

Modeling and Predicting Financial Market Movements Using Probabilistic Graphical Models: Incorporating Economic Indicators and External Factors

Vivekanand R (Draft/Sample Report Only)

Submitted on November 3, 2024
for the Unknown Course
in the WS 2024/25

Abstract Numerous large-scale projects and prominent financial institutions focus on predicting financial markets using advanced Artificial Intelligence (AI) and Machine Learning (ML) approach, leading to publicly available stock indices data sets. In this report, we will model, integrate the financial market movements using probabilistic graphical model approach and will incorporate economic indicators using external factors.

This probabilistic approach incorporating a wide range of external factors that influence market behavior, such as macroeconomic trends, economic indicators, geopolitical events, technical indicators[7], regional and sector-specific developments.

Additionally, we will deploy this model in a dynamic testing environment designed to simulate real-world market conditions[8]. This will enable us to generate live, dynamic probabilistic tables[11] that provide a continuously updated forecast of market movements. Here Hidden Markov Models (HMM's)[11] and Bayesian switching models will serve as a decision-support tool[9], offering a probabilistic view of potential future scenarios and aiding in the formulation of more informed investment strategies[11].

Keywords Probabilistic Models, Kalman Filter, Conditional Probabilities, Financial Market

1 Introduction

In recent years, AI has made a significant progress in many fields areas including financial stock prediction. Predicting market movement is still challenging due to various reasons such as high dimensionality, data noise, non stationary. The stock market is influenced by countless variables, including economic indicators, company performance, geopolitical events, investor sentiment, and even natural disasters. This high dimensional means there are numerous potential predictors, many of which are noisy or have weak predictive power.

Often Stock prices do not follow a steady or consistent, predictable pattern over time. They result in non-stationarity, meaning their statistical properties (mean, variance, etc.) change over time due to various economic cycles, regulatory changes, and market sentiment shifts.

Also, market movements are significantly driven by human emotions and behaviors, such as fear, greed, and panic. These are not easily quantifiable and can lead to irrational market behaviors that are difficult to model using conventional methods. For example, Unpredictable events (e.g., the COVID-19 pandemic, financial crises) can cause significant market disruptions. These events are rare and often outside the scope of historical data, making them challenging to predict using traditional models or even complex AI models.

Probabilistic models offer unique advantages in handling certain aspects of stock market prediction. Stock prices due to non-stationarity, can change regimes (e.g., bull and bear markets). Such Hidden Markov Models (HMMs)[9] and Bayesian switching models will be used to detect regime changes and model time series data with different statistical properties

across different periods. This allows for better adaptation to changing market conditions.

Below model will be explored and trained based on 7 major indices which driven by 955 major companies in five major economies. And, finally such probabilistic models will also be used or trained on individual stocks for exploration purpose.

2 Overview

As noted in the introduction, the process involves below steps:

- 2.1 Data Sources and Collection
- 2.2 Pre-processing Steps
- 2.3 Methodology
- 2.4 Probabilistic Models
- 2.5 Validation and Output

2.1 Data Sources and Collection

Data used in this project were collected mainly from yahoo finance and government open source platforms. Major 7 indices were considered in this study in three different regions such as US, Europe and Asia. We use python API which will pull historical data from yahoo finance (Indices data) and world bank (wbdata for economic indicators) and store it in google cloud for model development.

Some of major indices were considers such as:

- SP 500 (Standard and Poor's 500) – USA: Represents 500 of the largest publicly traded companies in the United States.
- Dow Jones Industrial Average (DJIA) – USA: Includes 30 major, large-cap companies across various industries.
- NASDAQ Composite – USA: Encompasses around 3,000 companies listed on the NASDAQ Stock Market.
- FTSE 100 (Financial Times Stock Exchange 100 Index) – UK: Tracks the 100 largest companies listed on the London

