Prediction of Standard and Poor's 500

UPSTAT 2018

Andrew Flannery,
George Kuliner,
and
Chase Yaeger

GENESEO STATE UNIVERSITY OF NEW YORK

What We Will Be Discussing

- Introduction to the Data Set
- Poor Performing Models
- Classifying Feed Forward Neural Network
- Forecasting using stochastic LSTM
- Forecasting using Exponential Smoothing
- Reviewing Model Performance

Introduction to the Data Set

The Predictors

- S and P 500 Close: the value we wanted to predict
- Dow Jones Industrial and Average Indices
- Effective Federal Funds Rate: gives quantitative data to “expectation” for the future of economy
- Bond Yields: gives quantitative data to “mood” of financial investors
- VIX: volatility indexes of market, S and P 100, S and P 500 over three months
- S and P 500: open, high, volume
- SPDR (SPY): ETF traded on the S and P 500

Collecting of Data

- Federal Reserve of Economic Data
- Yahoo Finance
- 22 Predictors
- Start Date: 4/14/08
- End Date: 3/27/18

Cleaning the Data

- Utilized VBS in Excel to merge data files
- Missing values cleaned with Macros created in VBS
- Chose to omit observations with no S and P index values
- Other predictors missing values filled in with previous days value

Poor Performing Models

Naive Forecasting

- This method involves using the last periods actual value as the current periods forecast without adjustment or causal factors
- We used all predictors for this model.

Results

- Very High Test Results
- RMSE: 6711.73
- MAPE: 125.86
- MPE: -34.56
- All of our other indicators of performance for our forecast were very high.

Forecasting Using a Linear Model

- For this model, we built a linear model in which we forecasted future values
- Our linear model consisted of all predictors except for SP500Adj.Close
- Adjusted Closing was found to be extremely correlated with our dependent variable, SP500close.

This model yielded good performance on our training data, but performed poorly on our test data.

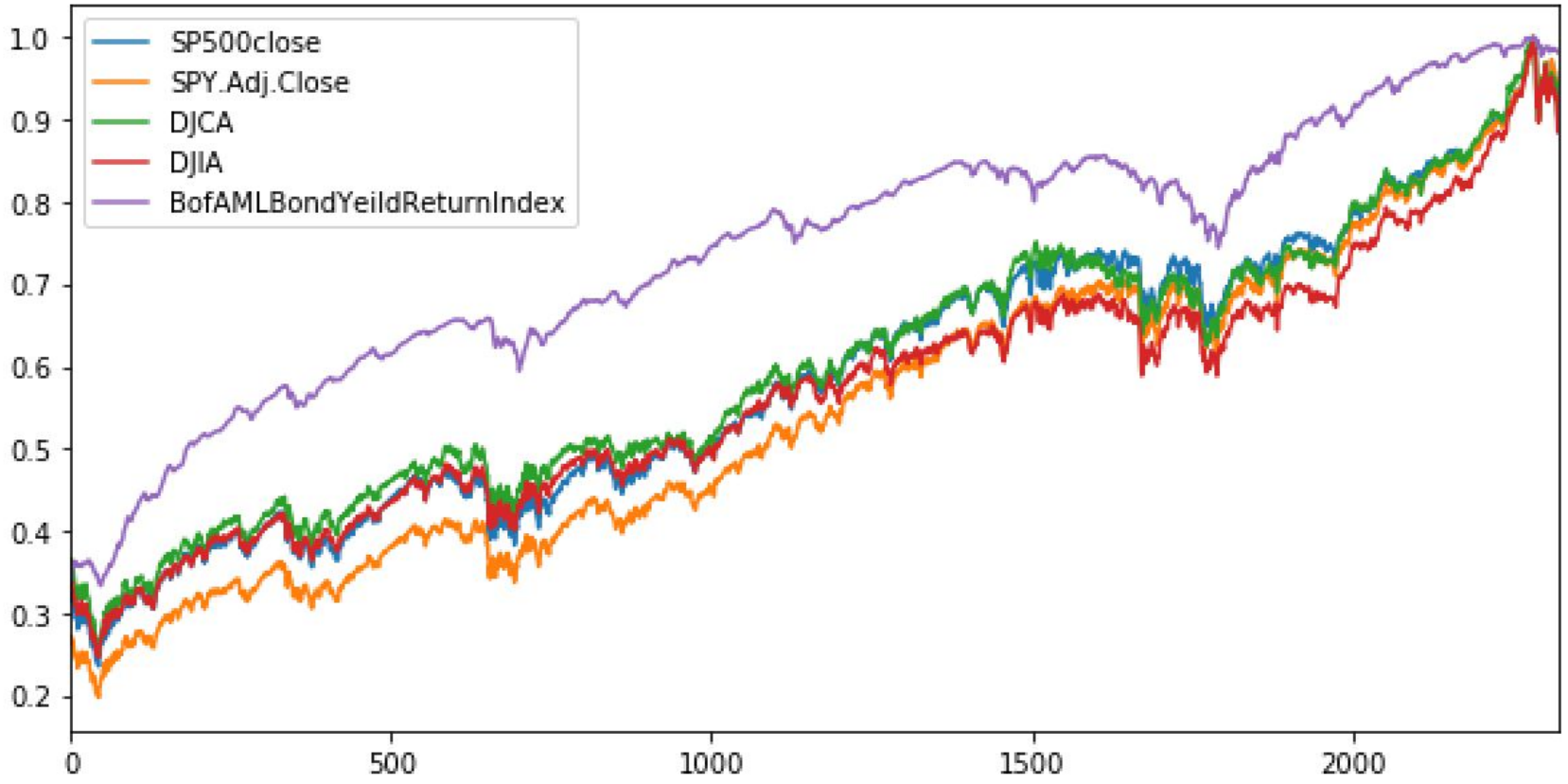
	ME	RMSE
Training set	0.00000000000000001068555639	1.743214958
Test set	1315.781388831481308443471789	1316.935797134
	MAE	MPE	MAPE
Training set	1.184959871	-0.0003794647529	0.08212434593
Test set	1315.781388831	48.56695005567	48.56695005567

