Forecasting with ML Methods

Yoonhyok Choi

Department of Economics, SNU

February 6, 2025

Introduction

- Forecasting : Estimating the future value of a variable Y_{t+h} $(h \geq 1)$ given a set of variables X_t that consists of a constant plus $Y_t, Y_{t-1}, ...,$ and Y_{t-m+1} $(0 \leq m \leq t+1)$. (Hamilton, 1994)

- Machine Learning(ML) : Developing algorithms that predict some variables given others. (Athley and Imbens, 2019)

- There are some time series models used for forecasting (*e.g.*, Auto-Regressive(AR) Models or Dynamic Factor Models(DFM))

- But why should we consider ML methods as alternatives?

- There are some gains (*e.g.*, accuracy) with ML methods.

Contents

Basics of ML

- Training sample fit vs. prediction (Athley and Imbens, 2019)

- For example, consider OLS (Y: outcome; X: regressor (feature); T_r: # of obeservations for model estimation(training)):

$$\hat{\theta}_{ols} = (\hat{\alpha}_{ols}, \hat{\boldsymbol{\beta}}_{ols}) = \underset{\alpha,\beta}{arg\,min} \sum_{t=1}^{T_r} (Y_t - \alpha - \beta' \mathbf{X_t})^2$$

- However, in ML, the goal is to make a 'prediction' for the outcome for 'new units'(*e.g.*, Y_{T_r+1}) on the basis of their feature values(*e.g.*, X_{T_r+1}). Take an example of linear predictor, then the prediction is:

$$\hat{Y}_{T_r+1} = \hat{\alpha} + \hat{\beta}' \mathbf{X}_{T_r+1}$$

for some $(\hat{\alpha}, \hat{\beta})$. With the associated 'loss function'.

Basics of ML(Cont.d)

- Researchers try to find a model that fit well, but not so well that out-of-sample prediction is compromised. $\Rightarrow$ The problem of 'overfitting' or 'high variance'
- How to solve the problem of overfitting? $\Rightarrow$ **1**. reduce the number of features by discarding some of them *or* **2**. do 'regularization' : use all the features but put a (size) restriction on βs (shrinkage method).
- Augmenting penalty term would lead to a bias, but considerably reduces the variance (bias-variance trade-off).
- Therefore, the problem of overfitting can be resolved. $\Rightarrow$ *e.g.*, Ridge Regression

Basics of ML(Cont.d)

- Some examples of ML methods (Kapetanios and Papailias, 2018 ; Hansen 2022)

- Shrinkage Methods (*e.g.*, Ridge Regression ($p = 2$), and LASSO ($p = 1$))

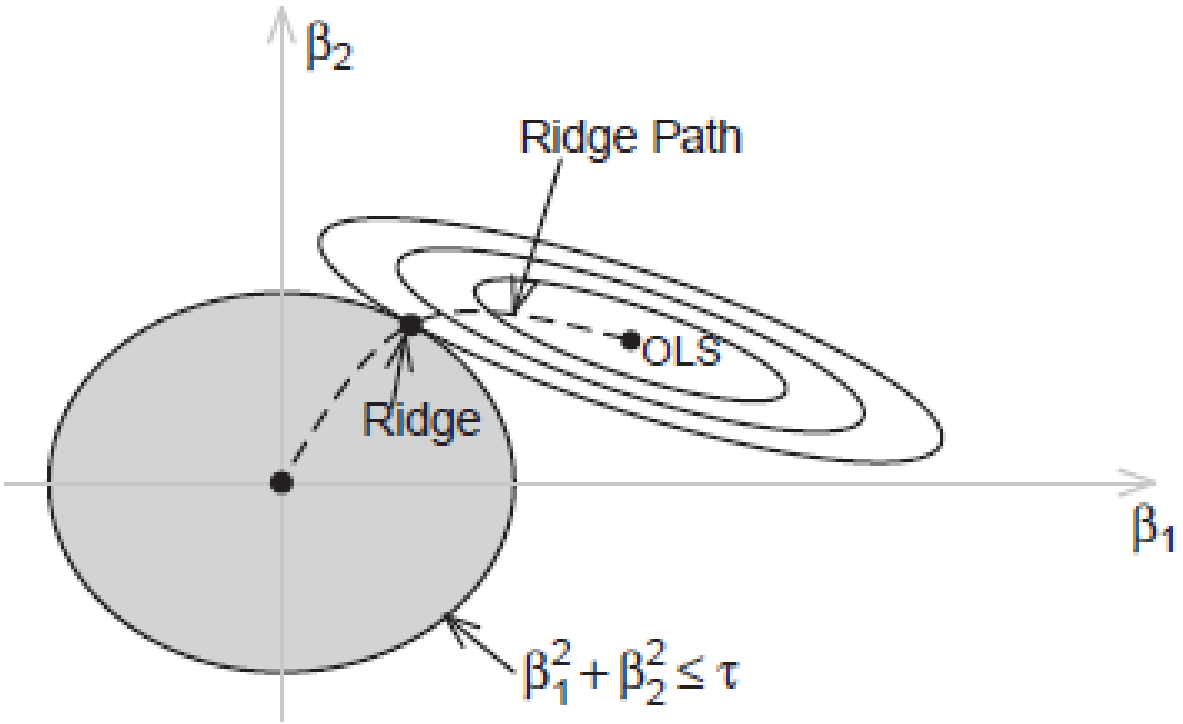
$$\hat{\beta} = \underset{\beta_N}{arg\,min} \sum_{t=1}^{T_r} (Y_t - \beta_N' X_{t,N})^2 + \lambda(\|\beta_N\|_p^p - \tau)$$

