

Comparing Different Cross Validation Techniques for Regression and Classification Problems in Finance

MENTOR: YUSUF DANISMAN

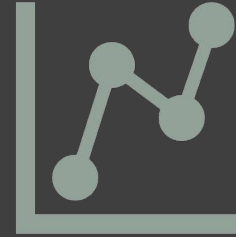
STUDENTS: SEAN HE, MAXIMILLAN YAM

CRSP, MA 905



Classification

Output: Categorical variables in groups, in our case we want to classify stock prices going up or down



Regression

Output: Continuous variable (percentage change, stock price), and use RMSE to measure the output

Goal

Reference Paper



*Data Science in Finance
and Economics*

DSFE, 1(1): 1–20.
DOI: 10.3934/DSFE.2021001
Received: 18 April 2021
Accepted: 26 May 2021
Published: 01 June 2021

<http://www.aimspress.com/journal/dsfe>

Research article

Cross-validation research based on RBF-SVR model for stock index prediction

Feite Zhou*

Department of Applied Mathematics, The Hong Kong Polytechnic University, 11 Yuk Choi Road, Hung Hom, Hong Kong

* **Correspondence:** Email: jupiterzhou@foxmail.com.

Abstract: The ups and downs of stock indexes are one of the most concerned issues for investors in the stock market. To improve the accuracy of stock index prediction, this paper compares traditional K-Fold Cross-Validation (KCV) and three cross-validation methods in the Radial Basis Function Support Vector Regression (RBF-SVR) model. They are named as Abandon Tail Cross-Validation (ATCV), Sequential Division Cross-Validation (SCV) and Gap Sequential Division Cross-Validation (GSCV). It is found that KCV has very limited validation ability for stock indexes time series data with no certain relevance. However, SCV and GSCV with small gap perform better, with high accuracy about 88% and small error about 2%. This research shows that the establishment of time series forecasting models for stock indexes needs to pay more attention to cross-validation methods, which cannot randomly dividing training set and test set. It is strongly recommended to use SCV and GSCV instead of KCV. In addition, the choice of the penalty parameter C and the radial basis kernel function parameter γ largely determines the accuracy and reliability of RBF-SVR stock index prediction model.

Steps

- 1) Data Preparation
 - Import the data
 - Prepare the data
 - Scaling
 - Lag

Data Preparation

Amazon		Amazon								
2000-01-03	4.47	2000-01-04	-0.09							
2000-01-04	4.10	2000-01-05	-0.16							
2000-01-05	3.49	2000-01-06	-0.06							
2000-01-06	3.28	2000-01-07	0.06							
2000-01-07	3.48	2000-01-10	-0.01							