Time Series Classification with Feed-Forward Neural Network

Goal

Predict whether the S&P500 close today will be higher or lower than the index the previous day.

Correlation - Time Series Graph



Predictors

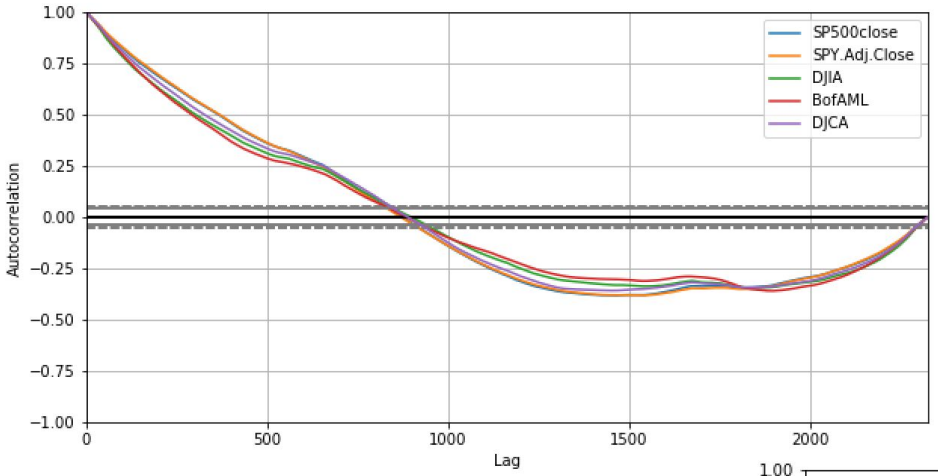
DJIA - Dow Jones Industrial Average

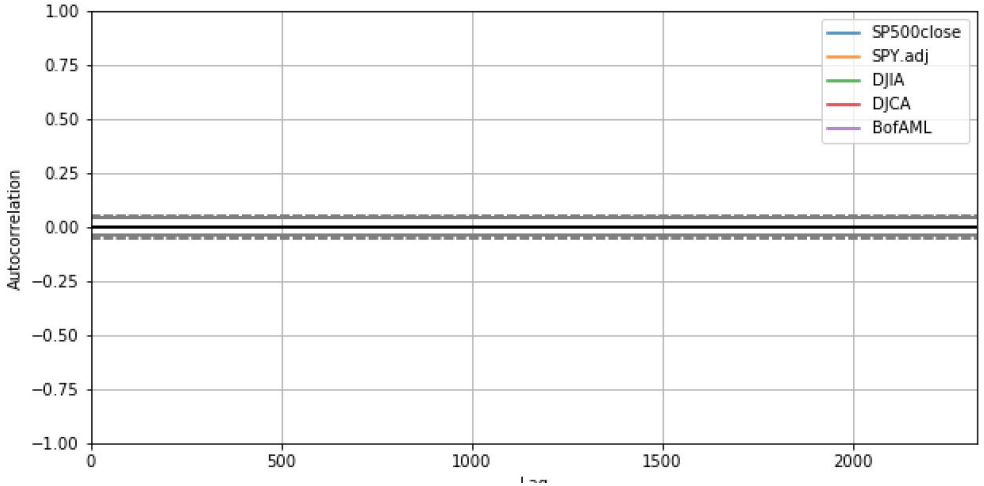
DJCA - Dow Jones Composite Average

BofAML - Bond Yields for Bank of America and Merryl Lynch

SPY.adj - SP500 adjusted for inflation

Stationary Data





Model

- Classification: predict log of return of Sp500 is positive or negative
- Training Data 80% of dataset
- Testing Data: 20% of dataset
- Binary Classification with TensorFlow
- Feed Forward Neural Network
- Weights as predictors, defined cost function

Judgment

Precision: Ability of classifier to not incorrectly label negative/positive

Recall: Ability of classifier to find all positive samples

F1 Score: A weighted Avg. of the precision and recall

Accuracy: percent correctly predicted in the test data

Binary Classification

Output:

```
1000 0.54633623
2000 0.54633623
3000 0.54741377
4000 0.54633623
5000 0.5479526
6000 0.54956895
7000 0.5538793
8000 0.5538793
9000 0.5549569
10000 0.5560345
Precision =  1.0
Recall =  0.01442307692307692 4
F1 Score =  0.028436018957345974
Accuracy =  0.5581896551724138
```

Feed-Forward Neural Network

Output:

```
1000 0.54633623
2000 0.54633623
3000 0.56681037
4000 0.8512931
5000 0.9612069
6000 0.96875
7000 0.9752155
8000 0.9779095
9000 0.9779095
10000 0.9784483
Precision =  0.9565217391304348
Recall =  0.9519230769230769
F1 Score =  0.9542168674698795
Accuracy =  0.959051724137931
```

