Model vs Modalities
CSE 512 - Project Proposal

Nilesh Gajwani (ngajwani@cs.stonybrook.edu)
Nikunj Mittal (nimittal@cs.stonybrook.edu)
Nishit Jain (nisjain@cs.stonybrook.edu)
Omkar Kanade (okanade@cs.stonybrook.edu)

I. Motivation

Multi-dimensional time-series data, a.k.a. panel data, is collected by observing particular variables over a period of time at a regular frequency. This data can help experts establish trends, make correlations and forecast future values of itself, which is a crucial problem in the field of statistics and machine learning with many real-life applications like Stock Market Analysis, Workload Projections, and Predictive Maintenance among others. Owing to the high impact of the result of such analyses, there has been a growing body of research applying everything from linear, non-linear and deep learning models to this task.

Modality of data is dependent on the way data has been collected, which in turn influences the statistical and behavioral properties of the data. For panel data, one factor that affects modality is the time elapsed between the collection of contiguous observations. Hence, panel data can be divided into short-term and long-term data.

We believe that certain models capture information for certain data modalities better than other models, for which we want to explore the performance of different models with respect to short-term financial dataset (intraday stock prices) and long-term financial dataset (GDP). Both these data sources are considered to be real and continuously evolving complex dynamic systems and based on our research (Section II), we have chosen Vector AutoRegression (captures linear dynamics), Dynamic Mode Decomposition (captures non-linear dynamics through linear approximation) and Long Short-Term Memory (captures non-linear dynamics) to compare forecasting performance.

II. Literature Review

A. Vector AutoRegression

Vector AutoRegression (VAR) is a multivariate forecasting algorithm that is used when two or more time series influence each other. VAR models the interdependence between multiple time series by treating all variables symmetrically and including for each variable an equation explaining its evolution based on its own lags (auto correlation) and the lags of all the other variables in the model. As an example suppose that we have three different time series variables, denoted by $x_{t,1}, x_{t,2}$, and $x_{t,3}$. Taking all three variables as endogenous to our system, we get the following VAR equations:

$$\begin{aligned}
x_{t,1} &= \alpha_1 + \phi_{11} x_{t-1,1} + \phi_{12} x_{t-1,2} + \phi_{13} x_{t-1,3} + w_{t,1} \\
x_{t,2} &= \alpha_2 + \phi_{21} x_{t-1,1} + \phi_{22} x_{t-1,2} + \phi_{23} x_{t-1,3} + w_{t,2} \\
x_{t,3} &= \alpha_3 + \phi_{31} x_{t-1,1} + \phi_{32} x_{t-1,2} + \phi_{33} x_{t-1,3} + w_{t,3}
\end{aligned}$$

VAR allows us to model contemporaneous shocks on one of our endogenous variable and its effect on other endogenous variables. This is a key to accurate forecast of long-term data such as GDP where we have explicit relationships between variables based on economic theory. We also have Benjamin and Lundgren(2) who have successfully used VAR model as baseline model for comparison with other short-term trading models where we might not have proven theoretical relationships between endogenous variables.

B. Dynamic Mode Decomposition (DMD)

DMD is an equation free, data-driven method to approximate the non-linear dynamics of a complex system which can be used to predict the future state of the system. (3)(5)

According to Mann and Kutz (6), DMD can be interpreted as a combination of spatial dimensionality-reduction technique (like proper orthogonal decomposition) and Fourier transforms in time. Hence, it decomposes the data matrices into low-rank spacio-temporal structures.

Let X be a data matrix with each column x_i^T, $i = 1, 2, ..., m$, representing a snapshot of data at ith time-step and X' be X shifted by Δt such that $x_i^{\prime T} = x_{i+\Delta t}^T$, $i = 1, 2, ..., m$. We want to find a linear operator (known as Koopman operator) A such that $X' = AX$. However, since we are working with a high dimensional data, finding the A matrix becomes computationally infeasible. To overcome this, DMD approximates the dominant eigenvalues and eigenvectors of A matrix without actually computing A matrix.

$$\text{SVD with rank-r truncation} : X = U_r \Sigma_r V_r^* \tag{1}$$

To find A, we need to multiply X' with pseudo-inverse of X denoted as X^+. However, in place of A, we will compute rank-r truncated $\tilde{A}$

$$A = X' X^+ = X' V_r \Sigma_r^{-1} U_r^* \tag{2}$$

$$\tilde{A} = U_r^* A U_r = U_r^* X' V_r \Sigma_r^{-1} \tag{3}$$

We can then find the eigenvalues and eigenvectors of $\tilde{A}$.

$$\tilde{A}W = W\Lambda \tag{4}$$

The eigenvectors of $\tilde{A}$ can be reshaped into eigen flow-fields (called dynamic modes) which are coherent spatial structures, while their eigenvalues correspond to coherent dynamics of the evolution of these modes in time (in the form of sin, cosine and other basis functions). We then extend the time dynamics and recombine them with the modes to predict the future state of the system.