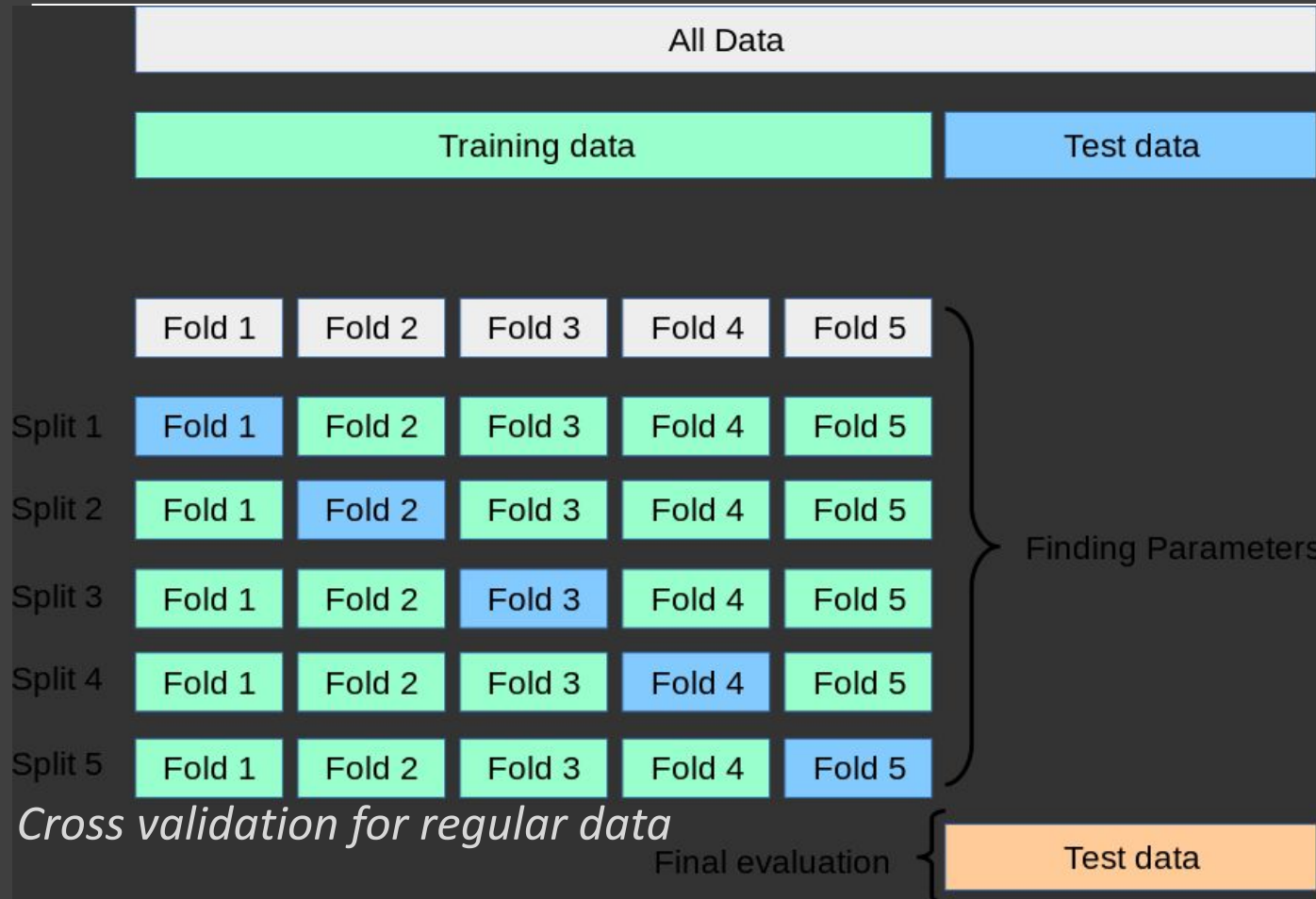
	Amazon	lag_1	lag_2	lag_3	lag_4	lag_5
2000-01-11	-0.035866	-0.005405	0.059222	-0.061914	-0.161039	-0.086884
2000-01-12	-0.048931	-0.035866	-0.005405	0.059222	-0.061914	-0.161039
2000-01-13	0.036684	-0.048931	-0.035866	-0.005405	0.059222	-0.061914
2000-01-14	-0.025926	0.036684	-0.048931	-0.035866	-0.005405	0.059222
2000-01-18	-0.001947	-0.025926	0.036684	-0.048931	-0.035866	-0.005405

Data is converted to log returns to normalize and remove trends.
Lag is then applied to create the features.

Steps

- 2) Split the data into training and test set
- 3) Apply cross validation and grid search

Training, Validation, and Test sets



Time series data is characterized by each sample's respect to time.

Classical cross validation techniques fail to account for this and may use future data to predict the past.

