Faculty of Economic Sciences | Master's Thesis Predefence | Moscow, 2024

Forecasting Real Estate Prices with US Market Data Using Machine Learning Methods

Author: A.A. Lebedev, MSF222
Scientific supervisor: Candidate of Sciences S.V. Kurochkin, Associate Professor

1. Research Problem, Motivation and Tasks
2. Literature Review
3. Data Sources and Variables
4. Pearson and Spearman Correlations
5. Feature Generation and Scaling
6. Research Hypotheses
7. Final Model Architecture
8. Prediction Models and Interpretation with Shapley Values
9. Results of the Modelling
10. Shapley Values
11. Conclusions
12. Questions of the Reviewer

Research problem:
Determine the feasibility of employing datasets containing weekly residential real estate price indices to predict their values using machine learning techniques on a large (in terms of market size) statistical areas of the USA for the 1, 3 and 6 months in advance (4, 13 and 26 weeks accordingly)

Motivation:
Need for a weekly models, which would be able to forecast real estate prices from the US regional markets on the horizon from one to six month with an adequate level of accuracy and explanatory power

Tasks:
- Ascertain which data could be used for the prediction of the real estate market in the US, both as an independent variable and as a target variable
- Define which machine learning methods can be applied for the forecasting of weekly time series and which methods can be used to prepare data for modelling based on previous research in finance, economics and computer science
- Estimate the models on the available data, compare their performance on the different prediction horizons with OLS benchmark and understand how predictions depend on the macro data

Implications:
- For scientific researchers - it can be demonstrated that weekly data may be employed for the highly detailed and regional prediction of the markets with lower levels of liquidity, such as regional real estate
- For analytics in the real estate sphere - the results demonstrate that weekly regional indices can be highly beneficial and worthy of calculation when real estate aggregators are available for specific countries
- For businesses - the findings provide significant assistance to businesses in the development of company risk management procedures and in the optimisation of decision-making processes for real estate deals

Data / Models	Group	Sources	References
Data	Past property sales, including sale prices, property features, and sale dates	Local private or governmental real estate aggregators	Kuşan et al., 2010; Park & Bae, 2015; Walthert & Sigrist, 2019; Ma et al., 2018; Chiu et al., 2021
	Location features such as neighbourhood, proximity to amenities, transportation infrastructure, and local economic conditions	Geoinformation systems	Park & Bae, 2015; Dimopoulos et al., 2018; Lee et al., 2022; Mubarak et al., 2022
	Information about the large districts or the whole regions	Regional private or governmental real estate aggregators	Case & Shiller, 1990, 1993; Schindler, 2011; Li & Chu, 2017; Alfaro-Navarro et al., 2020; Chou et al., 2022
	Economic indicators such as interest rates, inflation rates, unemployment rates, GDP growth, and consumer confidence can influence real estate markets	Private or governmental statistical aggregators	Park & Bae, 2015; Meharie et al., 2021
	Market trends, such as inventory levels, days on market, housing supply and demand dynamics, and sentiment indicators	Private or governmental real estate aggregators	Wang et al., 2014; Hausler et al., 2018; Ma et al., 2018
	Demographic data such as population growth, household income levels, age distribution, and migration patterns	Private or governmental statistical aggregators	Reed, 2016
	Information on seasonal and cyclical patterns, with prices and sales activity varying throughout the year and over longer economic cycles	Private or governmental statistical aggregators	Lee, et al., 2022
Models	Gradient boosting		Kuşan et al., 2010; Dimopoulos et al., 2018; Bentéjac et al., 2020
	Recurrent neural networks		Sutskever et al., 2014; Salehinejad et al., 2018; Lee et al., 2022
	Deep neural networks		Canziani et al., 2017; Li & Chu, 2017
	Support vector machines		Wang et al., 2014; Chou et al., 2022
	Stacking		Pavlyshenko, 2018; Alfaro-Navarro et al., 2020; Meharie et al., 2021

FRED ECONOMIC DATA | ST. LOUIS FED

KeyRate - Federal Funds Effective Rate

CPI - US CPI

VIX - US VIX by CBOE

PPI - US PPI

MortgageRate30 - US 30-year mortgage rate

Electricity, Water, Plywood, Steel, Glass, Concrete - US building materials indices

Unemployment - US unemployment rate

Yield10Y - Return for 10-year treasuries

Case-Shiller* - Household price index for the US - target variable for the additional test

yahoo! finance

DJI - Dow Jones index

S&P500 - S&P500 index

HAUS

CHPI* - Household price index** for top-100 statareas of the US - target variable

*CHPI - main target variable, Case-Shiller - target variable for testing via transfer learning

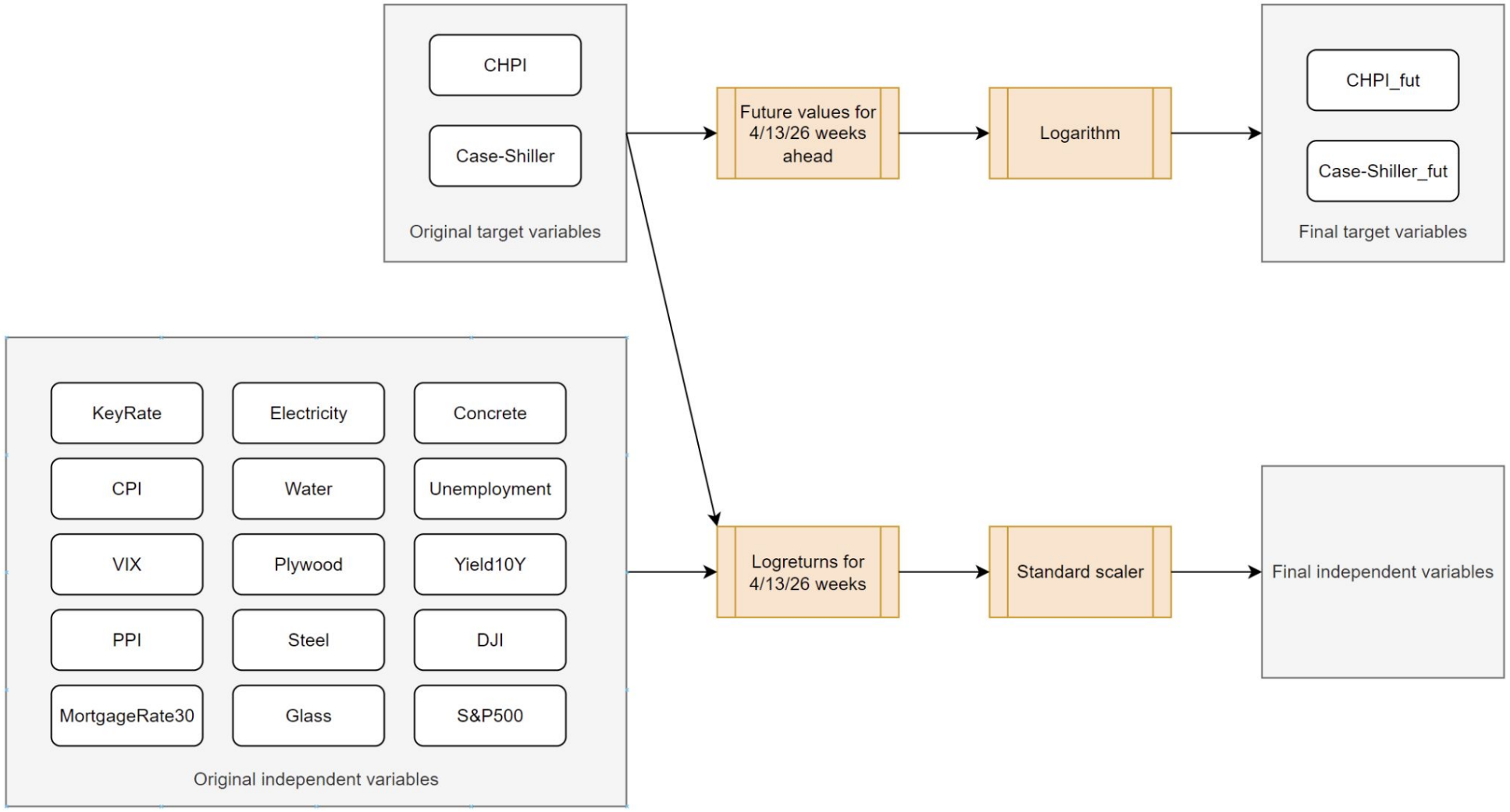
** Standard home for the index - three-bedroom, two-bathroom, 1,500-square-foot (140 square meters) home built in 1977 on a quarter-acre (1012 square meters) lot

Pearson Correlation

	CHPI	KeyRate	CPI	VIX	PPI	MortgageRate30	Electricity	Water	Plywood	Steel	Glass	Concrete	Unemployment	Yield10Y	DJI	S&P500
CHPI	1	0.41	0.95	0.43	0.88	0.11	0.85	0.98	0.78	0.66	0.97	0.97	-0.4	-0.05	0.97	0.96
KeyRate	0.41	1	0.45	-0.02	0.12	0.63	0.46	0.46	0.04	0.07	0.45	0.5	-0.6	0.54	0.39	0.29
CPI	0.95	0.45	1	0.44	0.9	0.28	0.95	0.94	0.79	0.76	0.98	0.99	-0.42	0.1	0.92	0.93
VIX	0.43	-0.02	0.44	1	0.42	0	0.38	0.42	0.32	0.26	0.43	0.44	0.31	-0.29	0.3	0.34
PPI	0.88	0.12	0.9	0.42	1	0.05	0.82	0.83	0.95	0.84	0.89	0.86	-0.26	-0.07	0.87	0.91
MortgageRate30	0.11	0.63	0.28	0	0.05	1	0.48	0.12	0.03	0.17	0.2	0.28	-0.4	0.9	0.04	0
Electricity	0.85	0.46	0.95	0.38	0.82	0.48	1	0.84	0.72	0.73	0.91	0.93	-0.46	0.29	0.79	0.81
Water	0.98	0.46	0.94	0.42	0.83	0.12	0.84	1	0.71	0.58	0.97	0.97	-0.44	-0.05	0.96	0.95
Plywood	0.78	0.04	0.79	0.32	0.95	0.03	0.72	0.71	1	0.82	0.78	0.74	-0.23	-0.02	0.79	0.82
Steel	0.66	0.07	0.76	0.26	0.84	0.17	0.73	0.58	0.82	1	0.72	0.67	-0.25	0.12	0.69	0.74
Glass	0.97	0.45	0.98	0.43	0.89	0.2	0.91	0.97	0.78	0.72	1	0.99	-0.45	0.03	0.95	0.95
Concrete	0.97	0.5	0.99	0.44	0.86	0.28	0.93	0.97	0.74	0.67	0.99	1	-0.45	0.08	0.92	0.93
Unemployment	-0.4	-0.6	-0.42	0.31	-0.26	-0.4	-0.46	-0.44	-0.23	-0.25	-0.45	-0.45	1	-0.51	-0.45	-0.41
Yield10Y	-0.05	0.54	0.1	-0.29	-0.07	0.9	0.29	-0.05	-0.02	0.12	0.03	0.08	-0.51	1	-0.06	-0.1
DJI	0.97	0.39	0.92	0.3	0.87	0.04	0.79	0.96	0.79	0.69	0.95	0.92	-0.45	-0.06	1	0.99
S&P500	0.96	0.29	0.93	0.34	0.91	0	0.81	0.95	0.82	0.74	0.95	0.93	-0.41	-0.1	0.99	1