The prediction of state of the system at any time t is done in the form of a reconstruction through linear combination of DMD modes Φ, eigenvalues Λ and an initial amplitude b.

$$\Phi = X'V_r\Sigma_r^{-1}W \tag{5}$$

$$\Omega = \frac{log(\Lambda)}{\Delta t} \tag{6}$$

$$X(t) = \Phi e^{\Omega t} b \tag{7}$$

Theoretically, DMD gives accurate forecasts only for shorter forecasting horizons because $X(t)$ is computed using eigenvalues which can exponentially blow up as we go further forward in time. However, since DMD is formed from least square fit of low-rank matrices, we can efficiently fit new models as the dynamics of our system change.

C. Long Short Term Memory (LSTM)

LSTMs are generally used for sequential prediction tasks due to their ability to store past information. (2) uses a two-layered LSTM with a dropout of 0.1 to train over intraday time series data at one minute intervals for all stocks in the S&P 500. The model is evaluated on simple 0-1 accuracy and Sharpe Ratio, which is a metric of return versus risk. The LSTM was found to have an accuracy of 51.6% which was the highest amongst the models discussed in this paper.
(9) does GDP forecasting using LSTM. It involves a Co-integration test and a causality relationship test, along with a threshold model between CPI growth rate and GDP growth rate in Threshold model, where the threshold is referred as an indicator of different CPI fluctuation states. The conclusion was that the LSTM had the best accuracy and consistency by detecting useful historical information within a suitable time window.

D. VAR vs DMD vs LSTM

VAR estimates exact coefficient matrix that can model the linear dynamics of a complex system using autoregressive equations of order p. Moreover, it models contemporaneous relationships between endogenous variables in our model.

DMD is closely related to $VAR(1)$ model but approximates the coefficient matrix using spacio-temporal decomposition of low-rank data matrix. This decomposition allows it to estimate the non-linear dynamics of a complex system through a linear relationship.

LSTM models the non-linear dynamics of a complex system.

III. Data

A. Short-Term Data

We will use Exchange Traded Funds (ETFs) for short-term analysis. ETFs are collections of dozens, sometimes hundreds, of stocks. An ETF's return is the weighted average of all its holdings, which makes them more diversified and less volatile compared to individual stocks. We will shortlist five ETFs from within a sector (sector as defined here (4)) according to the following criteria:

- Three-month average volume is greater than $1M
- Average correlation between any two ETFs is greater than 0.9

The five ETFs that we have selected are SPY, QQQ, VXX, DIA and EEM (7).

B. Long-Term Data

We will use quarterly Global GDP data (8) across the period 1960 Q1 to 2012 Q4 for our long-term analysis. Each data point is the GDP over a three month duration and is dependent on various factors such as Consumer Price Index (CPI), unemployment rate, import/export, etc.

IV. Methodology

A. Data Preprocessing

We will first apply Co-integration test on different variables in the ETFs and GDP datasets. Co-integration helps to establish the presence of a long run, statistically significant relationship between two or more time series, which helps validate modelling hypothesis. Co-integration test looks for a stationary linear combination of non-stationary random variables. We will use Johanesen Test (12) to test co-integration. Time series data need to be made stationary (mean and variance does not change over time) for which we propose using Augmented Dickey-Fuller test of stationarity and using differencing to make the time series stationary.

B. Model Training and Selection

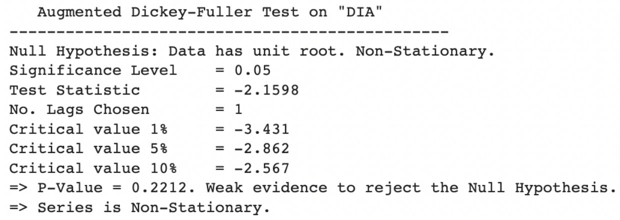
We split our dataset into the standard partition of 70/15/15 for training/validation/testing. We train all proposed models on each dataset to compare performance. For time series datasets, model performance depends on on the choice of lag-length, p. Lag-length decides how far back in time we look for values that can help forecast for future time-steps. We use validation datasets to tune the value of P based on Akaike Information Criterion values, which quantifies the tradeoff between the goodness of fit of the model and the complexity of the model.