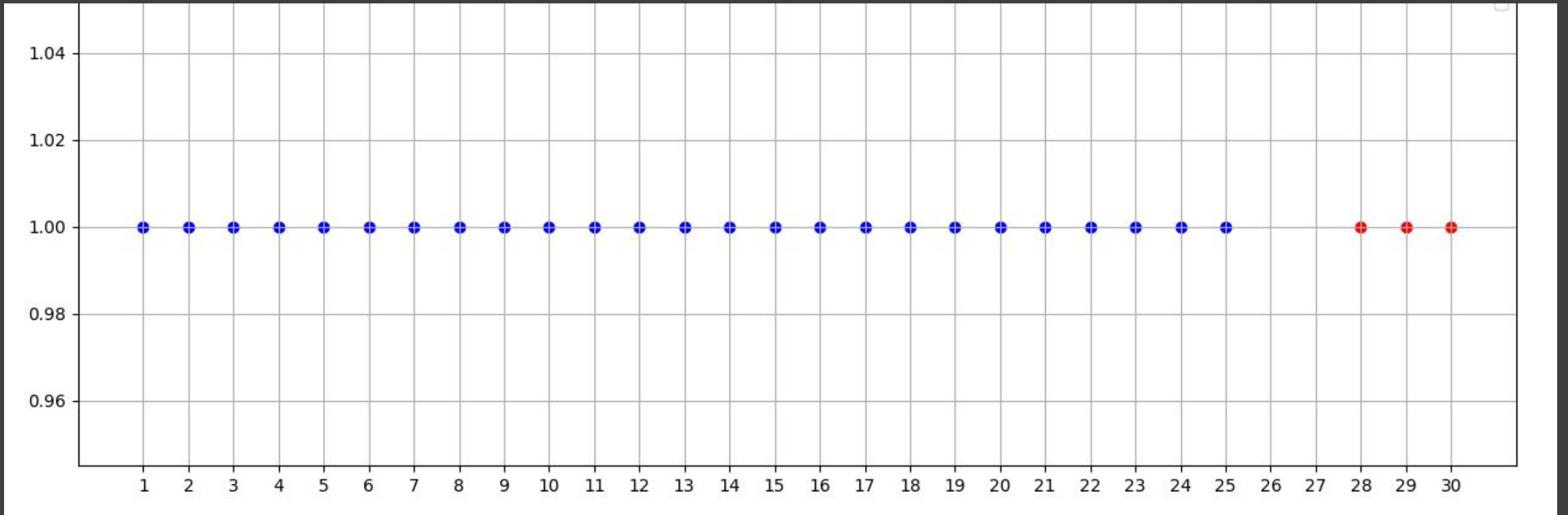
Cross Validation Methods

Simple Fold

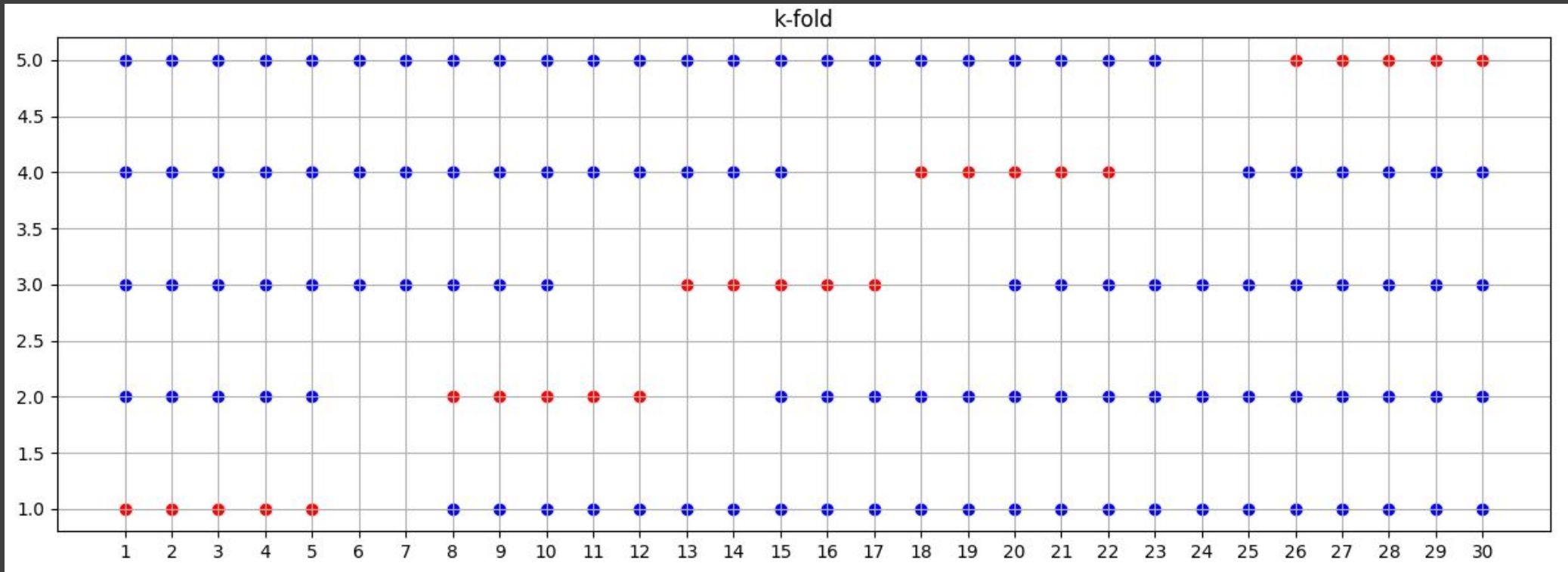
K-Fold

Time Series Split

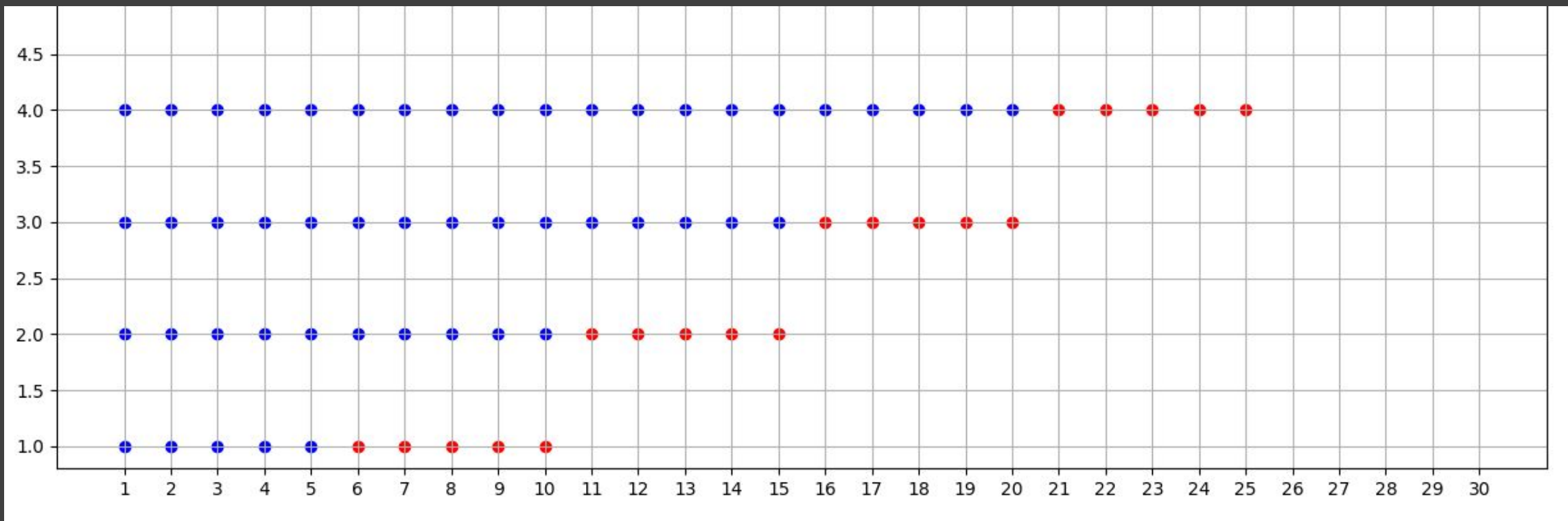
Built In Time Series Split



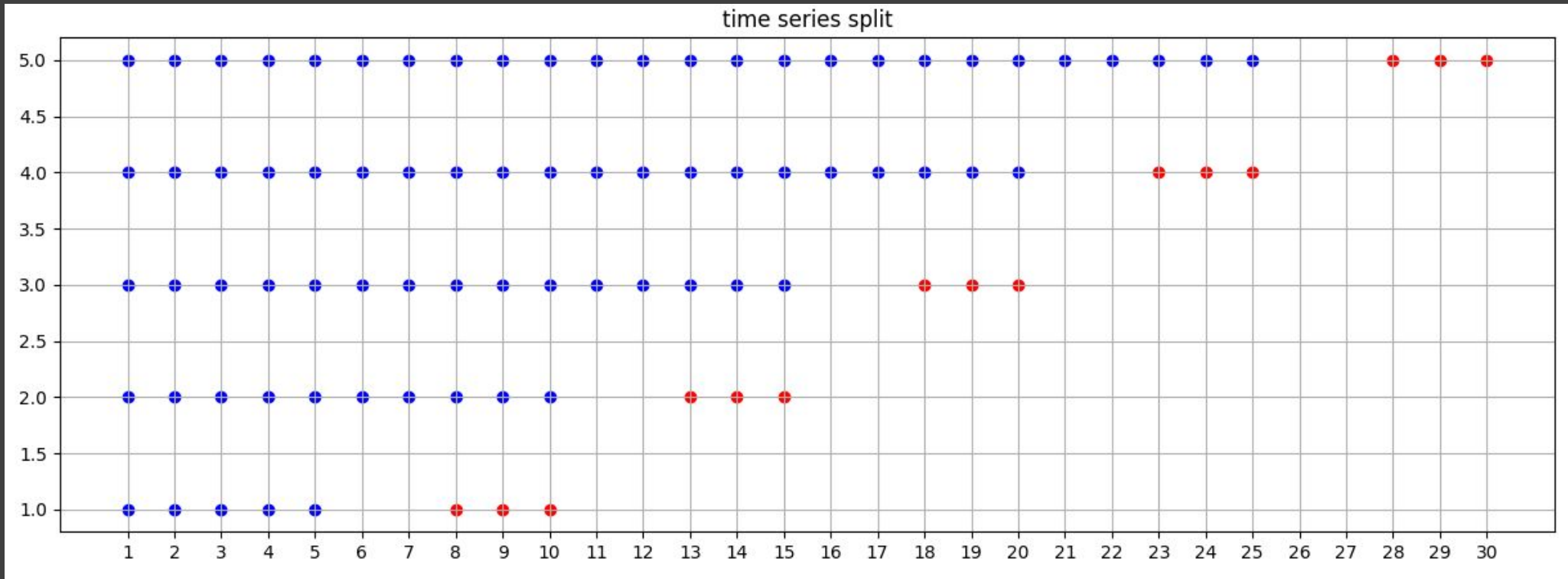
Simple Fold



K-Fold



Built In Time Series Split



Time Series Split (with gaps)

Steps

- 4) Apply classic machine learning algorithms
 - Random Forest Classifier
 - Random Forest Regressor



Random Forest Classifier



Works on a tree based model, each tree will give an categorical variable, and the forest will output the label that is most common.

