National University of Singapore

Forecasting 20-year bond yields of Singapore

EC4308 Machine Learning and Economic Forecasting AY 2023/2024 Sem 1

Dr Vu Thanh Hai

Name	Matriculation Number
Chew Shi Zhan	A0216042W
Chen Liangyan	A0201881L
Lim Zhi Yuan	A0201781M
Lee Jun Yi Aaron	A0217895M
Siti Maryani Binte Siful Bahri	A0222983Y

1. Introduction

In this report, we forecast Singapore's 20-year (20Y) monthly yield, a fundamental financial metric that plays a pivotal role in shaping the country's economic outlook using data from January 2009 to October 2023.

1.1 Background

Importance of Forecasting Bond Yield

Compared to the returns of other asset classes, fixed-income instruments had not been the subject of much study in recent decades. This may be due to the extremely low rates in the years following the 2009 Great Recession and the COVID-19 Pandemic that occurred shortly after, resulting in the There Is No Alternative (to equity) (TINA) regime in recent years.

However, starting in 2022, many Central Banks around the world had begun hiking interest rates to rein in possible inflation influenced by supply-chain shortages and the prior low-rate regimes. Furthermore, the bond markets in the recent period have been the subject of much discussion, especially with the recent rout in the bond market. The US 10-year treasury yields touched 4.9% p.a. in the second half of October, its highest level since July 2007. 2022 was the worst year in terms of returns for US bond investors since 1870 (Ferguson, 2023).

Therefore, there seems to be good motivation for studying the forecastability of interest rates. In this paper, we attempt to use machine learning methods to forecast Singapore government bond yields using macroeconomic factors. We will then compare the performance of these models against two commonly used benchmarks – random walk and AR(1) – in econometrics.

1.2 Interest rate models

Interest rates and yields have been modelled in generally two categories of methods:

I. Directly modelling spot rates across different periods using time-series analysis (yields-only models)

II. Fitting a term structure to model spot rates in each period (term-structure models).

For instance, the first category of method includes modelling 20Y yields using an AR(1) model:

$$Y_{t+1} = \Phi Y_t + e_t$$

Equation 1: AR(1) Model

The second category would include the Nelson-Siegel model:

$$y(\tau) = \beta_1 + \beta_2 \left(\frac{1 - e^{-\tau\lambda}}{\tau\lambda} \right) + \beta_3 \left(\frac{1 - e^{-\tau\lambda}}{\tau\lambda} - e^{-\tau\lambda} \right)$$

Equation 2: Nelson-Siegel Model

In our case here, we chose to use the first method and look to forecast only the yields of 20Y SGS bills. There are several reasons for this, but the largest factor is practical. A term structure model's purpose is to fit a curve onto the spot rates with respect to time-to-maturity in a single period. Its purpose is thus not so much about forecasting future yields. In fact, it has been shown that simple AR(1) yields-only models outperform dynamic Nelson-Siegel models in out-of-sample forecasts (Tsui et al., 2022).

Since term structure models generally fit the yields very well in a given period, there is little economic justification for predicting the (largely uncorrelated) residuals in these models using macroeconomic factors. Instead, these factors (**X**) must be used to explain the parameters of the yield curve models $\boldsymbol{\beta}_t = (\beta_1, \beta_2, \beta_3)^T$, which is often done with state-space models.

$$\boldsymbol{y}_t = \mathbf{N}\boldsymbol{\beta}_t + \boldsymbol{\epsilon}_t$$

$$\boldsymbol{\beta}_t = \mathbf{A}\boldsymbol{\beta}_{t-1} + \mathbf{B}\mathbf{X}_{t-1} + \boldsymbol{\omega}_t$$

Equation 3: State-Space Model example

However, such an approach would be impractical for our purposes, which involve high dimensions of macroeconomic factors. The state-space models will be too complex and thus computationally costly. In any case, if the objective was to instead predict term structures in the future, it is possible to apply our chosen method to predict future spot rates of varying maturities

1.3 The choice of 20Y yield

There are three main reasons we chose to study and forecast specifically the 20Y yields. Firstly, long-maturity bonds are subject to higher interest rate risk, which is caused by and contributed to the recent bond rout. Secondly, long-maturity bonds are generally more illiquid, disqualifying many models that rely heavily on arbitrage assumptions, but not necessarily those that rely on macroeconomic factors. Lastly, because of the above reasons, we are interested in yields of the longest maturity possible, and 20Y is the longest maturity SGS bond that has data from 2009-2023.

1.4 Macroeconomic factors

It has been shown that among Asia-Pacific countries, bond yields and macroeconomic factors are related (Chernov et al., 2019) within a country. Thus, there is a theoretical basis for the inclusion of certain macroeconomic factors for forecasting purposes, namely economic growth, exchange rate depreciation, and inflation rate of Singapore.