Spearman Correlation

	CHPI	KeyRate	CPI	VIX	PPI	MortgageRate30	Electricity	Water	Plywood	Steel	Glass	Concrete	Unemployment	Yield10Y	DJI	S&P500
CHPI	1	0.22	0.99	0.39	0.94	-0.06	0.95	0.99	0.72	0.32	0.99	0.99	-0.6	-0.08	0.97	0.98
KeyRate	0.22	1	0.24	-0.17	0.18	0.45	0.18	0.23	-0.02	-0.07	0.23	0.23	-0.71	0.43	0.2	0.18
CPI	0.99	0.24	1	0.37	0.94	-0.05	0.96	1	0.72	0.32	0.99	0.99	-0.63	-0.07	0.97	0.98
VIX	0.39	-0.17	0.37	1	0.35	-0.18	0.37	0.4	0.29	0.27	0.4	0.4	0.07	-0.26	0.28	0.31
PPI	0.94	0.18	0.94	0.35	1	-0.11	0.91	0.95	0.77	0.35	0.95	0.95	-0.54	-0.11	0.94	0.95
MortgageRate30	-0.06	0.45	-0.05	-0.18	-0.11	1	-0.02	-0.07	0.03	0.23	-0.07	-0.07	-0.43	0.96	-0.09	-0.11
Electricity	0.95	0.18	0.96	0.37	0.91	-0.02	1	0.96	0.79	0.38	0.96	0.96	-0.6	-0.02	0.96	0.96
Water	0.99	0.23	1	0.4	0.95	-0.07	0.96	1	0.72	0.3	1	1	-0.62	-0.08	0.97	0.98
Plywood	0.72	-0.02	0.72	0.29	0.77	0.03	0.79	0.72	1	0.58	0.72	0.72	-0.33	0.09	0.74	0.74
Steel	0.32	-0.07	0.32	0.27	0.35	0.23	0.38	0.3	0.58	1	0.3	0.3	-0.15	0.25	0.35	0.33
Glass	0.99	0.23	0.99	0.4	0.95	-0.07	0.96	1	0.72	0.3	1	1	-0.62	-0.09	0.97	0.98
Concrete	0.99	0.23	0.99	0.4	0.95	-0.07	0.96	1	0.72	0.3	1	1	-0.62	-0.09	0.97	0.98
Unemployment	-0.6	-0.71	-0.63	0.07	-0.54	-0.43	-0.6	-0.62	-0.33	-0.15	-0.62	-0.62	1	-0.42	-0.64	-0.61
Yield10Y	-0.08	0.43	-0.07	-0.26	-0.11	0.96	-0.02	-0.08	0.09	0.25	-0.09	-0.09	-0.42	1	-0.08	-0.11
DJI	0.97	0.2	0.97	0.28	0.94	-0.09	0.96	0.97	0.74	0.35	0.97	0.97	-0.64	-0.08	1	1
S&P500	0.98	0.18	0.98	0.31	0.95	-0.11	0.96	0.98	0.74	0.33	0.98	0.98	-0.61	-0.11	1	1


CHPI
Case-Shiller
Original target variables
Future values for 4/13/26 weeks ahead
Logarithm
CHPI_fut
Case-Shiller_fut
Final target variables
KeyRate
Electricity
Concrete
CPI
Water
Unemployment
VIX
Plywood
Yield10Y
PPI
Steel
DJI
MortgageRate30
Glass
S&P500
Original independent variables
Logreturns for 4/13/26 weeks
Standard scaler
Final independent variables

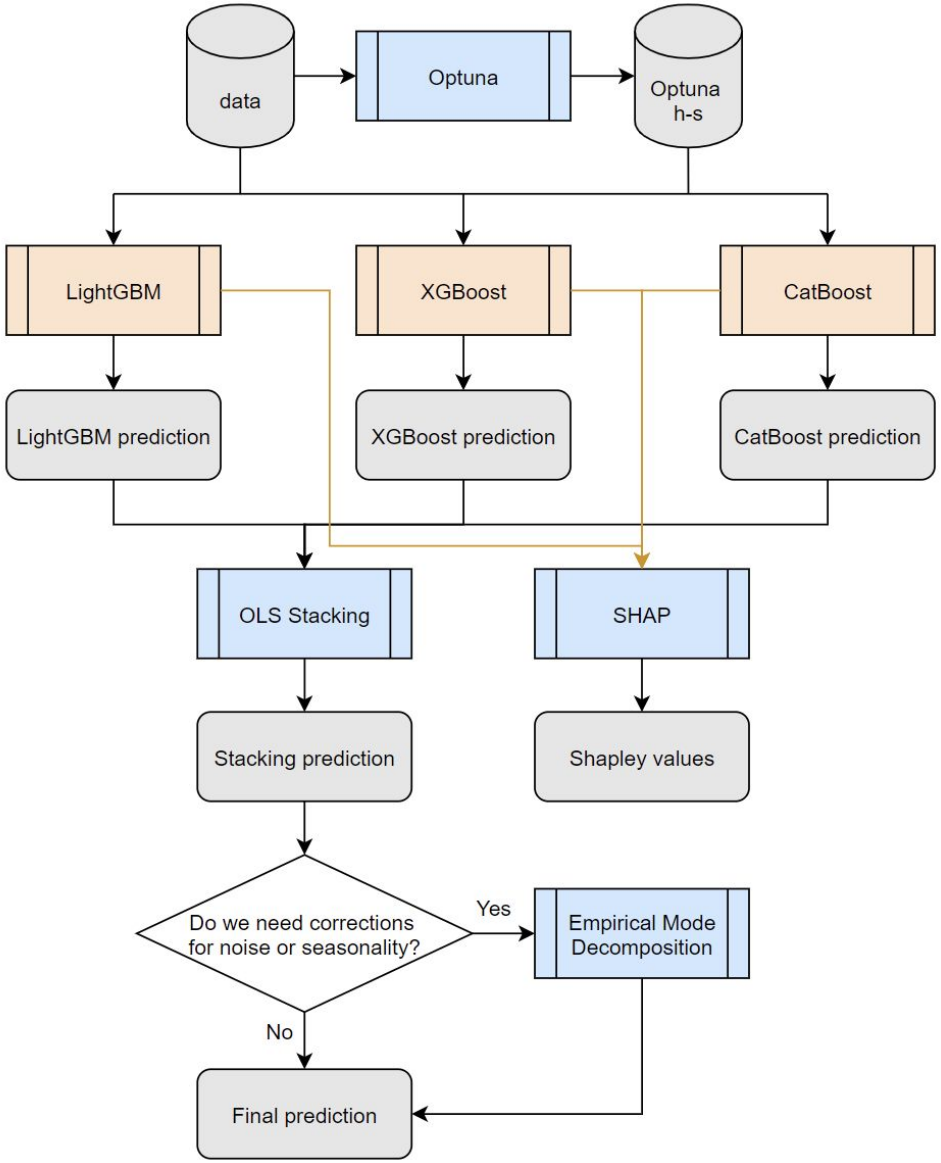
1. Past prices would have the most determining effect on the model predictions, given the low degree of chaos and relatively high degree of persistence
2. Building material indices would be the second most valuable variables.
3. The predictive power of the independent variables would increase with the horizon, as the autoregressive effect of the target variable would decrease over time and be replaced by less powerful predictors.
4. The predictive power of logarithmic returns would be higher compared to static variables for the smaller horizons, as a result of the same autoregressive effect.
5. Statistical areas with lower prices would have better prediction errors than those with higher prices (mainly because there are only a few high-price areas in the sample).
6. The error of the models would increase significantly between the 4 and 13 week forecasts, but the difference between the 13 and 26 week forecasts would be less dramatic. This is because the decreasing autoregressive effect would be partly compensated by the other variables.
7. Stacking the gradient boosting models would allow us to outperform the OLS benchmark both on the original data and on the Case-Shiller index. This is because of the ability to capture more patterns in the data.

Models:

- Hyperparameters of the models are prepared with Optuna
- Base predictions are made with three gradient boosting models
- Gradient boosting models return not only predictions but also Shapley values for variable impact interpretation
- Predictions are brought together via stacking which is trained on the validation data
- Stacked model is tested on the test sample and on the Case-Shiller index (from 2004 to 2024), for which predictions are additionally cleared from noise and seasonality
- Main accuracy metric - RMSE, additional - MAE and MAPE

Train-validation-test split with shuffle:

- Train set: 70%
- Validation set: 15%
- Test set: 15%



Model	Group	Description
OPTUNA	Pretraining	Automatic parameter selection tool for machine learning models and for other applications that use heavy calculations Main feature - Bayesian method of hyperparameter selection
LightGBM	Gradient boosting	Main features - dependent and independent variable are broken down by quantiles, single leaf error is minimized
dmlc XGBoost		Main feature - minimizes the error of the loss function approximated using Taylor series
CatBoost		Main features - symmetric trees, built-in handling of categorical variables
Shap	Model interpretation	Tool to evaluate the impact of input features on the output of a machine learning model on a per-instance basis
Empirical Mode Decomposition	Time series cleaning	Tool to decompose a signal into physically meaningful components Main feature - ability to separate stochastic and deterministic components of the time series based on the threshold

Main results

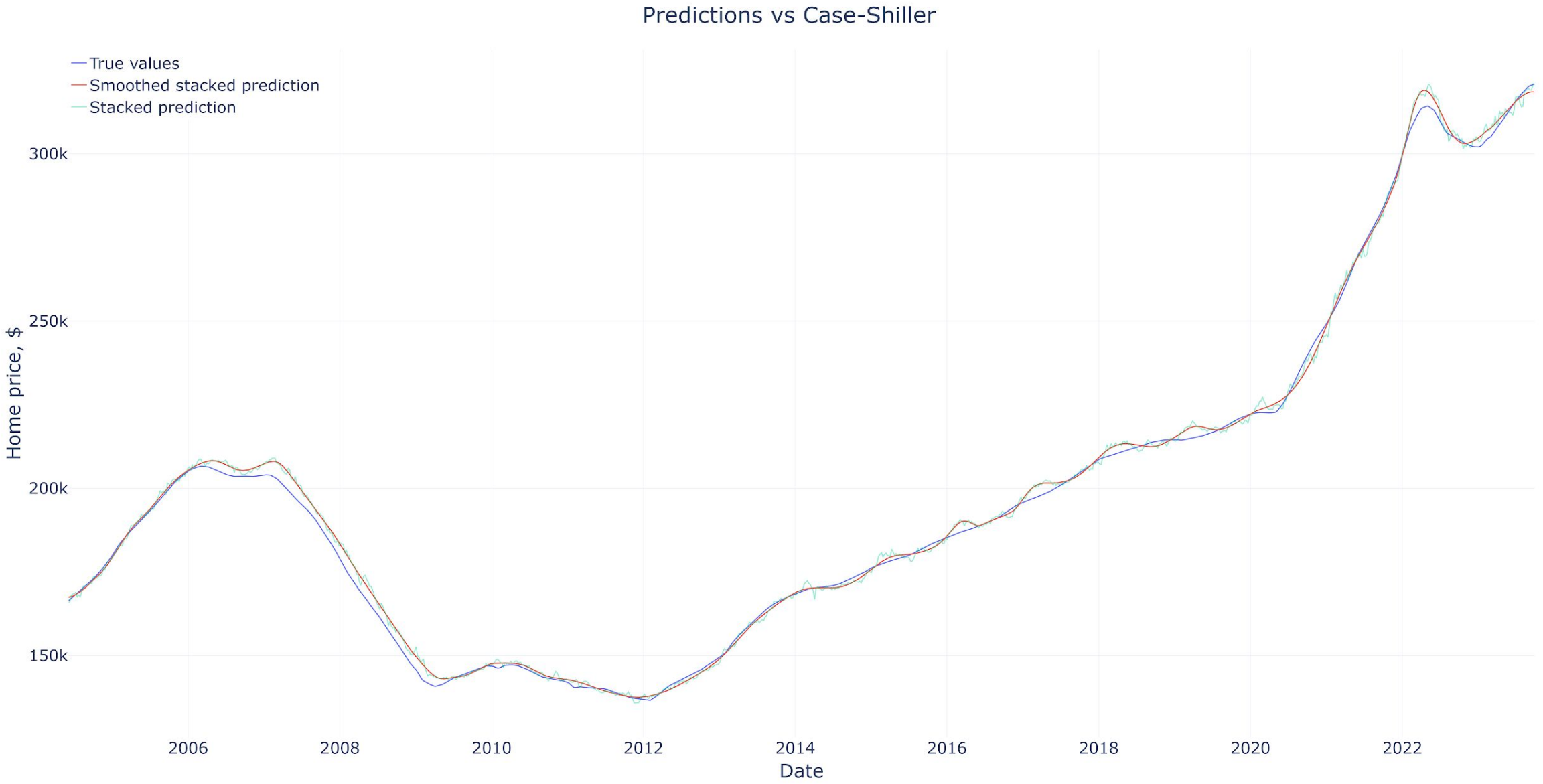
Data	Models	Validation RMSE / MAPE (%)	Test RMSE / MAPE (%)
Haus	LightGBM	5 697.991 / 0.097	5 815.907 / 0.098
	XGBoost	5 544.043 / 0.093	5 574.208 / 0.095
	CatBoost	8 270.735 / 0.136	8 763.188 / 0.138
	Stacked model	**5 265.810 / 0.091**	**5 274.117 / 0.093**
	Benchmark (OLS)	6 176.287 / 1.241	6 407.752 / 1.273
Case-Shiller	Stacked model	2 242.050 / 0.856	
	Smoothed stacked model	**2 009.940 / 0.760**	
	Benchmark (OLS)	1 820.272 / 0.711	