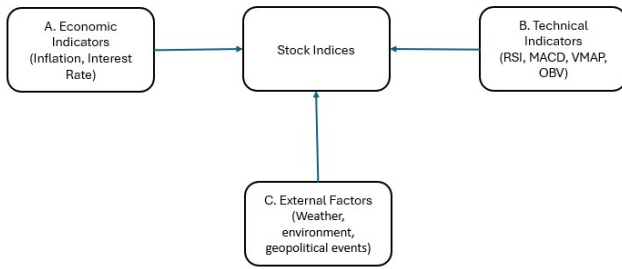


Fig. 1: Probabilistic Model Flow Chart

Table 1: List of Major Stock Indices Considered

Indices	Frequency Estimates	No. of Companies
SP 500	Standard and Poor's 500	500
Dow Jones	Dow Jones Industrial Average	30
NASDAQ	NASDAQ Composite	30
FTSE 100	Financial Times Stock Exchange	100
Nikkei 225	Nikkei 225	225
DAX	Deutscher Aktienindex	40
Sensex	Sensitive Index	30

Stock Exchange.

- Nikkei 225 – Japan: Consists of 225 blue-chip companies listed on the Tokyo Stock Exchange.
- DAX (Deutscher Aktienindex) – Germany: It tracks 40 of the largest and most liquid companies on the Frankfurt Stock Exchange.
- Sensex (Sensitive Index) – India: Sensex tracks 30 well-established and financially sound companies listed on the Bombay Stock Exchange.

2.2 Pre-processing Steps

Out of collected indices data from yahoo finance, data wrangling will be performed along with economic indicators like unemployment rate, GDP growth, inflation rate and central bank interest rate. And, then widely used technical indicators will be calculated based on the closing price.

Below are the list of pre-processing Steps:

First the daily indices data will be pre-processed and below indicators will be calculated based on the close price. Then data wrangling (lookup) will be performed to merge other datasets to convert it into single matrices.

2.2.1 Technical Indicators

- Relative Strength Index (RSI)
- Moving Average convergence and divergence (MACD)
- Volume Weighted average price (VMAP)
- On Balance Volume (OBV)

Date	Open	High	Low	Close	Adj Close	Volume
2024-08-26	5639.660156	5651.620117	5602.339844	5616.839844	5616.839844	2938570000
2024-08-27	5602.890137	5631.180176	5593.479980	5625.799805	5625.799805	2798990000
2024-08-28	5624.509766	5627.029785	5560.950195	5592.180176	5592.180176	3053450000
2024-08-29	5607.299805	5646.950195	5583.709961	5591.959961	5591.959961	3065640000
2024-08-30	5612.740234	5651.370117	5581.790039	5648.399902	5648.399902	4185850000

Fig. 2: Indices Datasets bottom 10 View

country	date	GDP (Current US\$)	GDP Growth (Annual %)	Inflation, consumer prices (annual %)	Unemployment Rate (%)
Germany	2023-01-01	4.46	-0.30	131.89	3.04
India	2023-01-01	3.55	7.58	216.86	4.17
United States	2023-01-01	27.36	2.54	139.74	3.62
Japan	2023-01-01	4.21	1.92	111.36	2.58
United Kingdom	2023-01-01	3.34	0.10	142.74	4.06
United States	2022-01-01	25.74	1.94	134.21	3.65
Germany	2022-01-01	4.08	1.81	124.49	3.14

Fig. 3: GDP Growth, Inflation and Unemployment Rate Table

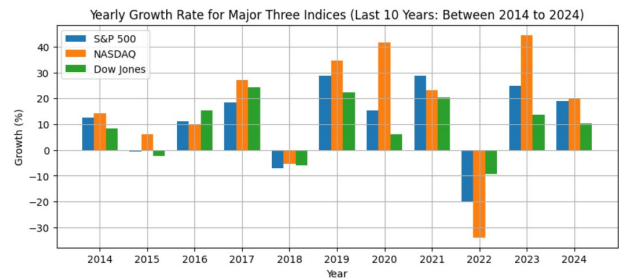


Fig. 4: Top 3 Indices: Last 10 years growth year on year (YoY). Overall Average Growth for Each Index, Between 2014 to 2024: S&P 500: 11.91%, NASDAQ: 16.55%, Dow Jones: 9.34%

