機器學習與計量財務
Final Presentation

MIS 109306091 蘇希甫

Table of Contents

01

Intro and experiment idea

Experiment Idea

- 目標：我們的主要目標是使用Lagged return作為預測變量來預測加元/日元匯率的回報。最終目標是在限制的條件內找出最有效和最高效的模型來預測這些回報。

- 背景：財務預測是金融界的一項重要活動。它對於學習長期投資是很關鍵的。從相較好理解的匯率下手，更能讓我透徹理解財務金融這個領域的知識。對於這個項目，我們關注 CAD/JPY 匯率回報，因為這些會影響從個人投資者到跨國公司的各種利益相關者。

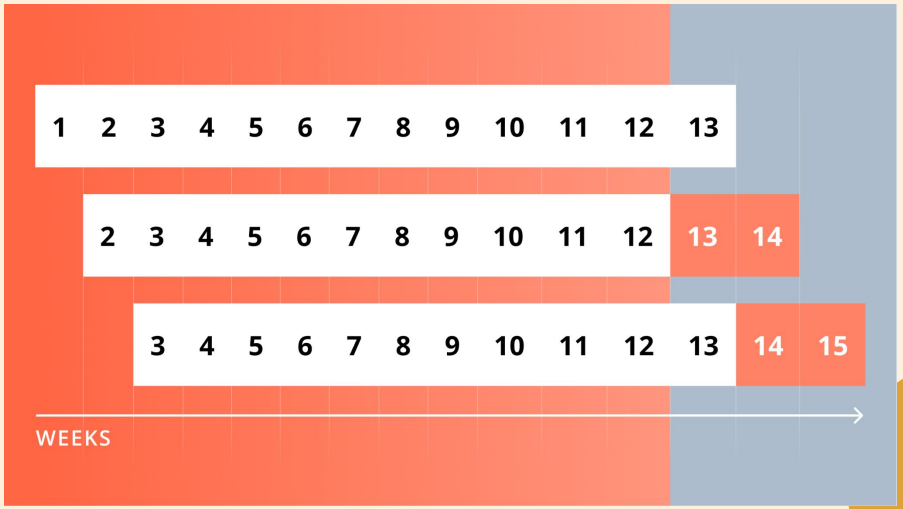
- 數據：數據包括 2013 年至 2022 年期間的每週加元/日元匯率。

Times Series Forecasting

時間序列預測的重要性

加元/日元匯率回報表現出時間依賴性，這意味著過去的回報會影響未來的回報。時間序列預測模型可以捕獲這些時間依賴性，使其非常適合這項任務。

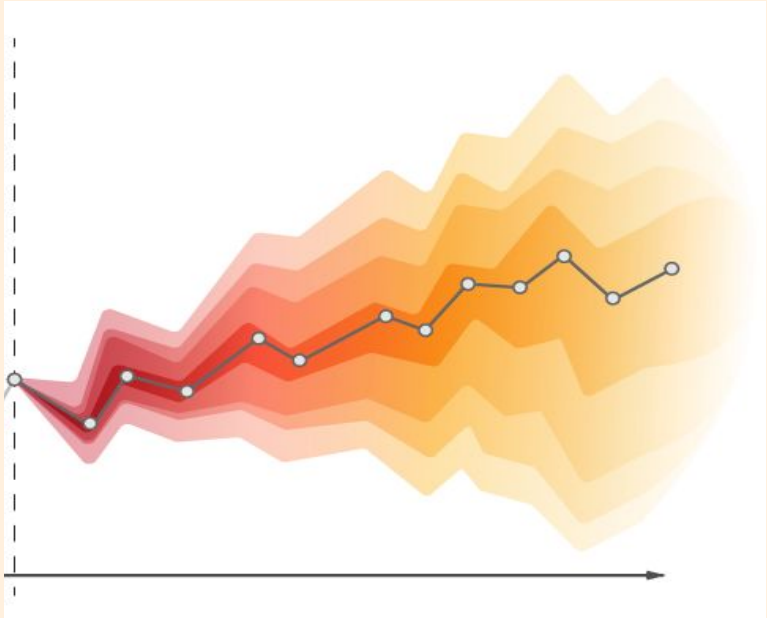
時間序列模型還可以處理時間序列數據的獨特特徵，例如趨勢、季節性和自相關，這有助於提高我們預測的準確性。



Challenges in time series prediction

時間序列預測中的挑戰：

- Non-Stationarity：金融時間序列數據通常表現出非平穩性，這違反了許多統計模型的假設，並可能導致不可靠的預測。
- Autocorrelation: Autocorrelation 代表 correlation between a time series and its own lagged version.自相關錯誤可能導致回歸分析中的誤導性測試統計。這在我們的模型中通過使用滯後回報作為預測變量來解決。
- Overfitting：在時間序列預測中，過度擬合發生在模型與歷史數據過於接近時，捕獲了噪聲和信號。這可能導致對新數據的預測性能不佳。我們稍後將討論的正則化技術，例如 Ridge 和 Lasso 回歸，用於防止過度擬合。



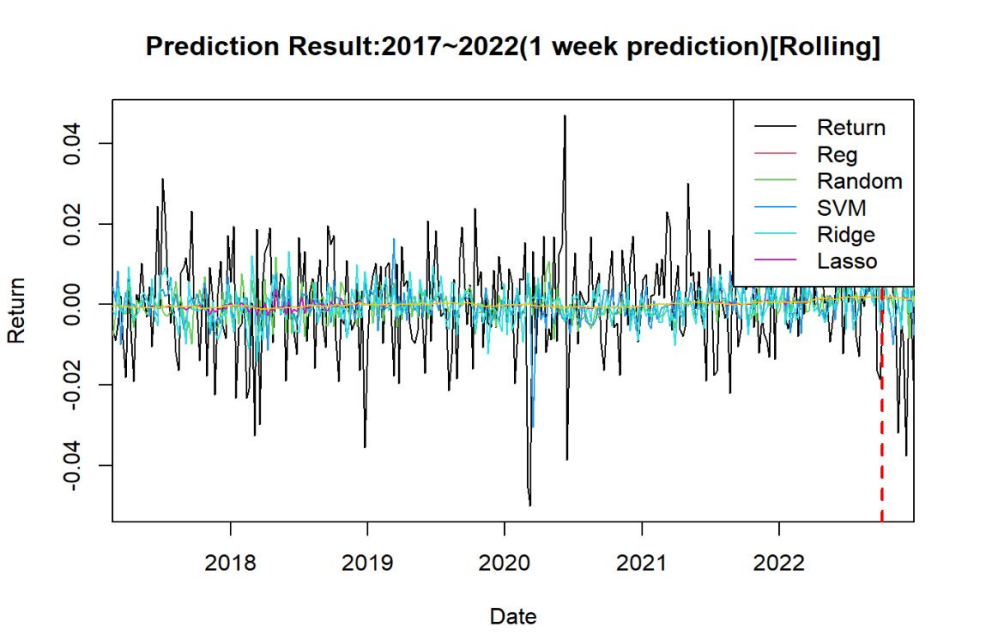
Intro

數據導入和清理:數據是從包含 Yahoo Finance 上下載的, 我們清理了數據並構建了新特徵像轉成Weekly data, 還有像one week prediction and 12 weeks prediction。x是使用過去12周的報酬當作解釋變數。y為當期的return, 為被解釋變數。

樣本外 1 週的rolling prediction:我們使用 rolling window的預測方法。我們在 150 週的 window內訓練我們的模型, 然後預測下週的回報。重複此過程, 每次將 window向前移動一周。

Model的部分, 包括:

1. **Linear Regression**
2. **Random Forest**
3. **Support Vector Machines**
4. **Ridge Regression**
5. **Lasso Regression**

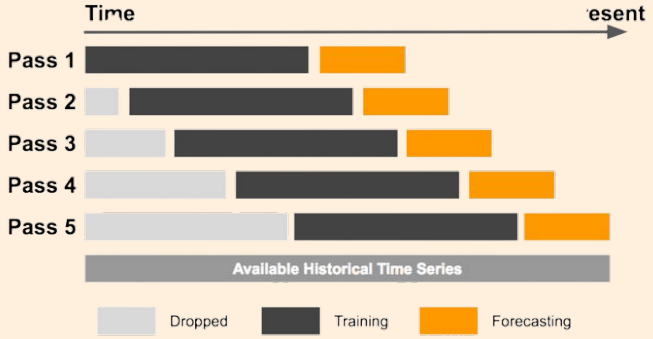


For ridge and lasso regressions, we use cross-validation to select the optimal regularization parameter (lambda).