Detalisation for the stacked model on Haus data

Separate model	Coefficient	Standard Error	P-value
LightGBM	0.3427	0.027	0.000
XGBoost	0.6574	0.027	0.000
CatBoost	-	-	> 0.05

Segmentation for the stacked model on Case-Shiller data

Period	RMSE ($) / MAPE (%)	RMSE ($) / MAPE (%) with Smoothening
Whole	2 242.050 / 0.856	2 009.940 / 0.76
Before 2010-01-01	2 975.273 / 1.302	2 856.586 / 1.262
2010-01-01 - 2018-01-01	1 375.758 / 0.630	1 110.665 / 0.507
After 2018-01-01	2 378.139 / 0.734	1 981.697 / 0.623


Predictions vs Case-Shiller
True values
Smoothed stacked prediction
Stacked prediction
300k
250k
200k
150k
Home price, $
2006
2008
2010
2012
2014
2016
2018
2020
2022
Date

Main results

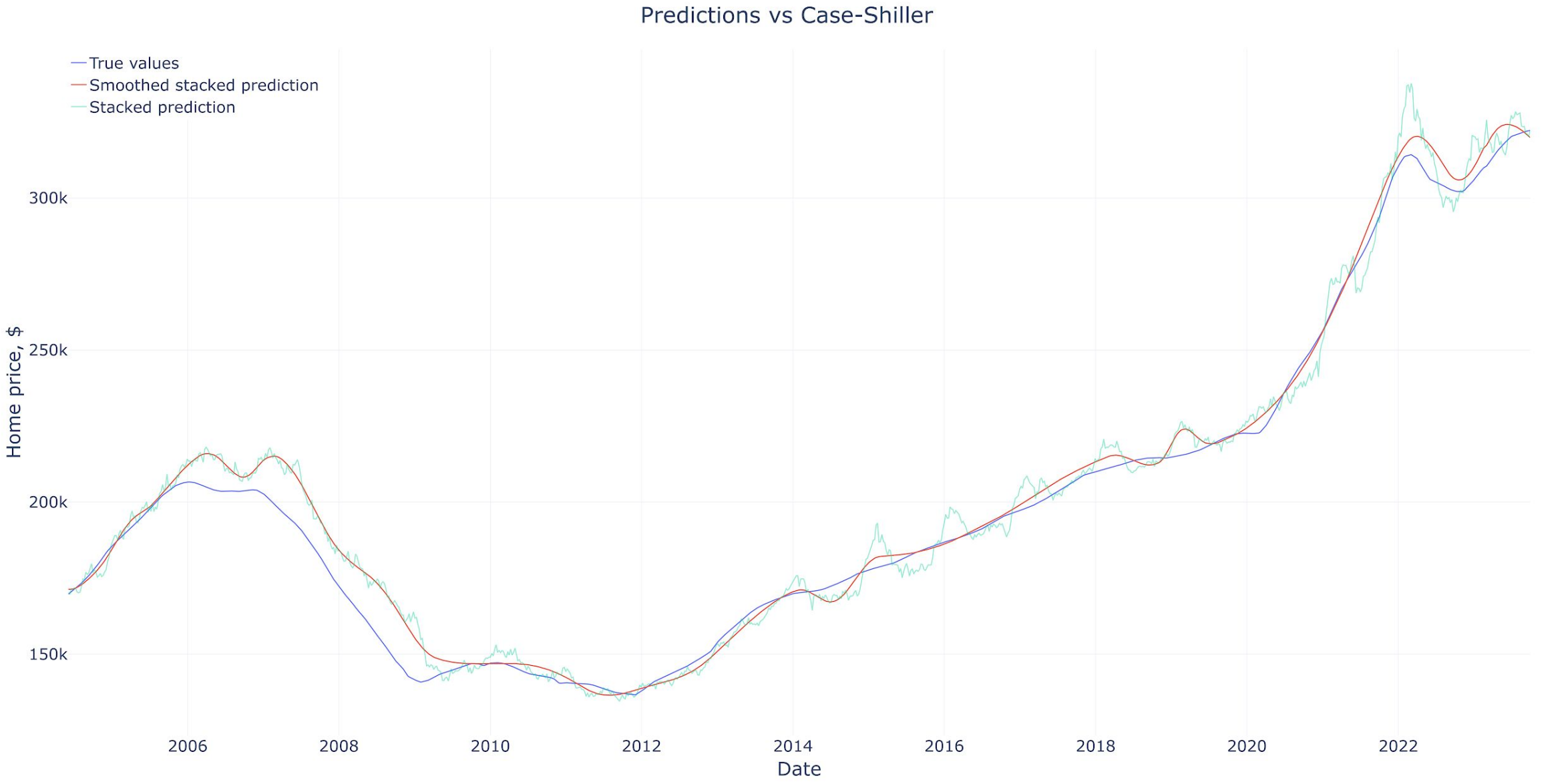
Data	Models	Validation RMSE / MAPE (%)	Test RMSE / MAPE (%)
Haus	LightGBM	8 416.799 / 0.156	8 479.791 / 0.153
	XGBoost	7 803.822 / 0.147	7 801.324 / 0.146
	CatBoost	10 782.573 / 0.185	11 283.988 / 0.187
	Stacked model	**7 633.267 / 0.145**	**7 517.742 / 0.143**
	Benchmark (OLS)	11 630.265 / 2.425	12 375.693 / 2.448
Case-Shiller	Stacked model	7 182.020 / 2.607	
	Smoothed stacked model	**6 044.992 / 2.065**	
	Benchmark (OLS)	8 275.517 / 3.331	

Detalisation for the stacked model on Haus data

Separate model	Coefficient	Standard Error	P-value
LightGBM	0.2971	0.025	0.000
XGBoost	0.7029	0.025	0.000
CatBoost	-	-	> 0.05

Segmentation for the stacked model on Case-Shiller data

Period	RMSE ($) / MAPE (%)	RMSE ($) / MAPE (%) with Smoothening
Whole	7 182.020 / 2.607	6 044.992 / 2.065
Before 2010-01-01	10 380.332 / 4.365	9 974.363 / 4.234
2010-01-01 - 2018-01-01	3 963.906 / 1.803	2 210.441 / 1.102
After 2018-01-01	6 834.538 / 2.010	4 382.027 / 1.286


Predictions vs Case-Shiller
True values
Smoothed stacked prediction
Stacked prediction
300k
250k
200k
150k
Home price, $
2006
2008
2010
2012
2014
2016
2018
2020
2022
Date

Main results

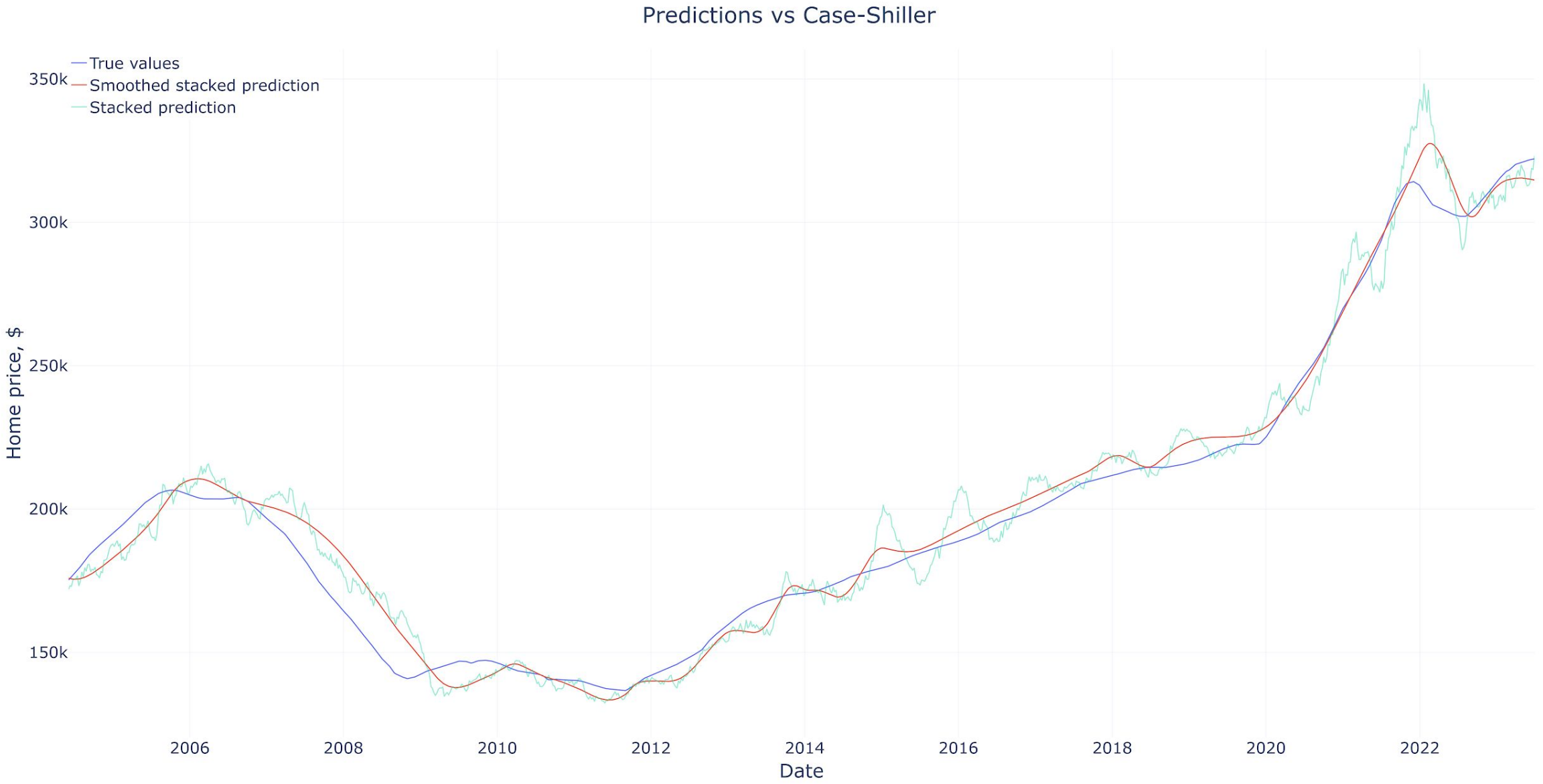
Data	Models	Validation RMSE / MAPE (%)	Test RMSE / MAPE (%)
Haus	LightGBM	10 771.676 / 0.193	10 495.897 / 0.189
	XGBoost	10 271.295 / 0.197	10 720.080 / 0.199
	CatBoost	13 566.801 / 0.240	13 183.759 / 0.236
	Stacked model	**9 968.613 / 0.187**	**10 041.655 / 0.185**
	Benchmark (OLS)	15 375.112 / 3.355	16 404.234 / 3.417
Case-Shiller	Stacked model	8 952.761 / 3.447	
	Smoothed stacked model	**6 947.070 / 2.723**	
	Benchmark (OLS)	27 636.263 / 11.767	