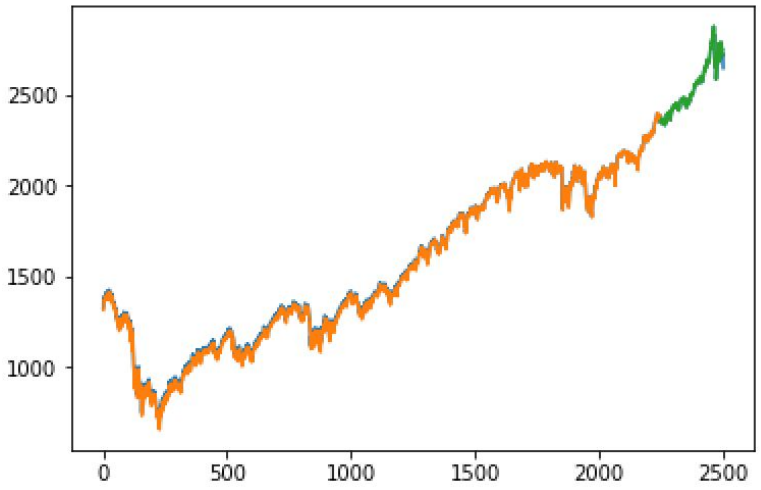
Stochastic Modeling Using LSTM RNN

How the Models Were Made

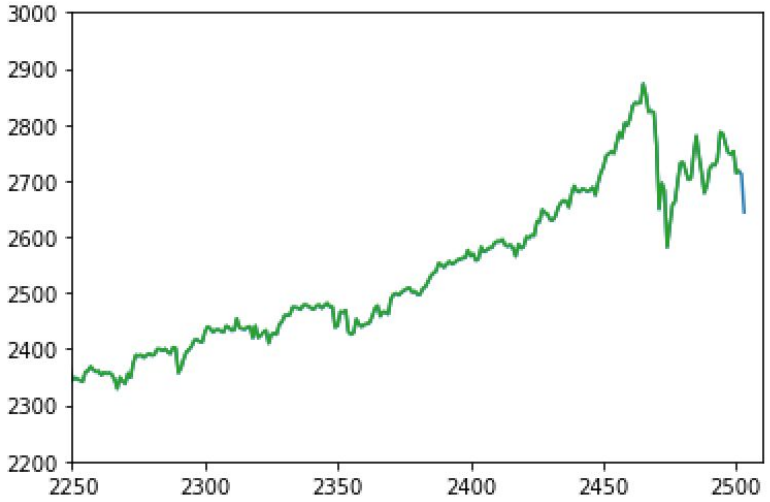
- Python 2.7
- Using Keras, Python, SKlearn, Panda, PyPlot, Math, and Numpy Libraries
- Shifted Data, then Normalized
- RMSE were computed after inverse normalizing the data
- Goal: Predict Index Value

Using Previous Day as Predictor

Full Data



Test Data

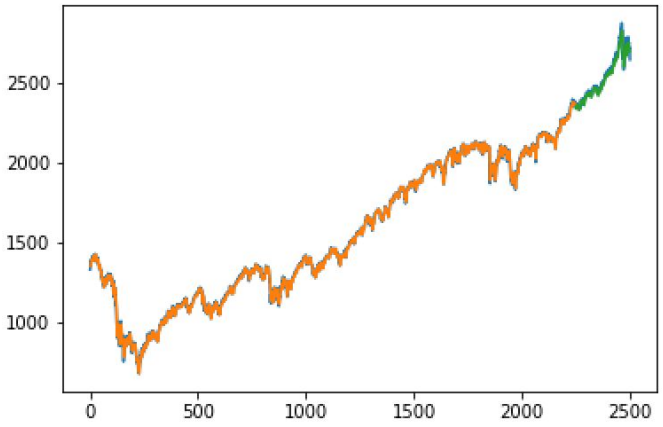


- Trained on 10 epochs
- 100 Nodes used
- Adam Optimizer

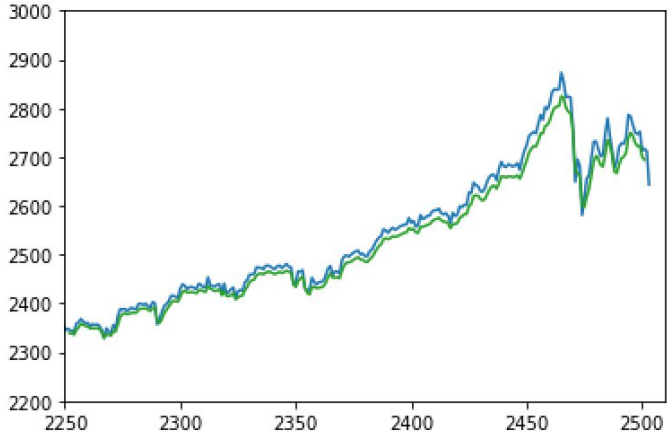
- Training RMSE: 20.25
- Testing RMSE: 17.04