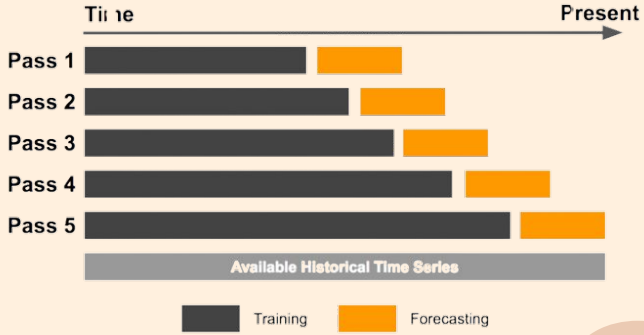
Intro

我們總共做了四個不同的預測方法：

1. 1 week rolling prediction
2. 12 weeks rolling prediction
3. 1 week expanding rolling prediction
4. 12 week expanding rolling prediction



*expanding:start with a certain window size and continually add more data to the training set as you move

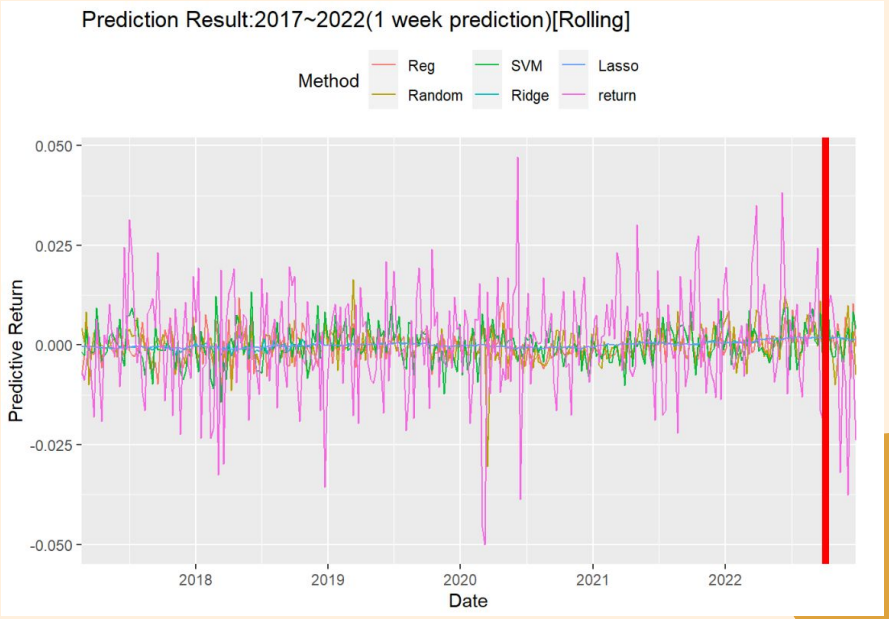


Linear Regression

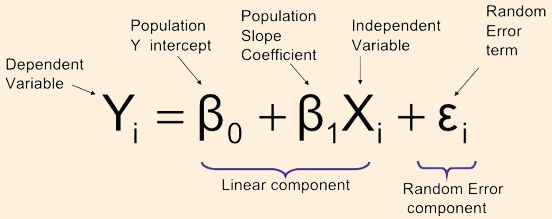
線性回歸是一種簡單的算法，但它有一些局限性：

- 它假設特徵和目標之間存在線性關係，現實世界可能並不成立。
- 對outliers特別敏感。
- 如果數據集包含高度相關的預測變量，它可能會出現多重共線性。

儘管簡單，使用線性回歸作為模型之一來預測 CAD/JPY 的回報可以作為建立基線預測的良好起點。



Reg
MAE: 0.01076
MAPE: 144.30437
SMAPE: -2.46249
MSE: 0.00020
RMSE: 0.01427



- 平均絕對誤差 (MAE)：MAE 測量一組預測中誤差的平均幅度，而不考慮它們的方向。它是預測 值和實際值之間絕對差值的平均值。如果目標變量的範圍是 0-1，則 0.01 的 MAE 將被認為是好的。在您的情況下，MAE 為 0.01076。

- 平均絕對百分比誤差 (MAPE)：MAPE 是實際值和預測值之間的平均絕對百分比差值。低於 10% 的 MAPE 通常被認為是表現好的。我的 MAPE 相當高，為 144.30437，這表明平均而言，模型的預測與實際值的偏差約為 144.30%。

- 對稱平均絕對百分比誤差 (SMAPE)：當實際值中存在零時，SMAPE 是 MAPE 的替代方法，因為它考慮了兩個方向上的絕對差異。小於 10% 的 SMAPE 通常被認為是表現良好的。我的 SMAPE 值為 1.55586，這表明您的模型預測與實際 值的偏差約為 1.56%。

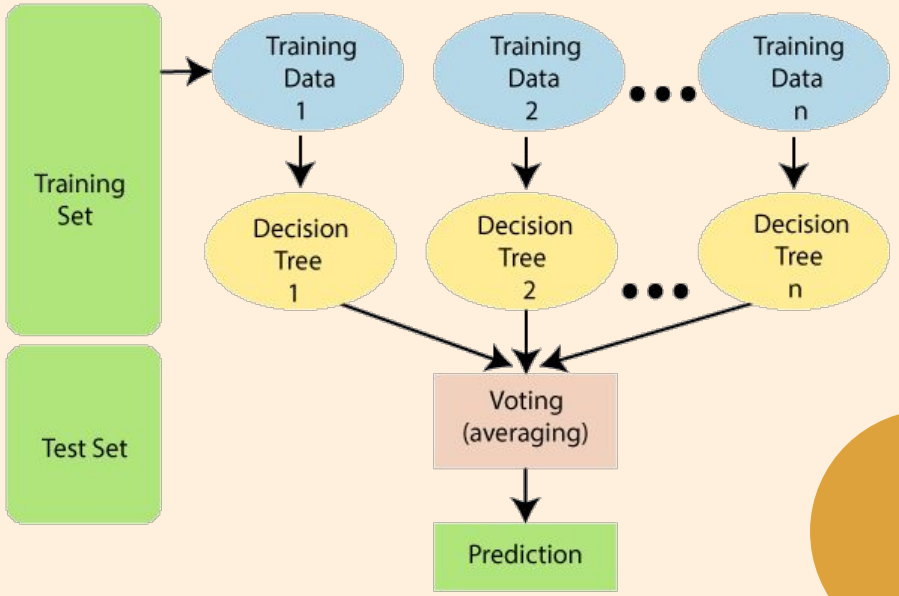
- 均方誤差 (MSE)：MSE 是預測值和實際值之間的平方差的平均 值。它是回歸問題的常用損失函數。較低的 MSE 表明模型性能較好。我的 MSE 是 0.00020，說明你的模型對數據的擬合比較好。

- 均方根誤差 (RMSE)：RMSE 是 MSE 的平方根。因為 MSE 是平方的，所以它可以過度懲罰大的錯誤。與其他指標一樣，RMSE 越低表示模型性能越好。 我的 RMSE 為 0.01427，這表明平均而言，模型的預測與實際 值的偏差約為 0.01427 個單位。

Random Forests

Random forest method：創建一個bootsrap dataset。它通過隨機抽取數據樣本，創建一個大小相同但組成不同的新數據集，創建 decision tree。不同的是，它考慮了樹中每個分裂的所有特徵，基於 dataset創建的隨機性和特徵進行分割，這些樹中的每一個都會有所不同。

- data = na.omit(train_data[,-(1:10)]) 指定刪除具有缺失值的行並排除前 10 列後的訓練數據集。
- mtry = ncol(na.omit(train_data[,-(1:10)]))-1 指定每次拆分時要考慮的隨機選擇的預測變量的數量。它被設置為訓練數據集中的列數減一。
- subset = sample(1:nrow(na.omit(train_data[,-(1:10)]))) 指定用於構建森林中每棵樹的隨機行子集。
  ntree = 300 指定要在隨機森林中生長的樹的數量。



- 平均絕對誤差 (MAE)：預測值和實際值之間的平均絕對差為 0.01080。這意味著，平均而言，模型的預測與實際值相差 0.01080 個單位。根據目標變量的規模，這可能是一個不錯的結果。

- 平均絕對百分比誤差 (MAPE)：MAPE 為 161.54624%。這個值看起來相當高，表明該模型可能表現不佳。

- 對稱平均絕對百分比誤差 (SMAPE)：SMAPE 為 1.55965。此值可能表明準確性適中。

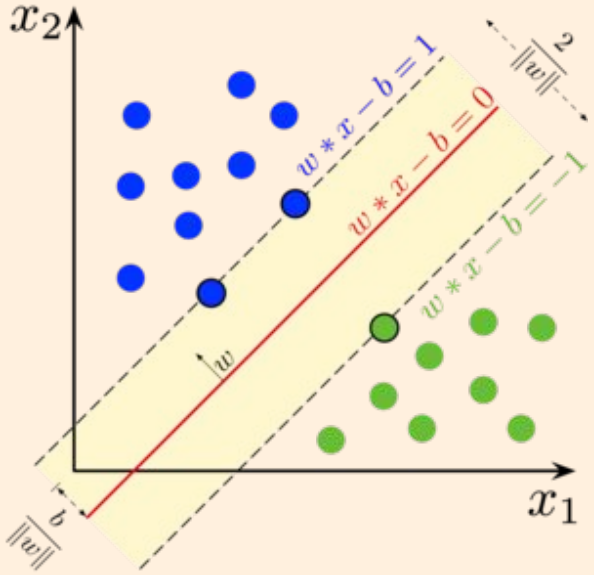
- 均方誤差 (MSE)：此模型的 MSE 為 0.00020，與回歸模型相同。同樣，這可能是一個很好的結果，具體取決於目標變量的規模。