Detalisation for the stacked model on Haus data

Separate model	Coefficient	Standard Error	P-value
LightGBM	0.5355	0.026	0.000
XGBoost	0.4644	0.026	0.000
CatBoost	-	-	> 0.05

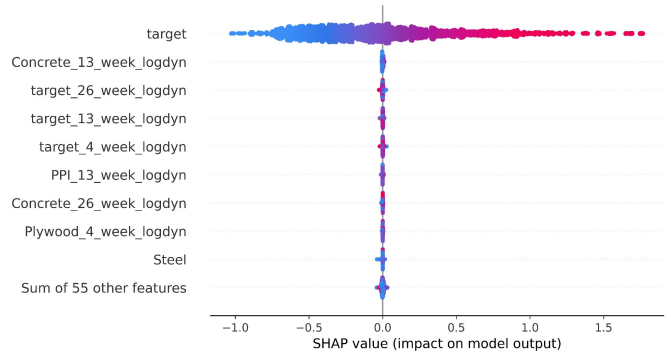
Segmentation for the stacked model on Case-Shiller data

Period	RMSE ($) / MAPE (%)	RMSE ($) / MAPE (%) with Smoothening
Whole	8 952.761 / 3.447	6 947.070 / 2.723
Before 2010-01-01	10 850.050 / 5.289	10 319.326 / 4.915
2010-01-01 - 2018-01-01	6 205.332 / 2.669	3 904.350 / 1.983
After 2018-01-01	10 120.035 / 2.700	6 077.45 / 1.566

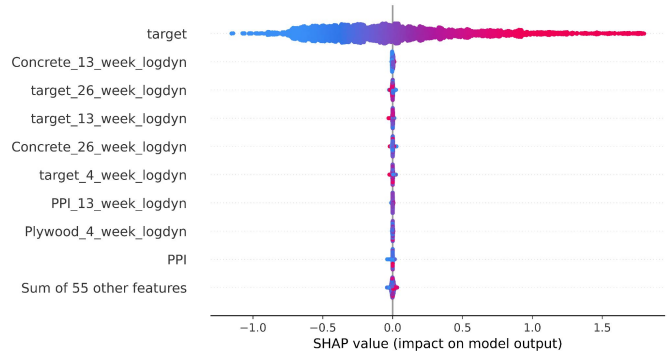

Predictions vs Case-Shiller
True values
Smoothed stacked prediction
Stacked prediction
350k
300k
250k
200k
150k
Home price, $
2006
2008
2010
2012
2014
2016
2018
2020
2022
Date

3 weeks ahead

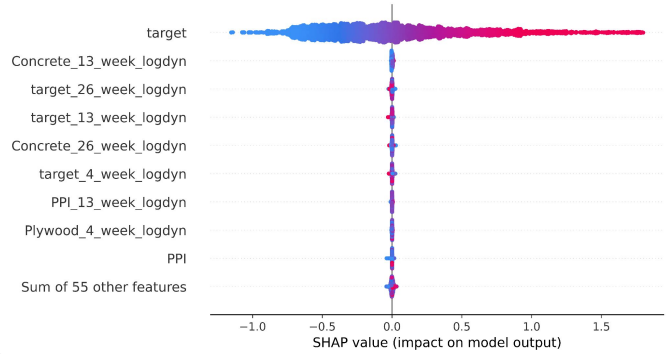
LightGBM



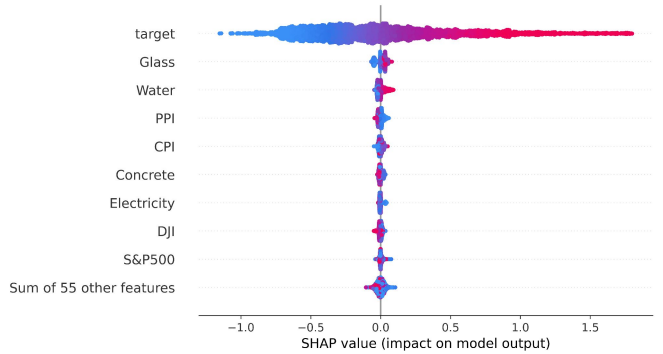
dmlc XGBoost

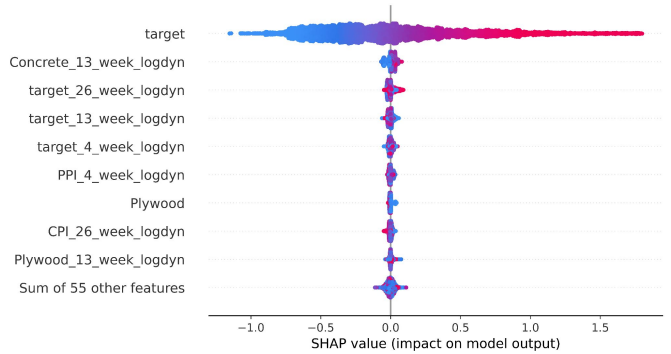


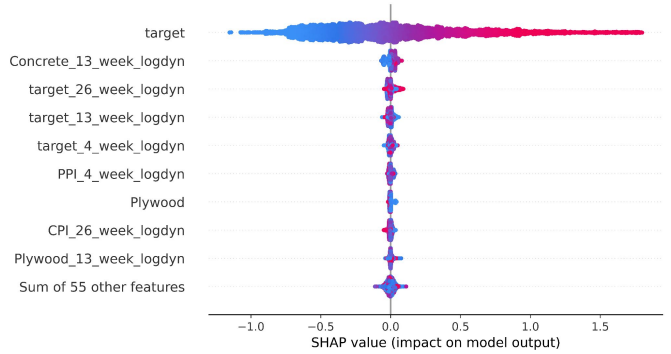
CatBoost



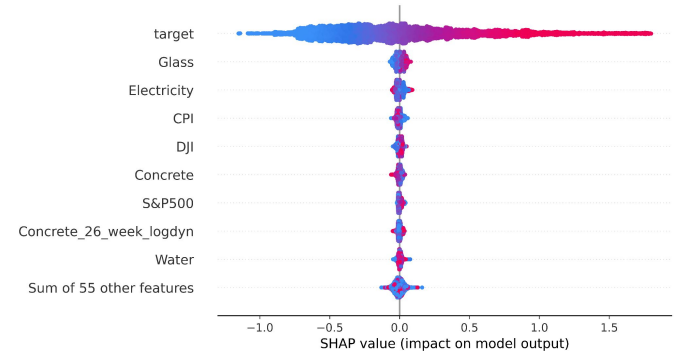
13 weeks ahead

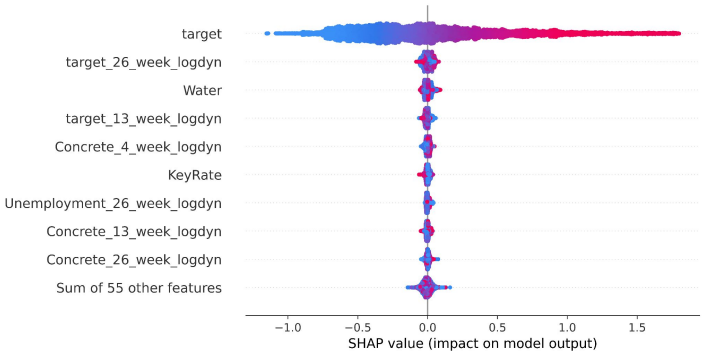


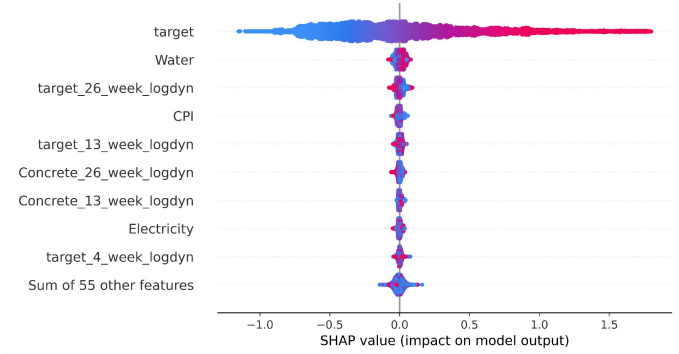




26 weeks ahead







Conclusions for the research hypotheses:

1. Past prices would have the most determining effect on the model predictions - This is correct, as these prices have the defining impact on the predictions and their quality.
2. Building material indices would be the second most descriptive - This is partially correct, as they have shown much larger impact for all of the horizons and models in comparison with pure macroeconomic variables but market indices managed to get relatively high Shapley values.
3. The predictive power of the independent variables would increase with the horizon - This is correct, as the autoregressive part is no longer capable of describing all of the changes in the dataset.
4. The predictive power of logarithmic returns would be higher compared to static variables for the smaller horizons - This is correct, as dynamic variables are losing their power on the longer horizons.
5. Statistical areas with lower prices would have better prediction errors than those with higher prices - This is correct, even after considering that in percentage terms they are actually at the same level of error.
6. The error of the models would increase significantly between the 4 and 13 week forecasts, but the difference between the 13 and 26 week forecasts would be less dramatic - This is correct, as building material indices allow to mitigate part of the decrease in the accuracy.
7. Stacking the gradient boosting models would allow us to outperform the OLS benchmark both on the original data and on the Case-Shiller index. This hypothesis is partially correct, with the exception of the four-week prediction on the Case-Shiller data.

Other conclusions:

8. We have successfully created a stacked model of three gradient boosting models that can predict weekly real estate prices for standard households in 100 of the largest statistical areas in the US.
9. This model can be used for transfer learning on different real estate indices, which adds flexibility for the real implementation in research and business purposes.

R1: Andrey should comment on the novelty of his study, which was not mentioned.

A1: This is the first prediction model for regional real estate prices based on the weekly data for the US (and which can be used for transfer learning on the other real estate markets)

R2: He needs to comment on one of his intended purposes of the study: "For scientific researchers, it can be demonstrated that weekly data may be employed for the prediction of the markets with lower levels of liquidity, such as those pertaining to real estate." It's not obvious what exactly Andrey implied here as weekly data has been used by researchers for ages.

A2: Agree. Added "highly detailed and regional" to the first principal implication.

R3: The "Research Hypothesis" stated "The data collection and feature generation process yielded a set of seven hypotheses, as follows" – it was not clear for me how Andrey came up with those 7 hypotheses and it needs more clarification on how they aligned with the goals of the study.