Generally, models with macroeconomic factors have been shown to improve predictability of interest rates, regardless of the nature of the underlying time series model. This is especially the case when there are recessions, and such macro models have outperformed random walk. In particular, AR models with macroeconomic factors performed better than simple time series and other models with macroeconomic factors in the short forecast horizon (De Pooter et al., 2010).

Factors from other economies are included because the globalisation of financial markets should intuitively increase the influence of other countries' spot rates on Singapore's bond yields. This was evident for Japanese yields (Suimon et al., 2019). Given that other countries' yields are likely affected by their own macroeconomic factors, they are included as well. In particular, we included the data of four countries that have economic relationships with Singapore and also the largest bond markets in the world – China, Japan, the UK, and the US.

Equity performance of each economy is included since it can be indicative of the economy's health and growth. Furthermore, bond yields and equity yields are often correlated because investors often have portfolios containing assets of both classes. The prices of the major stock indices are used - Shanghai Stock Exchange Composite, Nikkei 225, FTSE 100, S&P 500, and Straits Times Index.

2. Data

Our data set includes the interest rates of Singapore, China, Japan, UK and the US, and then macroeconomic factors of these economies. All yield data are obtained from the respective Central Bank and ChinaBond websites. Exchange rates and price of major stock indices (to measure equity performance) are obtained from Yahoo Finance and Wall Street Journal. Industrial Production Indices (used as measures of economic growth by Chernov et al., 2019) are obtained from the OECD database for Japan, UK, and the US, while Singapore’s is obtained from SingStat. China does not use an Industrial Production Index, but the OECD database includes China’s manufacturing production indicator, which we will use as an equivalent. We also included Singapore Merchandise Trade, which is published by MTI on SingStat to measure Singapore’s trade performance.

Consumer Price Indices are obtained from the OECD database for Japan, US, and UK, while Singapore’s CPI is obtained from SingStat and China’s CPI is obtained from the website of its National Bureau of Statistics.

2.1 Data Cleaning and Transformation

Not all data available in the above mentioned sources are used in our research. While the Bank of England publishes a very comprehensive yield curve (with 100 maturities from 1-month to 25-year), this is done using their own yield curve fitting methodology. Aside from the obvious impracticality in using all 100, the data also contained many missing values for the very short end of the curve. Thus, we only included maturities for which there are actual UK bonds sold.

Similarly, there are many missing values for Singapore’s bonds, given the frequent changes in the type of SGS bonds available on the market, so we removed all maturities that have incomplete bond yields from 2009-2023.

To ensure the congruence between the dependent and independent variables, we use growth rates of macroeconomic factors to predict bond yields (which can be seen as a type of “growth rate”). Thus, these data are transformed simply by taking:

$$\Delta X_t = (X_t - X_{t-1})/X_{t-1}$$

Lastly, given the goal here is 1-step ahead forecasting, all features used are from one period prior to the time period of the response variable. All features are lagged by 1 period, with the exception of the Industrial Production Index. Industrial production data are usually published one month later, e.g., the industrial production of the month of September is published in mid- to late-October. Thus, these will be lagged by 2 periods.

2.2 Selected features and response variable

In short, the following are the features used in the study,

Macroeconomic factors

Economic Indicator	China	Japan	UK	US	Singapore
Economic growth (lagged 2 periods)	*prcipgrl2*	*jpigrl2*	*ukipgrl2*	*usipgrl2*	*sgipgrl2*
Inflation rate (lagged 1 period)	*prcinfl1*	*jpinfl1*	*ukinfl1*	*usinfl1*	*sginfl1*
Depreciation rate (lagged 1 period)	*prcdeprl1*	*jpdeprl1*	*ukdeprl1*	-	*sgdeprl1*
Equity returns (lagged 1 period)	*prcequityl1*	*jpequityl1*	*ukequityl1*	*usequityl1*	*sgequityl1*
Trade growth (lagged 1 period)	*sgtradegrl1*	-	-	-	-

Table 1: Summary of economic indicators used as features

Bond Yields

Duration	China	Japan	UK	US	Singapore
<1 year	*prc3ml1* *prc6ml1* *prc1y1l*	*jp1yl1*	*uk1yl1*	*us1ml1* *us3ml1* *us6ml1* *us1yl1*	*sg1yl1*
1-10 year	*prc3yl1* *prc5yl1* *prc7yl1*	*jp2yl1* *jp3yl1* *jp5yl1* *jp7yl1* *jp10yl1*	*uk3yl1* *uk5yl1* *uk7yl1* *uk10yl1*	*us2yl1* *us3yl1* *us5yl1* *us7yl1* *us10yl1*	*sg5yl1* *sg10yl1*
>10 year	*prc10yl1* *prc30yl1*	*jp20yl1* *jp30yl1*	*uk25yl1*	*us20yl1* *us30yl1*	*sg15yll* *sg20yl1*

Table 2: Summary of bond yields used as features

There are in total 58 features. The response variable would be Singapore 20Y yield of that time period, *sg20y*.

