Machine Learning Models in Algorithmic Trading

Abuubakry Ali, David Savage, Elizabeth Morris, Jessica Carmichael, Daniel Lynes

Introduction

Algorithmic trading has become a buzzword more and more people seem to be entering this fast-growing domain, but what makes a good trading algorithm? In this project, we will seek to compare two models to see which is better for predicting the market. We will use fundamental economic indicators in one model which we will call our "trend" model and technical indicators for the other model, which we will refer to as our "trade" model.

Our Project Goals

- To compare an algo trading strategy built on fundamental economic features (trend) and an algo trading strategy built on technical features (trade).
- Building a ML model using fundamental economic inputs to predict the monthly returns of S&P500

Our Core Message

Predicting stock market returns using economic indicators creates a more reliable model than a model based on technical indicators.

Trend Model

Project Goal: to build a Machine Learning Model using fundamental economic inputs to predict monthly S&P500 returns. Then using this model, we will build an algorithmic trading strategy aiming to outperform the S&P500 return.

Model Features:

Date	Monthly Return	Initial_Claims_YoY	Capacity_Utilization_YoY	US_Real_DPI_YoY	M2_YoY	Inflation Rate	Inversion Signal	Vol Signal
2020-04-30	1.0	879.522498	-6.174855	0.630061	13.860213	1.98	0.0	0.0
2020-05-31	1.0	2238.863109	-17.629041	16.482196	19.313314	1.87	0.0	1.0
2020-06-30	1.0	1312.256586	-16.834414	10.724800	22.058693	1.92	0.0	1.0
2020-07-31	1.0	442.624434	-11.586505	8.388189	23.215818	1.87	0.0	0.0
2020-08-31	1.0	540.350877	-8.116476	8.250507	22.685123	1.86	0.0	1.0

Trend Model - Volatility and Inversion Signal

Inversion Signal:

Date	10yr minus 2yr yield	Inversion Signal
2019-03-31	0.14	False
2019-04-30	0.24	False
2019-05-31	0.19	False
2019-06-30	0.25	False
2019-07-31	0.13	False
2019-08-31	0.00	True
2019-09-30	0.05	False
2019-10-31	0.17	False

Volatility Signal:

Date	Adj Close	VIX Monthly Return	Vol Signal
2019-03-31	13.710000	-7.239511	1.0
2019-04-30	13.120000	-4.303429	1.0
2019-05-31	18.709999	42.606702	0.0
2019-06-30	15.080000	-19.401386	1.0
2019-07-31	16.120001	6.896558	0.0
2019-08-31	18.980000	17.741927	0.0
2019-09-30	16.240000	-14.436248	1.0
2019-10-31	13.220000	-18.596056	1.0

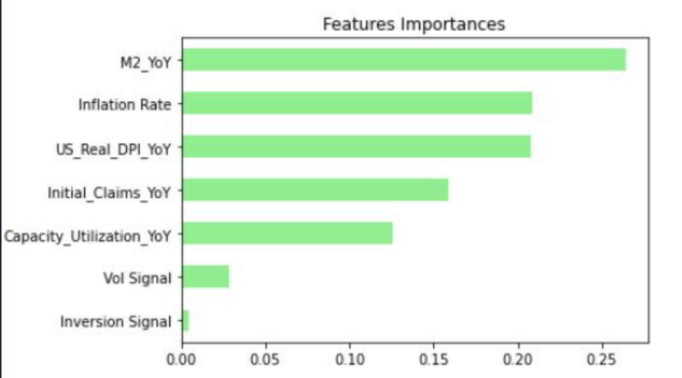
Trend Model - Random Forest

The accuracy score overall for the model was 72%, however as you can see from the confusion matrix the model was heavily biased towards predicting positive monthly returns.

Confusion matrix:

	Predicted Negative	Predicted Positive
Negative	2	23
Positive	4	69

Feature Importance:



Trend Model - Gradient Boosting

The accuracy score overall for the Gradient Boosting model was 61%. Similar to the Random Forest model, the model struggled to predict the minority class (negative return months), scoring just slightly better with a 16% recall score for the minority class. However this came at the expense of significantly poorer score for predicting the majority class (positive return months) with a recall score of 77% vs 95% for the Random Forest model.