Rolling Window

Full Data



Test Data



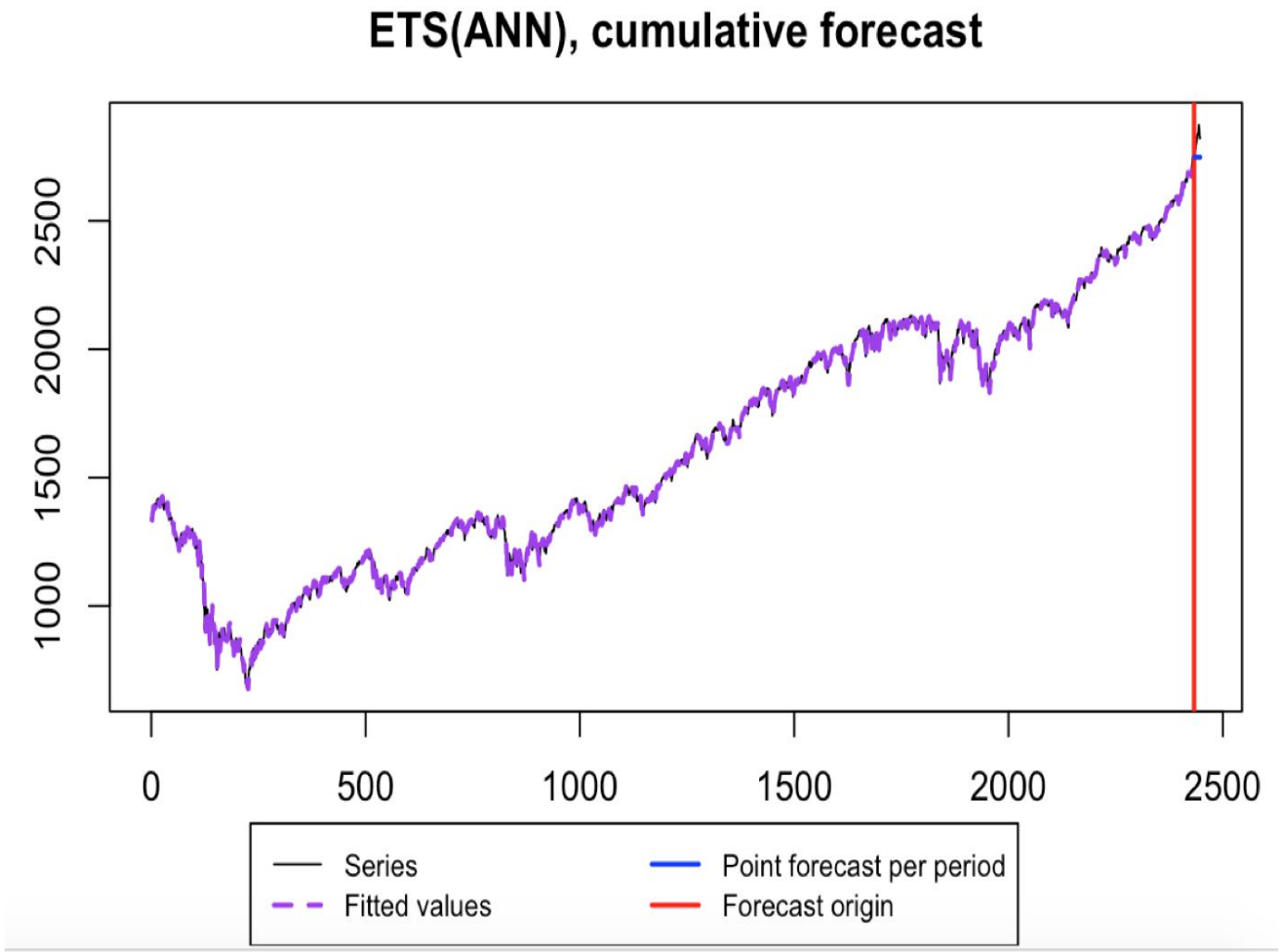
- Used lag of 3 days only using index values
- Trained on 15 Epochs
- 150 Nodes used