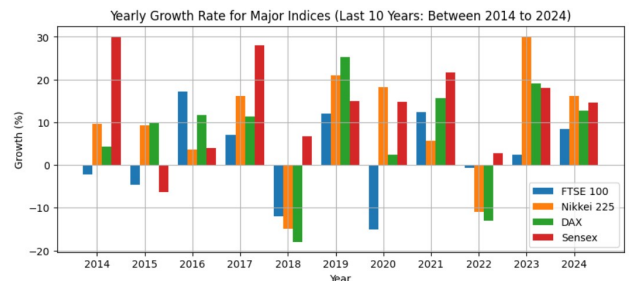


Fig. 5: Next 4 Indices: Overall Average Growth for Each Index (Between 2014 to 2024): FTSE 100: 2.26%, Nikkei 225: 9.46%, DAX: 7.41%, Sensex: 13.56%

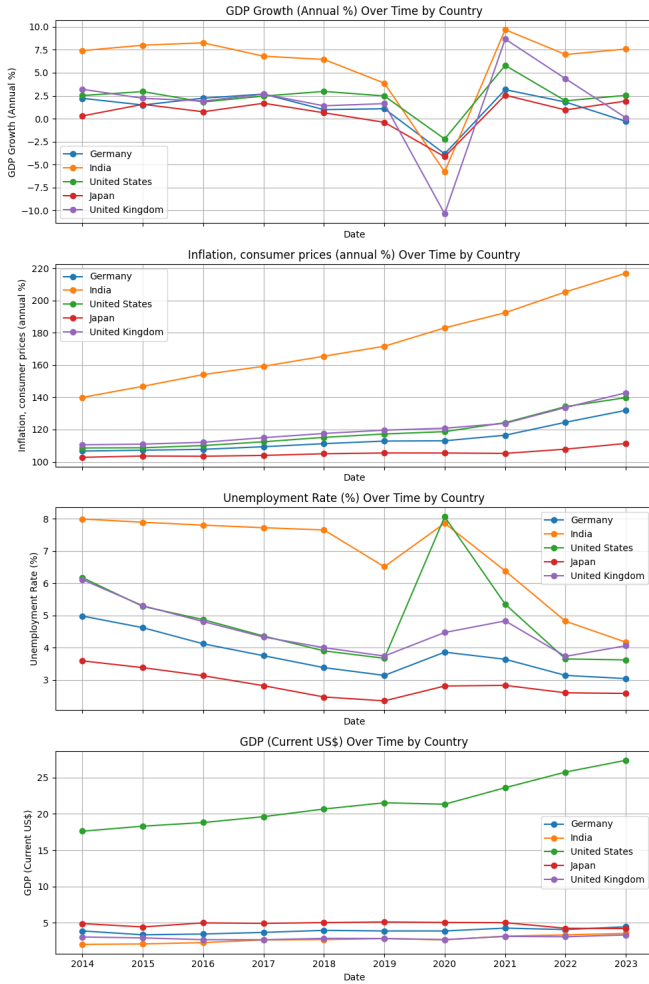


Fig. 6: List of Economic Indicators Charts

RSI: Relative Strength Index

$$RSI = 100 - \left(\frac{100}{1 + \frac{\text{Average Gain}}{\text{Average Loss}}} \right)$$

MACD: Moving Average convergence and divergence

$$MACD = EMA_{12} - EMA_{26}$$

$$\text{Signal Line} = EMA_9(MACD)$$

VWAP: Volume Weighted average price

$$VWAP = \frac{\sum(\text{Price}_i \times \text{Volume}_i)}{\sum \text{Volume}_i}$$

OBV: On Balance Volume

$$OBV = OBV_{\text{prev.}} + \begin{cases} \text{Volume,} & \text{if current close} > \text{prev. close} \\ -\text{Volume,} & \text{if current close} < \text{prev. close} \\ 0, & \text{if the current close} = \text{prev. close} \end{cases}$$

2.2.2 Inflation Rate

In simple terms, Inflation rate measures how much prices for certain goods or services increases over time. Refer below fig.6 for more information.