- 均方根誤差 (RMSE)：RMSE 為 0.01417，略低於回歸模型。這表明，就預測 值與實際值的偏差程度而言，隨機森林模型的表現略好於回歸模型。

Support Vector Machine

SVM是一種supervised learning，用統計風險最小化的原則來估計一個分類的 hyperplane，就是找到一個決策邊界(decision boundary)讓兩類之間的邊界(margins)最大化，使其可以完美區隔開來。

x：預測變量（train_data[,-(1:11)]）的矩陣或數據框，不包括前 11 列，假定為用於預測的特徵。
y：響應變量 (train_data[,11])，它是 train_data 數據幀中的第 11 列。
然後使用提供的訓練數據訓練 SVM 模型。



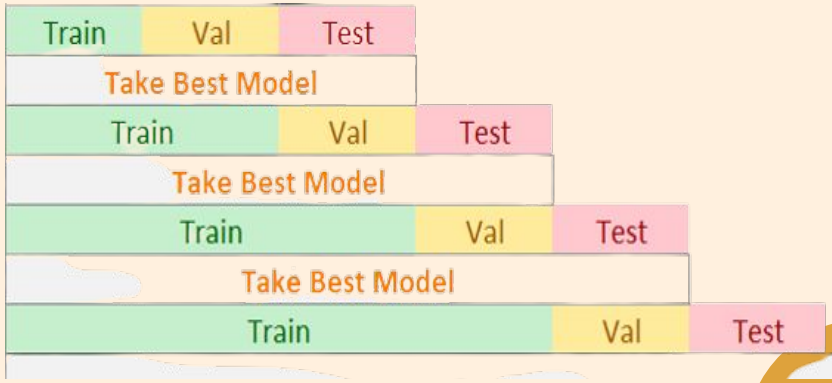
- 平均絕對誤差（MAE）：MAE為0.01081，與回歸模型和隨機森林模型相似。這表明 SVM 模型的預測平均偏離實際值大約 0.01081 個單位。
- 平均絕對百分比誤差 (MAPE)：MAPE 為 161.10927%。這略低於隨機森林模型的 MAPE，但仍然很高，這可能表明模型性能存在潛在問題。
- 對稱平均絕對百分比誤差 (SMAPE)：SMAPE 為 1.56490。這與其他模型的值非常接近，表明性能水平相似。
- 均方誤差 (MSE)：SVM 模型的 MSE 為 0.00020，與其他兩個模型的 MSE 值相同。這表明 SVM 模型的誤差平均與回歸和隨機森林模型的誤差相似。
- 均方根誤差 (RMSE)：RMSE 為 0.01417，與隨機森林模型的 RMSE 相等，略小於回歸模型。這表明 SVM 模型的預測與實際值的偏差與其他模型的偏差大致相同。

Cross-Validation and Ridge, Lasso

Ridge 和 Lasso 回歸中的交叉驗證：

Cross validation用於確定 lambda (λ) 的最佳值，就是Lasso 跟 Ridge 的Regularization parameters。

Cross Validation的原理是將Training set分成多個子集。然後，在除其中一個子集之外的所有子集上對模型進行訓練，並在其餘子集上進行測試。對每個子集重複此過程。



Ridge Regression：
Ridge在 linear regression cost function中添加了一個懲罰項，使係數向零收縮，從而去排除掉比較沒用的變數。cv.glmnet 函數用於通過交叉驗證Ridge以確定lambda 的最佳值。選擇 (mod1$lambda.min) 的 lambda 值，然後使用 glmnet 函數用所選的 lambda 值擬合 (mod_ridge)。

Lasso Regression：
Lasso 回歸與嶺回歸類似，懲罰向的估計方法不同，且使用的是 L1 Regularization 而不是 L2 Regularization。

Cross-Validation and Ridge, Lasso

```
> coef_ridge
13 x 1 sparse Matrix of class "dgCMatrix"
                      s1
(Intercept)   1.419279e-03
return_lag1   7.680455e-04
return_lag2  -4.401022e-04
return_lag3  -5.082445e-04
return_lag4   7.629611e-04
return_lag5  -5.611009e-04
return_lag6   1.616211e-04
return_lag7  -6.738653e-04
return_lag8   1.026775e-03
return_lag9  -3.419553e-04
return_lag10  6.843374e-05
return_lag11  2.278946e-04
return_lag12  3.194344e-04
```

```
> coef_lasso
13 x 1 sparse Matrix of class "dgCMatrix"
                      s1
(Intercept)  0.001420834
return_lag1  0.000000000
return_lag2  .
return_lag3  .
return_lag4  .
return_lag5  .
return_lag6  .
return_lag7  .
return_lag8  .
return_lag9  .
return_lag10 .
return_lag11 .
return_lag12 .
```

Lasso: Lag2-Lag12都是趨近於0，所有解釋變數其實表現都不太好
Ridge:也都是趨近於0，數字非常小

- 平均絕對誤差 (MAE)：Ridge 模型的 MAE 為 0.01018，Lasso 模型的 MAE 為 0.01016，這比您之前測試過的模型的 MAE 略好（即更低）。這表明，平均而言，Ridge 和 Lasso 模型的預測更接近實際值。

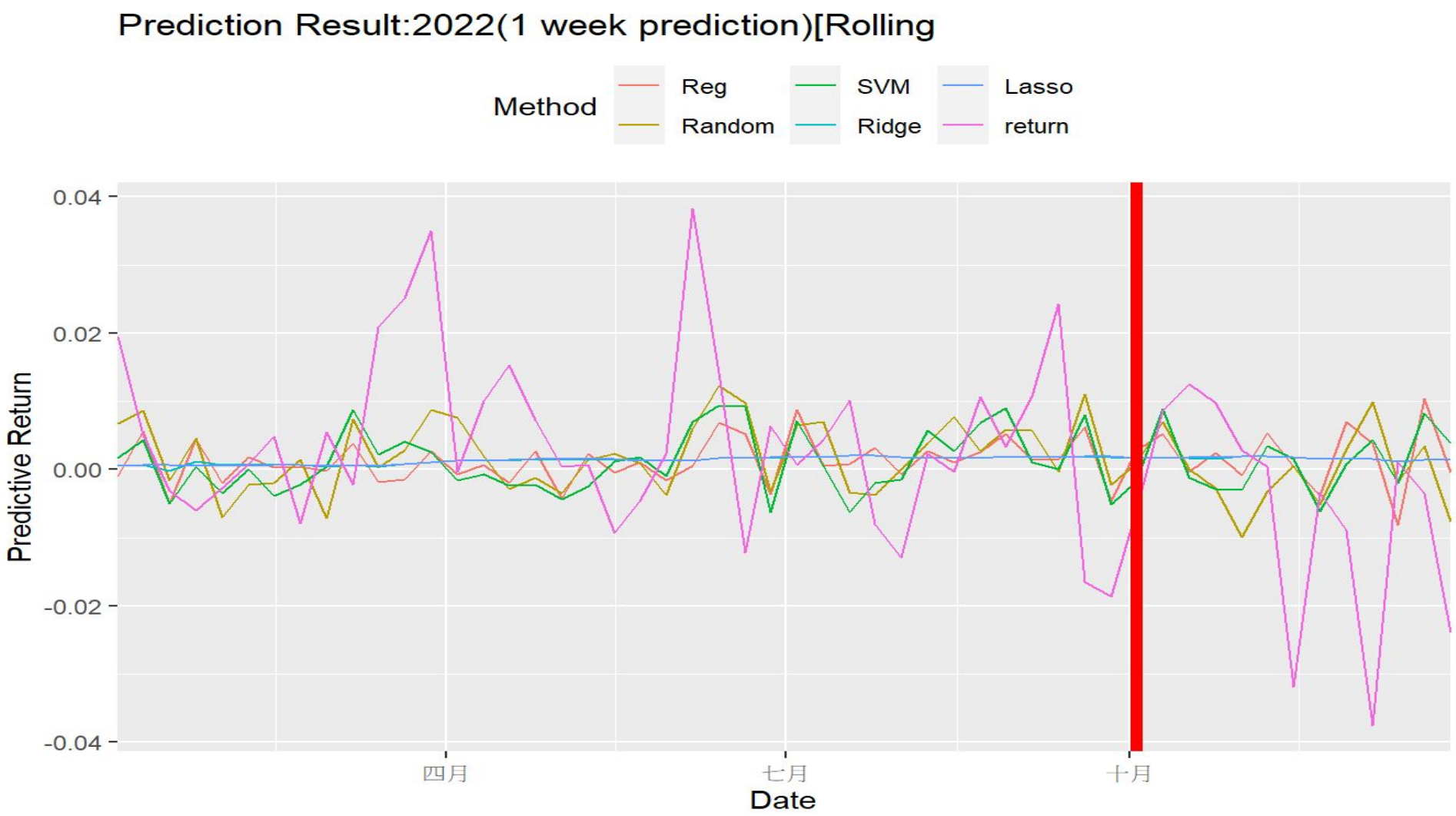
- 平均絕對百分比誤差 (MAPE)：Ridge的 MAPE 為 106.21176%，Lasso 模型為 107.03701%。這些明顯優於（即低於）之前模型的 MAPE 值，表明 Ridge 和 Lasso 模型具有較低的相對錯誤率。