Confusion matrix:

	Predicted Negative	Predicted Positive
Negative	4	21
Positive	17	56

Trend Model - Algo Trading in Action

The strategy is simply to buy (or hold) the SPY etf if the model indicates next month's return will be positive and to short the etf if the model predicts that next month's return will be negative. Below is the last 12 predictions for each of the models, as well as the actual percentage return over those periods.

Date	Monthly Return	Random Forest Predicted Value	Percentage Return	SMOTE Predicted Value	Gradient Boost Predicted Value
2019-09-30	1.0	1.0	0.019458	-1.0	1.0
2019-10-31	1.0	1.0	0.022105	1.0	1.0
2019-11-30	1.0	1.0	0.036198	1.0	-1.0
2019-12-31	1.0	-1.0	0.029055	-1.0	1.0
2020-01-31	0.0	1.0	-0.000404	-1.0	-1.0
2020-02-29	0.0	-1.0	-0.079166	-1.0	1.0
2020-03-31	0.0	-1.0	-0.124871	-1.0	-1.0
2020-04-30	1.0	1.0	0.126984	-1.0	1.0
2020-05-31	1.0	1.0	0.047645	1.0	-1.0
2020-06-30	1.0	1.0	0.017734	-1.0	-1.0
2020-07-31	1.0	-1.0	0.058892	-1.0	-1.0
2020-08-31	1.0	1.0	0.069797	-1.0	-1.0

Trend Model - Algo Trading in Action

Trend Model Takeaways:

- RF Algo strategy outperformed the benchmark!
- Beware of small sample size!
- Divergence of economic fundamentals and equity returns in 2020 created interesting outcomes for our models.
- How closely related are S&P500 returns and economic fundamentals?
- Would our features work better on other assets (i.e Russell 2000)?

Data Cleanup - Trade Model

Download stock data from from Yahoo finance

Ticker “SPY”

Period : 1993 - Present

No major cleaning required

Saved to CSV file

Extract required column for “Closing Price”

Calculated “Daily Return”- using pct change

Constructing the Model

<u>Features</u>

Features are based on technical indicators

(1) **Exponential Moving Average of Closing Price**

(2) **Exponential Moving Average of Daily Return Volatility**

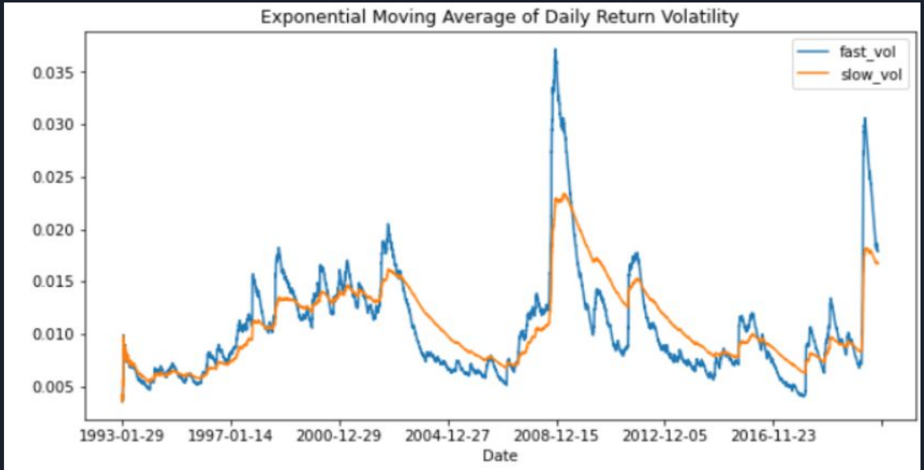
(3) **Bollinger Band**

Both EMA were constructed using a short window of 50 and long window 200

Bollinger window was 20

Visualizations


Exponential Moving Average of Closing Prices
Close
fast_close
slow_close
350
300
250
200
150
100
50
1993-01-29
1997-01-14
2000-12-29
2004-12-27
2008-12-15
2012-12-05
2016-11-23
Date