2.2.3 GDP Growth Rate

This is the important driving factor which tells the rate of economic growth. Refer below fig.6 for more information.

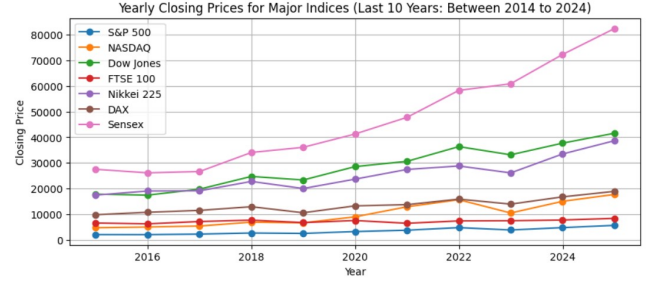


Fig. 7: Major Indices Closing Price - Combined View

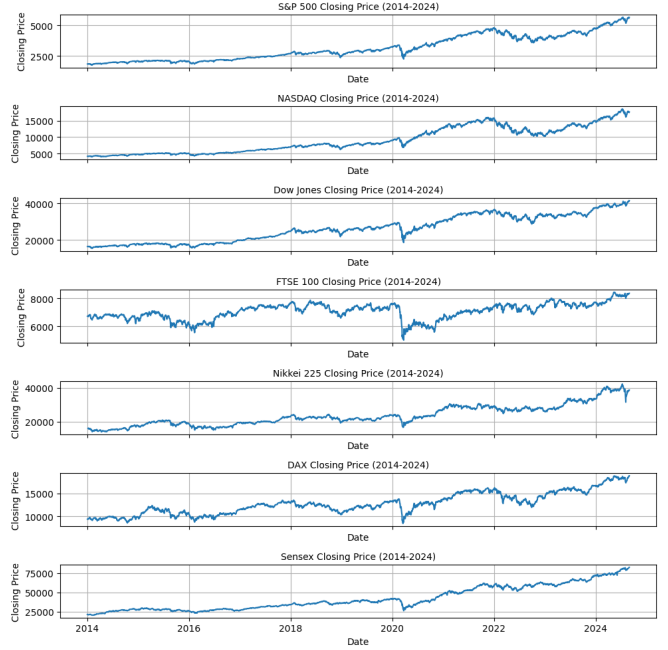


Fig. 8: Major Indices Closing Prices - Splitted View

2.2.4 Central Bank Interest Rate

This plays crucial role in economic decision making. Interest rates set by the Central Banks (like Fed) directly influence the cost of borrowing money, lower interest encourages more business/people participation in countries financial transactions.

2.3 Methodology

Data Wrangling: Used data Wrangling techniques to integrate the multi-omics data into a single data matrix. This integration allows us to create a probabilistic model that not only considers historical data but also adapts to changing economic conditions in real-time, enhancing its predictive accuracy and relevance.

Federal Reserve Interest Rates, 1954-Present has been downloaded from kaggle which helps on identifying economic indicators. Also, live API has been implemented to update latest economic indicators from world bank for all the listed countries.

Model Building: To be added

Feature Importance: To be added

2.4 Probabilistic Model Explainability

In this model, we will explore and will develop Probabilistic model for:

- Each Indices (7 Indices Independently)

