diff std**: rolling standard deviation of HPI_{chg} in last 24 months
- **HPI chg diff ewma**: exponentially weighted moving average of **HPI chg diff** with a half life decay window of 1 month
- **HPI chg diff ewma std**: exponentially weighted moving standard deviation of **HPI chg diff** with a half life decay window of 3 month