Exponential Moving Average of Daily Return Volatility
fast_vol
slow_vol
0.035
0.030
0.025
0.020
0.015
0.010
0.005
1993-01-29
1997-01-14
2000-12-29
2004-12-27
2008-12-15
2012-12-05
2016-11-23
Date


Bollinger Bands for SP500 closing prices
Close
bollinger_mid_band
bollinger_upper_band
bollinger_lower_band
350
300
250
200
150
100
50
1993-02-01
1997-01-15
2001-01-02
2004-12-28
2008-12-16
2012-12-06
2016-11-25
Date

Final Features used to create the models

These are used to calculate the signals, which are then used as the features shown below.

(1) **Exponential Moving Average of Closing Price → Crossover_signal**

(2) **Exponential Moving Average of Daily Return Volatility → Vol_trend_signal**

(3) **Bollinger Band → Bollinger_signal**

Date	crossover_signal	vol_trend_signal	bollinger_signal
2012-06-06	-1.0	-1.0	1.0
2012-06-07	-1.0	-1.0	0.0
2012-06-08	-1.0	-1.0	0.0
2012-06-11	-1.0	1.0	0.0
2012-06-12	-1.0	-1.0	0.0
2012-06-13	-1.0	-1.0	0.0
2012-06-14	-1.0	-1.0	0.0
2012-06-15	-1.0	1.0	-1.0
2012-06-18	1.0	1.0	-1.0
2012-06-19	1.0	1.0	-1.0

Dependent Variable

What are we predicting?

- Whether the market will have a POSITIVE return

Positive Return	Negative Return
1	0

Use the values in the “Daily return” column to construct the Y variable

Daily Return >0 → Y =1

Daily Return < 0 → Y = 0

Training

Split Data

70 % Training, 30 % Testing (train on 19 yrs of data, Test on 8 years of data)

Models used

Random Forrest Classifier - RFC: (why- good for classification, good at avoiding overfitting)

Gradient Boosting Classifier- GBC : (why ? GBC is boosting, so hence contrast with bagging in RFC)

Support Vector Machine -SVM : (Why ? good for classification)

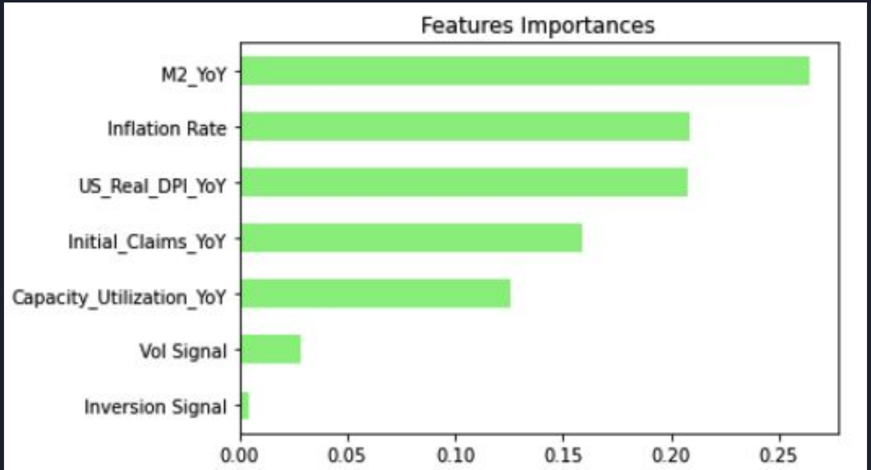
Model Evaluation - Trend Model Techniques Used

Random Forest Model - Confusion Matrix

	Predicted Negative	Predicted Positive
Negative	2	23
Positive	4	69

Gradient Boosting Classifier - Confusion Matrix

	Predicted Negative	Predicted Positive
Negative	4	21
Positive	17	56



Model Evaluation - Trend Model Techniques Used

The Matthews Correlation Coefficient

Random Forest MCC Score
= 0.0458

Gradient Boosting MCC Score
= -0.0774

Random Forest Imbalanced Classification Report

	pre	rec	spe	f1	geo	iba	sup
0.0	0.33	0.08	0.95	0.13	0.27	0.07	25
1.0	0.75	0.95	0.08	0.84	0.27	0.08	73
avg / total	0.64	0.72	0.30	0.66	0.27	0.08	98