Classic Machine Learning Models

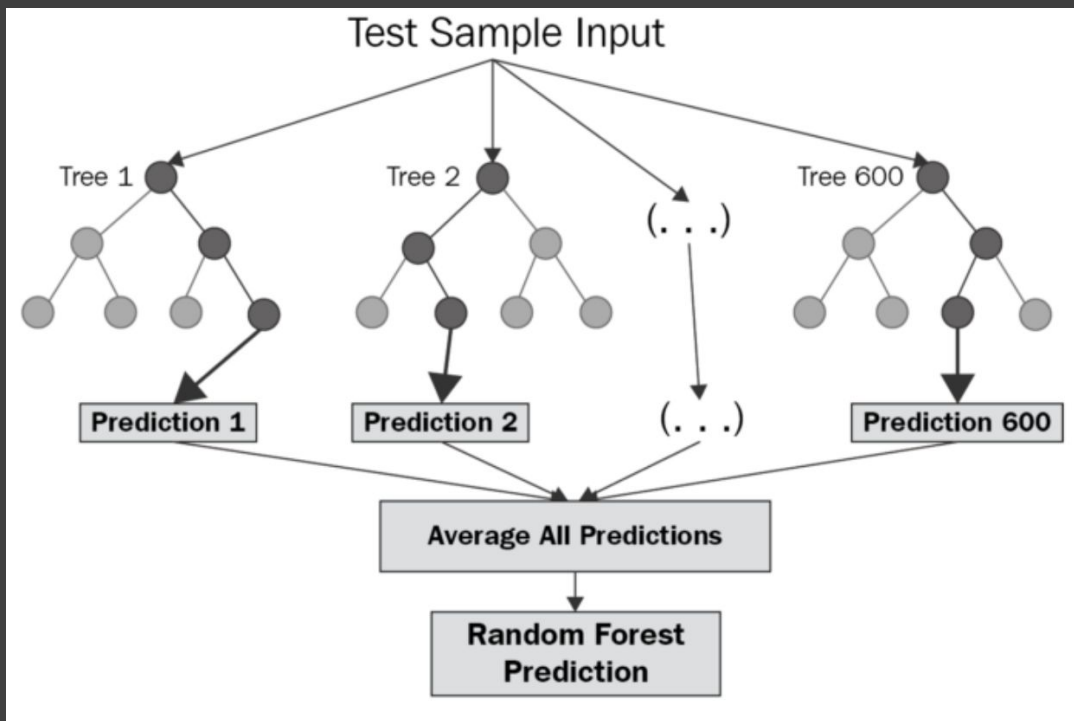


Random Forest Regressor

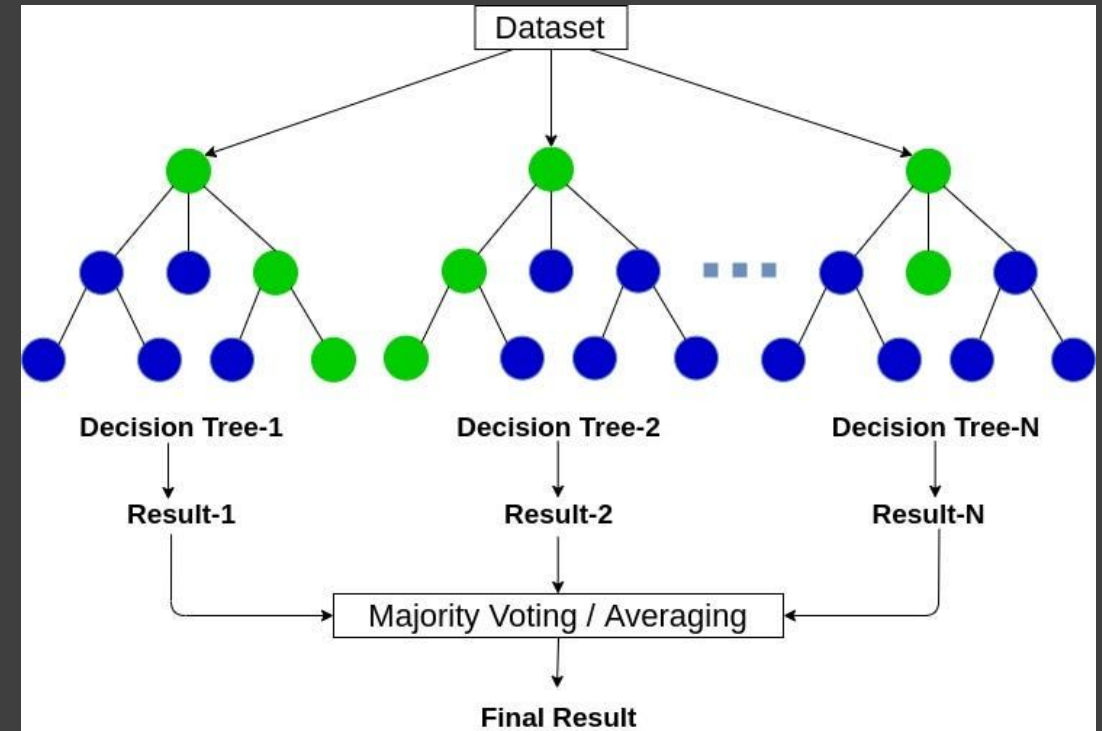


Works on a tree base model that takes the average of all outputs to make a final prediction