where there are N features, so that $\beta_N = (\beta_1, \beta_2, ..., \beta_N)'$, $X_N = (X_1, X_2, ..., X_N)'$, and τ (≥ 0) is a constant.

- The λ is a hyperparameter that regulates the relative importance of the penalty term, which is estimated to minimize the forecast error in cross-validation.

Basics of ML(Cont.d)

- Ridge Regression (Hansen(2022) Figure 29.1)



Basics of ML(Cont.d)

- Ridge Regression

$$\hat{\beta}^{Ridge} = \underset{\beta_N}{arg\,min} \sum_{t=1}^{T_r} (Y_t - \beta_N' X_{t,N})^2 + \lambda(\|\beta_N\|_2^2 - \tau) \tag{1}$$

where $\|\beta_N\|_2^2 = \sum_{i=1}^{N} \beta_i^2$, or equivalently,

$$\hat{\beta}^{Ridge} = \underset{\beta}{arg\,min} (\mathbf{Y} - \mathbf{X}\beta)'(\mathbf{Y} - \mathbf{X}\beta) + \lambda(\beta'\beta - \tau)$$

- Also, the followings hold :

$$\hat{\beta}^{Ridge} = (\mathbf{X}'\mathbf{X} + \lambda I_N)^{-1}\mathbf{X}'\mathbf{Y}$$

$$\hat{\beta}^{Ridge} \to \hat{\beta}^{OLS} \ as \ \lambda \to 0; \ \hat{\beta}^{Ridge} \to 0 \ as \ \lambda \to \infty$$

Basics of ML(Cont.d)

$$bias[\hat{\beta}^{Ridge}|\mathbf{X}] = -\lambda(\mathbf{X}'\mathbf{X} + \lambda I_N)^{-1}\beta$$

$$var[\hat{\beta}^{Ridge}|\mathbf{X}] = (\mathbf{X}'\mathbf{X} + \lambda I_N)^{-1}(\mathbf{X}'\mathbf{D}\mathbf{X})(\mathbf{X}'\mathbf{X} + \lambda I_N)^{-1}$$

where $\mathbf{D} = diag\{\sigma^2(X_1), ..., \sigma^2(X_{T_r})\}$.