Finally the models are compared on basis of Root Mean Squared Percentage Error, so we can compare across different units of data.

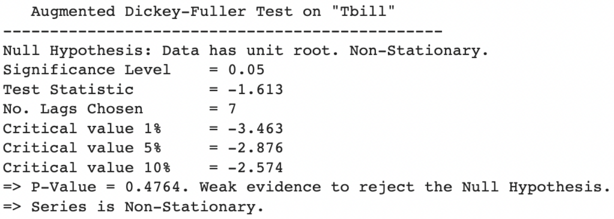
V. Results

A. Preprocessing Tests

VAR requires endogenous variables to be stationary (10) and co-integrated (11). We performed Augmented Dickey-Fuller test of stationarity to find variables in both GDP and ETF data to be non-stationary:

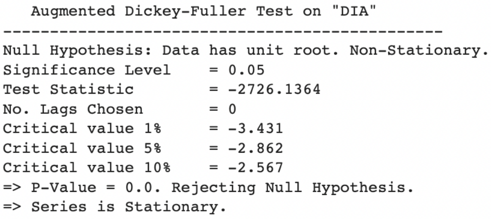


```
    Augmented Dickey-Fuller Test on "DIA"
-----------------------------------------------
Null Hypothesis: Data has unit root. Non-Stationary.
Significance Level    = 0.05
Test Statistic        = -2.1598
No. Lags Chosen       = 1
Critical value 1%     = -3.431
Critical value 5%     = -2.862
Critical value 10%    = -2.567
=> P-Value = 0.2212. Weak evidence to reject the Null Hypothesis.
=> Series is Non-Stationary.
```

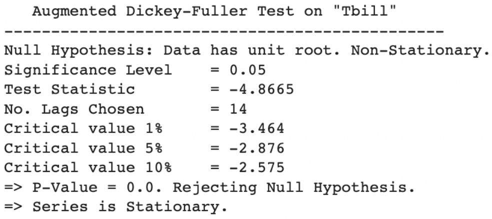


```
    Augmented Dickey-Fuller Test on "Tbill"
-----------------------------------------------
Null Hypothesis: Data has unit root. Non-Stationary.
Significance Level    = 0.05
Test Statistic        = -1.613
No. Lags Chosen       = 7
Critical value 1%     = -3.463
Critical value 5%     = -2.876
Critical value 10%    = -2.574
=> P-Value = 0.4764. Weak evidence to reject the Null Hypothesis.
=> Series is Non-Stationary.
```

ETF data was integrated I(1) and GDP data was integrated with I(2). We made the datasets stationary by differencing them once and twice respectively.

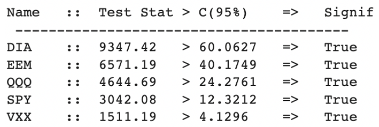


```
    Augmented Dickey-Fuller Test on "DIA"
-----------------------------------------------
Null Hypothesis: Data has unit root. Non-Stationary.
Significance Level    = 0.05
Test Statistic        = -2726.1364
No. Lags Chosen       = 0
Critical value 1%     = -3.431
Critical value 5%     = -2.862
Critical value 10%    = -2.567
=> P-Value = 0.0. Rejecting Null Hypothesis.
=> Series is Stationary.
```

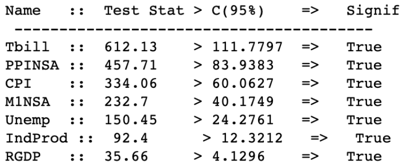


```
    Augmented Dickey-Fuller Test on "Tbill"
-----------------------------------------------
Null Hypothesis: Data has unit root. Non-Stationary.
Significance Level    = 0.05
Test Statistic        = -4.8665
No. Lags Chosen       = 14
Critical value 1%     = -3.464
Critical value 5%     = -2.876
Critical value 10%    = -2.575
=> P-Value = 0.0. Rejecting Null Hypothesis.
=> Series is Stationary.
```

For Co-Integration, we ran the Johansen Co-Integration test (12). This is a hypothesis test which computes test statistic and compares at 95% confidence to identify significance. True significance indicates presence of co-integration.