Classic
Machine
Learning
Models



Random Forest Classifier



Random Forest Regressor


```
RandomForestClassifier --- Apple          --- Accuracy ---> Training Score: 1.000000 --- Test Score: 0.534783
RandomForestClassifier --- Microsoft      --- Accuracy ---> Training Score: 1.000000 --- Test Score: 0.528986
RandomForestClassifier --- Visa           --- Accuracy ---> Training Score: 1.000000 --- Test Score: 0.507246
RandomForestClassifier --- Amazon         --- Accuracy ---> Training Score: 1.000000 --- Test Score: 0.523188
RandomForestClassifier --- SP500          --- Accuracy ---> Training Score: 1.000000 --- Test Score: 0.523188
RandomForestClassifier --- Gold_ETF       --- Accuracy ---> Training Score: 1.000000 --- Test Score: 0.486957
RandomForestClassifier --- Goldman_Sachs --- Accuracy ---> Training Score: 1.000000 --- Test Score: 0.507246
RandomForestClassifier --- Ford           --- Accuracy ---> Training Score: 1.000000 --- Test Score: 0.497101
```

Random Forest Classification Results

BASE LINE

RandomForestClassifier	---	Apple	---	Accuracy	---	Training Score: 0.555394	---	Test Score: 0.543478
RandomForestClassifier	---	Microsoft	---	Accuracy	---	Training Score: 0.590711	---	Test Score: 0.507246
RandomForestClassifier	---	Visa	---	Accuracy	---	Training Score: 0.564586	---	Test Score: 0.565217
RandomForestClassifier	---	Amazon	---	Accuracy	---	Training Score: 0.607160	---	Test Score: 0.504348
RandomForestClassifier	---	SP500	---	Accuracy	---	Training Score: 0.562651	---	Test Score: 0.560870
RandomForestClassifier	---	Gold ETF	---	Accuracy	---	Training Score: 0.536043	---	Test Score: 0.472464
RandomForestClassifier	---	Goldman_Sachs	---	Accuracy	---	Training Score: 0.562167	---	Test Score: 0.494203
RandomForestClassifier	---	Ford	---	Accuracy	---	Training Score: 0.617804	---	Test Score: 0.510145

Random Forest Classification Results

AFTER GRID SEARCH

RandomForestRegressor	---	Apple	---	RMSE	---	>	Training Score: 0.006043	---	Test Score: 0.022442
RandomForestRegressor	---	Microsoft	---	RMSE	---	>	Training Score: 0.005423	---	Test Score: 0.020514
RandomForestRegressor	---	Visa	---	RMSE	---	>	Training Score: 0.005658	---	Test Score: 0.020108
RandomForestRegressor	---	Amazon	---	RMSE	---	>	Training Score: 0.007392	---	Test Score: 0.021287
RandomForestRegressor	---	SP500	---	RMSE	---	>	Training Score: 0.003545	---	Test Score: 0.015603
RandomForestRegressor	---	Gold_ETF	---	RMSE	---	>	Training Score: 0.009081	---	Test Score: 0.026664
RandomForestRegressor	---	Goldman_Sachs	---	RMSE	---	>	Training Score: 0.006479	---	Test Score: 0.023899
RandomForestRegressor	---	Ford	---	RMSE	---	>	Training Score: 0.006625	---	Test Score: 0.025688

Random Forest Regressor Results

BASE LINE

RandomForestRegressor	---	Apple	---	RMSE	---	Training Score: 0.015863	---	Test Score: 0.022408
RandomForestRegressor	---	Microsoft	---	RMSE	---	Training Score: 0.013903	---	Test Score: 0.020353
RandomForestRegressor	---	Visa	---	RMSE	---	Training Score: 0.014503	---	Test Score: 0.019449
RandomForestRegressor	---	Amazon	---	RMSE	---	Training Score: 0.019489	---	Test Score: 0.020876
RandomForestRegressor	---	SP500	---	RMSE	---	Training Score: 0.009184	---	Test Score: 0.014923
RandomForestRegressor	---	Gold_ETF	---	RMSE	---	Training Score: 0.023400	---	Test Score: 0.026594
RandomForestRegressor	---	Goldman_Sachs	---	RMSE	---	Training Score: 0.016258	---	Test Score: 0.023636
RandomForestRegressor	---	Ford	---	RMSE	---	Training Score: 0.017061	---	Test Score: 0.025496

Random Forest Regressor Results

AFTER GRID SEARCH

Steps

5) Apply deep learning machine learning algorithms

- LSTM

- GRU

Neural Networks



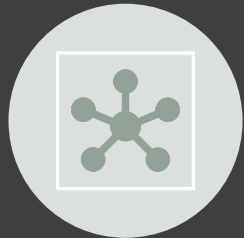
What are Neural Networks



Modeled after the human brain



Consist of layers of nodes(neurons)