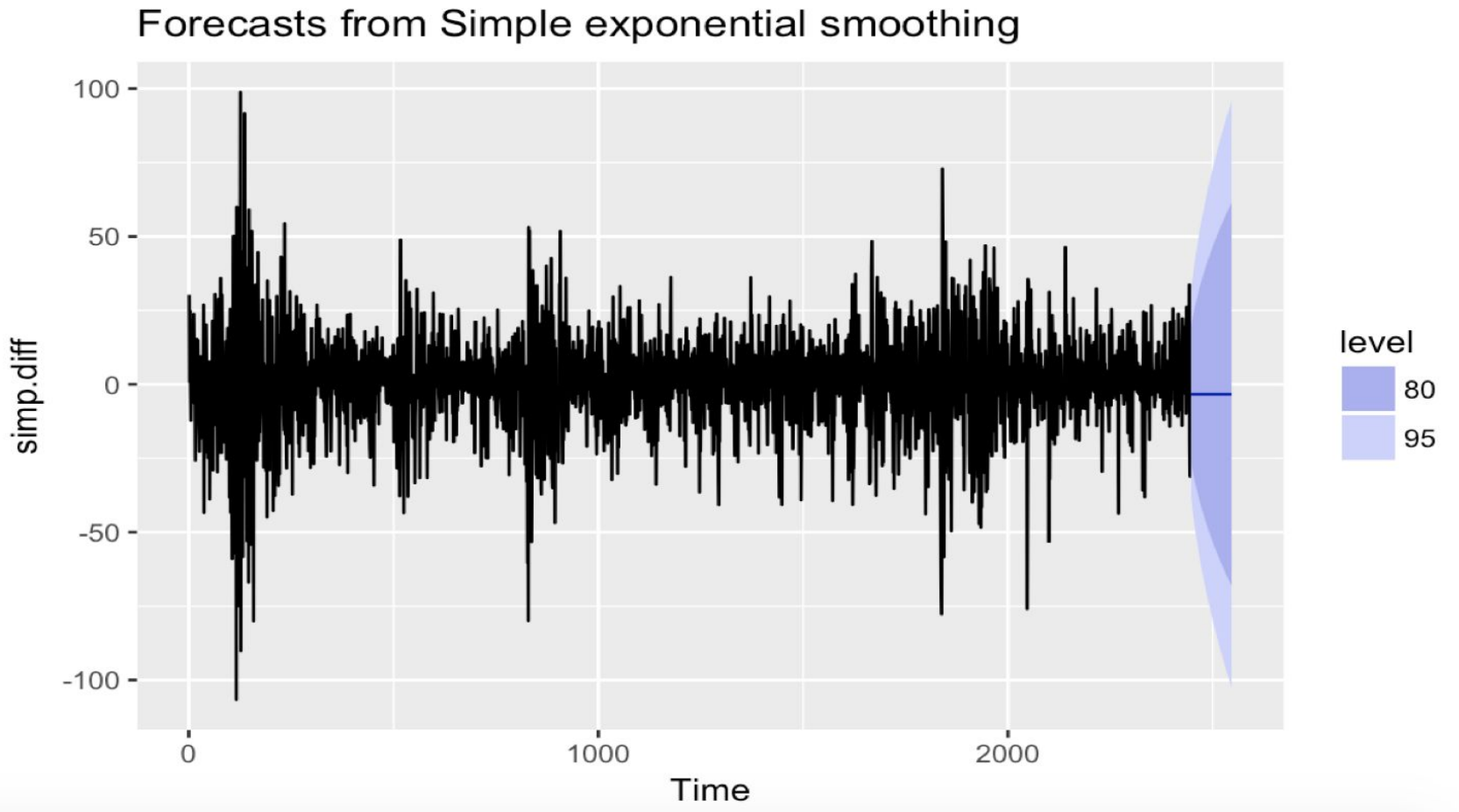
- Training RMSE: 17.52
- Testing RMSE: 26.25

Forecasting Using Exponential Smoothing

- For this method, we used the data from 4/14/08-12/31/17 as our training data set
- Data from 1/2/18-3/27/18 as our test data set

