Predicting tuna prices using machine learning

Rodney Beard

Transitional artificial intelligence research group
→ Pingala Institute (new nonprofit research institute)

July 15, 2024

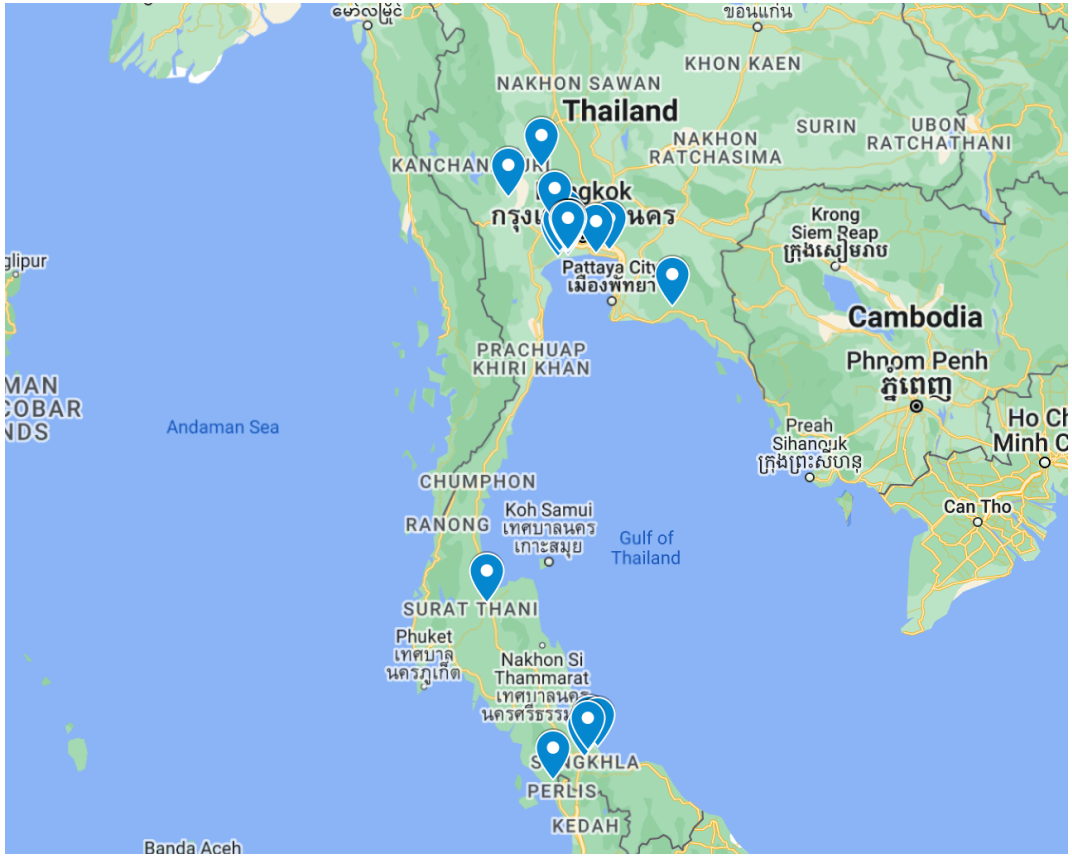
Background

- Western and Central Pacific Ocean
- World's largest Tuna fishery
- Tuna is a $ 4 billion annual industry
- Thailand is the main export market for Skipjack

Why is there interest in price prediction?