- 對稱平均絕對百分比誤差 (SMAPE)：Ridge 模型的 SMAPE 為 1.77866，Lasso 模型為 1.80458。這些值高於之前模型的 SMAPE 值，表明平均相對絕對誤差略高。

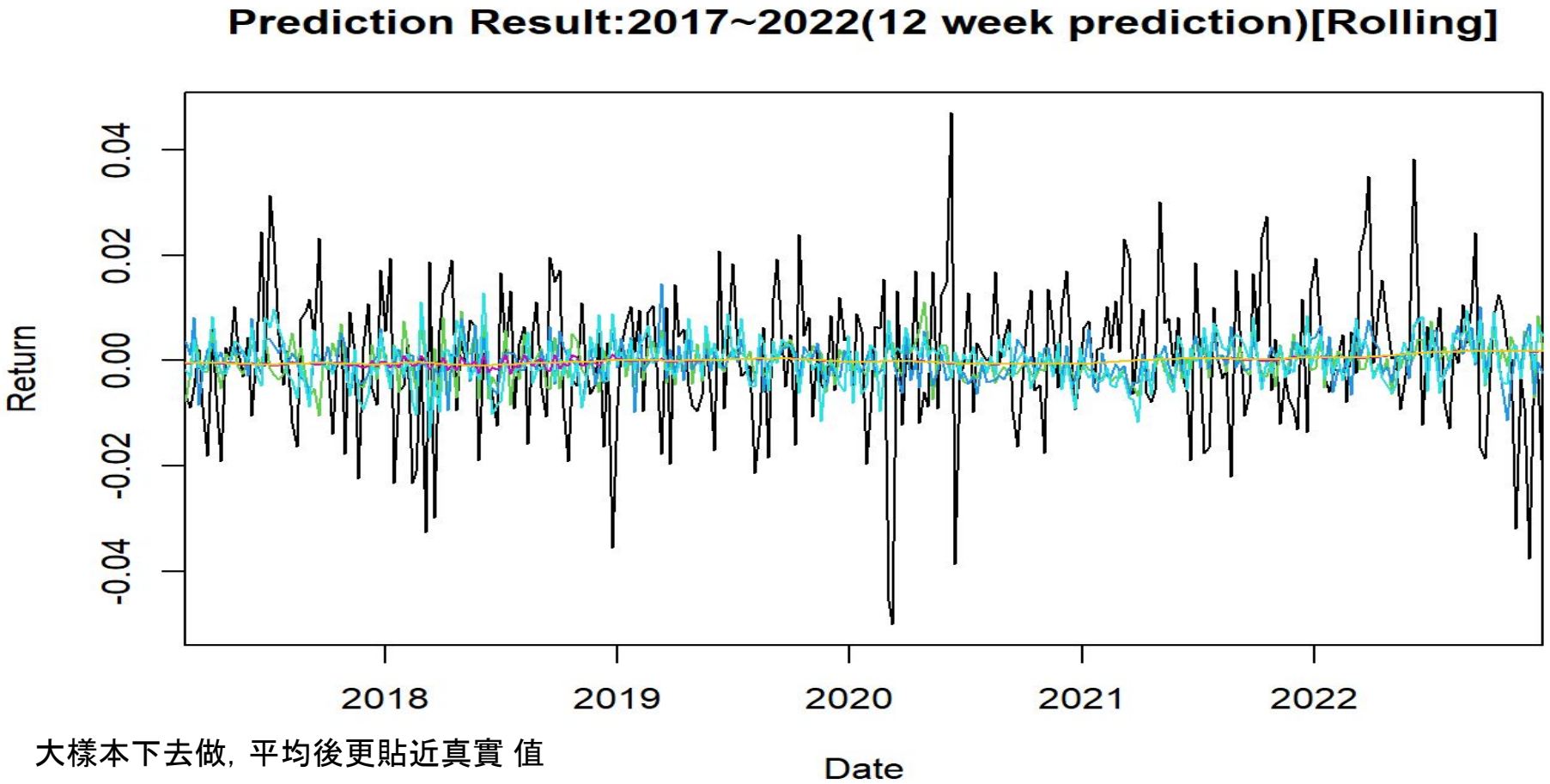
- 均方誤差 (MSE)：Ridge 和 Lasso 模型的 MSE 均為 0.00018，略低於其他模型的 MSE。這表明 Ridge 和 Lasso 模型的預測值和實際值之間的平方差平均略小。

- 均方根誤差 (RMSE)：Ridge 模型的 RMSE 為 0.01339，Lasso 模型的 RMSE 為 0.01334，均低於之前模型的 RMSE。這表明 Ridge 和 Lasso 模型的預測與實際值的平均偏差略小。

- 綜上所述，Lasso 在 MAE 和 RMSE 方面表現稍好，而 Ridge 在 MAPE 和 SMAPE 方面表現稍好。但是，差異非常小，這表明這兩個模型在您的數據集上的表現相似。


Prediction Result:2022(1 week prediction)[Rolling
Method
Reg
Random
SVM
Ridge
Lasso
return
0.04
0.02
0.00
-0.02
-0.04
Predictive Return
四月
七月
十月
Date

12 weeks rolling prediction



大樣本下去做，平均後更貼近真實 值

12 weeks rolling prediction

平均絕對誤差 (MAE)：對於所有模型，12 週預測的 MAE 略低於 1 週預測。然而，差異非常小。這表明平均而言，與預測 1 週相比，模型預測 12 週後的絕對誤差略小。

平均絕對百分比誤差 (MAPE)：同樣，所有模型的 12 週預測的 MAPE 都較低。這表明模型在預測未來 12 週時的百分比誤差較小。

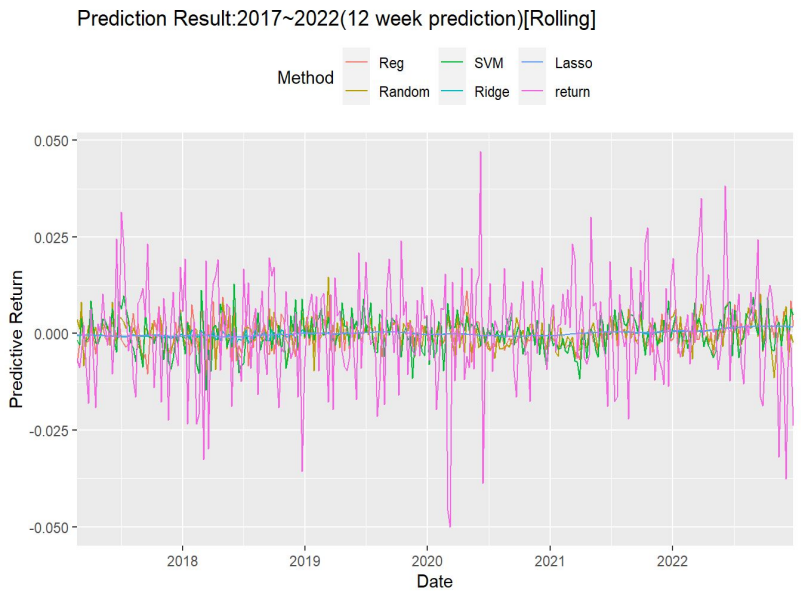
對稱平均絕對百分比誤差 (SMAPE)：回歸和 SVM 模型的 12 週預測的 SMAPE 略低，而 Ridge 和 Lasso 模型的 SMAPE 略高。這表明在 12 週預測中，以百分比為基礎，回歸和 SVM 的預測誤差對稱性有所改善，但 Ridge 和 Lasso 的預測誤差對稱性有所惡化。