A3: These hypotheses were chosen based on the exploratory data analysis as they are the most technical and so can be quantitatively checked with the available methods. 7 were chosen in order to fully evaluate performance at different time horizons and for different independent variables and to complete the third task of the research.

R4: Discussion of comparison between LightGBM, XGBoost, and CatBoost, and potential issues like overfitting for LightGBM could be valuable for future readers.

A4: Base discussion was mentioned in the chapter II.1. More detailed comparison of the technologies was decided to leave out of the paper in order not to overcomplicate it.

R5: A clearer description of how the final models were tested, including how the learning sample and test sample were divided.

A5: Agree about samples division - information on the train-validation-test split was added on the slide "Final Model Architecture". Bucket comparison as an instrument of final testing was described in the chapter III, but I agree that there might be a little bit more details added.

R6: I would like to see Andrey's thoughts on the applicability of the approaches used in his work to other markets, particularly the real estate market in Russia.

A6: At this moment we can try to use this model on the weekly national index from "Domclick" with adjustment for different currencies and inflation rates. However, due to the much more severe nature of the economic shocks in Russia current 389 observations may not be enough to fully adjust model to our market. Still, this model can be freely used for other countries with aforementioned adjustments if their markets are somehow similar to the US because gradient boosting models are limited to the original training data and cannot predict out of sample.


ВЫСШАЯ • ШКОЛА • ЭКОНОМИКИ •

Appendix

1. Statistical Areas that were Used in Modelling
2. Stationarity and Measures of Chaos for Variables
3. Accuracy Metrics
4. Empirical Mode Decomposition
5. Prediction, Target and Error distributions

Code	City	State	Minimal date	Maximum date	Number of entries
10420	Akron	OH	2011-01-07	2023-03-24	638
10580	Albany-Schenectady-Troy	NY	2010-03-12	2023-03-24	681
10740	Albuquerque	NM	2010-01-22	2023-03-24	688
10900	Allentown-Bethlehem-Easton	PA-NJ	2010-01-22	2023-03-24	688
11244	Anaheim-Santa Ana-Irvine	CA	2010-01-22	2023-03-24	688
12060	Atlanta-Sandy Springs-Roswell	GA	2010-06-18	2023-03-24	585
12420	Austin-Round Rock	TX	2010-01-22	2023-03-24	688
12580	Baltimore-Columbia-Towson	MD	2010-01-22	2023-03-24	688
12940	Baton Rouge	LA	2010-09-17	2023-03-24	654
13820	Birmingham-Hoover	AL	2012-02-10	2023-03-24	581
14454	Boston	MA	2010-01-22	2023-03-24	556
14860	Bridgeport-Stamford-Norwalk	CT	2010-01-22	2023-03-24	688
15380	Buffalo-Cheektowaga-Niagara Falls	NY	2010-07-09	2023-03-24	664
15764	Cambridge-Newton-Framingham	MA	2010-01-22	2023-03-24	557
15804	Camden	NJ	2010-01-22	2023-03-24	688
15980	Cape Coral-Fort Myers	FL	2010-01-22	2023-03-24	688
16700	Charleston-North Charleston	SC	2010-01-22	2023-03-24	531
16740	Charlotte-Concord-Gastonia	NC-SC	2010-01-22	2023-03-24	688
16974	Chicago-Naperville-Arlington Heights	IL	2010-01-22	2023-03-24	688
17140	Cincinnati	OH-KY-IN	2010-01-22	2023-03-24	688
17460	Cleveland-Elyria	OH	2010-01-22	2023-03-24	688
17820	Colorado Springs	CO	2010-01-22	2023-03-24	607
17900	Columbia	SC	2010-01-29	2023-03-24	536
18140	Columbus	OH	2010-01-22	2023-03-24	688
19124	Dallas-Plano-Irving	TX	2010-01-22	2023-03-24	688
19380	Dayton	OH	2010-07-09	2023-03-24	664
19660	Deltona-Daytona Beach-Ormond Beach	FL	2010-07-23	2023-03-24	662
19740	Denver-Aurora-Lakewood	CO	2010-01-22	2023-03-24	688
19804	Detroit-Dearborn-Livonia	MI	2010-01-22	2023-03-24	688
21340	El Paso	TX	2010-07-09	2023-03-24	664
22744	Fort Lauderdale-Pompano Beach-Deerfield Beach	FL	2010-01-22	2023-03-24	688
23104	Fort Worth-Arlington	TX	2010-01-22	2023-03-24	688
23844	Gary	IN	2010-01-22	2023-03-24	688
24340	Grand Rapids-Wyoming	MI	2010-02-05	2023-03-24	666
24660	Greensboro-High Point	NC	2010-04-23	2023-03-24	665
24860	Greenville-Anderson-Mauldin	SC	2010-01-29	2023-03-24	687
25540	Hartford-West Hartford-East Hartford	CT	2010-01-22	2023-03-24	688
26420	Houston-The Woodlands-Sugar Land	TX	2010-01-22	2023-03-24	688
26900	Indianapolis-Carmel-Anderson	IN	2010-01-22	2023-03-24	688
27260	Jacksonville	FL	2010-01-22	2023-03-24	627
28140	Kansas City	MO-KS	2010-01-22	2023-03-24	688
28940	Knoxville	TN	2010-01-22	2023-03-24	688
29404	Lake County-Kenosha County	IL-WI	2010-01-22	2023-03-24	688
29460	Lakeland-Winter Haven	FL	2010-01-22	2023-03-24	684
29820	Las Vegas-Henderson-Paradise	NV	2010-01-22	2023-03-24	688
30780	Little Rock-North Little Rock-Conway	AR	2010-08-06	2023-03-24	660
31084	Los Angeles-Long Beach-Glendale	CA	2010-01-22	2023-03-24	688
31140	Louisville/Jefferson County	KY-IN	2010-01-22	2023-03-24	688
32820	Memphis	TN-MS-AR	2010-01-22	2023-03-24	688
33124	Miami-Miami Beach-Kendall	FL	2010-01-22	2023-03-24	688

Code	City	State	Minimal date	Maximum date	Number of entries
33340	Milwaukee-Waukesha-West Allis	WI	2010-01-22	2023-03-24	688
33460	Minneapolis-St. Paul-Bloomington	MN-WI	2010-01-22	2023-03-24	688
33874	Montgomery County-Bucks County-Chester County	PA	2010-01-22	2023-03-24	688
34980	Nashville-Davidson--Murfreesboro--Franklin	TN	2010-01-22	2023-03-24	688
35004	Nassau County-Suffolk County	NY	2010-01-22	2023-03-24	688
35084	Newark	NJ-PA	2010-01-22	2023-03-24	688
35300	New Haven-Milford	CT	2010-01-22	2023-03-24	688
35380	New Orleans-Metairie	LA	2010-01-22	2023-03-24	688
35614	New York-Jersey City-White Plains	NY-NJ	2010-01-22	2023-03-24	688
35840	North Port-Sarasota-Bradenton	FL	2010-01-22	2023-03-24	688
36084	Oakland-Hayward-Berkeley	CA	2010-01-22	2023-03-24	688
36420	Oklahoma City	OK	2010-01-22	2023-03-24	688
36540	Omaha-Council Bluffs	NE-IA	2010-07-16	2023-03-24	663
36740	Orlando-Kissimmee-Sanford	FL	2010-01-22	2023-03-24	687
37100	Oxnard-Thousand Oaks-Ventura	CA	2010-01-22	2023-03-24	631
37340	Palm Bay-Melbourne-Titusville	FL	2010-07-23	2023-03-24	662
37964	Philadelphia	PA	2010-01-22	2023-03-24	688
38060	Phoenix-Mesa-Scottsdale	AZ	2010-01-22	2023-03-24	688
38300	Pittsburgh	PA	2010-01-22	2023-03-24	676
38900	Portland-Vancouver-Hillsboro	OR-WA	2010-01-22	2023-03-24	688
39300	Providence-Warwick	RI-MA	2010-01-22	2023-03-24	688
39580	Raleigh	NC	2010-01-22	2023-03-24	688
40060	Richmond	VA	2010-01-22	2023-03-24	688
40140	Riverside-San Bernardino-Ontario	CA	2010-01-22	2023-03-24	688
40380	Rochester	NY	2010-01-29	2023-03-24	670
40900	Sacramento--Roseville--Arden-Arcade	CA	2010-01-22	2023-03-24	688
41180	St. Louis	MO-IL	2010-01-22	2023-03-24	670
41620	Salt Lake City	UT	2010-01-22	2023-03-24	688
41700	San Antonio-New Braunfels	TX	2014-04-04	2023-03-24	467
41740	San Diego-Carlsbad	CA	2010-01-22	2023-03-24	688
41884	San Francisco-Redwood City-South San Francisco	CA	2010-01-22	2023-03-24	688
41940	San Jose-Sunnyvale-Santa Clara	CA	2010-01-22	2023-03-24	688
42644	Seattle-Bellevue-Everett	WA	2010-01-22	2023-03-24	688
43524	Silver Spring-Frederick-Rockville	MD	2010-01-22	2023-03-24	688
45060	Syracuse	NY	2010-04-23	2023-03-24	662
45104	Tacoma-Lakewood	WA	2010-01-22	2023-03-24	688
45300	Tampa-St. Petersburg-Clearwater	FL	2010-01-22	2023-03-24	688
45780	Toledo	OH	2010-04-30	2023-03-24	674
46060	Tucson	AZ	2010-01-22	2023-03-24	603
46140	Tulsa	OK	2010-02-05	2023-03-24	653
46520	Urban Honolulu	HI	2010-01-22	2023-03-24	685
47260	Virginia Beach-Norfolk-Newport News	VA-NC	2010-01-22	2023-03-24	688
47664	Warren-Troy-Farmington Hills	MI	2010-01-22	2023-03-24	688
47894	Washington-Arlington-Alexandria	DC-VA-MD-WV	2010-01-22	2023-03-24	687
48424	West Palm Beach-Boca Raton-Delray Beach	FL	2010-01-22	2023-03-24	688
48620	Wichita	KS	2010-03-26	2023-03-24	679
48864	Wilmington	DE-MD-NJ	2010-01-22	2023-03-24	688
49180	Winston-Salem	NC	2010-03-12	2023-03-24	670
49340	Worcester	MA-CT	2010-04-09	2023-03-24	669
49660	Youngstown-Warren-Boardman	OH-PA	2010-09-17	2023-03-24	654

Stationarity and Measures of Chaos for Variables

Variable	ADF statistics	ADF p-value	Largest Lyapunov Exponent	Hurst Exponent
CHPI	-0.386	0.912	0.002	0.892
KeyRate	-0.819	0.814	0.003	0.946
CPI	2.961	1.000	0.004	0.910
VIX	-3.346	0.013	-0.009	0.773
PPI	-0.370	0.915	0.014	0.860
MortgageRate30	-0.531	0.886	0.011	0.914
Electricity	0.704	0.990	0.008	0.880
Water	2.178	0.999	0.000	0.910
Plywood	-1.458	0.554	0.022	0.901
Steel	-2.224	0.198	0.009	0.903
Glass	2.377	0.999	0.003	0.920
Concrete	2.901	1.000	0.002	0.906
Unemployment	-2.648	0.083	-0.012	0.849
Yield10Y	-1.465	0.551	-0.000	0.935
DJI	-0.903	0.787	0.002	0.860
S&P500	-0.911	0.784	0.004	0.886
KeyRate_4_week_logdyn	-4.881	0.000	0.025	0.590
CPI_4_week_logdyn	-2.828	0.054	0.008	0.779
VIX_4_week_logdyn	-7.290	0.000	0.018	0.438
PPI_4_week_logdyn	-7.377	0.000	0.024	0.737
MortgageRate30_4_week_logdyn	-4.188	0.001	0.029	0.692
Electricity_4_week_logdyn	-2.457	0.126	0.028	0.689
Water_4_week_logdyn	-4.135	0.001	-0.014	0.704
Plywood_4_week_logdyn	-6.202	0.000	0.019	0.806
Steel_4_week_logdyn	-2.853	0.051	0.016	0.805
Glass_4_week_logdyn	-3.636	0.005	0.018	0.698
Concrete_4_week_logdyn	-2.091	0.248	0.000	0.743
Unemployment_4_week_logdyn	-5.462	0.000	0.030	0.671
Yield10Y_4_week_logdyn	-4.947	0.000	0.031	0.665
DJI_4_week_logdyn	-5.896	0.000	0.030	0.623
S&P500_4_week_logdyn	-5.714	0.000	0.039	0.637