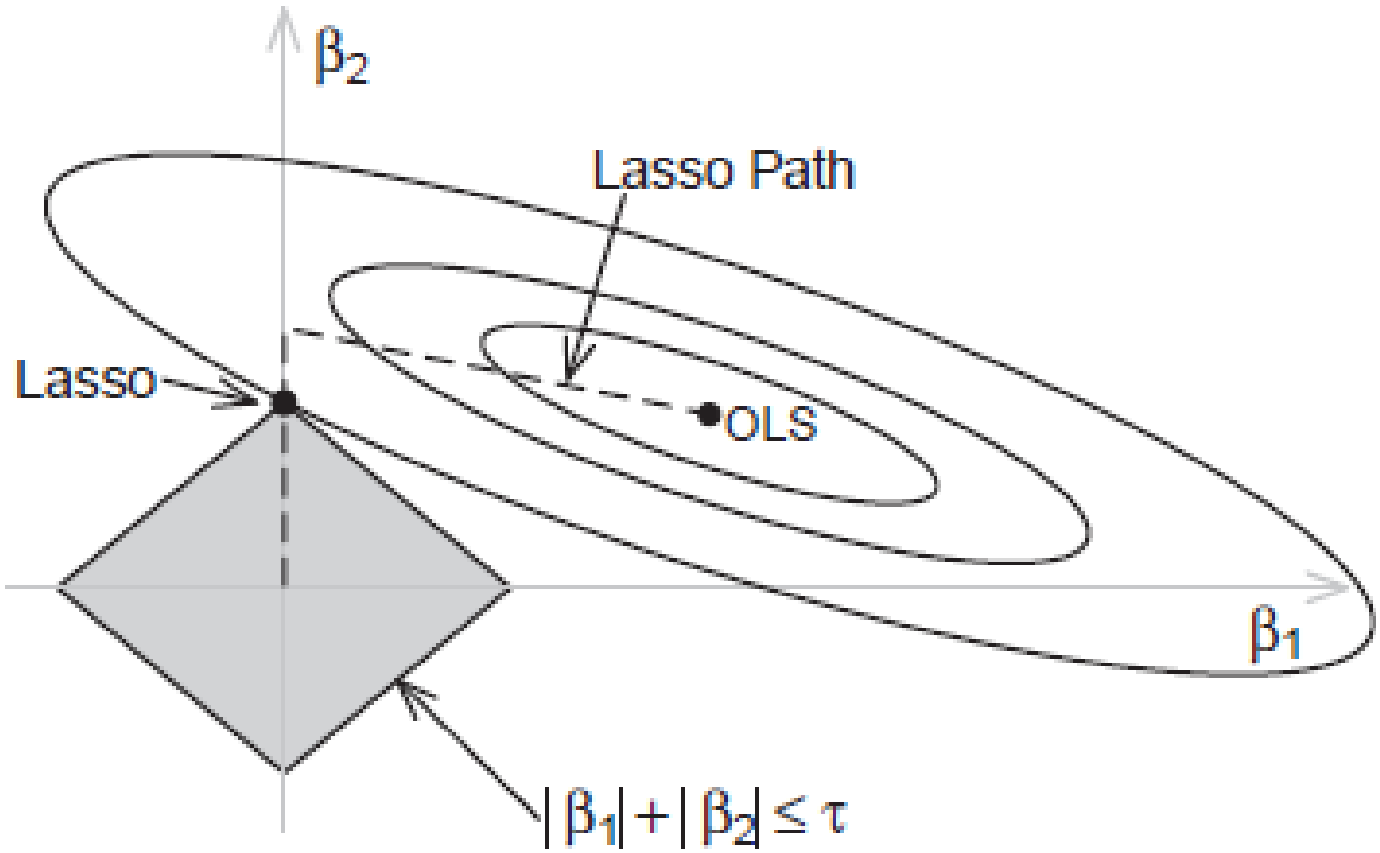
- According to Theorem 29.2 of Hansen(2022), the mean squared error(MSE) is smaller for Ridge Regression than OLS given a range of λ.

$$mse[\hat{\beta}^{Ridge}|\mathbf{X}] < mse[\hat{\beta}^{OLS}|\mathbf{X}]$$

where $mse[\hat{\beta}|X] = E[(\hat{\beta} - \beta)(\hat{\beta} - \beta)'|\mathbf{X}] \Rightarrow$ bias-variance trade-off

Basics of ML(Cont.d)

- Least Absolute Shrinkage and Selection Operator (LASSO) (Hansen(2022) Figure 29.3)



Basics of ML(Cont.d)

- Least Absolute Shrinkage and Selection Operator (LASSO)

$$\hat{\beta}^{LASSO} = \underset{\beta_N}{arg\,min} \sum_{t=1}^{T_r} (Y_t - \beta_N' X_{t,N})^2 + \lambda(\|\beta_N\|_1 - \tau) \tag{2}$$

where $\|\beta_N\|_1 = \sum_{i=1}^{N} |\beta_i|$

- Because of the nature of L_1-norm constraint, LASSO tends to produce some coefficients that are exactly 0 and hence gives more interpretable models than Rigde regression.