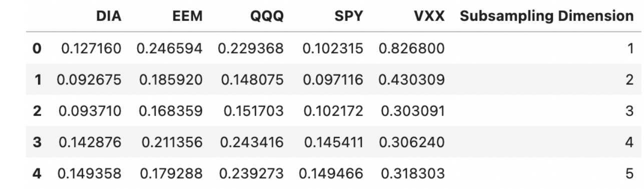
```
Name   ::  Test Stat > C(95%)    =>   Signif
 ---------------------------------------
DIA    ::  9347.42   > 60.0627   =>   True
EEM    ::  6571.19   > 40.1749   =>   True
QQQ    ::  4644.69   > 24.2761   =>   True
SPY    ::  3042.08   > 12.3212   =>   True
VXX    ::  1511.19   > 4.1296    =>   True
```



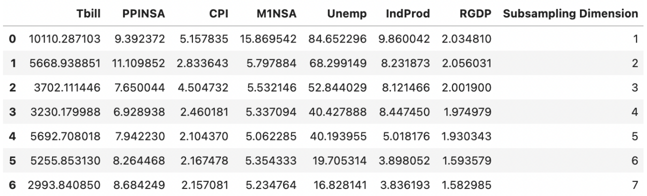
```
Name   ::  Test Stat > C(95%)    =>   Signif
 ----------------------------------------
Tbill  ::  612.13    > 111.7797  =>   True
PPINSA ::  457.71    > 83.9383   =>   True
CPI    ::  334.06    > 60.0627   =>   True
M1NSA  ::  232.7     > 40.1749   =>   True
Unemp  ::  150.45    > 24.2761   =>   True
IndProd ::  92.4      > 12.3212   =>   True
RGDP   ::  35.66     > 4.1296    =>   True
```

B. Model Selection

DMD has only one parameter, r, which is the sub-sampling dimension for SVD. This is chosen based on the root mean squared percentage error.



	DIA	EEM	QQQ	SPY	VXX	Subsampling Dimension
0	0.127160	0.246594	0.229368	0.102315	0.826800	1
1	0.092675	0.185920	0.148075	0.097116	0.430309	2
2	0.093710	0.168359	0.151703	0.102172	0.303091	3
3	0.142876	0.211356	0.243416	0.145411	0.306240	4
4	0.149358	0.179288	0.239273	0.149466	0.318303	5

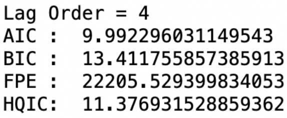


	Tbill	PPINSA	CPI	M1NSA	Unemp	IndProd	RGDP	Subsampling Dimension
0	10110.287103	9.392372	5.157835	15.869542	84.652296	9.860042	2.034810	1
1	5668.938851	11.109852	2.833643	5.797884	68.299149	8.231873	2.056031	2
2	3702.111446	7.650044	4.504732	5.532146	52.844029	8.121466	2.001900	3
3	3230.179988	6.928938	2.460181	5.337094	40.427888	8.447450	1.974979	4
4	5692.708018	7.942230	2.104370	5.062285	40.193955	5.018176	1.930343	5
5	5255.853130	8.264468	2.167478	5.354333	19.705314	3.898052	1.593579	6
6	2993.840850	8.684249	2.157081	5.234764	16.828141	3.836193	1.582985	7

We see that for both the datasets, we have minimum root mean squared percentage error for $r = 2$.

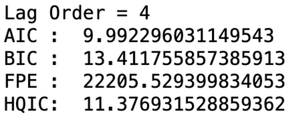
For VAR and LSTM, we use AIC (13) to choose lag order $p = 4$ that captures maximum information with minimum model parameters.

GDP:



```
Lag Order = 4
AIC :  9.992296031149543
BIC :  13.411755857385913
FPE :  22205.529399834053
HQIC:  11.376931528859362
```

ETF:

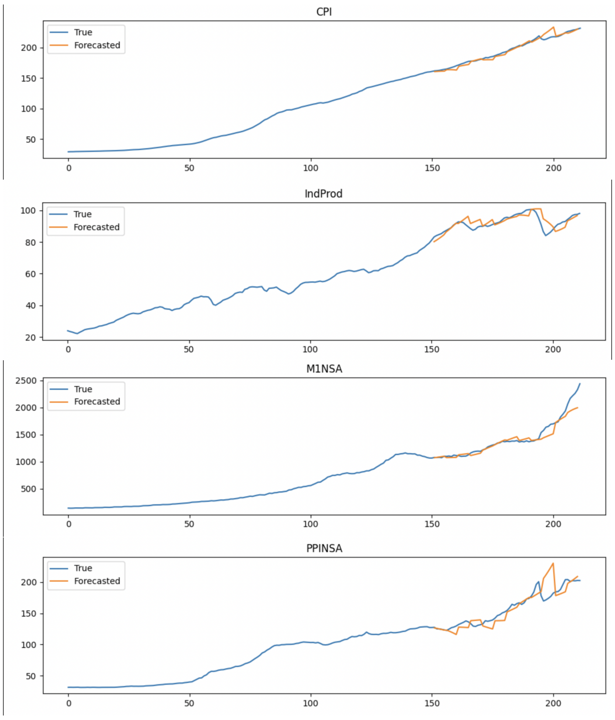


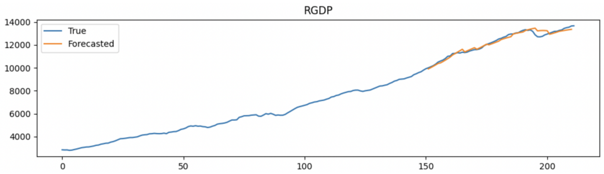
```
Lag Order = 4
AIC :  9.992296031149543
BIC :  13.411755857385913
FPE :  22205.529399834053
HQIC:  11.376931528859362
```

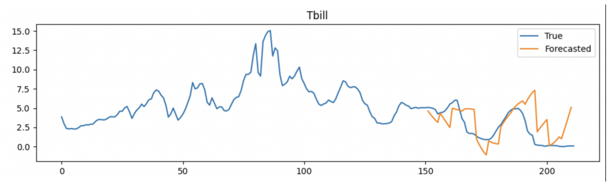
C. Forecasts

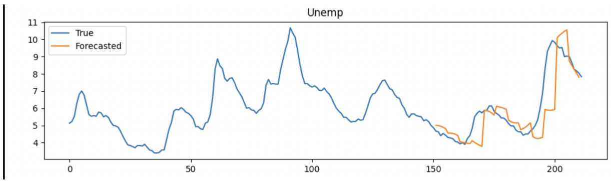
1. DMD:

GDP:

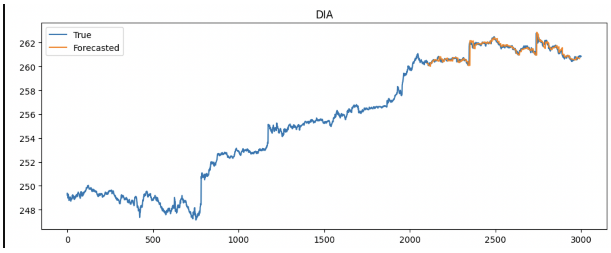


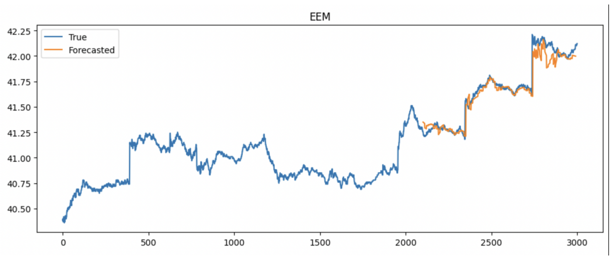


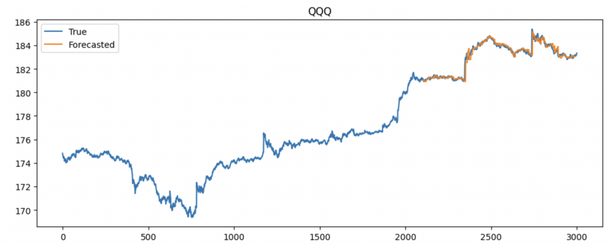


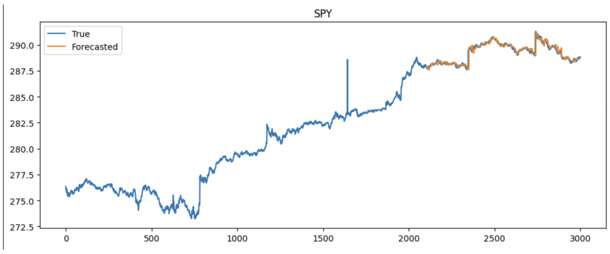


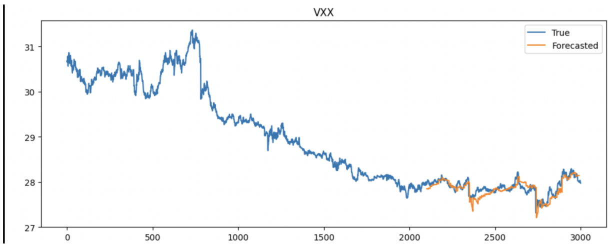
ETF:





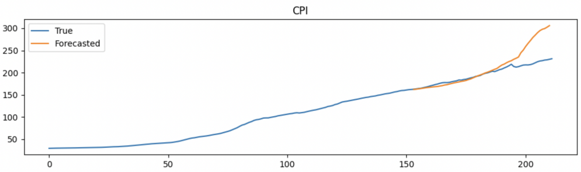


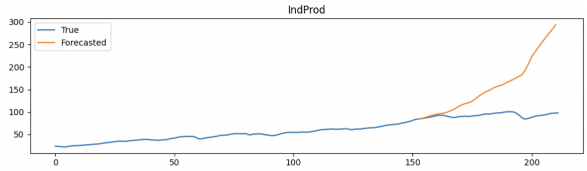


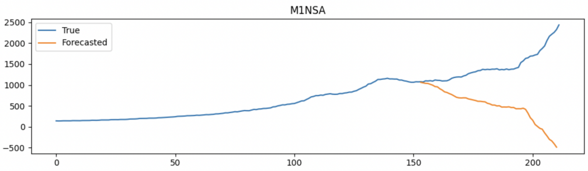


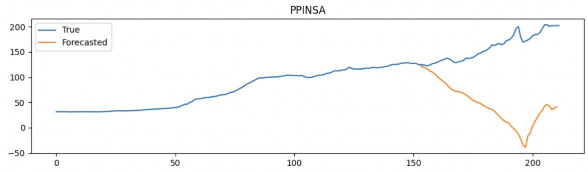
2. VAR

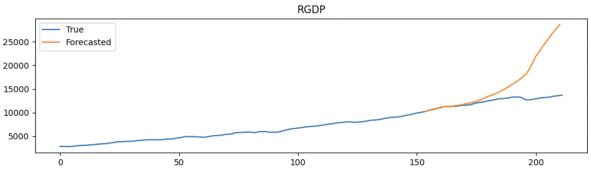
GDP:

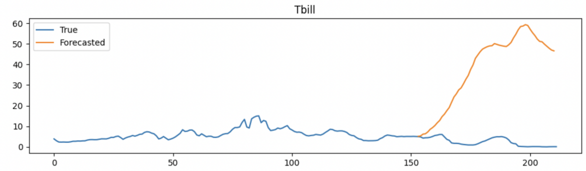


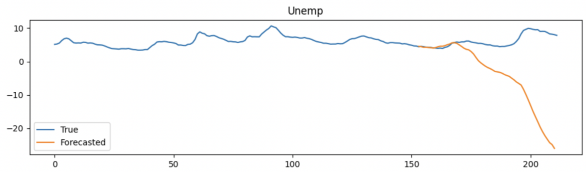




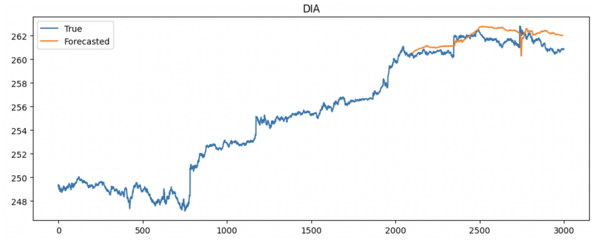


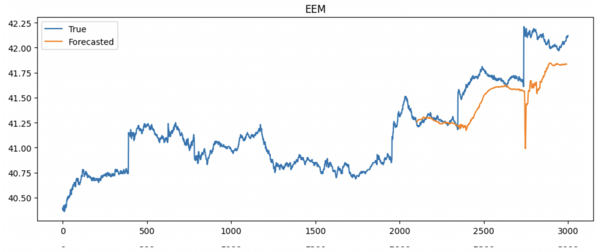


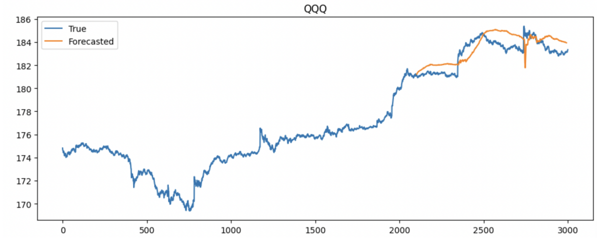


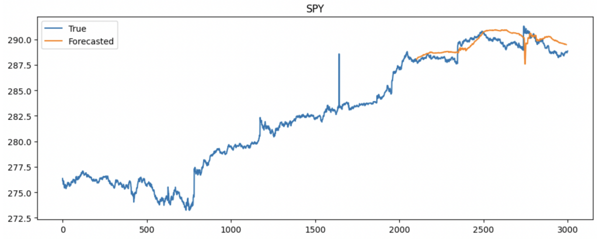


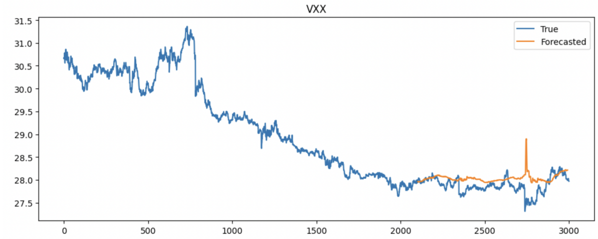
ETF:





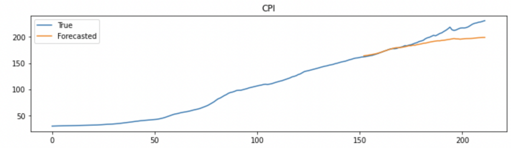


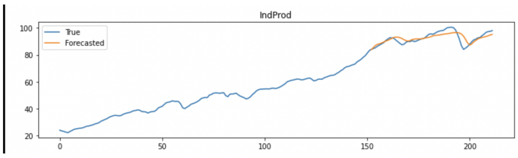


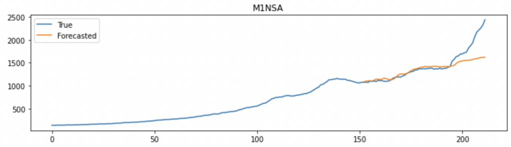


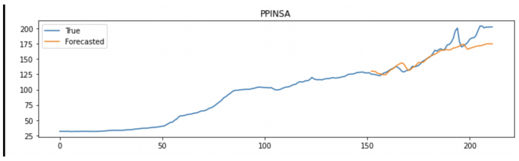
2. LSTM

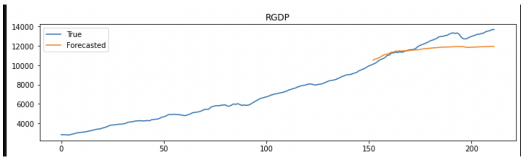
GDP:

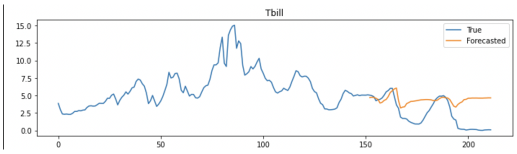


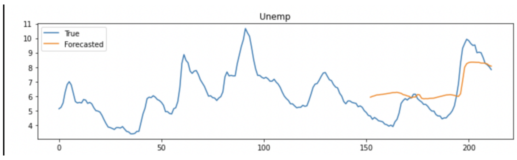




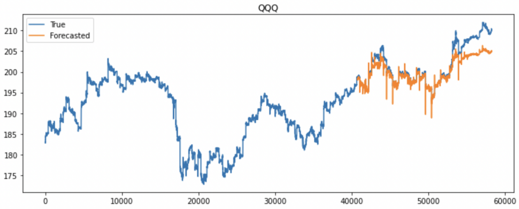


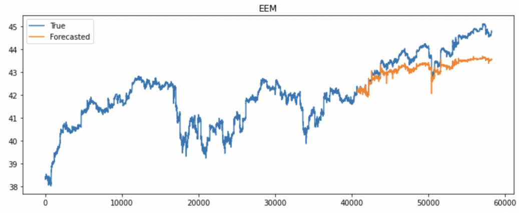


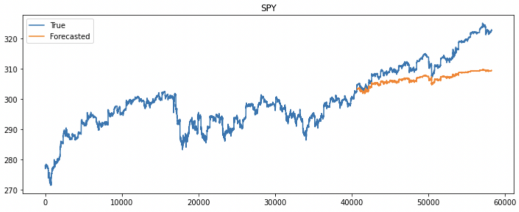


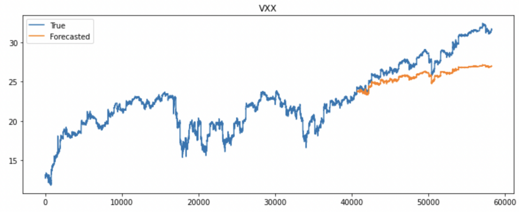


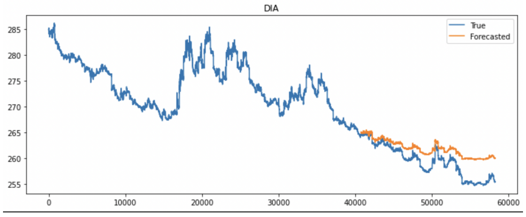
ETF:











D. Model Evaluation

In the tables shown, we have compared the Root Mean Squared Percentage Error across different variables for our models on long and short-term time-series datasets.