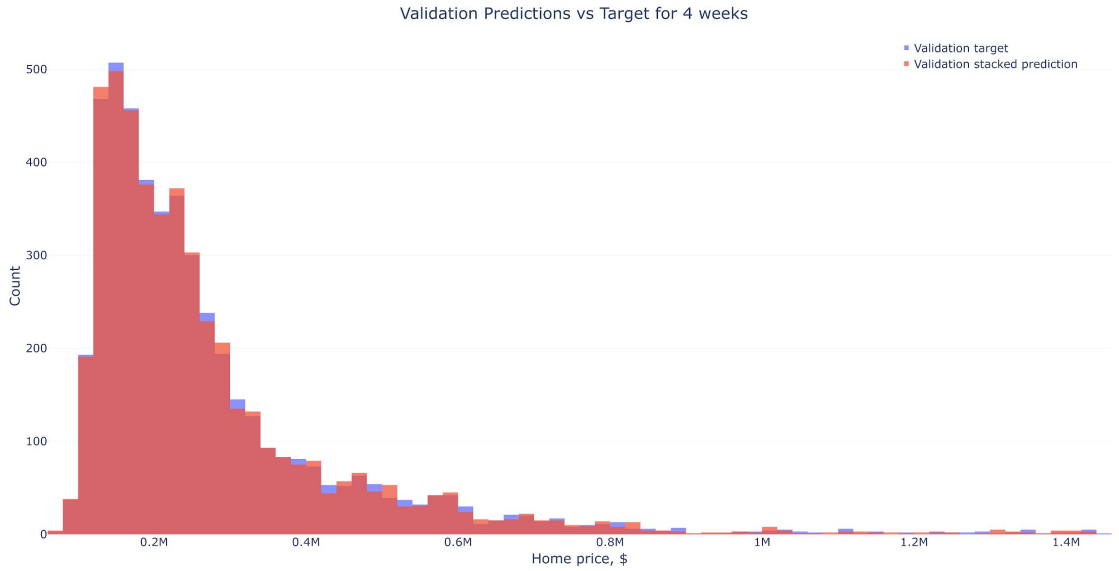
Variable	ADF statistics	ADF p-value	Largest Lyapunov Exponent	Hurst Exponent
KeyRate_13_week_logdyn	-3.835	0.003	0.027	0.921
CPI_13_week_logdyn	-2.723	0.070	0.035	0.844
VIX_13_week_logdyn	-6.230	0.000	0.015	0.819
PPI_13_week_logdyn	-4.946	0.000	0.019	0.844
MortgageRate30_13_week_logdyn	-2.973	0.038	0.017	0.882
Electricity_13_week_logdyn	-2.459	0.126	0.020	0.843
Water_13_week_logdyn	-3.476	0.009	-0.006	0.823
Plywood_13_week_logdyn	-5.612	0.000	0.019	0.869
Steel_13_week_logdyn	-3.356	0.013	0.014	0.902
Glass_13_week_logdyn	-3.594	0.006	0.018	0.886
Concrete_13_week_logdyn	-2.819	0.056	0.014	0.909
Unemployment_13_week_logdyn	-5.555	0.000	0.014	0.789
Yield10Y_13_week_logdyn	-3.323	0.014	0.011	0.879
DJI_13_week_logdyn	-5.015	0.000	0.021	0.856
S&P500_13_week_logdyn	-4.650	0.000	0.018	0.806
KeyRate_26_week_logdyn	-3.030	0.032	0.012	0.964
CPI_26_week_logdyn	-2.345	0.158	0.032	0.877
VIX_26_week_logdyn	-4.909	0.000	0.013	0.827
PPI_26_week_logdyn	-4.274	0.000	0.025	0.853
MortgageRate30_26_week_logdyn	-3.772	0.003	0.007	0.937
Electricity_26_week_logdyn	-3.194	0.020	0.013	0.855
Water_26_week_logdyn	-3.890	0.002	0.020	0.868
Plywood_26_week_logdyn	-3.567	0.006	0.013	0.909
Steel_26_week_logdyn	-2.778	0.062	0.001	0.908
Glass_26_week_logdyn	-3.746	0.004	0.003	0.908
Concrete_26_week_logdyn	-2.542	0.106	0.018	0.909
Unemployment_26_week_logdyn	-4.929	0.000	0.005	0.850
Yield10Y_26_week_logdyn	-3.789	0.003	0.008	0.930
DJI_26_week_logdyn	-4.086	0.001	0.010	0.871
S&P500_26_week_logdyn	-3.995	0.001	0.006	0.872

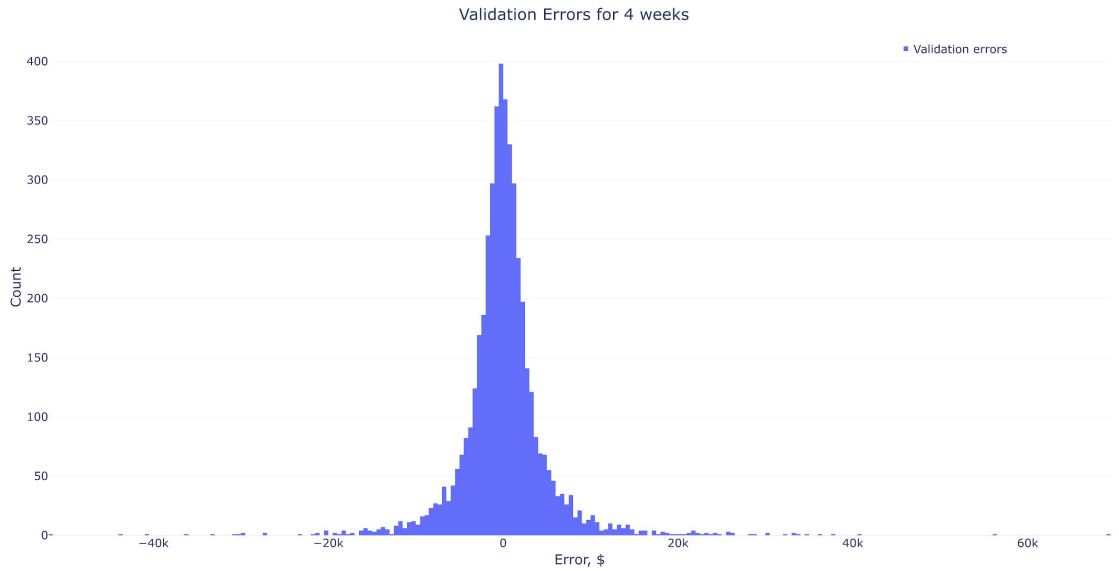
$$RMSE(y, \hat{y}) = \sqrt{\frac{1}{n_{samples}} \sum_{i=0}^{n_{samples}-1} (y_i - \hat{y}_i)^2}$$

$$MAE(y, \hat{y}) = \frac{1}{n_{samples}} \sum_{i=0}^{n_{samples}-1} y_i - \hat{y}_i$$

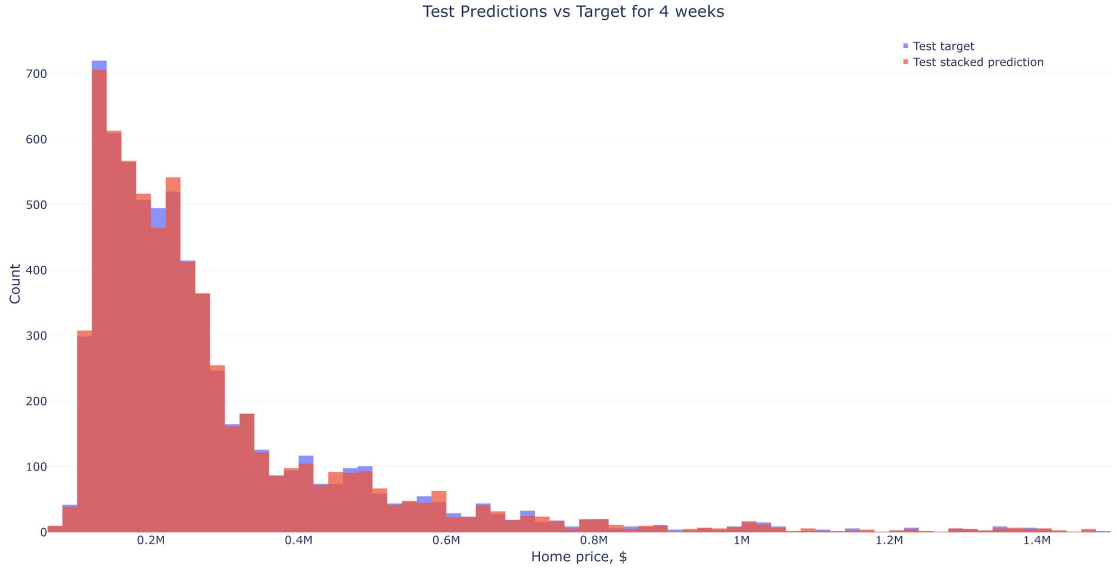
$$MAPE(y, \hat{y}) = \frac{100\%}{n_{samples}} \sum_{i=0}^{n_{samples}-1} \left| \frac{y_i - \hat{y}_i}{y_i} \right|$$

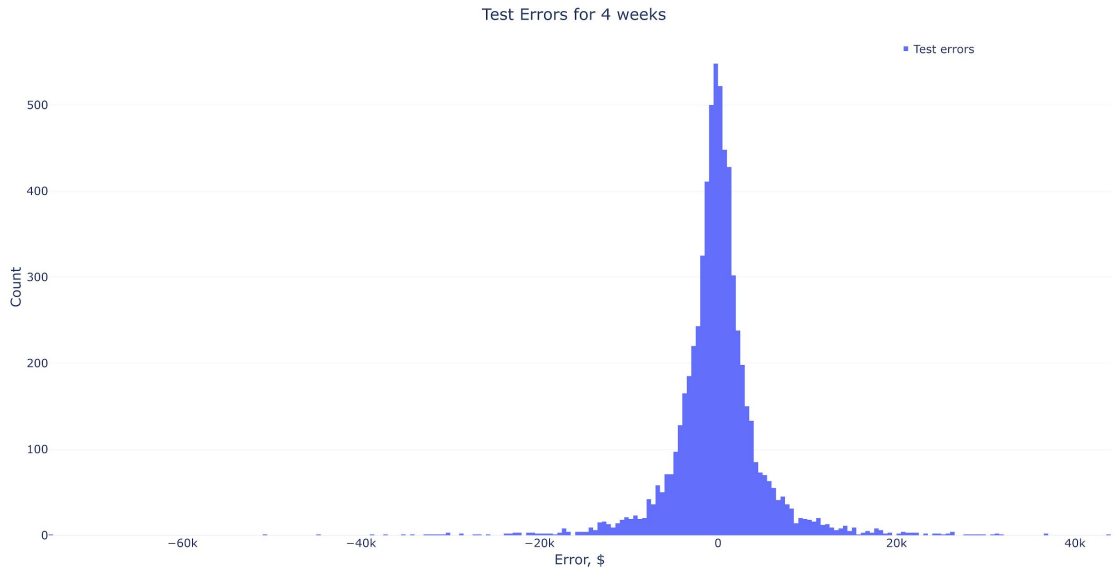
1. Identify the local extrema of the signal. These extrema can be conceptualised as the "envelopes" that encapsulate the oscillatory components of different scales in the signal.
2. A mean envelope is created by interpolating between the maxima and minima at each point in the signal. This mean envelope serves as a reference for extracting oscillatory modes.
3. The IMF is then obtained, which is the difference between the original signal and the mean envelope. This IMF represents the finest scale oscillatory mode present in the signal. If the IMF is monotonic, the subsequent step is to be taken; otherwise, the first three steps must be repeated.
4. The IMFs are to be separated into deterministic and stochastic components based on the mutual information criteria. The higher the criteria, the fewer IMFs are included in the deterministic component.