- Compared to Ridge Regression, LASSO has worse performance when there are collinearities between features, but have the advantage of clearly selecting which features to be included.

Basics of ML(Cont.d)

- Elastic Net

$$\hat{\beta}^{EN} = \left\{ \underset{\beta_N}{arg\,min} \sum_{t=1}^{T_r} (Y_t - \beta_N' X_{t,N})^2 + \lambda[(1-\alpha)\|\beta_N\|_2^2 + \alpha\|\beta_N\|_1] \right\} \quad (3)$$

- Zou and Hastie(2005) added an additional squared L_2-norm penalty because when there are two strongly correlated covariates, the LASSO may select one but typically not both of them.

ML Methods as Alternative Forecasting Models

- What advantage do ML models(methods) possess compared to traditional forecasting models (*e.g.*, AR models, DFMs) ?
- Compared to AR models, ML models handle overfitting problem by adding penalty terms to the objective function.
- When the number of features exceed the number of observation (*i.e.*, $N >> T_r$), ML methods provide a good way to handle the problem.
- Non-linear ML models tend to outperform other models in terms of forecast accuracy. (Coulombe *et al.*, (2022))
- However, the relative performance between ML vs. DFM is not clear : ML not superior(Coulombe *et al.*, (2022)) or some ML superior(Richardson *et al.*, (2021))

ML Applied in Nowcasting

- Nowcasting : Now + (Fore)casting (Bok *et al.* (2018))
- Many economists as well as central bank decision makers want to know the current state of the economy, in order to come up with proper policies.
- However, the 'real-time' data in macroeconomics is not guaranteed for all the major variables. (*e.g.*, GDP (Quarterly), Price Level/Inflation(Monthly))
- Also, the availability of big data poses a trade off between timeliness and quality (*e.g.*, google search log, credit card usage data).
- Dealing with big data and achieving higher forecast accuracy through complex models require overcoming the curse of dimensionality (*i.e.*, $N >> T_r$). $\Rightarrow$ ML ?

ML Applied in Nowcasting(Cont.d)

- Nowcasting real GDP (RGDP) growth of New Zealand (Richardson *et al.*(2021, International Journal of Forecasting)
- In this lecture, focus on the performance of some ML methods (Ridge regression, LASSO, and Elastic Net) against AR or DFM.
- 'Some' ML methods outperform AR of DFM in nowcasting RGDP growth (2009 Q1 - 2019 Q1) for New Zealand.

Forecasting Models

- Auto-Regressive(AR) model
- As a benchmark, use a univariate AR model of order 1 (AR(1)) for quarterly GDP growth (y_t):

$$y_t = \alpha_0 + \alpha_1 y_{t-1} + u_t \tag{4}$$

where α_0 and α_1 ($|\alpha_1| < 1$) are parameters, and u_t is the residual term with i.i.d., zero mean, and a positive variance.

Forecasting Models(Cont.d)

- Dynamic Factor Model (DFM) (Stock and Watson, 2002)
- An outcome (y_t) can be explained by the (unobservable) few factors (F_t), which captures the common component of predictors (X_t) $\Rightarrow$ Dimension Reduction

$$y_{t+1} = \beta' F_t + \gamma(L) y_t + \epsilon_{t+1} \tag{5-1}$$

$$X_t = \Lambda F_t + e_t \tag{5-2}$$

where X_t is an N-dimensional multiple time series of predictor variables,
$e_t = (e_{1t}, ..., e_{Nt})'$ is the $N \times 1$ idiosyncratic disturbance,
$F_t = (f_t', ...f_{t-q}')'$ is the matrix of p common factors with lags q , and
$\gamma(L) = \sum_{j=0}^{r-1} \gamma_j L^j$ (AR lag of r)

Forecasting Models(Cont.d)

- Since N is large (especially, $N >> T_r$), use Principal Component Analysis (PCA) to estimate $\hat{F}_t$ in (5-2)
- PCA : Summarizing data (dimensionality reduction) with several principal components that maximizes the variance of the data. (cov $\rightarrow$ eigval decomp. $\rightarrow$ p eigvec corrsp. p largest eigval)
- Then, using $\hat{F}_t$, estimate (5-1), just as in AR(1) model in (4).
- In this exercise, we estimate the model with four factors ($p = 4$), five factor lags ($q = 5$) and one lag of the target variables ($r = 1$).
- ML models (Ridge regression, LASSO, and Elastic Net) as explained earlier.