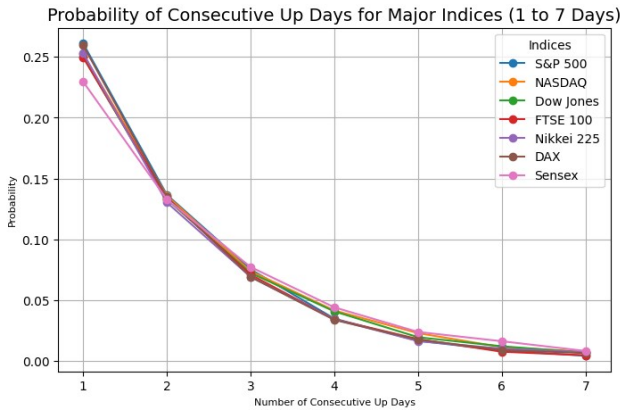


Fig. 9: Probability of stock to move up in 7 consecutive Days

	S&P 500	NASDAQ	Dow Jones	FTSE 100	Nikkei 225	DAX	Sensex
1	0.261047	0.252135	0.249907	0.249723	0.253058	0.259573	0.229627
2	0.136701	0.135958	0.132987	0.133975	0.130402	0.134807	0.132952
3	0.075437	0.073950	0.072464	0.071085	0.069243	0.069270	0.077363
4	0.034944	0.041636	0.040892	0.034444	0.034443	0.033911	0.044224
5	0.016735	0.023057	0.019710	0.018155	0.016462	0.017699	0.024027
6	0.008929	0.011533	0.012277	0.007784	0.010341	0.009222	0.016406
7	0.004466	0.007815	0.007071	0.004820	0.007280	0.006642	0.008397

Fig. 10: Consecutive 7 Days Up Table

- Study the relation one indices with another
- Re-apply it for individual stock for exploration purpose
- Automate the process to generate live conditional probabilistic tables

For example, SP 500 has a 26.10 percent chance of having 1 consecutive up day, and this decreases to 0.44 percent for 7 consecutive days up. The 26.10 percent probability refers to the chance of having exactly 1 consecutive up day – meaning the market went up for one day, but did not go up on the second day.

Similarly, NASDAQ has a 25.21 percent chance of having 1 consecutive up day, decreasing to 0.78 percent for 7 consecutive days up.

1. Each Indices (7 Indices Independently):

To be added, generate Conditional Probability Tables, Apply Bayes' Rule, Kalman Filter, Hidden Markov Models (HMM's) and Bayesian switching models.

These probabilities are crucial for:

Inference: Using Bayesian Network to compute posterior probabilities of growth increasing or decreasing given new observations about the parent states.

Prediction: Understanding how likely it is for Growth to change given different market conditions represented by the parent variables (like RSI, MACD, VWAP, OBV).

2. Study the relation one indices with another:

To be added, generate Conditional Probability Tables, Apply Bayes' Rule, Kalman Filter, Hidden Markov Models (HMM's) and Bayesian switching models.

For example, the probability of stock goes up (i.e. bull market) on any give day is given below.

3. Re-apply the model for individual stock for exploration purpose

Conditional Probability, Bayes' Rule, Hidden Markov Models (HMM's) and Bayesian switching models

$$P(A | B) = \frac{P(B | A) \cdot P(A)}{P(B)}$$

CPD for RSI:

RSI(Oversold)	0.0122699
RSI(Neutral)	0.922875
RSI(Overbought)	0.0648554

CPD for MACD:

MACD(Bearish)	0.502191
MACD(Bullish)	0.497809

CPD for VWAP:

VWAP(Below)	0.503067
VWAP(Above)	0.496933

CPD for OBV:

OBV(Decreasing)	0.000876424
OBV(Increasing)	0.999124

Fig. 11: CPT Table for Technical Indicators for one ticker

The Chain Rule

$$P(A_1, A_2, \dots, A_n) = P(A_1) \cdot P(A_2 | A_1) \cdot P(A_3 | A_1, A_2) \dots P(A_n | A_1, A_2, \dots, A_{n-1})$$

The Law of Total Probability

Kalman Filter: A state observation model described by one or several continuous variable

1. Bayesian Network Structure: The Bayesian Network (BN) which we constructed, models the probability of Google's stock (GOOGL) increasing or decreasing by 5% in the next 15 days, given several factors or "parents". These factors include: S&P 500 Index (SP500) Economic Indicator (e.g., Good or Bad) Technical indicators for GOOGL like RSI (Relative Strength Index), MACD (Moving Average Convergence Divergence), VWAP (Volume Weighted Average Price), and OBV (On-Balance Volume)
2. Conditional Probability Distribution (CPD):