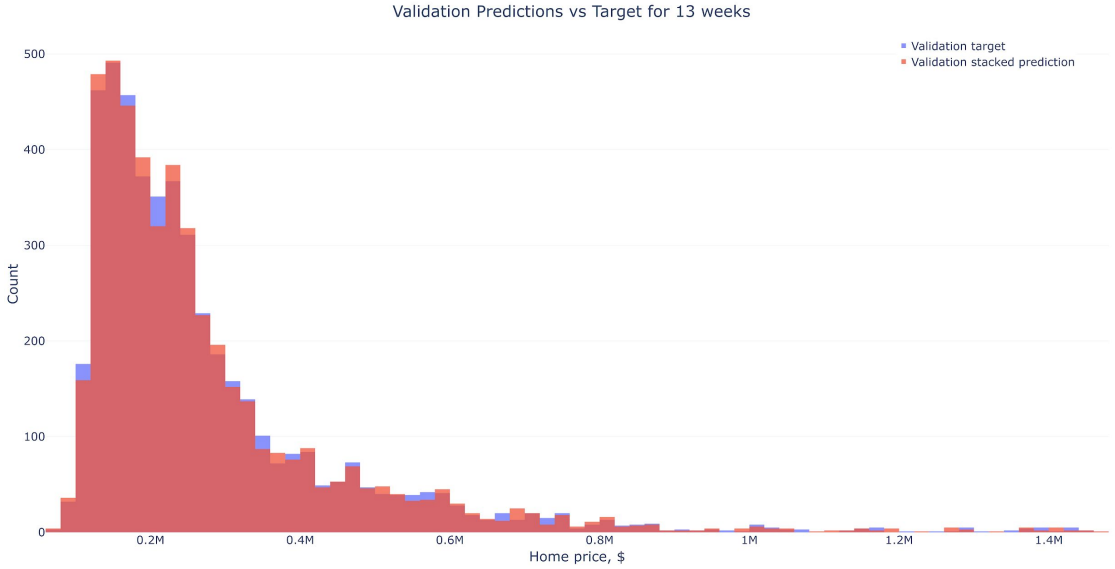


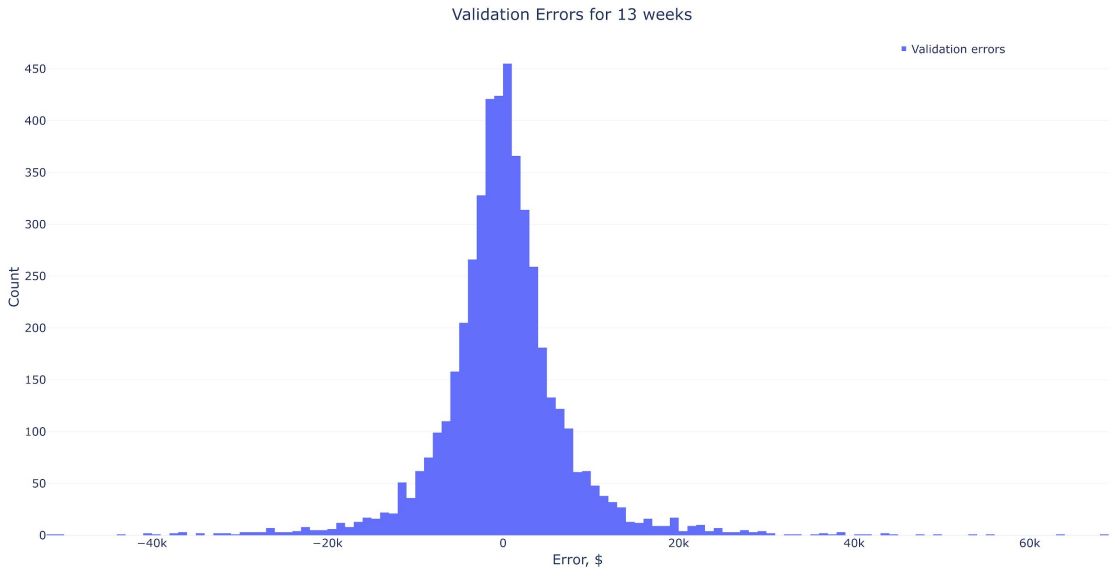
Bucket	Lower	Upper	Number	RMSE	MAE	MAPE (%)
1	74 623.23	211 646.52	2 254	2 608.93	1 809.77	1.138
2	211 646.52	348 669.81	1 548	4 228.82	2 949.84	1.107
3	348 669.81	485 693.09	480	6 198.43	4 236.59	1.026
4	485 693.09	622 716.38	256	8 538.65	6 198.49	1.114
5	622 716.38	759 739.67	104	11 558.97	8 894.96	1.287
6	759 739.67	896 762.95	56	14 087.18	11 270.37	1.370
7	896 762.95	1 033 786.24	18	13 415.71	11 437.12	1.172
8	1 033 786.24	1 170 809.53	16	15 346.56	11 514.28	1.061
9	1 170 809.53	1 307 832.81	13	25 217.84	18 845.96	1.491
10	1 307 832.81	1 444 856.10	18	25 448.50	20 990.76	1.521



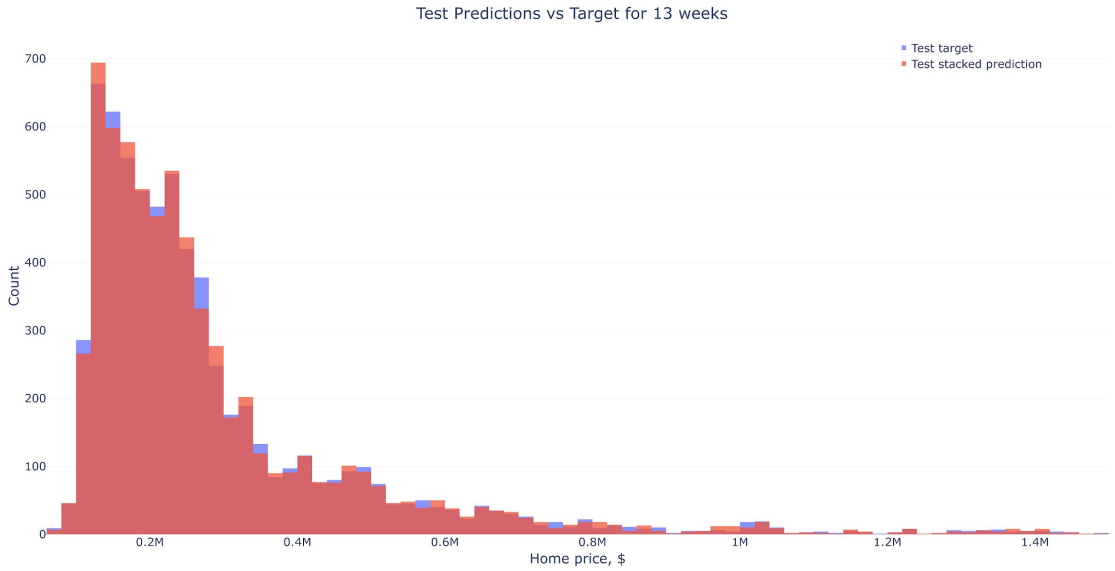


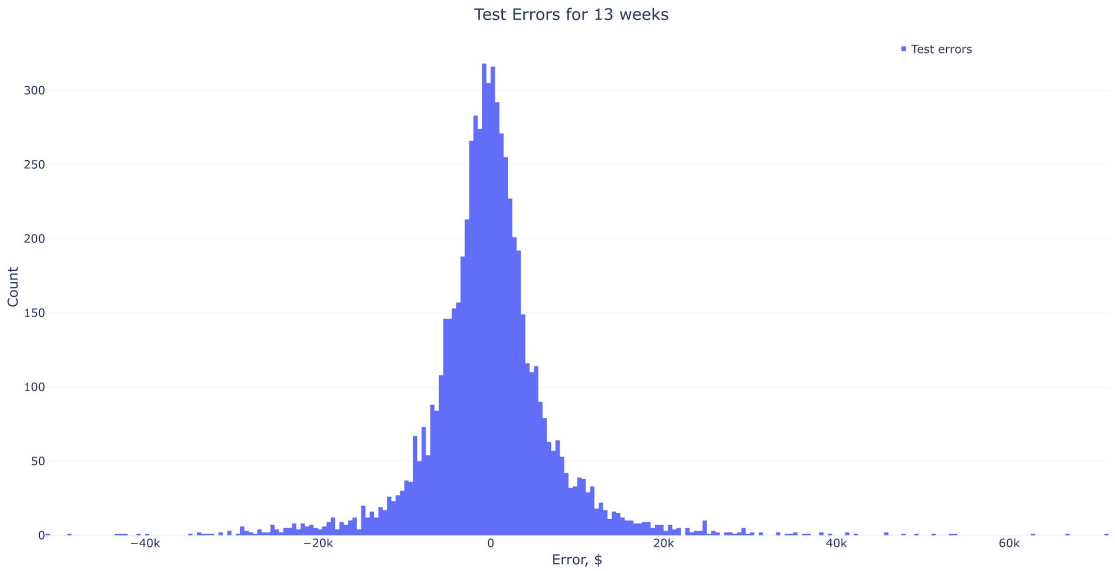
Bucket	Lower	Upper	Number	RMSE	MAE	MAPE (%)
1	71 697.97	213 548.93	3 075	2 654.98	1 834.34	1.162
2	213 548.93	355 399.90	2 158	4 558.06	3 104.20	1.163
3	355 399.90	497 250.86	666	6 472.75	4 631.05	1.086
4	497 250.86	639 101.82	310	8 271.75	5 991.56	1.062
5	639 101.82	780 952.78	171	10 258.85	7 852.57	1.129
6	780 952.78	922 803.75	77	11 834.03	8 696.62	1.036
7	922 803.75	1 064 654.71	61	17 761.32	12 486.21	1.242
8	1 064 654.71	1 206 505.67	17	14 705.66	11 923.17	1.048
9	1 206 505.67	1 348 356.64	27	16 205.56	11 333.50	0.890
10	1 348 356.64	1 490 207.60	31	17 018.71	13 039.02	0.934