Data

- Numerous macroeconomic and financial data in New Zealand (from 1995 Q1) used.
- Including New Zealand business surveys, consumer and producer prices, general domestic activity indicators, domestic trade statistics, international macroeconomic variables, and international & domestic financial market variables.
- Mean of daily or monthly data for transforming higher into lower frequency.
- 41 real-time vintages with 532 to 634 series in each vintage.
- 41 vintages correspond to the number of quarters between 2009 Q1-2019 Q1. To avoid look-ahead bias in real-time forecast, use different vintages to predict each quarterly GDP growth.

Model Estimation and Hyperparameter Optimization

- In order to estimate ML models and choose optimal hyperparameters (*e.g.*, λ), we devide the predictors (X) and target data (y) into *training*, *validation*, and *test sets*.
- The model is estimated using the data in the training set.
- Optimal hyperparameters are then chosen by finding the parameters that minimize the forecast error in the validation set.
- We reserve the test set to assess the out-of-sample predictive accuracy of the fitted model, once hyperparameter optimization is complete.

Model Estimation and Hyperparameter Optimization (Cont'd)

- In practice, the training and validation sets are combined, and we use k-fold cross-validation to optimize parameters.

1. We devide the data into five random partitions, and one partition is held back to be the validation set.
2. The model is fitted on the remaining four partitions.
3. We then specify the hyperparameters and the range to explore (to avoid overfitting) for each hyperparameter.
4. We then use a search algorithm to find the set of hyperparameter values that minimizes forecast errors on the partition that we held back.
5. This process is repeated, so each partition is used as a validation set.

k-fold cross validation

- Usually, the data set have to be divided into *training*, *validation*, and *test sets*.
- However, k-fold cross-validation first divides the whole dataset into 'training/validation set' and 'test set' and then further divides the former set into k-distinct folds. $\Rightarrow$ We can use more diverse data set to estimate the model and hyperparameters.
- Assign one fold as a 'validation set' and the other $k-1$ folds as 'training set'.
- Estimate the model using the 'training set' and get the optimal hyperparameters that minimies the forecast error using the 'validation set'.
- Repeat the same process while replacing the old 'validation set' by the new 'validation set' from the previous 'training set', until all the folds are used as 'validation set'.

k-fold cross validation (Cont'd)