均方誤差 (MSE)：所有模型的 1 周和 12 週預測的 MSE 相同，表明兩種情況下的平均平方誤差相同。

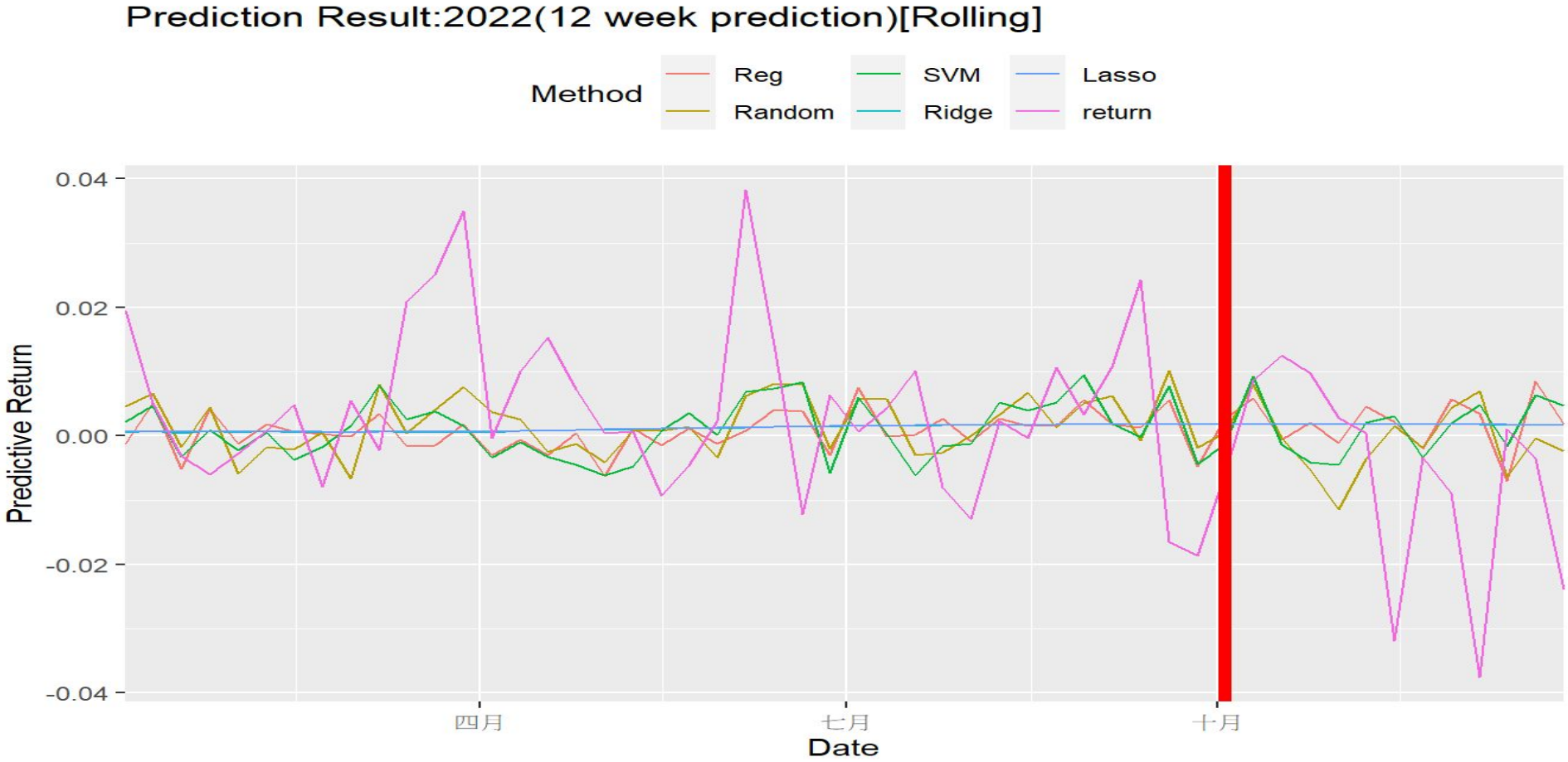
均方根誤差 (RMSE)：所有模型的 12 週預測的 RMSE 略低。這表明與提前 1 週相比，預測未來 12 週時模型的預測誤差略小。

總之，與 1 週預測相比，模型對 12 週預測的表現略好，SMAPE 除外。然而，這些差異非常小，表明模型的性能在不同的預測範圍內相當穩定。

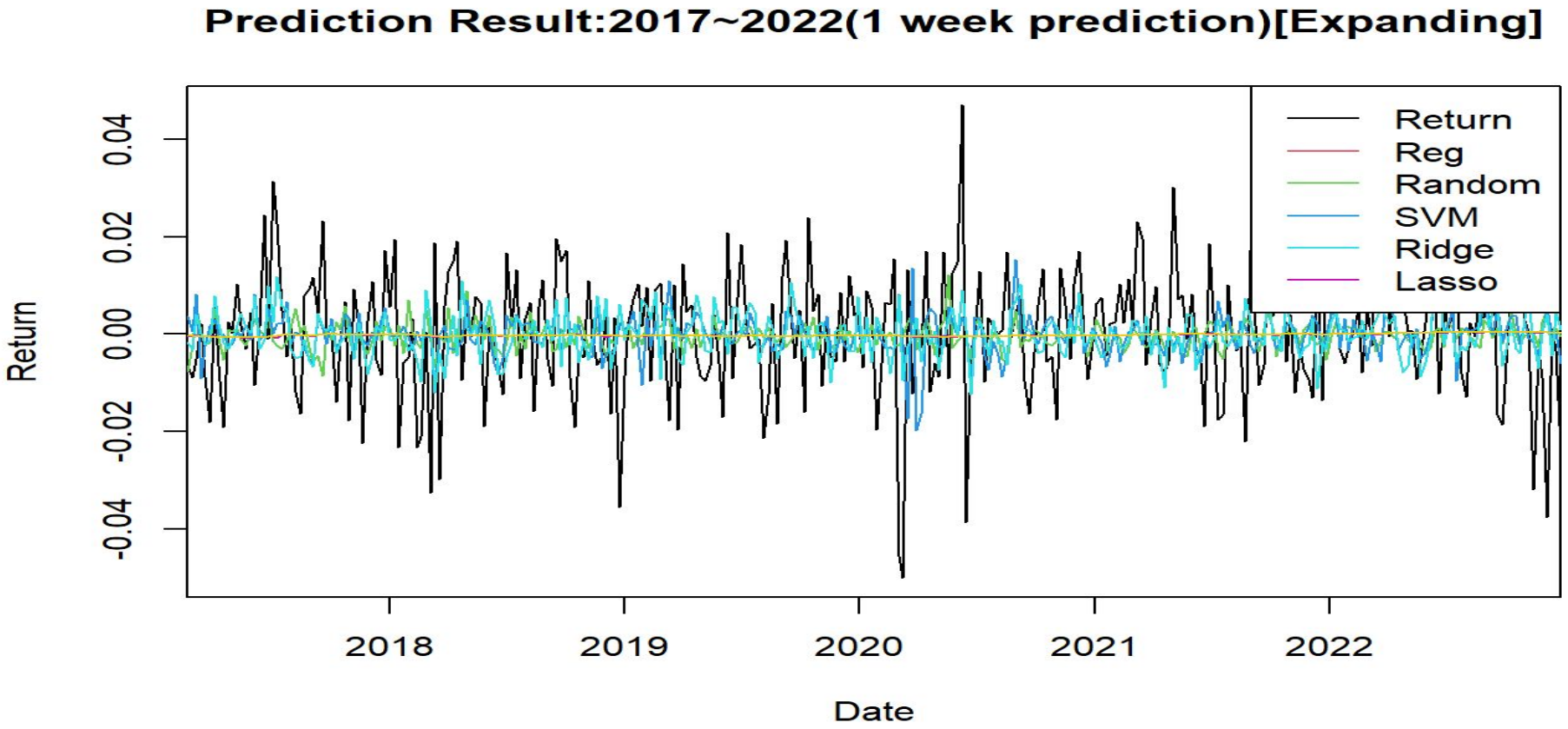
	Reg	Random	SVM	Ridge	Lasso
MAE	0.01065	0.01076	0.01083	0.01015	0.01015
MAPE	143.43143	157.46812	163.08905	103.68575	104.96504
SMAPE	1.53603	1.56993	1.54529	1.78190	1.80815
MSE	0.00020	0.00020	0.00020	0.00018	0.00018
RMSE	0.01418	0.01413	0.01421	0.01336	0.01333