- Using the caret, tseries, and smooth package, we built a forecasting model by taking the lagged differences in our time series model
- We then found our optimal value for α to be 0.27. Finally, we used this model to forecast our future SP500 values


ETS(ANN), cumulative forecast
1000
1500
2000
2500
0
500
1000
1500
2000
2500
Series
Fitted values
Point forecast per period
Forecast origin


Forecasts from Simple exponential smoothing
100
50
0
-50
-100
simp.diff
0
1000
2000
Time
level
80
95

This model gives us the following results:

	ME	RMSE	MAE
Training set	-0.0226279648	17.64117666	12.45616309
Test set	-2.2989678652	40.41192862	30.87292518
	MPE	MAPE	MASE
Training set	-Inf	Inf	0.7360151045
Test set	102.607125	113.9991492	1.8242326372

As we can see, we achieved a model with a very low training and test RMSE, but fairly high test MPE and MAPE

Reviewing Model Performance

Error Rates of Each Model

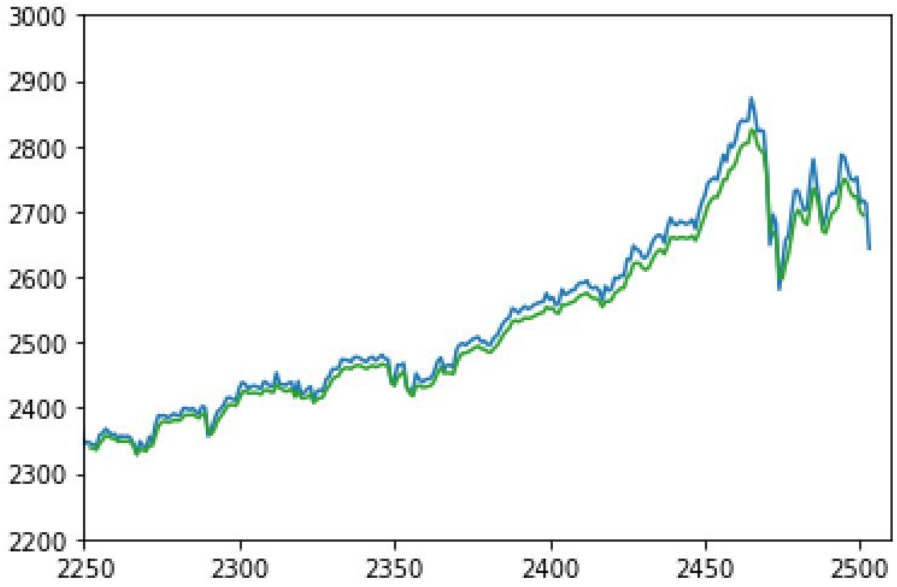
Model	Data Set	Accuracy	RMSE	MAPE	MPE
Naive Forecasting	Training	N/A	921865750	2509.13	-2474.98
	Test	N/A	6711.72	125.86	-34.56
Forecasting with Linear Model	Training	N/A	1.74	0.08	-0.0003
	Test	N/A	1316.93	48.56	48.56
Binary Classification	Training	.556	N/A	N/A	N/A
	Test	.558	N/A	N/A	N/A
Feed-Forward Neural Network	Training	.978	N/A	N/A	N/A
	Test	.959	N/A	N/A	N/A

Model	Data Set	Accuracy	RMSE	MAPE	MPE
LSTM (previous day)	Training	N/A	20.25	N/A	N/A
	Test	N/A	17.04	N/A	N/A
LSTM Rolling Window	Training	N/A	17.52	N/A	N/A
	Test	N/A	26.25	N/A	N/A
Exponential Smoothing	Training	N/A	17.64	INF	-INF
	Test	N/A	40.42	113.99	102.61

The Best Model

Stochastic LSTM Rolling Window

- Higher novelty by underpredicting
- Able to predict troughs and peaks without lag
- LSTM allowed model to learn cycles of S and P 500



References

[1]-https://fred.stlouisfed.org/

[2]-https://finance.yahoo.com/

Questions?