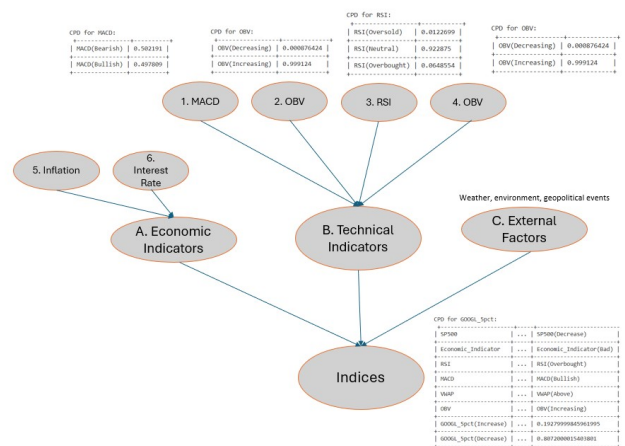


Fig. 12: Final CPT Prediction Table

CPD for GOOGL_5pct:

SP500	...	SP500(Decrease)
Economic_Indicator	...	Economic_Indicator(Bad)
RSI	...	RSI(Overbought)
MACD	...	MACD(Bullish)
VWAP	...	VWAP(Above)
OBV	...	OBV(Increasing)
GOOGL_5pct(Increase)	...	0.1927999845961995
GOOGL_5pct(Decrease)	...	0.8072000015403801

Fig. 13: Final Prediction Table which tells us the probability of reaching 5 percent stock growth in next 15 days is very less i.e. 19.2 percent due to bad economic indicator and overbought RSI score.

The CPD for GOOGL_5pct represents the conditional probabilities of Google's stock increasing or decreasing by 5% given the different combinations of parent node states (e.g., the state of the SP 500, the economic indicator, RSI levels, etc.).

The CPD is created by considering the probabilities of all possible combinations of the parent nodes' states.

3. CPD for GOOGL_5pct:

The CPD table for GOOGL_5pct has probabilities for each possible outcome (Increase or Decrease) of Google's stock price, conditional on different states of the parent variables. The values for GOOGL_5pct(Increase) and GOOGL_5pct(Decrease) are generated based on historical data, probabilities derived from those data, and potentially some random sampling or normalization to sum up to 1 for each column.

Hidden Markov Model: Statistical model used to represent systems that follow certain processes over time, where the system's exact state is not directly observable (or basically its "hidden"), but we can observe some outputs or results that give us clues about the state.

HMM consists of states, observations, transition probabilities, emission probabilities, and initial state distribution.

States: "Hidden" parts of the system, where each state represents a condition or situation that we cannot directly observe it.

Observations: We can see or measure, this give us indirect information about the hidden states.

Transition Probabilities: Likelihood of moving from one hidden state to another state. It's a matrix where each entry gives the probability of transitioning from one state to another.

Emission Probabilities: Probability of observing a particular output given the hidden state.

Initial State Distribution: Probability of starting in any particular state at the beginning of the process.

The HMM uses these above components to calculate the most likely sequence of hidden states that could produce the observed data, often using algorithms like the Viterbi algorithm.

Summary of the Key Formulas:

A. Transition probability:

$$a_{ij} = P(S_{t+1} = S_j | S_t = S_i)$$

Transition Matrix:

```
[[7.94803497e-01 4.76121404e-05 1.31996841e-01 7.31520498e-02]
[1.21179488e-08 2.89882402e-01 7.85684241e-17 7.10117586e-01]
[1.16769670e-01 1.31367180e-11 8.83229055e-01 1.27533506e-06]
[2.41402387e-01 7.02310708e-01 4.92764568e-02 7.01044800e-03]]
Means of each hidden state:
[[-0.15296565 0.35548432]
[-1.59755972 0.96234209]
[0.68372935 -0.64804762]
[-1.42931475 1.05402981]]
Covariances of each hidden state:
[[[0.20912866 0.11404722]
[0.11404722 0.34695282]]
[[0.12997968 0.11478177]
[0.11478177 1.53005218]]
[[0.47630648 0.2544723 ]
[0.2544723 0.48779895]]
[[0.67917252 0.04544666]
[0.04544666 1.05744715]]]
```