12 weeks rolling prediction result



1 week expanding rolling prediction



1 week expanded rolling prediction

平均絕對誤差 (MAE)：1 week expanded prediction顯示 MAE 略低於回歸的 1 周和 12 週預測，以及隨機、SVM、Ridge 和 Lasso 模型的類似值。這表明當對回歸模型使用擴展的滾動窗口時，模型的平均絕對誤差略小。

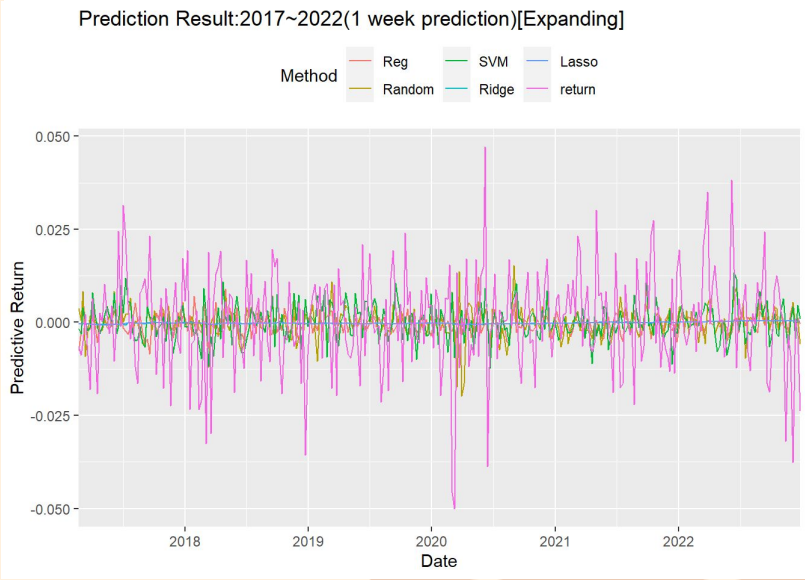
平均絕對百分比誤差 (MAPE)：1 week expanded prediction的 MAPE 低於所有模型的 1 周和 12 週預測。這表明使用擴展滾動窗口時模型的百分比誤差更小。

對稱平均絕對百分比誤差 (SMAPE)：Ridge 和 Lasso 模型的 1 week expanded prediction的 SMAPE 略高，但回歸、隨機和 SVM 模型的 SMAPE 較低。當使用擴展的滾動窗口時，回歸模型、隨機模型和 SVM 模型的預測誤差對稱性得到改善，但 Ridge 和 Lasso 模型的預測誤差對稱性惡化。

均方誤差 (MSE)：對於回歸和 SVM 模型的 1 week expanded prediction，MSE 略低，對於隨機、Ridge和Lasso模型也是如此。這表明使用擴展滾動窗口時平均平方誤差略小或相同。

均方根誤差 (RMSE)：回歸和 SVM 模型的 1 week expanded prediction的 RMSE 較低，隨機、Ridge和Lasso模型的相似。這表明使用擴展滾動窗口時模型的預測誤差更小或相同。

	Reg	Random	SVM	Ridge	Lasso
MAE	0.01036	0.01082	0.01075	0.01014	0.01015
MAPE	120.49858	145.90938	154.79317	101.31599	101.52905
SMAPE	1.60003	1.58026	1.53952	1.88351	1.88598
MSE	0.00019	0.00020	0.00019	0.00018	0.00018
RMSE	0.01375	0.01414	0.01388	0.01331	0.01331



1 week expanded rolling prediction

平均絕對誤差 (MAE)：1 week expanded prediction顯示 MAE 略低於回歸的 1 周和 12 週預測，以及隨機、SVM、Ridge 和 Lasso 模型的類似值。這表明當對回歸模型使用擴展的滾動窗口時，模型的平均絕對誤差略小。

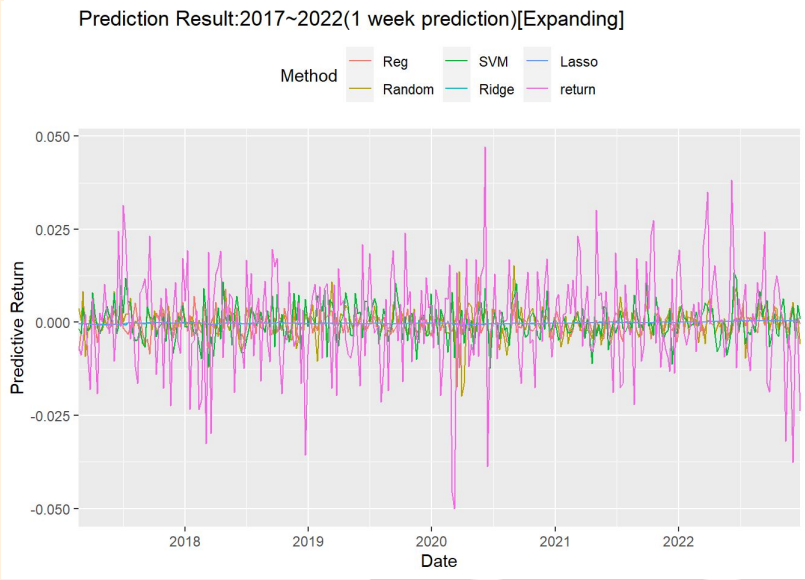
平均絕對百分比誤差 (MAPE)：1 week expanded prediction的 MAPE 低於所有模型的 1 周和 12 週預測。這表明使用擴展滾動窗口時模型的百分比誤差更小。

對稱平均絕對百分比誤差 (SMAPE)：Ridge 和 Lasso 模型的 1 week expanded prediction的 SMAPE 略高，但回歸、隨機和 SVM 模型的 SMAPE 較低。當使用擴展的滾動窗口時，回歸模型、隨機模型和 SVM 模型的預測誤差對稱性得到改善，但 Ridge 和 Lasso 模型的預測誤差對稱性惡化。

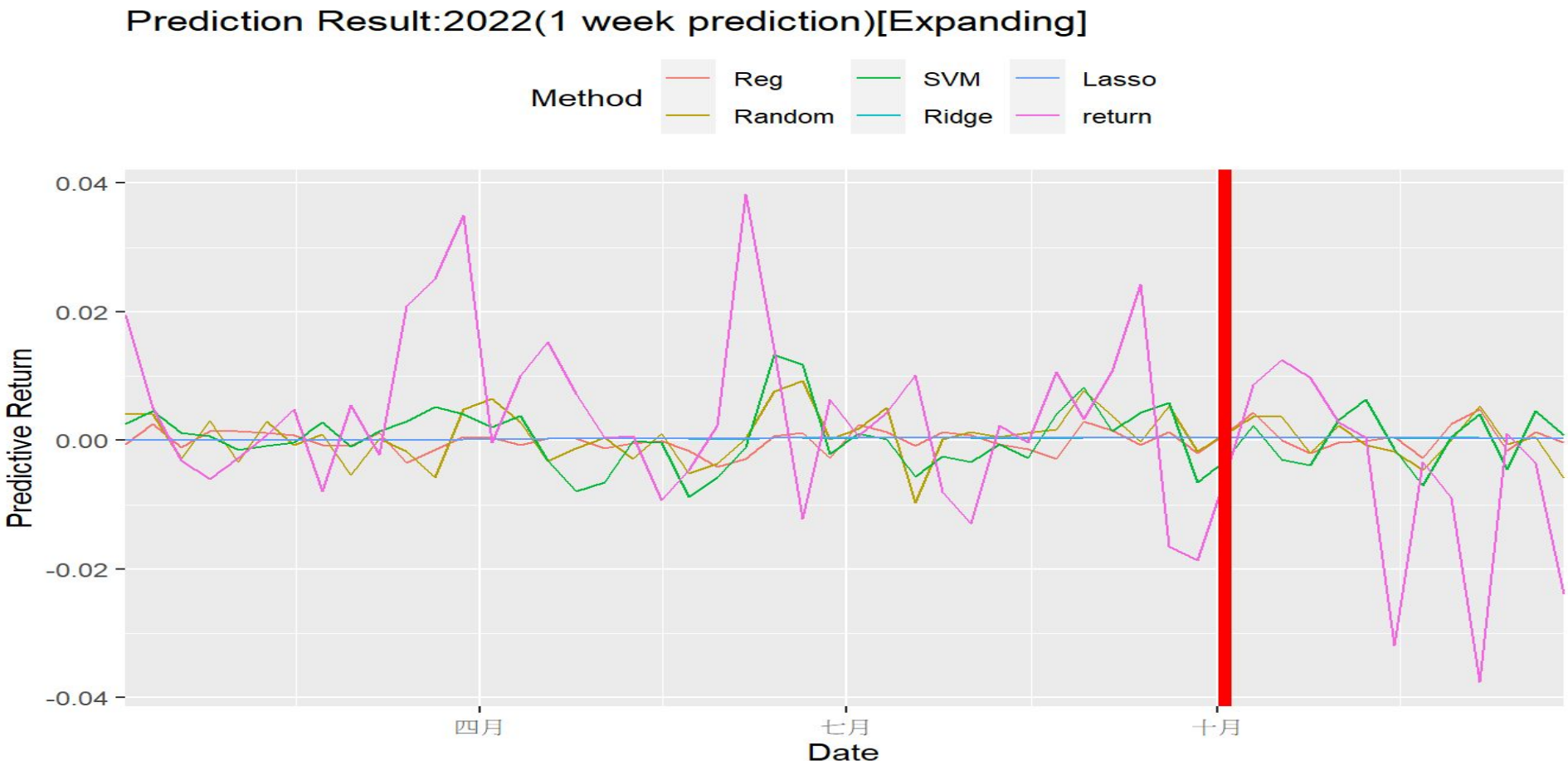
均方誤差 (MSE)：對於回歸和 SVM 模型的 1 week expanded prediction，MSE 略低，對於隨機、Ridge和Lasso模型也是如此。這表明使用擴展滾動窗口時平均平方誤差略小或相同。