2.3 Exploratory Data Analysis

Exploratory data analysis is done on the training sample. We first looked at the time series characteristics of our response variable, *sg20y*. As shown in Figure 1, the PACF shows significant autocorrelation for lag order of 1, but insignificant for higher lag orders. This is the main reason why we did not include features of higher lag orders.

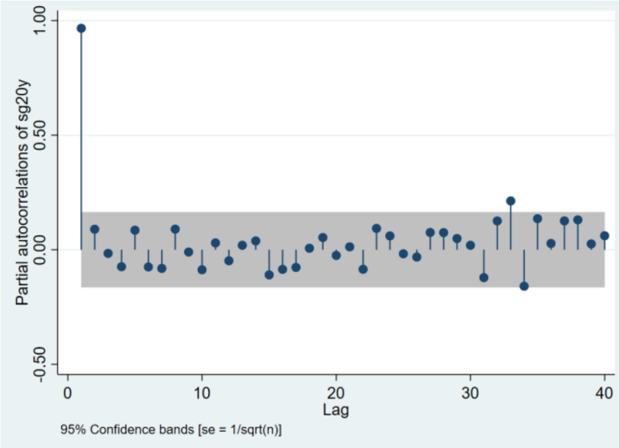
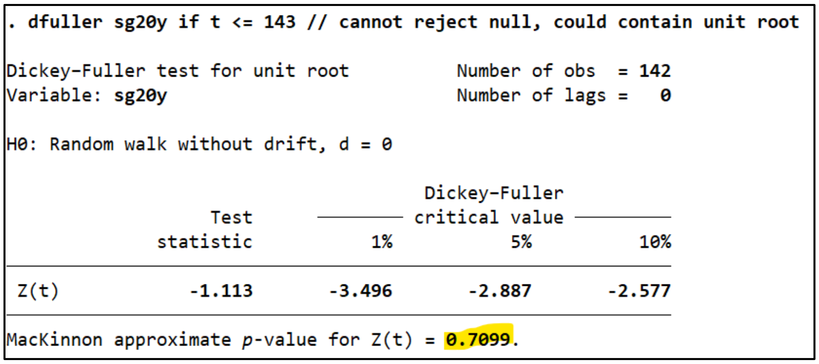


Figure 1: PACF of training data

Using the Dickey-Fuller test for unit root on the response variable also showed that it likely follows a unit root process. First differencing the 20Y yields restores its stationarity.



```
. dfuller sg20y if t <= 143 // cannot reject null, could contain unit root

Dickey-Fuller test for unit root                Number of obs   = 142
Variable: sg20y                                 Number of lags  =   0

H0: Random walk without drift, d = 0

                                       Dickey-Fuller
                   Test         ─────── critical value ───────
              statistic           1%           5%          10%
──────────────────────────────────────────────────────────────
 Z(t)            -1.113       -3.496       -2.887       -2.577
──────────────────────────────────────────────────────────────
MacKinnon approximate p-value for Z(t) = 0.7099.
```

Figure 2: Dickey-Fuller test for unit root on training data

The correlation heatmap (Figure 3) shows some potential relationships between the lagged macro-economic factors and the response variable. As illustrated, there is a need for our forecast methods to choose predictors that are important and weed out unimportant ones.

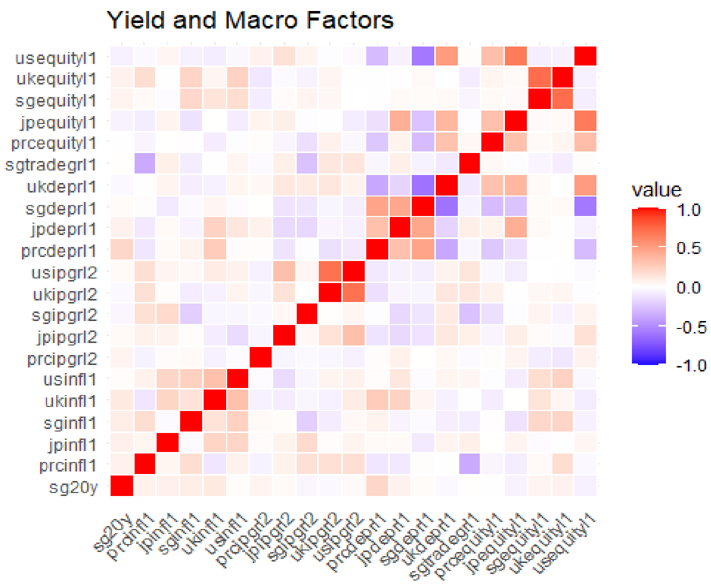


Figure 3: Correlation heatmap of sg20y and macro factors

3. Baseline Models

We begin our forecasts with baseline machine learning methods of three classes: Subset Selection, Penalised Regression, and Artificial Neural Network. We also include simple time series models as benchmark models for comparison.