Nodes will process and transmit information



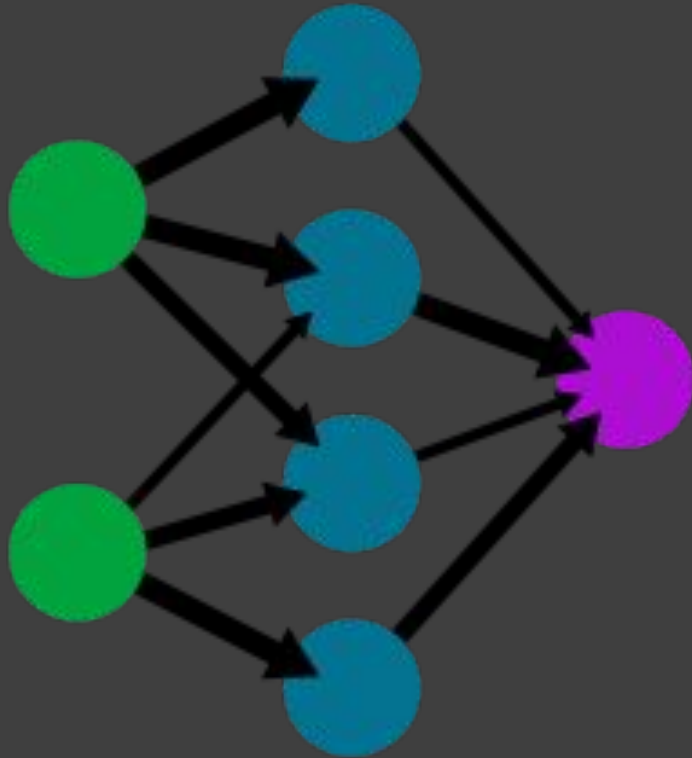
Three main layers, input layer, hidden layers, output layer

A simple neural network

input
layer

hidden
layer

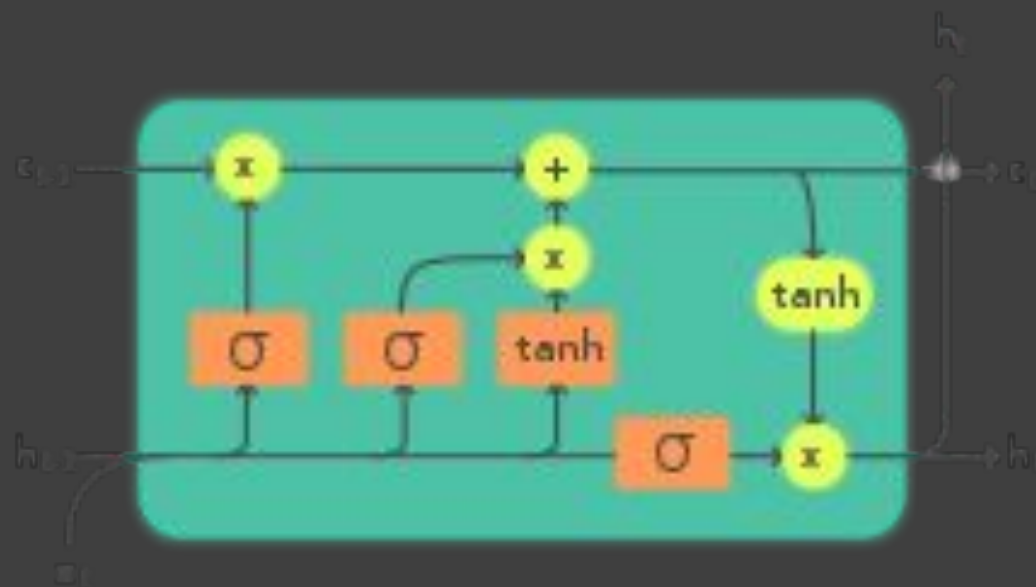
output
layer



Neural Networks

LSTM

	How much new info will be added
	How much info will be removed
	How much info will be outputted



Legend:




Componentwise Copy



Concatenate

LSTM

GRU

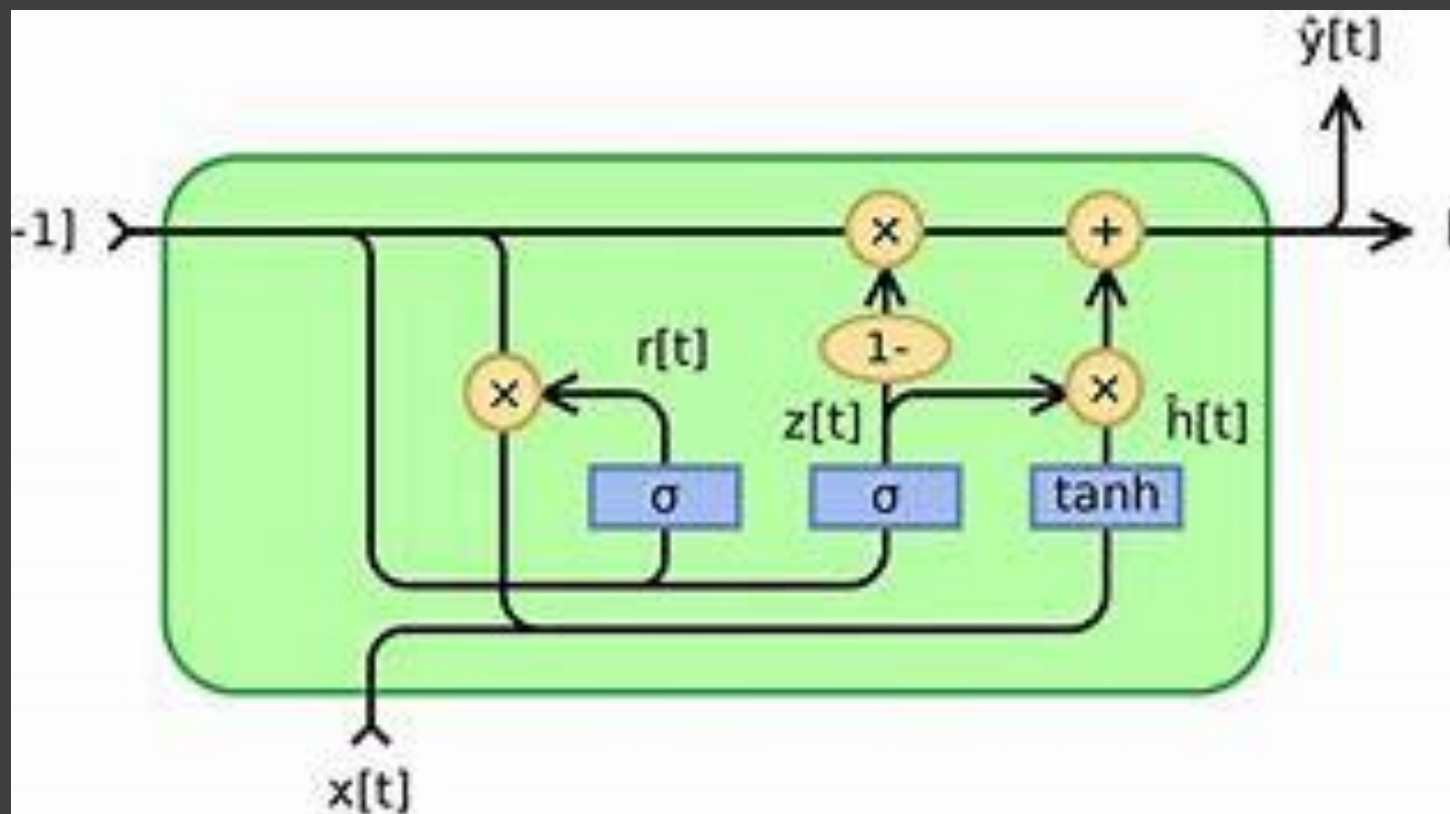


The diagram illustrates the three components of a Gated Recurrent Unit (GRU). It features a vertical stack of three rows. Each row consists of a white square on the left and a teal rectangle on the right. A thin vertical line is positioned to the left of the white squares. The text for each component is located within its respective teal rectangle.

How much info will be removed

How much new info will be added

The more simple form of LSTM



GRU

	CV	n_nodes	BL_train_val	BL_test	Train_Val Acc score	Test Acc Score	Train_Val RMSE score	Test RMSE Score
0	Simple	1	0.530272	0.557971	0.530272	0.554348	0.016331	0.028351
1	kFold	1	0.530272	0.557971	0.530272	0.557971	0.016393	0.028259
2	TSS	1	0.530272	0.557971	0.491670	0.500000	0.016609	0.028592
3	Blocked	1	0.530272	0.557971	0.530272	0.550725	0.016326	0.028286

LSTM Results

	CV	n_nodes	BL_train_val	BL_test	Train_Val Acc score	Test Acc Score	Train_Val RMSE score	Test RMSE Score
0	Simple	1	0.530272	0.557971	0.530272	0.561594	0.016424	0.028279
1	kFold	1	0.530272	0.557971	0.530272	0.557971	0.016317	0.028329
2	TSS	1	0.530272	0.557971	0.530272	0.557971	0.016325	0.028268
3	Blocked	6	0.530272	0.557971	0.529053	0.557971	0.016769	0.028530

GRU Results

Future Work



Use a larger grid of hyperparameters (nodes, layers, etc.)



Add more days of lag per sample



Use larger dataset as input



Acknowledgements



Questions