總之，使用擴展的滾動窗口進行預測會略微提高某些模型（Reg 和 SVM）的性能，而不會顯著改變其他模型的性能。這些改進在 MAPE 方面更為明顯。

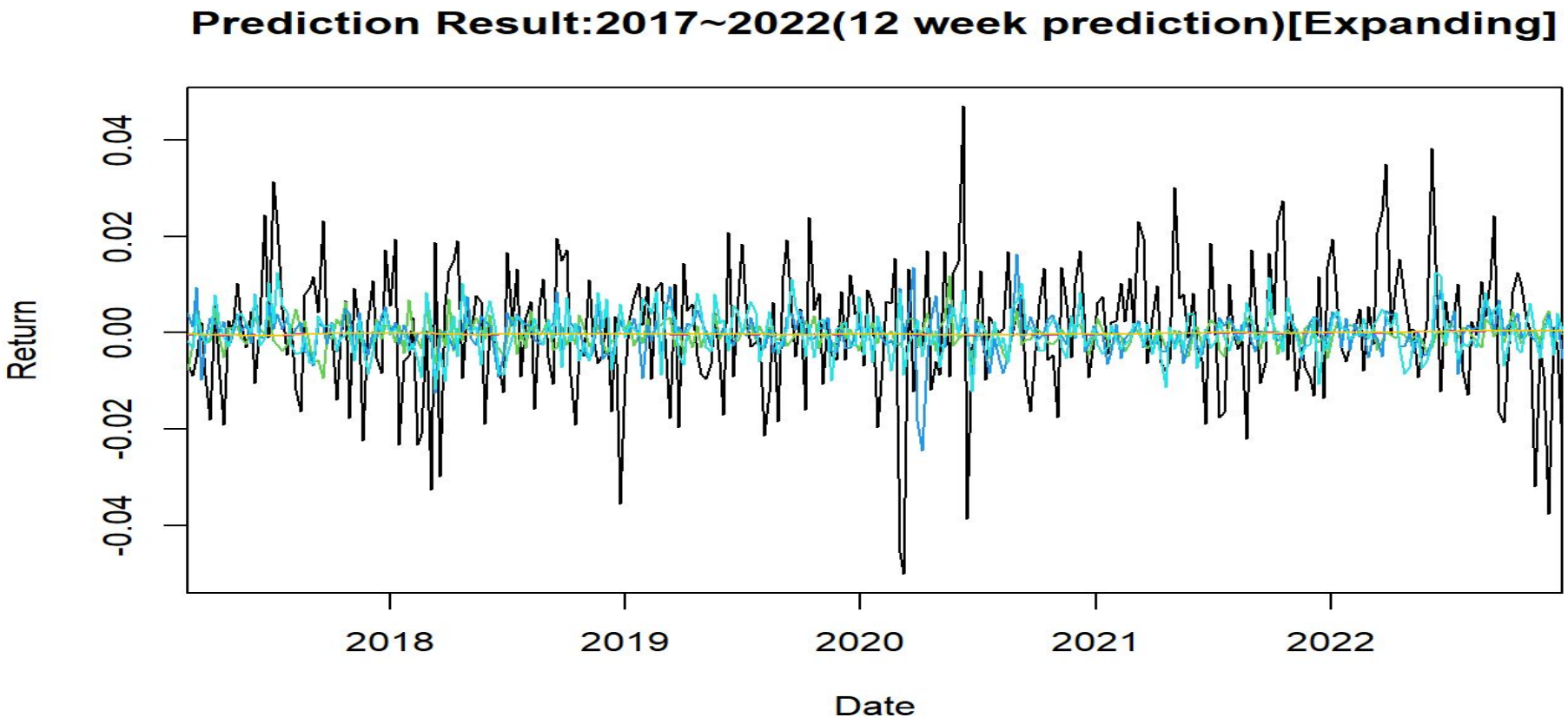
	Reg	Random	SVM	Ridge	Lasso
MAE	0.01036	0.01082	0.01075	0.01014	0.01015
MAPE	120.49858	145.90938	154.79317	101.31599	101.52905
SMAPE	1.60003	1.58026	1.53952	1.88351	1.88598
MSE	0.00019	0.00020	0.00019	0.00018	0.00018
RMSE	0.01375	0.01414	0.01388	0.01331	0.01331



1 week expanded rolling prediction result



12 weeks expanding rolling prediction



12 weeks expanding rolling prediction

平均絕對誤差 (MAE)：12 weeks expanding rolling prediction中所有模型的 MAE 與 1 週擴展預測中的 MAE 相似。這表明將滾動預測窗口從 1 週擴大到 12 週不會顯著改變模型的平均絕對誤差。

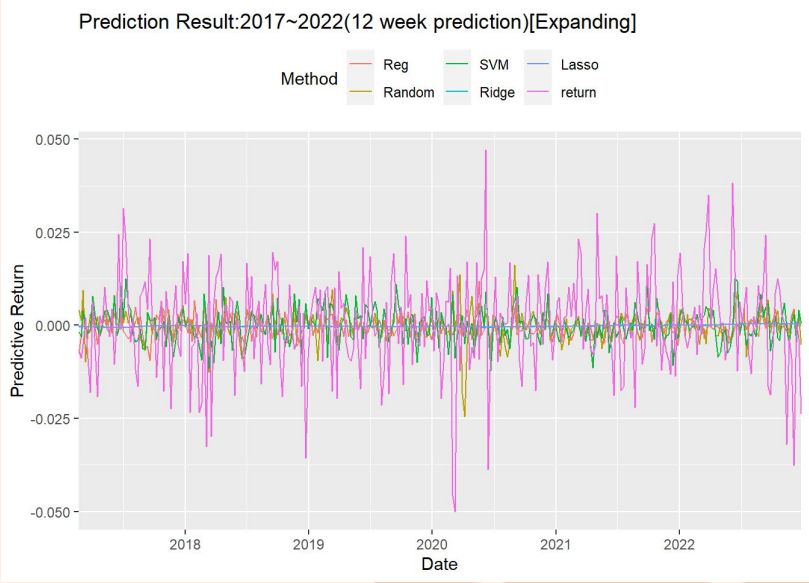
平均絕對百分比誤差 (MAPE)：對於所有模型，12 weeks expanding rolling prediction中的 MAPE 略高於 1 週擴展預測，但仍低於 1 周和 12 週預測。這表明，與 1 週擴展滾動窗口相比，12 週擴展滾動窗口的模型誤差百分比略有增加，但仍低於未擴展窗口。

對稱平均絕對百分比誤差 (SMAPE)：12 weeks expanding rolling prediction中所有模型的 SMAPE 均略低於 1 週擴展預測。這表明，在百分比基礎上，與 1 週擴展滾動窗口相比，使用 12 週擴展滾動窗口時預測誤差對稱性略有改善。

均方誤差 (MSE)：12 weeks expanding rolling prediction中所有模型的 MSE 與 1 週擴展預測中的 MSE 相同。這表明將滾動預測窗口從 1 週擴大到 12 週不會改變模型的平均平方誤差。

均方根誤差 (RMSE)：12 weeks expanding rolling prediction中所有模型的 RMSE 與 1 週擴展預測中的 RMSE 相似。這表明與使用 1 週擴展滾動窗口相比，使用 12 週擴展滾動窗口時模型的預測誤差相似。

	Reg	Random	SVM	Ridge	Lasso
MAE	0.01036	0.01086	0.01074	0.01014	0.01014
MAPE	123.42451	152.97043	156.03366	100.66814	100.67845
SMAPE	1.58914	1.58131	1.54301	1.87570	1.87709
MSE	0.00019	0.00020	0.00019	0.00018	0.00018
RMSE	0.01375	0.01418	0.01388	0.01332	0.01331