3.1 General Methodology

Data from January 2009 to December 2020 was used as the training data, and January 2021 to October 2023 as test data. We hoped for both the training and test data to include a few financial crises and recessions so we can test out the models' theoretical robustness to such shocks. 1-step-ahead forecasts are produced using an expanding window.

While the traditional k-fold CV method is usually avoided for time series models, there should be no practical issues in including them for testing overfitting models that produce uncorrelated errors (Bergmeir et al., 2018). Correlated errors generally only occur due to underfitting, which is tackled beforehand for our data using time-series analysis. With the large number of features chosen and small autocorrelation in the data, underfitting is unlikely.

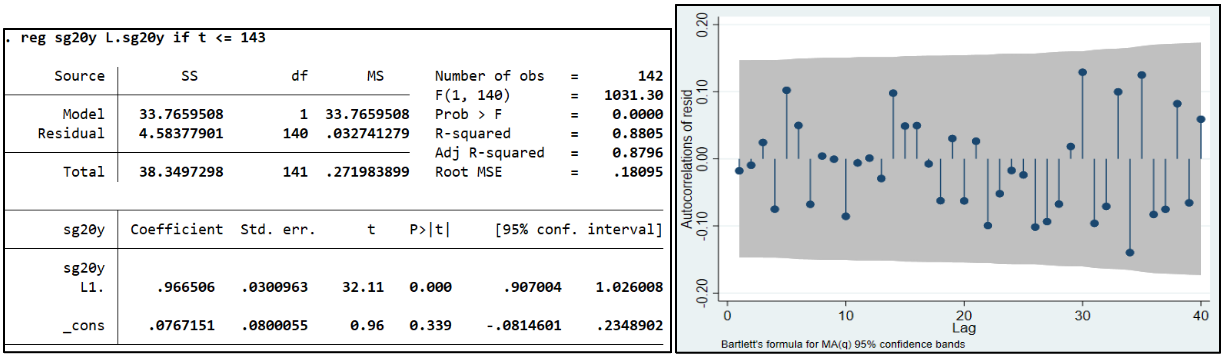
3.2 Benchmark Models

The benchmark models are very simple models which could be used for the justification of the larger and more sophisticated models. We have considered 2 benchmarks models which are the autoregressive model and the random walk model.

Autoregressive Model (AR(1))

We have considered the simple AR model due to the following reasons. Firstly, Stock & Watson (2004) has found that univariate forecasts in most cases do better than most forecasting models using predictors. Secondly, following the Box Jenkins methodology, the PACF of training data mentioned above would indicate that the AR process is likely AR(1).

We did a formal test by fitting the AR(1) model on the training data. From the results in Figure 4, the model fit is good with an adjusted R^2 value of 0.8796. Also, from the ACF plot of the residuals, the residuals are white noise which indicates that most of the components are compared by the AR(1) model. Lastly, from Figure 5, it has a low in-sample MSE which indicates a good model fit. Therefore, this justifies the AR(1) model being a good fit for our training data and a good benchmark model.



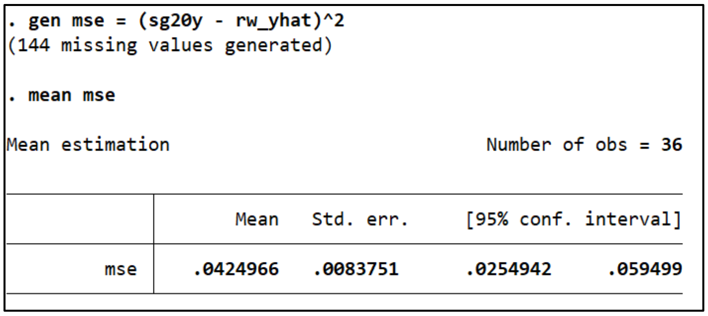
```
. reg sg20y L.sg20y if t <= 143

      Source |       SS           df       MS      Number of obs   =       142
-------------+----------------------------------   F(1, 140)       =   1031.30
       Model |  33.7659508         1  33.7659508   Prob > F        =    0.0000
    Residual |  4.58377901       140  .032741279   R-squared       =    0.8805
-------------+----------------------------------   Adj R-squared   =    0.8796
       Total |  38.3497298       141  .271983899   Root MSE        =    .18095

------------------------------------------------------------------------------
       sg20y | Coefficient  Std. err.      t    P>|t|     [95% conf. interval]
-------------+----------------------------------------------------------------
       sg20y |
         L1. |    .966506   .0300963    32.11   0.000      .907004    1.026008
             |
       _cons |   .0767151   .0800055     0.96   0.339    -.0814601    .2348902
------------------------------------------------------------------------------
```