Fig. 14: Sample Transition Matrix HMM for Nasdaq in last 1 year

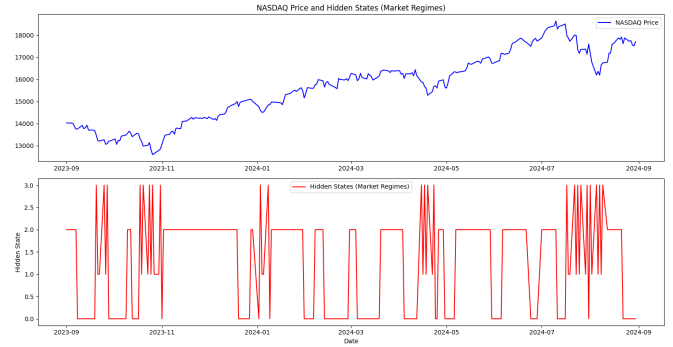


Fig. 15: Hidden Status for Nasdaq

B. Emission probability:

$$b_j(k) = P(O_t = O_k | S_t = S_j)$$

C. Initial state probability:

$$\pi_i = P(S_1 = S_i)$$

D. Forward algorithm:

$$\alpha_{t+1}(j) = \left(\sum_{i=1}^N \alpha_t(i) a_{ij} \right) b_j(O_{t+1})$$

E. Viterbi algorithm:

$$\delta_t(j) = \max_i (\delta_{t-1}(i) a_{ij}) b_j(O_t)$$

Here, We will be using the Viterbi algorithm to predict the most likely sequence of hidden states given the observed market index/stock price features. This will tell us whether the model believes the market is in a bull, bear, or sideways state.

Based on the current hidden state and the learned emission probabilities, we will be estimate the expected return or volatility for the next time step. For example, if the current state is a bull market, the model might predict a positive return with lower volatility.

* Transition Matrix Yet to be verified

The first row represents transitions starting from state 0: 79.48% chance of remaining in state 0.

13.19% chance of transitioning to state 2.

7.31% chance of transitioning to state 3.

A near-zero probability of transitioning to state 1 (very unlikely)

Interpretation:

State 0: slightly negative returns and relatively low volatility.

State 1: very negative returns and very high volatility, indicating a likely bearish market.

State 2: shows positive returns, potentially representing a bullish market with lower volatility.

State 3: negative returns and high volatility, likely another bearish or highly volatile market state.

We will also tweak the model and perform a backtesting with historical data for better risk management strategy.

Kalman Filter: Kalman Filter is an algorithm that provides estimates of the state of a system (in our case, its the market movement) over time, using measurements that may be inaccurate or noisy. In this report, we will be using to estimates the hidden state of a system that evolves over time and combines model predictions with noisy measurements to improve accuracy.

This works through a series of prediction and update steps, constantly refining its estimates.

Kalman Filter Equations

A. Prediction Step:

Predicted state estimate:

$$\hat{x}_{k|k-1} = A\hat{x}_{k-1} + Bu_k$$

Predicted covariance estimate:

$$P_{k|k-1} = AP_{k-1}A^T + Q$$

B. Update Step:

Innovation (measurement residual):

$$y_k = z_k - H\hat{x}_{k|k-1}$$

Innovation covariance:

$$S_k = HP_{k|k-1}H^T + R$$

Kalman Gain:

$$K_k = P_{k|k-1}H^TS_k^{-1}$$

Updated state estimate:

$$\hat{x}_k = \hat{x}_{k|k-1} + K_k y_k$$

Updated covariance estimate:

$$P_k = (I - K_k H)P_{k|k-1}$$

4. Automate the process to generate live conditional probabilistic tables Utilize google cloud GPU compute units, API, Generate probabilistic values, apply and generate the prediction market results in the tool.