- Price predictions are used in planning exercises by various regional organizations in the Pacific
- Price predictions are used as a reference in bilateral and multilateral access fee negotiations
- Price predictions might sometimes be used in bio-economic models to make licensing decisions

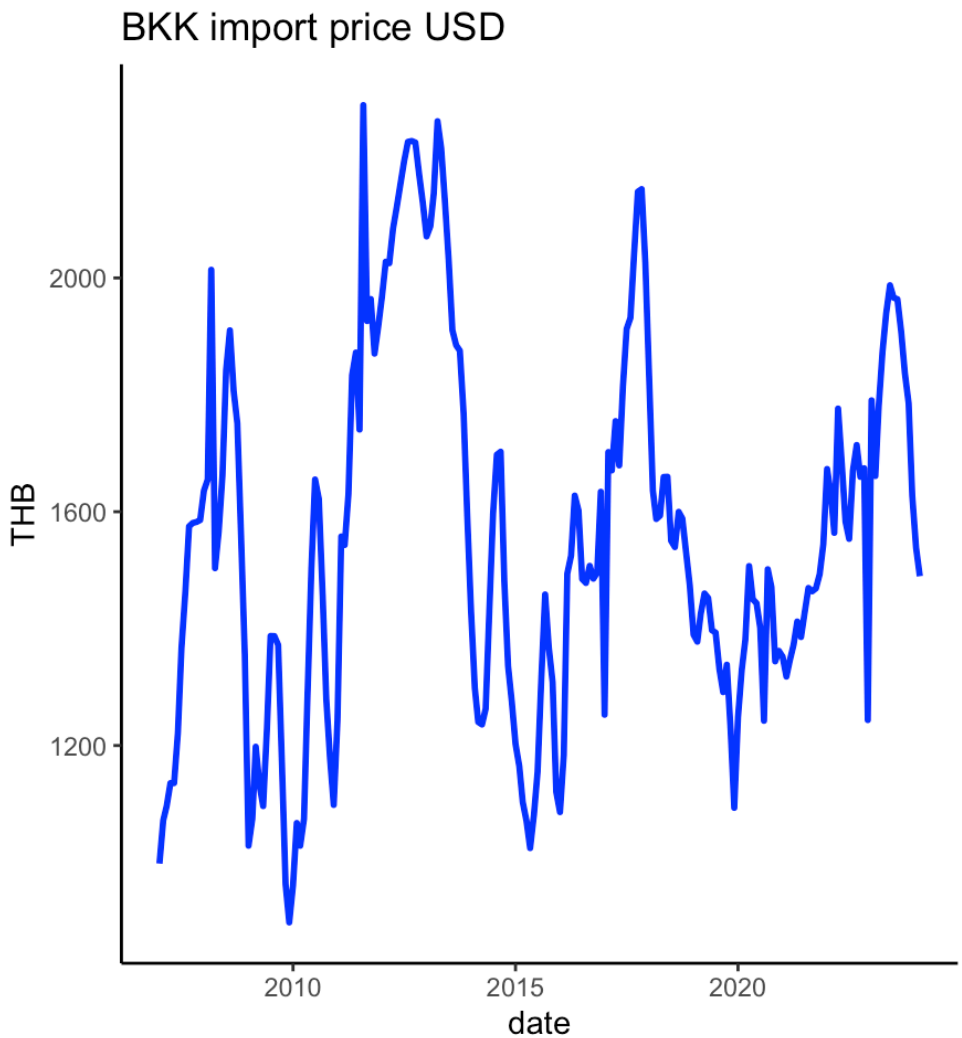
Price formation in Thailand: Thai canneries



Price data

- Survey of canneries (cost prohibits this)
- Thai union Skipjack data raw materials price (2011-present monthly and forward looking) Previously used this spliced with other data, questions about provenance of other data and how it was collected (data collection of this by FFA has been discontinued).
- Thai customs or Ministry of Commerce Import data (preferred source: Good monthly data back to 2007)

Thai import price series



Prior analyses 2018-2019

- We looked at price prediction using ARIMA initially (gives good in sample fit and short range predictions are good but not so good for longer range predictions)
- We then looked at using ARIMA-X and including some other variables however estimates wasnt consistent with an autoregressive process
- We also tried fitting a random forest model without success (poor out of sample performance)
- Finally, we tried support vector regression with some initial success which led us to pursue this approach (good out of sample performance)

Recent work at FFA and post FFA

- Exploration of price dispersion
- Experiments using Gradient Boosting methods (mostly LightGBM) 2022-2023
- Most recently have returned to looking at Support Vector Regression