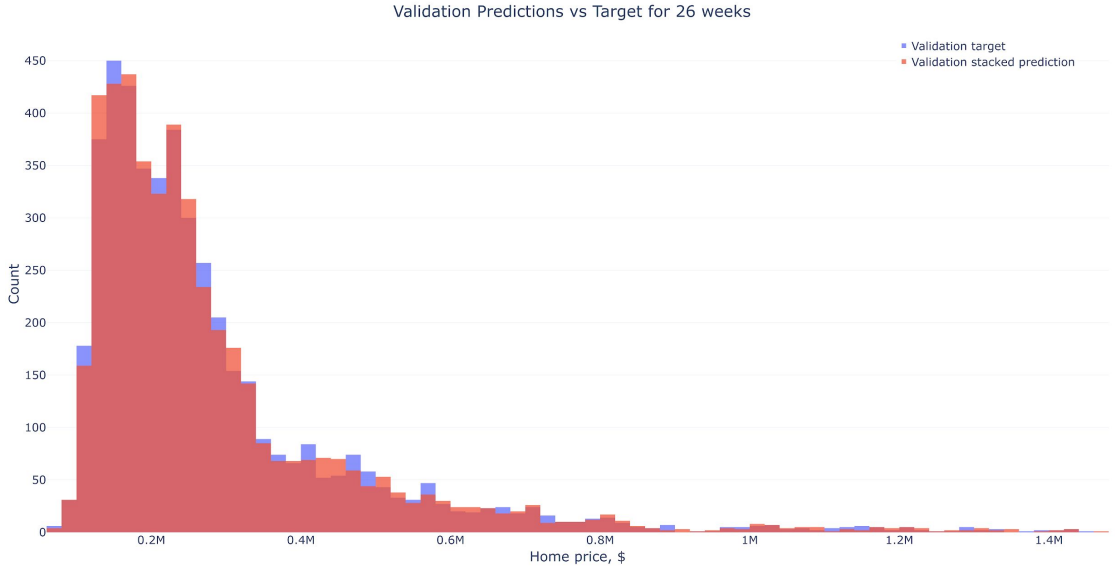


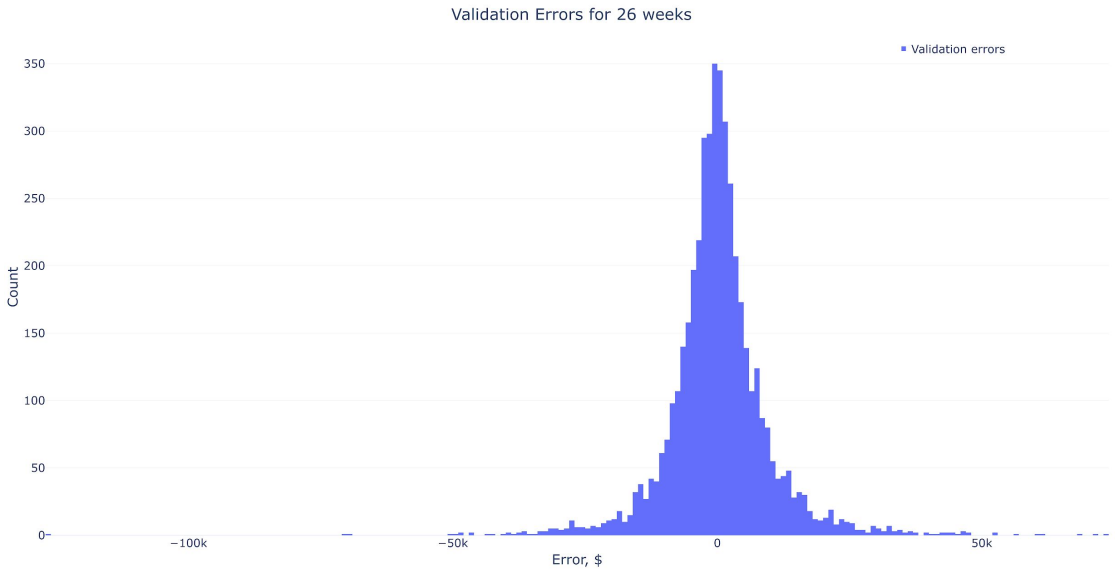
Bucket	Lower	Upper	Number	RMSE	MAE	MAPE (%)
1	71 940.71	208 808.34	2134	3 885.38	2 844.15	1.802
2	208 808.34	345 675.97	1635	6 545.16	4 806.00	1.815
3	345 675.97	482 543.60	485	9 404.34	6 999.32	1.715
4	482 543.60	619 411.23	269	12 292.54	8 987.82	1.643
5	619 411.23	756 278.86	118	17 112.45	13 374.09	1.923
6	756 278.86	893 146.48	53	19 522.15	13 986.48	1.686
7	893 146.48	1 030 014.11	21	18 369.57	13 537.09	1.399
8	1 030 014.11	1 166 881.74	17	27 960.95	23 721.49	2.182
9	1 166 881.74	1 303 749.37	11	19 942.16	17 029.94	1.379
10	1 303 749.37	1 440 617.00	20	24 718.41	19 663.45	1.420



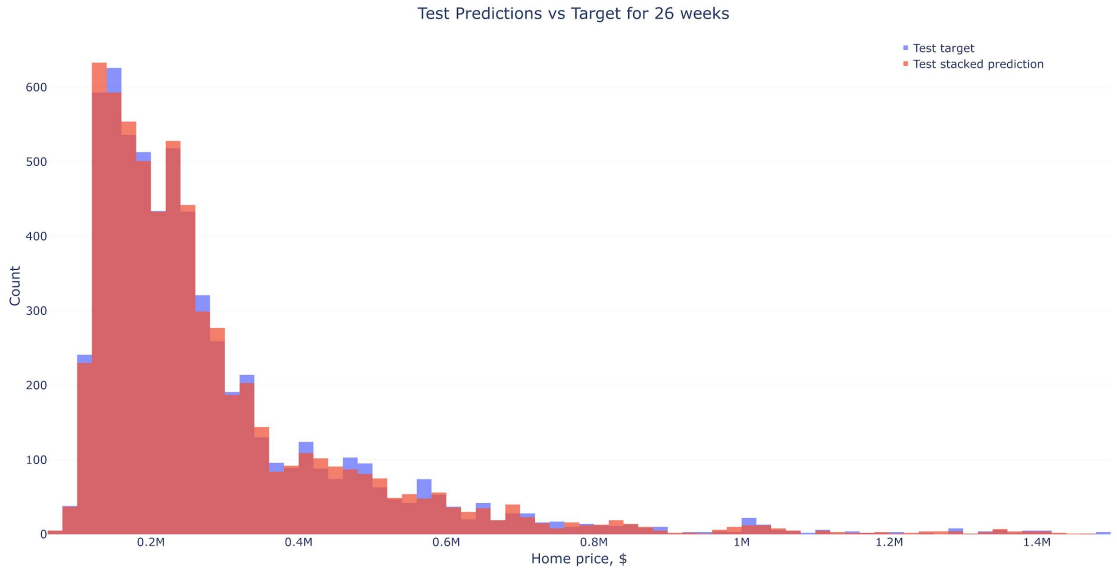


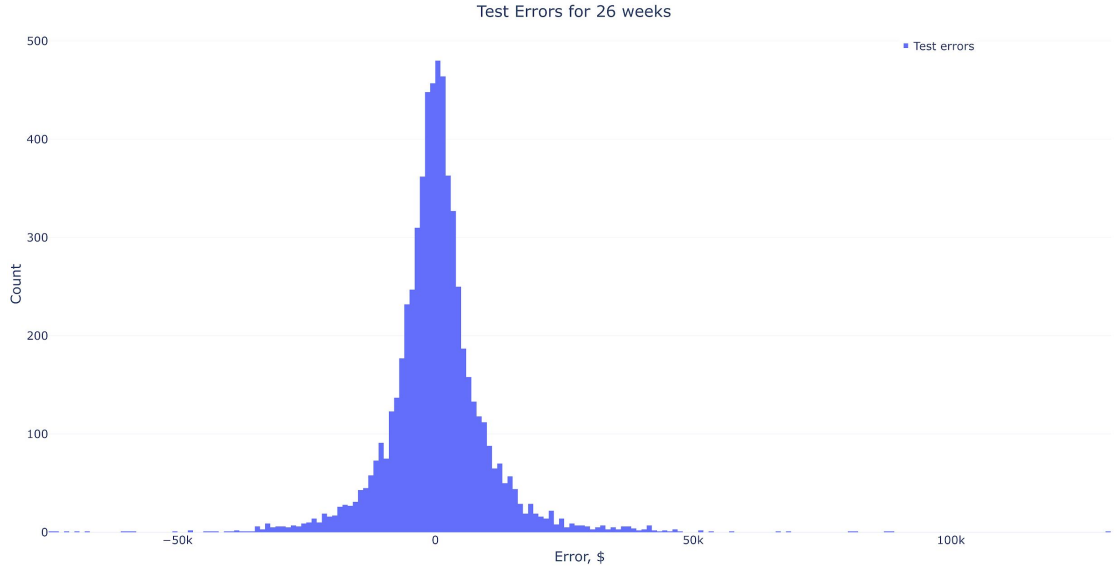
Bucket	Lower	Upper	Number	RMSE	MAE	MAPE (%)
1	71 808.14	213 231.28	3006	3 923.86	2 856.19	1.803
2	213 231.28	354 654.41	2196	6 644.10	4 893.08	1.831
3	354 654.41	496 077.55	673	9 480.41	7 070.48	1.658
4	496 077.55	637 500.68	320	10 875.83	8 259.94	1.483
5	637 500.68	778 923.82	177	13 461.65	9 937.24	1.424
6	778 923.82	920 346.96	75	19 098.24	14 854.79	1.772
7	920 346.96	1 061 770.09	67	19 535.86	14 656.46	1.450
8	1 061 770.09	1 203 193.23	20	21 294.32	16 519.76	1.455
9	1 203 193.23	1 344 616.36	30	24 917.96	20 457.48	1.581
10	1 344 616.36	1 486 039.50	29	25 157.16	20 588.26	1.460





Bucket	Lower	Upper	Number	RMSE	MAE	MAPE (%)
1	71 573.30	210 129.26	1972	4 788.13	3 566.69	2.255
2	210 129.26	348 685.22	1660	9 112.68	6 528.32	2.463
3	348 685.22	487 241.18	480	13 476.37	10 076.24	2.429
4	487 241.18	625 797.14	239	15 302.01	10 927.52	1.987
5	625 797.14	764 353.10	130	18 973.24	13 953.73	2.029
6	764 353.10	902 909.06	62	19 968.76	15 008.92	1.835
7	902 909.06	1 041 465.02	25	19 874.74	16 056.18	1.620
8	1 041 465.02	1 180 020.98	29	37 400.30	26 072.12	2.327
9	1 180 020.98	1 318 576.94	18	19 948.90	16 084.64	1.288
10	1 318 576.94	1 457 132.90	12	19 325.35	15 179.10	1.086





Bucket	Lower	Upper	Number	RMSE	MAE	MAPE (%)
1	71 808.14	213 394.63	2833	5 100.47	3 618.12	2.268
2	213 394.63	354 981.11	2182	8 968.12	6 569.51	2.446
3	354 981.11	496 567.60	689	12 333.60	9 370.73	2.205
4	496 567.60	638 154.08	350	16 029.55	11 593.84	2.068
5	638 154.08	779 740.57	163	20 440.70	14 961.85	2.144
6	779 740.57	921 327.06	70	18 519.91	13 882.00	1.658
7	921 327.06	1 062 913.54	56	32 207.30	20 504.45	2.059
8	1 062 913.54	1 204 500.03	21	39 304.83	32 319.29	2.852
9	1 204 500.03	1 346 086.51	21	20 503.36	15 542.85	1.198
10	1 346 086.51	1 487 673.00	19	23 328.11	20 431.10	1.441