Pseudocode The following pseudocode outlines the process of fetching and processing economic data for those multiple countries:

1. Initialized a list of economic indicators, countires to be fetched.

2. Defined the date range for data retrieval (from 2014 to the present).
3. Fetched economic data from the World Bank API for the specified countries and indicators.
4. Converted the fetched data into a usable data structure (e.g., DataFrame).
5. Extracted 'date' and 'country' from the fetched data structure.
6. Converted the 'date' field to a datetime format for proper filtering.
7. Filtered the data to include only records within the specified date range. Sorted the filtered data by date in descending order.
8. Round the GDP values to the nearest trillion. Round other numerical fields to two decimal places.
9. Removed the unnecessary columns (e.g., 'Real Interest Rate (%)').
10. Display and save the processed data to generate probabilistics tables.

3 Conclusion

Finally conclude that 7 major global stock indices data sets were collected from various sources and few probabilistic models has been developed to predict and to analyse the market movement with other economic indicators. While market is highly complex, and often risky these probabilistic tables could provide very valuable insights to define good investment strategies, to reduce unexpected losses, to build a better risk management, enhance risk management, showing efficiency during volatile periods [12].

This report supporting materials, coding has been documented here in this Github Page. (<https://github.com/Vivekanandr/Probabilistic-Model>)

In future, new studies from various finance/economic macro-economic parameters can also be employed with this paper where a set of probabilities actionable insights can follow up on.

References:

1. On the use of graphical models for valuation of financial assets: [<https://www.science.org/doi/10.1126/science.aat8127>]
2. Federal Reserve Interest Rates, 1954-Present. [<https://www.kaggle.com/datasets/federalreserve/interest-rates?n>]
3. On the use of graphical models for valuation of financial assets: [<https://www.cs.cmu.edu/~epxing/Class/10708-19/assets/project/1>]
4. Predicting stock and stock price index movement using Trend Deterministic Data Preparation and machine learning techniques [<https://www.sciencedirect.com/science/article/abs/pii/S09574174>]
5. xLSTM: Extended Long Short-Term Memory [https://www.researchgate.net/publication/381925940_xLSTM_E]
6. <https://data.worldbank.org/>
7. Forecasting the equity premium with frequency-decomposed technical indicators [<https://doi.org/10.1016/j.ijforecast.2022.12.001>]

7. Stock picking with machine learning
[<https://doi.org/10.1002/for.3021>]
8. Algorithmic Trading Strategies: Leveraging Machine Learning Models for Enhanced Performance in the US Stock Market [<https://doi.org/10.32996/jbms.2024.6.2.13>]
9. Theoretical Foundations and Application of Hidden Markov Models
[<https://doi.org/10.9734/jsrr/2024/v30i82303>]
10. PREDICTIVE MODELING FOR SHARE CLOSING PRICES THROUGH HIDDEN MARKOV MODELS WITH A SPECIAL REFERENCE TO THE NATIONAL STOCK EXCHANGE [<https://doi.org/10.17654/0972361724036>]
11. Dynamic Weighting Methods in Portfolio Construction: A Hidden Markov Model Approach
[<https://hdl.handle.net/2077/82236>]
12. AI-Powered Energy Algorithmic Trading: Integrating Hidden Markov Models with Neural Networks
[<https://doi.org/10.48550/arXiv.2407.19858>]

Acronyms:

GDP - Gross Domestic Product

HMM - Hidden Markov Models

CPD - Conditional Probability Distribution

CPT - Conditional Probability Table

RSI - Relative Strength Index

MACD - Moving Average Convergence And Divergence

VMAP - Volume Weighted Average Price

OBV - On Balance Volume

SP 500 (Standard and Poor's 500)

DJIA - Dow Jones Industrial Average

FTSE 100 - (Financial Times Stock Exchange 100 Index)

DAX - Deutscher Aktienindex

Sensex - Sensitive Index

¹

¹Note: This is a sample report, only for viewing and demonstration purpose.