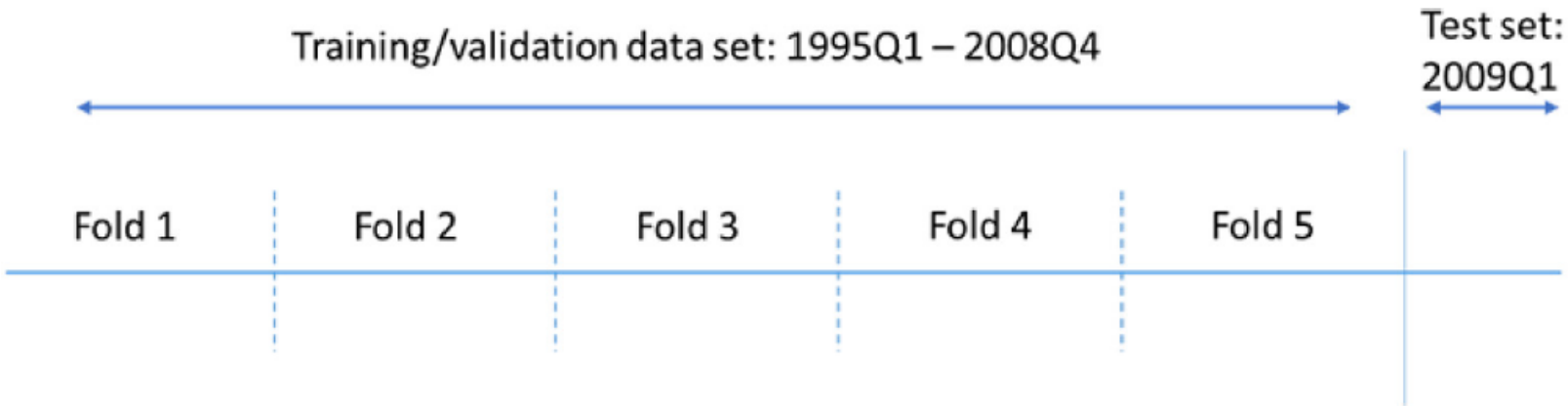
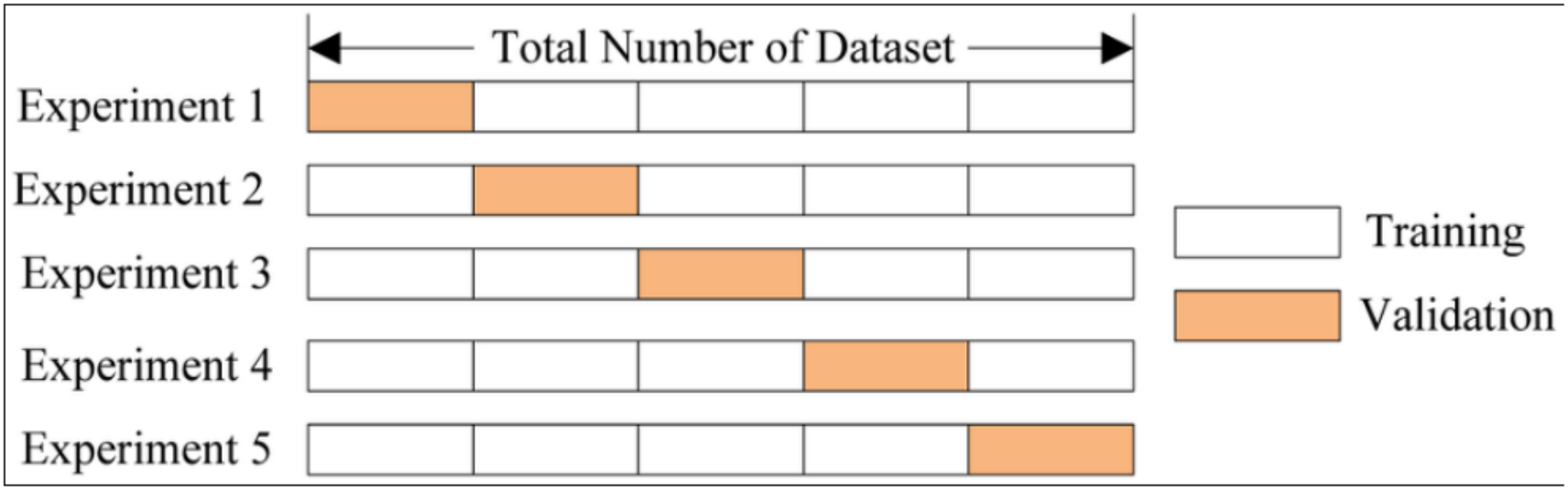


Fig. 1. Stylised representation of model estimation and forecast procedure..



Forecast Evaluation Methodology

- Root Mean Squared Error(RMSE) is used to measure the forecast accuracy of each model, defined as:

$$RMSE = \sqrt{\frac{\sum_{t=1}^{T}(y_t - \hat{y}_t)^2}{T}} \tag{5}$$

  where y_t and $\hat{y}_t$ are the actual and forecast values of GDP growth, respectively.

- Also, we use the Diebold-Mariano test (Diebold-Mariano, 1995) to determine whether the forecasts obtained from each ML model are significantly different than those from the AR model.

Diebold-Mariano test (Diebold and Mariano, 1995)

- A test of null hypothesis of no difference in the accuracy of two competing forecasts.
- Consider two forecasts, $\{\hat{y}_{it}\}_{t=1}^{T}$ and $\{\hat{y}_{jt}\}_{t=1}^{T}$, of time series $\{y_t\}_{t=1}^{T}$. Let the associated forecast errors be $\{e_{it}\}_{t=1}^{T}$ and $\{e_{jt}\}_{t=1}^{T}$.
- *WLOG*, let the loss from imprecise forecasting $(g(y_t, \hat{y}_{it}))$ to be a direct function of the forecast error $(g(e_{it}))$, *i.e.*, $g(y_t, \hat{y}_{it}) = g(e_{it})$. Also, let the differential d_t to be $d_t \equiv g(e_{it}) - g(e_{jt})$.
- Then the null hypothesis is a zero-median loss differential, *i.e.*, $H_0 : med(g(e_{it}) - g(e_{jt})) = 0$ (Then, $H_1 : med(g(e_{it}) - g(e_{jt})) \neq 0$)

Diebold-Mariano test (Cont'd)

- Assuming that the loss-functional series is *i.i.d.*, the number of positive loss-differential observations in a sample size of T : $S \sim Binomial(T, 1/2)$ under H_0.