Machine learning

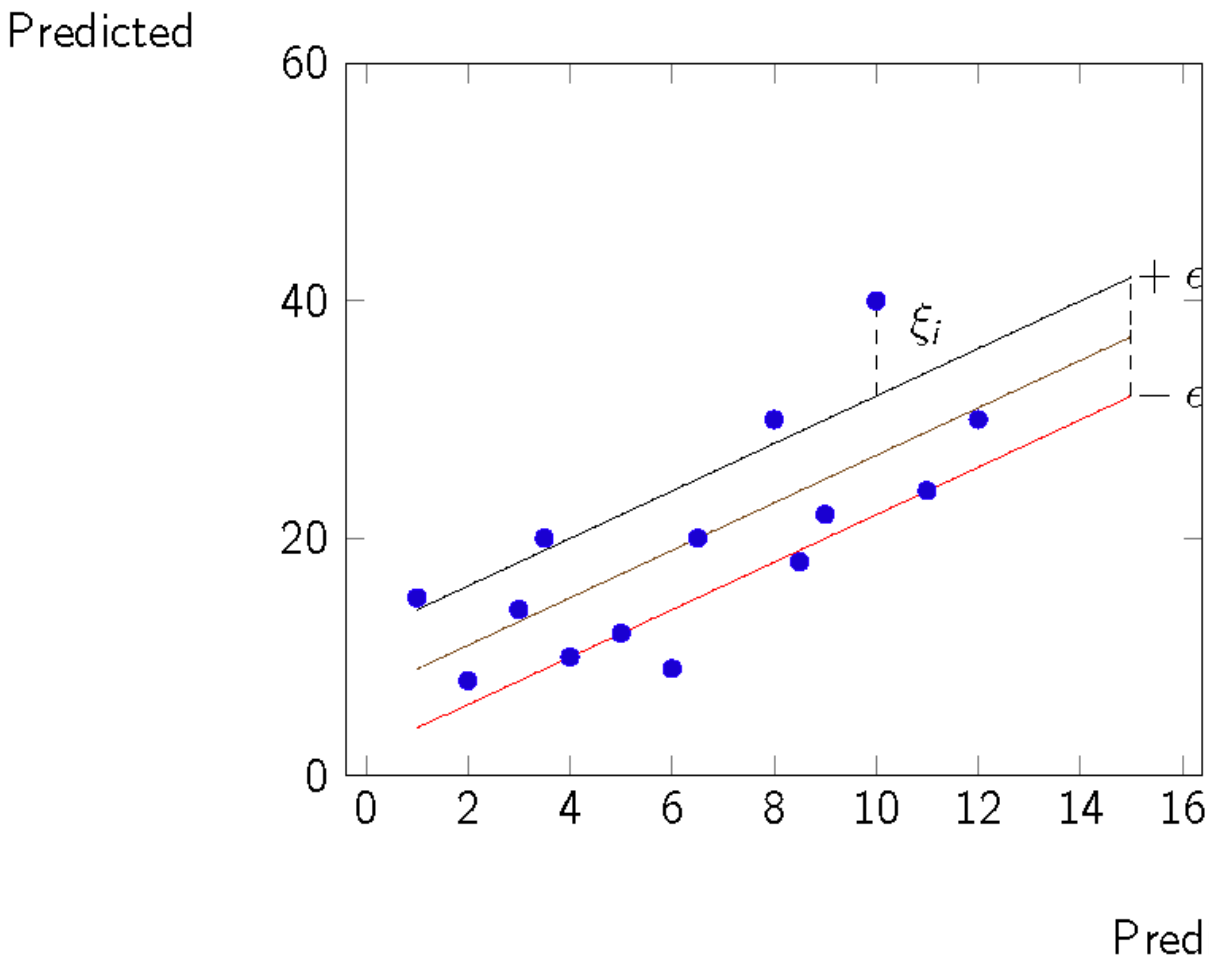
Traditionally econometrics aims to minimize bias whereas many machine learning methods trade off bias for low variance.