12 weeks expanding rolling prediction

平均絕對誤差 (MAE)：12 weeks expanding rolling prediction中所有模型的 MAE 與 1 週擴展預測中的 MAE 相似。這表明將滾動預測窗口從 1 週擴大到 12 週不會顯著改變模型的平均絕對誤差。

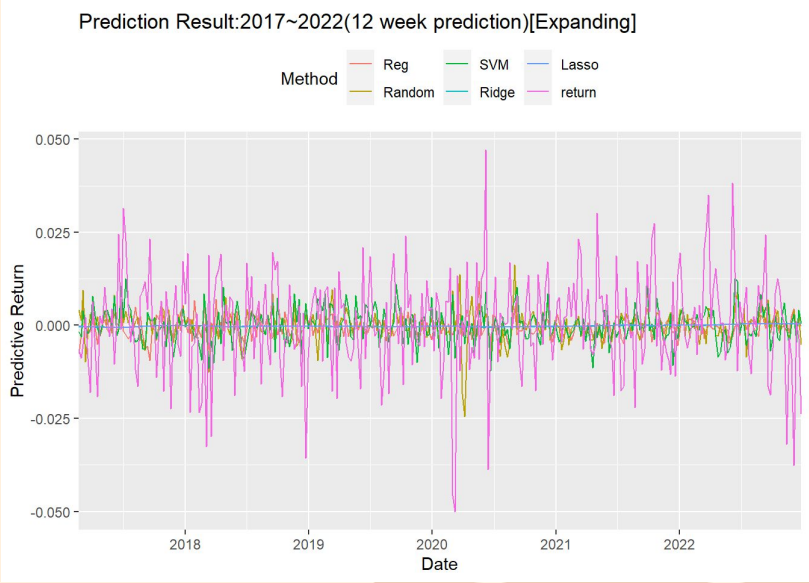
平均絕對百分比誤差 (MAPE)：對於所有模型，12 weeks expanding rolling prediction中的 MAPE 略高於 1 週擴展預測，但仍低於 1 周和 12 週預測。這表明，與 1 週擴展滾動窗口相比，12 週擴展滾動窗口的模型誤差百分比略有增加，但仍低於未擴展窗口。

對稱平均絕對百分比誤差 (SMAPE)：12 weeks expanding rolling prediction中所有模型的 SMAPE 均略低於 1 週擴展預測。這表明，在百分比基礎上，與 1 週擴展滾動窗口相比，使用 12 週擴展滾動窗口時預測誤差對稱性略有改善。

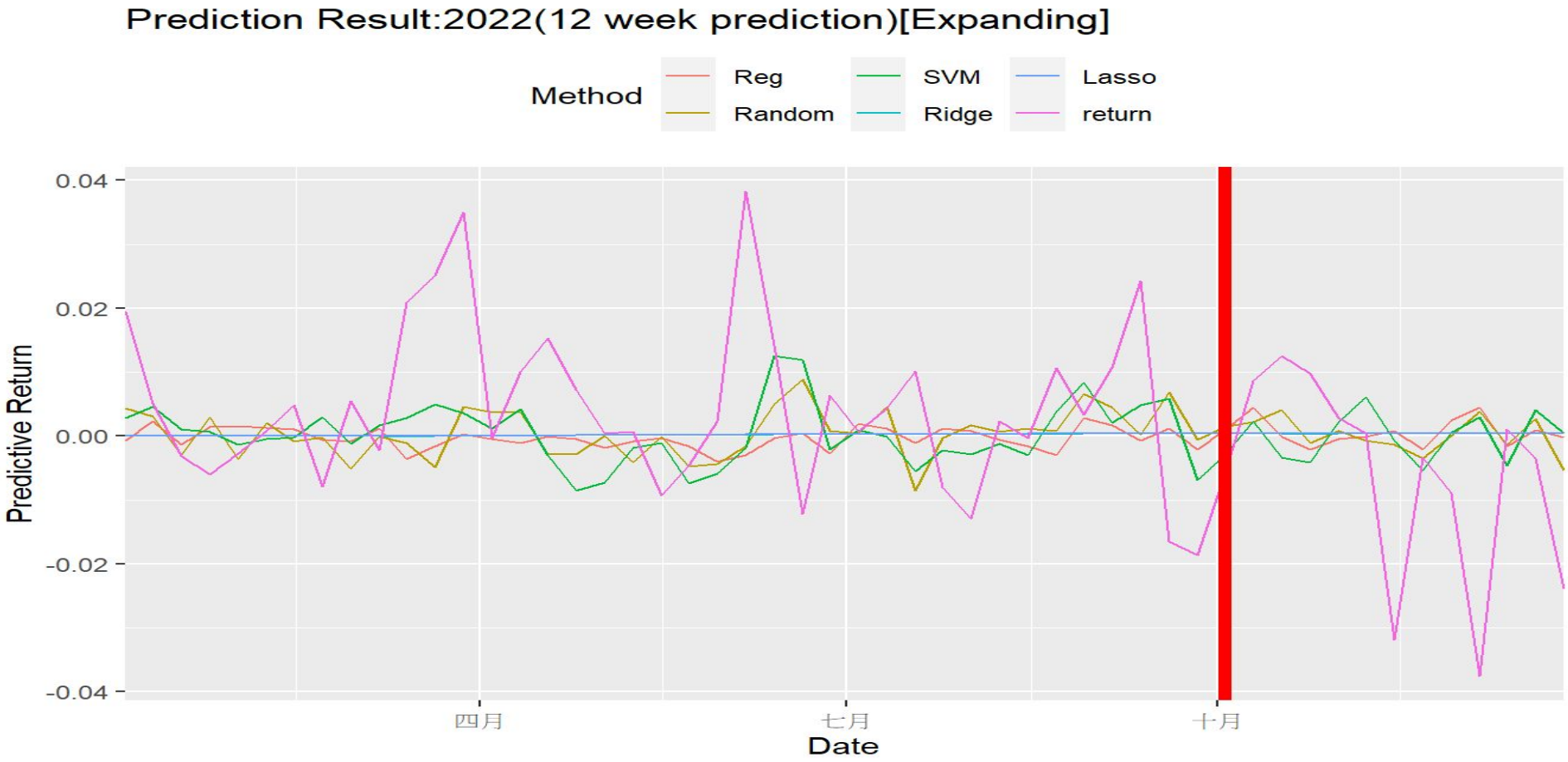
均方誤差 (MSE)：12 weeks expanding rolling prediction中所有模型的 MSE 與 1 week擴展預測中的 MSE 相同。這表明將滾動預測窗口從 1 週擴大到 12 週不會改變模型的平均平方誤差。

總之，將滾動預測窗口從 1 週擴大到 12 週不會顯著改變模型的性能。雖然 MAPE 略有增加，但 SMAPE 略有改善，而 MAE、MSE 和 RMSE 基本保持不變。這可能表明滾動窗口的時間尺度不會顯著影響模型的預測準確性。

	Reg	Random	SVM	Ridge	Lasso
MAE	0.01036	0.01086	0.01074	0.01014	0.01014
MAPE	123.42451	152.97043	156.03366	100.66814	100.67845
SMAPE	1.58914	1.58131	1.54301	1.87570	1.87709
MSE	0.00019	0.00020	0.00019	0.00018	0.00018
RMSE	0.01375	0.01418	0.01388	0.01332	0.01331



12 week expanded rolling prediction result



Conclusion

Ridge 和 Lasso 回歸：在所有實驗中，與其他模型相比，Ridge 和 Lasso 回歸始終具有較低的誤差指標（MAE、MAPE、MSE、RMSE）。

在 Ridge 和 Lasso 之間，Lasso 往往具有略低的 RMSE 和 MAE，這表明它在同時考慮平均誤差 (MAE) 和誤差分佈 (RMSE) 時表現更好。然而，差異非常小，Ridge回歸在某些情況下具有略低的 MAPE。

支持向量機 (SVM)、隨機森林和線性回歸：這些模型在不同實驗中往往比嶺回歸和套索回歸具有更高的誤差指標。在這三者中，很難宣布一個明顯的贏家，因為它們的表現因誤差指標和實驗而異。

Conclusion

比較實驗：在實驗方面，12 周和 1 週的擴展滾動預測實驗似乎比 1 周和 12 週的非擴展預測產生了更好的結果（較低的誤差指標），表明擴展滾動窗口可能有助於提高預測精度。然而，差異相對較小，“最佳”實驗可能取決於具體應用以及計算效率和預測準確性之間的權衡。

總之，Lasso 回歸模型，緊隨其後的是 Ridge 回歸，似乎在所有基於提供的誤差指標的實驗中表現最好。擴展的滾動預測實驗（1 周和 12 週）傾向於產生比非擴展的稍微好一些的結果。

Thanks For Listening!

Please keep this slide for attribution