TABEL I
RMSPE VALUES FOR INDIVIDUAL FEATURES OF GDP

Feature	VAR	DMD	LSTM
Tbill	76809.27	5668.94	7054.34
PPINSA	73.44	11.11	7.02
CPI	12.95	2.83	6.72
M1NSA	66.05	5.80	10.59
Unemp	181.98	68.30	26.81
IndProd	93.16	8.23	3.25
RGDP	43.68	2.05	7.14

TABEL II
RMSPE VALUES FOR INDIVIDUAL FEATURES OF ETF

Feature	VAR	DMD	LSTM
DIA	0.32	0.09	3.05
EEM	0.66	0.19	4.55
QQQ	0.52	0.15	3.58
SPY	0.30	0.09	2.87
VXX	0.89	0.43	15.45

VI. CONCLUSION

Based on our Literature Review (Section II), we had expected VAR to work well on GDP data (long-term) as VAR captures contemporaneous shocks on endogenous variables and their effects instantaneously, which is expected of GDP data. DMD, on the other hand, captures these effects with a delay, which is true for ETF data (short-term). LSTM, being a highly parameterized model, was expected to behave well on both modalities due to its ability to model non-linear dynamics of systems.

We observe that, though VAR captures shocks effectively, it does not work well on longer forecasts. It performs an ex-ante dynamic forecast (uses the forecast value of lagged dependent variables in place of the actual value of the lagged dependent variables) resulting in model errors accumulating over time. DMD is able to model both short and long-term data effectively, though it is modeling shocks in GDP with a delay. LSTM requires more data to truly learn patterns in the data and is ineffective on GDP as GDP is captured only once in a quarter and has only 200 data points. For ETF, we have data for every 1 minute over a year, so we have trained LSTM on all datapoints. LSTM is able to model the underlying process, but does capture the trend in the series.

VII. REFERENCES

(1) Mansoor Momeni, Mahmoud Dehghan Nayeri, Ali Faal Ghayoumi, Hoda Ghorbani. Robust Regression and its Application in Financial Data Analysis, 2010

(2) David Benjamin Lim, Justin Lundgren. Algorithmic Trading using LSTM-Models for Intraday Stock Predictions, 2019

(3) D. P. Kuttichira, E. A. Gopalakrishnan, V. K. Menon and K. P. Soman, ”Stock price prediction using dynamic mode decomposition,” 2017 International Conference on Advances in Computing, Communications and Informatics (ICACCI), 2017, pp. 55-60, doi: 10.1109/ICACCI.2017.8125816

(4) Finviz, Financial Visualization Tool https://finviz.com/map.ashx?t=etf

(5) J. Nathan Kutz, Steven L. Brunton, Bingni W. Brunton, and Joshua L. Proctor, ”Dynamic Mode Decomposition”, https://doi.org/10.1137/1.9781611974508

(6) Jordan Mann, J. Nathan Kutz, ”Dynamic Mode Decomposition for Financial Trading Strategies”, https://arxiv.org/abs/1508.04487

(7) Intra-day Trading Data. https://firstratedata.com/free-intraday-data

(8) Global GDP Data. http://time-series.net/yahoo_site_admin/assets/docs/ quarterly.7775706.xls

(9) China’s GDP forecasting using Long Short Term Memory Recurrent Neural Network and Hidden Markov Model. https://journals.plos.org/plosone/ article?id=10.1371/journal.pone.0269529

(10) Stock, J. H., and M. W. Watson. 2015. Introduction to Econometrics, Third Update, Global Edition. Pearson Education Limited.

(11) Stock, James, H., and Mark W. Watson. 2001. ”Vector Autoregressions.” Journal of Economic Perspectives, 15 (4): 101-115.

(12) Lütkepohl, Helmut, et al. “Maximum Eigenvalue versus Trace Tests for the Cointegrating Rank of a VAR Process.” The Econometrics Journal, vol. 4, no. 2, 2001, pp. 287–310. JSTOR, http://www.jstor.org/stable/23114982. Accessed 6 Dec. 2022.

(13) Akaike, H. (1974), ”A new look at the statistical model identification”, IEEE Transactions on Automatic Control, 19 (6): 716–723, Bibcode:1974ITAC...19..716A, doi:10.1109/TAC.1974.1100705, MR 0423716.