- Numerous approaches possible
- Neural networks (Deep learning methods - time consuming and not cost effective in a poorly resourced work environment)
- Tree based methods are popular in applied work for time series, Random forests, CATBoost, XGBoost, lightGBM
- Kernel based methods, e.g. Support vector regression (basically a more sophisticated version of quadratic programming)

Support vector regression

- Support vector regression is a machine learning technique Non-parametric and distribution free Originally developed for image classification and pattern recognition Bias-Variance trade-off.
- Method is basically quadratic programming

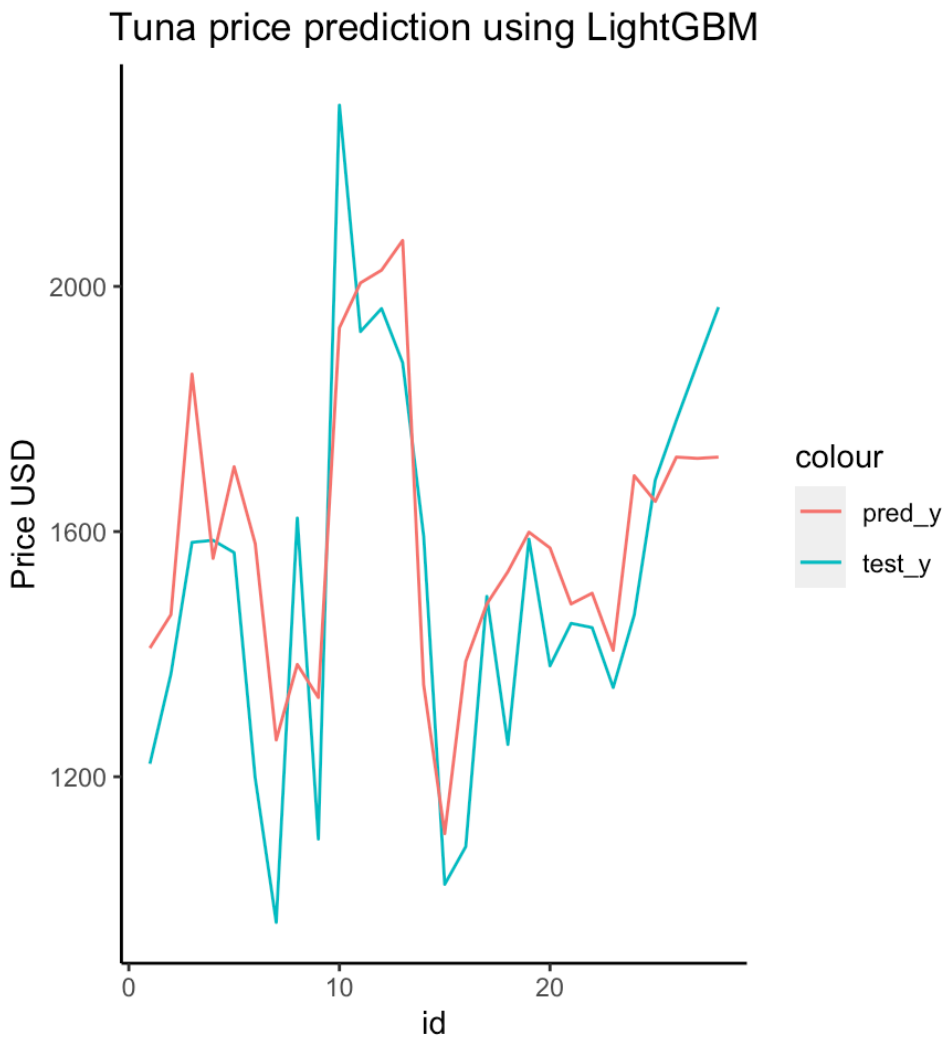
Support vector regression idea



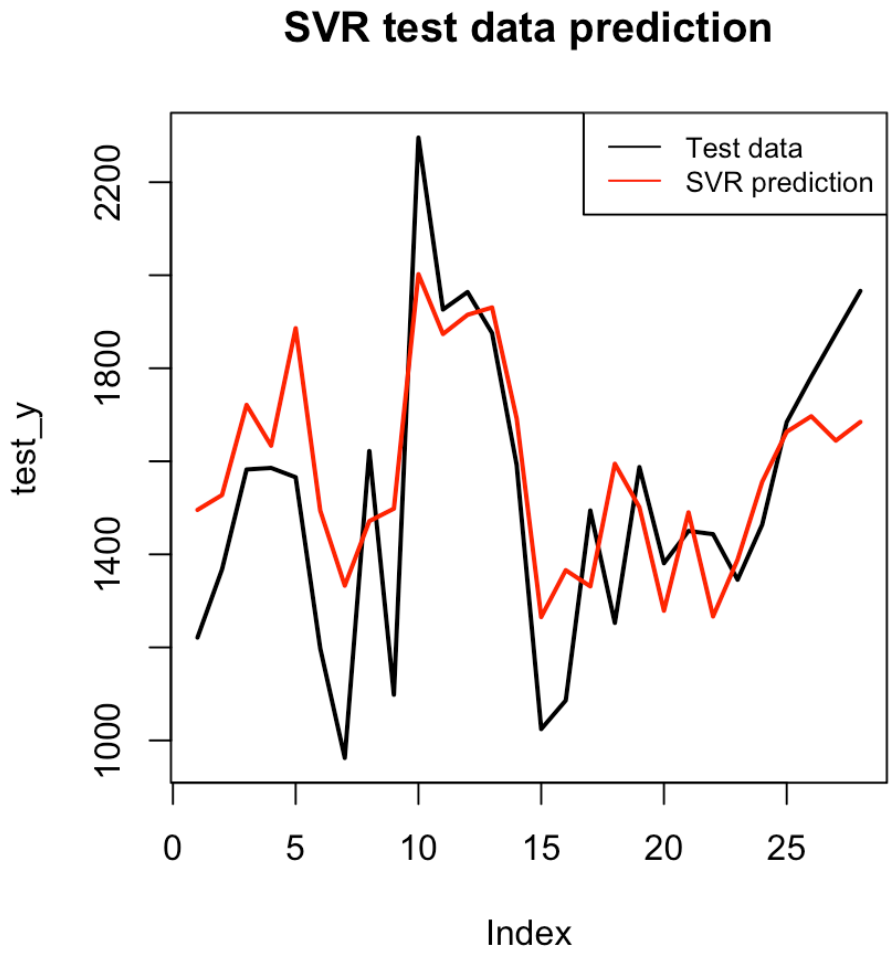
LightGBM

Light Gradient Boosting method: Simplified version of XGBoost (Extreme gradient boosting). LightGBM is an ensemble method that works by reweighting decision trees by varying terminal nodes (leaves), these essentially correspond to variables. The procedure adjusts weights by using steepest descent, a generalization of gradient descent methods (it doesn't fit parameters).

LightGBM results (out of sample prediction)



Support Vector Regression results (out of sample)



LightGBM, Support Vector Regression, ARIMA and ARIMA-X (out of sample)

(MAE), root mean square error (RMSE) and Accuracy %:

	MSE	MAE	RMSE	Accuracy (%)
LGBM	38944.41	163.61	197.34	0.89
SVR	43486.59	174.70	208.53	0.89
ARIMA	142306.31	310.56	377.24	0.81
ARIMA-X	141789.65	309.99	376.55	0.81

Summary

- LightGBM slightly better than SVR but similar accuracy around 90%
- All models point to moderate downward trend in prices for the remainder of the year.
- Multiple models not a single model (model stacking)
- Model designed to run off cloud database and APIs (semi-automated).
- Further work needed: comparing these models to CATBoost and XGBoost and possibly a deep learning model

Thanks for listening!

Rodney Beard
Rodney.M.Beard@gmail.com