Gradient Boosting Imbalanced Classification Report

	pre	rec	spe	f1	geo	iba	sup
0.0	0.19	0.16	0.77	0.17	0.35	0.12	25
1.0	0.73	0.77	0.16	0.75	0.35	0.13	73
avg / total	0.59	0.61	0.31	0.60	0.35	0.13	98

Model Evaluation - Trade Model Techniques Used

Random Forest Classifier & Gradient Boosting Classifier

Accuracy Score : 0.5437589670014347

Confusion Matrix

	Predicted 0	Predicted 1
Actual 0	124	805
Actual 1	149	1013

Classification Report

	precision	recall	f1-score	support
0.0	0.45	0.13	0.21	929
1.0	0.56	0.87	0.68	1162
accuracy			0.54	2091
macro avg	0.51	0.50	0.44	2091
weighted avg	0.51	0.54	0.47	2091

SVM Model

Accuracy Score : 0.555714968914395

Confusion Matrix

	Predicted 0	Predicted 1
Actual 0	0	929
Actual 1	0	1162

Classification Report

	precision	recall	f1-score	support
0.0	0.00	0.00	0.00	929
1.0	0.56	1.00	0.71	1162
accuracy			0.56	2091
macro avg	0.28	0.50	0.36	2091
weighted avg	0.31	0.56	0.40	2091

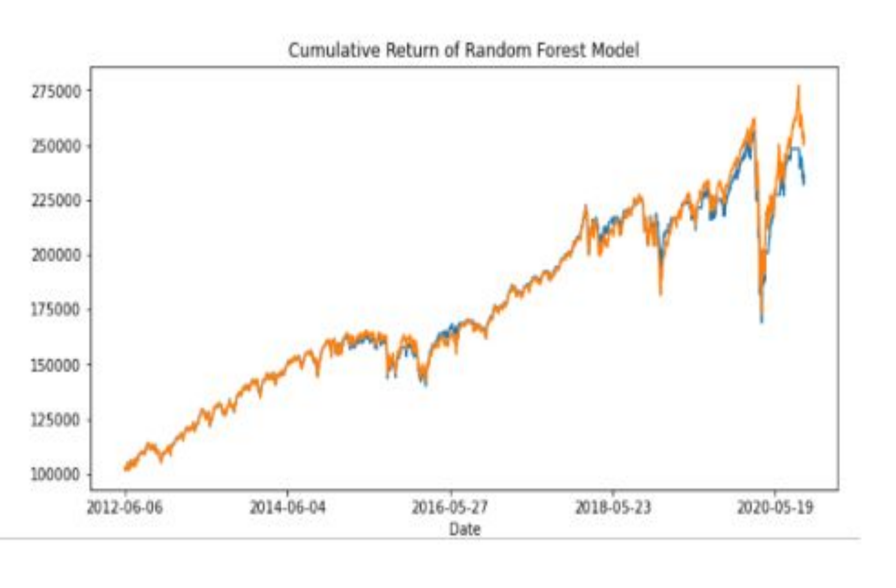
The Matthews Correlation Coefficient

Random Forest MCC Score = 0.07742

Gradient Boosting MCC Score = 0.00

Main Finding

If we compare the best performing "Trade" model with the best performing "Trend" model, we can conclude that the ***model based on economic indicators***, is better at predicting the direction of the market.



Best performing trade model

Best performing trend model

Limitations

- difficulties:
  - Same evaluation results in RFC & GBC models (trade)
  - Experimented with a new model, but the evaluation metrics were different so we were unable to compare
  - Limited data with respect to the economic factors (trend)

- moving forward, would
  - explore more features for economic models
  - try more models for the technical model
  - Consider survivorship bias

Questions?