Autocorrelations of resid
0.20
0.10
0.00
-0.10
-0.20
0
10
20
30
40
Lag
Bartlett's formula for MA(q) 95% confidence bands

Figure 4: AR(1) model fit and ACF of its residual



```
. gen mse = (sg20y - rw_yhat)^2
(144 missing values generated)

. mean mse

Mean estimation                              Number of obs = 36

--------------------------------------------------------------
             |       Mean   Std. err.     [95% conf. interval]
-------------+------------------------------------------------
         mse |   .0424966   .0083751      .0254942     .059499
--------------------------------------------------------------
```

Figure 5: In-sample MSE of AR(1)

Random Walk

The training data could also be modelled by the random walk model due to the following reasons. Firstly, from Figure 5 above, the coefficient of our AR(1) model has a value very close to 1 in magnitude and we cannot reject the null hypothesis of $H_0: \phi = 1$. Secondly, the Dickey-Fuller test mentioned above was not able to reject the null hypothesis of a unit root for level yields but this null hypothesis is rejected when the yields are first-differenced.

Therefore, this justifies the usage of the random walk model as a benchmark model. The random walk model will be similar to our AR(1) model with a slight difference of the coefficient being equal to 1 in magnitude. The equation of the random walk model is as follows:

$$Y_t = Y_{t-1} + \epsilon_t$$

Figure 6: Random walk model

3.3 Subset Selection Methods

Subset Selection Methods are a class of machine learning methods that are technically also penalised regression using L0 norm. However, it is different enough from the other penalised regression methods that we classed them separately for our purposes.

Best Subset Method

The best subset method was not performed due to computational reasons. It is not viable to apply best subset selection with a large number of predictors. In this case, the search space is very large with a number of 2^{58} possible models. Also, for our case, the number of predictors is large with respect to our number of observations (P = 58, N = 143). Therefore, the best subset selection is likely to face statistical problems such as overfitting and large variances of the coefficient estimates. Hence, the stepwise methods appear to be more attractive since a far more restricted set of models is explored. For forward stepwise selection and backward stepwise selection each, there will be 1712 possible models $(1 + \frac{58(1+58)}{2} = 1712)$.

Forward Stepwise Selection

We will be using AIC, BIC and the 5-fold Cross Validation (5-fold CV) methods to determine the best single model. For the information criterion, iterative variance is considered due to the fact that P/N is not small (58/143 = 0.41).

However, most of the models selected did not outperform the benchmark models in terms of test MSE even though these models included significant predictors. Here are some possible reasons. Firstly, the forward stepwise selection is a “greedy” algorithm where the best possible model may not be selected. It could be the case that our current model combination is less than ideal compared to the best possible model. Secondly, due to our large number of predictors with respect to our number of observations, it is more likely that the forward stepwise selection method would pick larger models which would overfit the training data. This is evident in our results where large models (P = 28 to 32) are picked. Overfitting would increase variance and result in poor test MSE results.

On the bright side, selection based on BIC did perform better than the benchmark models with a parsimonious model (P = 3) being selected. This is expected as BIC penalizes large models more as compared to AIC with log(143) > 2. Interestingly, the predictors selected by BIC are nested in all of the larger models considered by AIC. This could signal that the 3 predictors selected are strong predictors that we should consider. The predictors are *sg20yl1, jpdeprl1, ukequityl1.*

Backward Stepwise Selection

Similarly, AIC, BIC and the 5-fold Cross Validation (5-fold CV) methods will be used to determine the best single model. For the information criterion, iterative variance is considered since P/N is not small (58/143 = 0.41).

However, all the models selected did not outperform the benchmark models in terms of test MSE even though these models included significant predictors. Here are some possible reasons. Firstly, the backward stepwise selection is also a “greedy” algorithm where the best possible model may not be selected. It could be the case that our current model combination is less than ideal compared to the best possible model. Secondly, due to our large number of predictors with respect to our number of observations, it is more likely that the backward stepwise selection method would pick larger models which would overfit the training data. This is evident in our results where large models (P = 9 to 20) are picked. Overfitting would increase variance and result in poor test MSE results.

Interestingly, even though the models picked by the backward stepwise selection (P = 9 to 20) are more parsimonious as compared to the models picked by the forward stepwise selection (P = 28 to 32), the simpler models actually performed a lot worse in terms of test MSE. Since simpler models have a lower variance, this could mean that the simpler models picked are more biased, as stated in the bias-variance equation. Upon further inspection, the simpler models did not consider the predictor *sg20yl1* which is the first lag of *sg20y*. We believe that this could be the main reason for the increased bias due to the fact that the AR(1) and random walk models are good fits. The good fit implies that the predictor *sg20yl1* is likely a strong predictor for *sg20y*.