- The test statistic :

$$S = \sum_{t=1}^{T} I_+(d_t) \tag{6}$$

where

$$I_+(d_t) = 1 \ if \ d_t > 0$$
$$= 0 \ otherwise.$$

- Significance may be assessed with a table of the cumulative binomial distribution. In large samples, the studentized version of the sign-test statistic is standard normal.

Results

Table 1
Hyperparameter ranges and values.

Model	Hyperparameter	Range	Optimised value
SVM	C	0.8 to 1.2	0.56
	Epsilon	iqr(Y)/13.49 +/− 20 percent	0.05
	Kernel function	Linear/Polynomial/RBF	Polynomial
	Polynomial order	2 to 4	2
LSboost	No. of learning cycles	500 to 1500	1498
	Learn rate	0.001 to 0.15	0.0075
	Max no. of splits	9 to 13	9
Elastic net	Alpha	0.1 to 0.9	0.5
	Lambda	0.1 to 0.9	0.1
Lasso	Lambda	0.001 to 0.9	0.011
Ridge	Lambda	0.01 to 0.99	0.989

Results(Cont'd)

Table 2
Real-time nowcast performance of models and Reserve Bank (RMSE), 2009Q1–2019Q1.

Models	RMSE	RMSE (Rel. to AR)	*p*-value
LSBoost	0.3571	0.719	0.028
SVM	0.3796	0.764	0.049
Neural net	0.4070	0.819	0.127
Lasso	0.4387	0.883	0.310
Elastic net	0.4280	0.861	0.272
Ridge	0.4374	0.880	0.331
RBNZ	0.3673	0.739	0.037
Model average	0.3806	0.766	0.055
Factor	0.6010	1.209	0.178
AR	0.4970		

Notes: The fourth column refers to the p-values obtained from the Diebold–Mariano test of the significance of the forecast accuracy of each method versus that of the AR model.

Results(Cont'd)