Therefore, even though the backward selection did not produce any good models, its performance provides more evidence that the predictor *sg20yl1* is likely a strong predictor.

3.4 Penalised Regression Methods

Ridge

With Ridge, the lambda is selected using 5-fold CV MSE. Two lambdas are chosen, one which minimises the CV MSE (min. MSE) and another which is "1 standard error" away from the minimum lambda (1se).

The test MSEs generated from the minimum MSE and 1 standard error (1SE) lambda (0.04586375 and 0.04904615 respectively) do not consistent outperform those of the benchmark models. This method is evidently inferior to LASSO, as elaborated in the next portion. Unlike LASSO, with multiple correlated predictors, Ridge likely assigns similar coefficients to all of them, rather than selecting one and shrinking the coefficients of the others to zero. Ridge also tends to shrink the coefficients towards zero rather than exactly zero, retaining all features including the irrelevant, leading to a less sparse model than that in LASSO. This makes interpretation harder and reduces its generalization performance, as there may be many unnecessary features in our data.

LASSO

For LASSO, the test MSE generated using the minimum MSE lambda to regularise is also higher than the two benchmark models. However, the usage of 1SE lambda produces test MSE that beats the two benchmark models. Furthermore, LASSO's ability to shrink coefficients to 0 allows us to identify the more useful predictors while reducing variance in our model's coefficients, thus making LASSO effective. Based on our result, *sg20yl1*, *jpdeprl1* and *ukequityl1* are the more important features as they have the 3 highest coefficients. This matches what we have obtained so far with forward and backward stepwise selection. This matches our expectation as well since we have prior knowledge that *sg20yl1* is a very good predictor from AR(1) benchmark model.

Elastic Net

Elastic net regression strikes a balance between the Ridge and LASSO models. Based on iterative selection of parameters using 5-fold CV on training data, an optimal alpha = 0.78 was determined based on lowest MSE (slanting towards LASSO L1-norm regularization). As expected, the mixing ability of Elastic Net outperforms both the Ridge and LASSO models in out-of-sample MSE. This is because of the selection ability mimicking LASSO but with the added benefit of allowing more precise coefficient estimates similar to Ridge. Based on the simplified model using 1-standard-error lambda, s*g20yl1*, *jpdeprl1* and *ukequityl1* have the largest coefficient estimates, similar to LASSO and subset methods above, and in line with the AR(1) findings earlier. Using the optimal lambda, however, other predictors become more influential, such as *usinfl1*, *jpinfl1*, and *ukinfl1*.

3.5 Artificial Neural Network

Using ANN produces a test MSE that beats the two benchmark models. However, ANN models are a black box, and we are not able to interpret what is happening inside ANN. ANN models seem to be performing worse as compared to LASSO. ANN works better if there is non-linearity between the features as ANN can capture them while other regression models cannot capture. The lower test MSE in LASSO suggests that linear relationship better models the true relationship as compared to non-linearity.

3.6 Performance of Baseline Models

Method	Parameters	Test MSE
AR(1)	-	0.0429
Random Walk	-	0.0471
Forward Stepwise (BIC)	P = 3	**0.0297**
Forward Stepwise (AIC)	P = 28	0.0658

Forward Stepwise (AIC iv)	P = 32	0.0760
Forward Stepwise (5-Fold CV)	P = 26	0.0557
Backward Stepwise (BIC)	P = 9	0.1199
Backward Stepwise (AIC)	P = 19	0.0816
Backward Stepwise (AIC iv)	P = 20	0.0830
Backward Stepwise (5-Fold CV)	P = 11	0.0721
Ridge (min MSE)	λ = 0.0568	0.0424
Ridge (1se)	λ = 0.3828	0.0475
LASSO (min MSE)	λ = 0.0055	**0.0313**
LASSO (1se)	λ = 0.0440	0.0346
Elastic Net (min MSE)	α = 0.78 λ = 0.00672	0.0342
Elastic Net (1SE)	α = 0.78 λ = 0.04319	0.0382

ANN	Size = 10
	Decay = 0.1
		0.0389

Table 3: Summary of baseline models results. ***Bolded*** *test MSEs represent the best performing model of its class, while underlined values represent models that outperform either AR(1) or RW benchmark models.*

4. Ensemble Methods

Other than the baseline models, we also looked at 3 ensemble methods for the forecasting of Singapore's 20Y yields. It is expected that the tree methods we are using (Gradient Boosting and Random Forest) would incorporate interactions between the different features, which are not captured by the baseline models (other than ANN). Lastly, we will also use the Bates-Granger method to combine all machine learning forecasts into one single combined forecast.