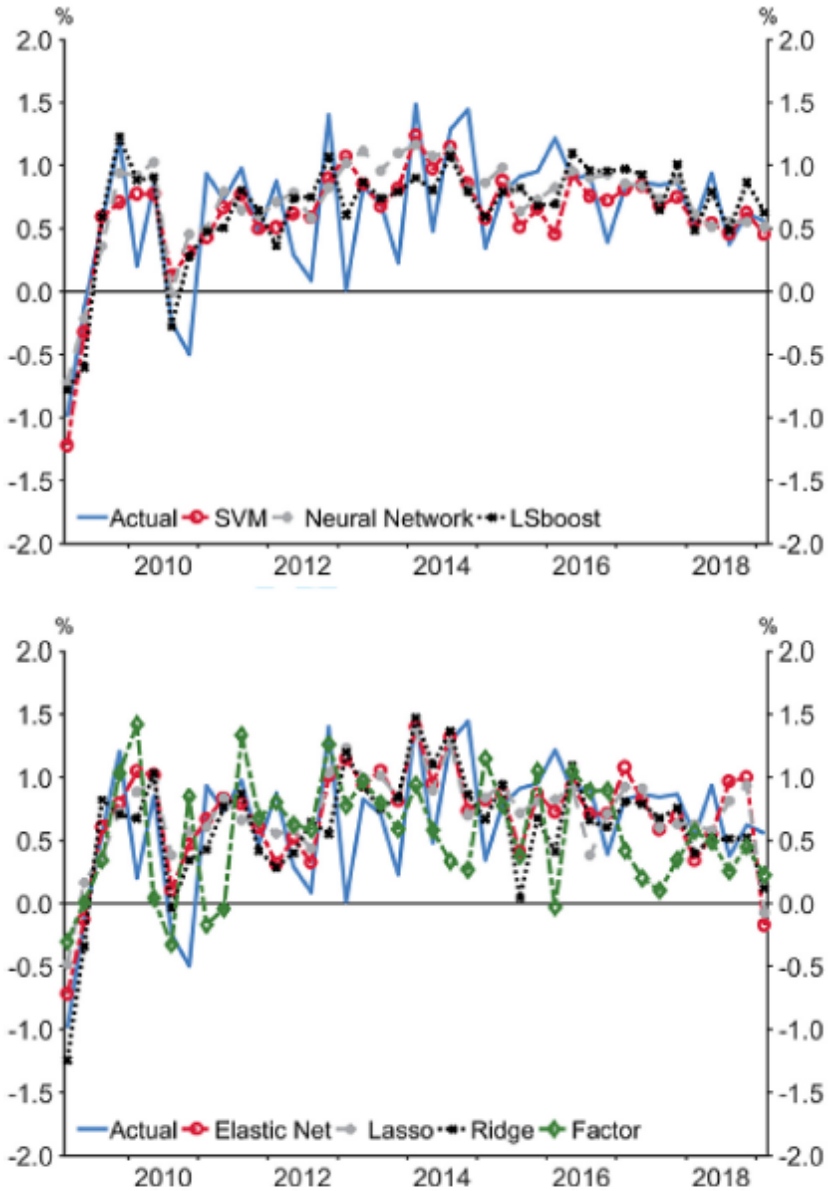


Fig. 2. Real-time nowcasts of quarterly GDP growth.

Comparison: Switzerland GDP(Arro-Cannarsa and Scheufele, 2024) (2009 Q3 - 2019 Q4)

Table 2: Results Overview

Model	RMSE	Benchmark comparison	
		AR	PCA
AR	0.421		25%
PCA	0.337	-20%**	
FSS	0.486	15%	44%
LASSO	0.315	-25%**	-6%
Ridge	0.310	-26%**	-8%*
Elastic Net	0.303	-28%***	-10%**
Bagging	0.354	-16%**	5%
Random Forest	0.353	-16%**	5%
Boosting	0.339	-19%**	1%
SVR	0.308	-27%***	-8%*

Notes: The table reports in column 2 the RMSE of models described in section 2. The models were estimated via five-fold cross-validation. Columns 3 and 4 show relative gains/losses in terms of RMSE compared to benchmark models AR and PCA, respectively. The stars represent the significance levels of one-sided Diebold-Mariano test: *** significant at 1%, ** significant at 5%, * significant at 10%.

Conclusion

- We have studied some applications of ML methods to forecasting GDP.
- ML methods can be an alternative to traditional methods of forecasting(AR, DFM) especially when there are many features ($N >> T_r$) or complex models are required.
- ML methods generally have smaller RMSE than traditional methods, but the superiority in accuracy may differ by specific ML models and the types of outcome variables.

The End

Thank You!

References

- Arro-Cannarsa, Milen, & Rolf Scheufele. (2024). "Nowcasting GDP: what are the gains from machine learning algorithms?" *Swiss National Bank(SNB) Working Papers 2024-06*.

- Athley, Susan., & Guido Imbens. (2019). "Machine Learning Methods That Economists Should Know About". *Annual Review of Economics*, 11(1): 685-725.

- Bok, Brandyn, Daniele Caratelli, Domenico Giannone, Argia M. Sbordone, & Andrea Tambalotti. (2018). "Macroeconomic Nowcasting and Forecasting with Big Data." *Annual Review of Economics*, 10: 615-643.

- Coulombe, Philippe Goulet, Maxime Leroux, Dalibor Stevanovic, & Stéphane Surprenant. (2022). "How is machine learning useful for macroeconomic forecasting?" *Journal of Applied Econometrics*, 37(5): 841-1090.

References (Cont'd)

- Diebold, Francis X., & Roberto S. Mariano. (1995). "Comparing predictive accuracy." *Journal of Business & Economic Statistics*, 13(3):253-263.
- Hamilton, James D. (1994). "Time Series Analysis." *Princeton University Press*.
- Hansen, Bruce. (2022). "Econometrics." *Princeton University Press*.
- Kapetanios, George, & Fotis Papailias. (2018). "Big Data & Macroeconomic Nowcasting: Methodological Review." *ESCoE Discussion Paper 2018-12*.

References (Cont'd)

- Stock, James H., & Mark W. Watson. (2002). "Macroeconomic Forecasting Using Diffusion Indexes." *Journal of Business & Economic Statistics*, 20(2): 147-162.

- Zou, Hui, & Trevor Hastie. (2005). "Regularization and Variable Selection Via the Elastic Net". *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 67(2): 301-320.