4.1 Gradient Boosting

Gradient boosting is an ensemble learning method that effectively combines many small models ("weak learners") into one single model used for forecasting. These learners are run sequentially based on a gradient descent algorithm over the RMSE in our case.

xgBoost (eXtreme Gradient Boosting) is a commonly used package for gradient boosting. Unlike traditional gradient boosting, the trees are built in parallel to ensure fast computation. The package includes features like subsampling regularisation (*subsample*), learning rate (*eta*), maximum depth of each tree (*max_depth*), and the number of trees to build (*nrounds*). To observe the behaviour of the algorithm, we ran XGBoost on the training data, repeatedly tweaking some parameters. Based on CV MSE, there was not much improvement past 150 trees and no need for pruning of trees. We then used the *caret* package to tune our hyperparameters of XGBoost over a range of suitable values for all parameters. The hyperparameters that produced the lowest CV MSE were chosen and used for the expanding window forecast.

The training set produced trees that placed great importance on the yield rates of other economies, but not much on macroeconomic factors, similar to the better performing subset selection models (see Figure 7). Based on the test MSE, XGBoost outperformed some baseline models, but did not outperform many of the penalised regression methods. This may show that there might not be much interaction between variables and the features are mostly linearly related. This may further prove that the subset selection methods' failures were likely due to the use of wrong features and not their inherent inability to capture nonlinear relationships.

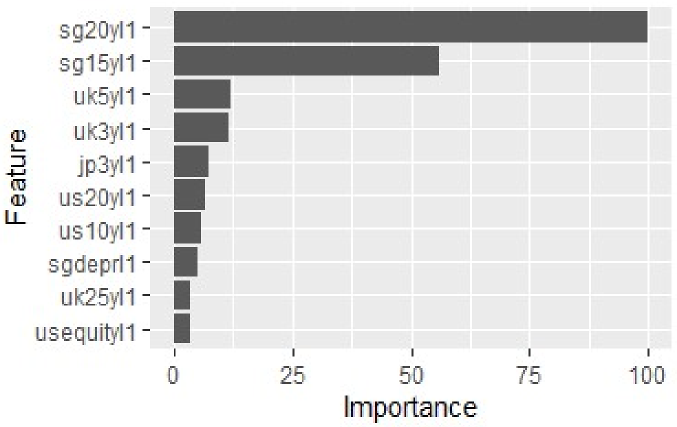


Figure 7: XGBoost feature importance

4.2 Random Forest

Unlike Gradient Boosting, Random Forest does not create trees sequentially and instead fit many small trees that are ideally uncorrelated with one another simultaneously, combining them into one single model later. Using the *randomForest* package, there are two parameters to tune - the number of trees (*ntree*) and the number of predictors used for each tree (*mtry*). Similarly, the *caret* package was used to tune these two hyperparameters based on the CV MSE of the training set.

Random Forest produced forecasts that outperformed Gradient Boosting, possibly because it is more difficult to overfit using Random Forest which forces more features to be used. In fact, as shown in Figure 8, Random Forest placed more importance on Singapore's own yields of different maturities, which may become even more important out-of-sample (after 2021).

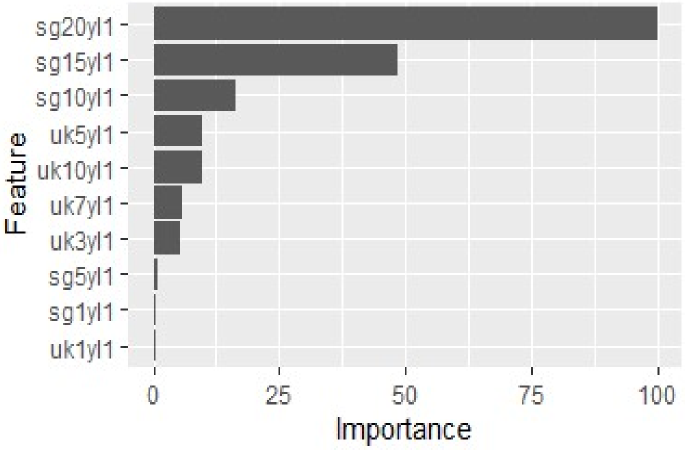


Figure 8: Random Forest feature importance

4.3 Bates-Granger Combination

Bates and Granger (1969) introduced the Bates-Granger method for combining forecasts based on out-of-sample forecast variances. While this method minimizes variance only for uncorrelated forecasts, it may still reduce variance for correlated forecasts. In our case, based on the assumption we made above about the validity of cross-validation in our time-series data, we used each model's 5-fold CV MSE as an estimate of its out-of-sample MSE. Thus, the Bates-Granger weights given to each method will be follow this formula:

$$w_i^{BG} = \frac{\sigma_{CV,i}^{-2}}{\Sigma\sigma_{CV,j}^{-2}}$$

$$\hat{y}_t^{BG} = (\,\hat{\mathbf{y}}_t)^T \mathbf{w}^{\mathrm{BG}}$$

Equation 4: Bates-Granger combination

Method	CV MSE
Forward Stepwise Selection (BIC)	0.02805243
Forward Stepwise Selection (AIC)	0.02759278
Forward Stepwise Selection (AIC iterative variance)	0.02683617
Forward Stepwise Selection (5-Fold CV)	0.0277076
Backward Stepwise Selection (BIC)	0.02795801
Backward Stepwise Selection (AIC)	0.02387185
Backward Stepwise Selection (AIC iterative variance)	0.02411753
Backward Stepwise Selection (5-Fold CV)	0.02824812
Ridge (min MSE)	0.03194331
Ridge (1se)	0.03624995
LASSO (min MSE)	0.03282932

LASSO (1se)	0.03484349
Elastic Net (min MSE)	0.03281803
Elastic Net (1se)	0.03470996
ANN	0.02980941
XGBoost	0.03104606
Random Forest	0.03266884

Table 4: CV MSE of ML forecasts

The *caret* package is again used to obtain CV MSE, mostly because it has been used for tuning earlier models and we wanted congruency in the way CV MSE is calculated between all models. Unfortunately, the CV MSEs calculated for many models were significantly different from their real test MSEs, leading to questions about our earlier assumption of using cross-validation for time series models. Nonetheless, the Bates-Granger combination forecasts outperformed many, but not all of the baseline models. In particular, Random Forest, Forward Stepwise Selection (BIC), LASSO, Elastic Net, and ANN outperformed the combination. Again, this is likely due to CV MSEs grossly underestimating the test MSEs of some models, especially the Backward Stepwise Selection models.

4.4 Performance of Ensemble Methods

Methods	Parameters	Test MSE
Stochastic Gradient Boosting	*nrounds* = 78 *max_depth* = 3 *eta* = 0.1 *gamma* = 0 *subsample* = 0.4	0.04173432
Random Forest	*ntree* = 900 *mtry* = 30	**0.03990928**

Bates-Granger Combination	-	0.04455819

Table 5: Summary of ensemble method results. ***Bolded*** *test MSEs represent the best performing model of its class, while underlined values represent models that outperform either AR(1) or RW benchmark models*

5. Conclusion

We found that Penalised Regression methods tend to work well in forecasting Singapore 20Y bond yields, along with more complex machine learning methods like ANN, Random Forest, and Gradient Boosting. Penalised regression is the only class of methods that consistently outperformed the AR(1) and Random Walk benchmarks. Subset selection seemed to perform quite bad, possibly due to the tendency to overfit past data. Nevertheless, their performances can be used to gauge the importance of different features.

However, future studies can improve on this research in a few ways. Firstly, a larger dataset may be used, instead of monthly data which is inherently small in size. While many macroeconomic indicators only exist in monthly frequency, there may be other proxies available with higher frequency. More frequent data would also make forecasts into further time horizons possible. Secondly, our results also cast doubt into our suspicions about the use of K-fold CV for our dataset. Future studies can opt for alternative validation methods like expanding window validation.

References

Bates, J.M., and Granger, C.W.J. (1969). The combination of forecasts. Operational Research Quarterly, vol. 20, no. 4: 451-468

Bergmeir, C., Hyndman, R.J., Koo, B. (2018). A note on the validity of cross-validation for evaluating autoregressive time series prediction. Computational Statistics & Data Analysis, vol. 20: 70-83

Chernov, M., Creal, D., and Hordahl, P. (2019). Determinants of Asia-Pacific government bond yields. Bank of International Settlements Papers no. 102

De Pooter, M., Ravazzolo, F., and van Dijk, D. (2010). Term structure forecasting using macro factors and forecast combination. Federal Reserve: International Finance Discussion Papers No. 993

Ferguson, N. (2023, October 9). Law of Unintended Consequences Caused the Great Bond Rout. Retrieved from Bloomberg: https://www.bloomberg.com/opinion/articles/2023-10-09/worst-bond-collapse-in-150-years-caused-by-unintended-consequences

Stock, J.H., and M.W. Watson (2004). Combination forecasts of output growth in a seven-country data set. Journal of Forecasting, vol. 23: 405-430

Y. Suimon, H. Sakaji, T. Shimada, K. Izumi and H. Matsushima, Extraction of Relationship between Japanese and US Interest Rates using Machine Learning Methods. 2019 8th International Congress on Advanced Applied Informatics