# Predicting Stock Market Index Direction Using ARIMA-Augmented Quantum Random Forest Models

**Integrating Quantum Computing, Machine Learning and Classical Time Series Modeling**

Don Krapohl (Advisor: Dr. Shusen Pu)

2025-10-09

## Table of contents

# 1 Introduction

## 1.1 Background and Motivation

Short-term stock market forecasting is a common challenge engaged by many millions of analysts and investors daily. Stock market data is frequently non-linear and is influenced by not only financial drivers but also geopolitical and macroeconomic policies and events. Random Forest has demonstrated the ability to handle non-linear, heterogeneous features while being explainable and resistant to overfitting. One basic issue with Random Forest models are that they do not intrinsically have memory and so can miss opportunities that are based on time-based influences in variables.

## 1.2 Link to Project Code

The Jupyter Notebook is available from **https://github.com/dkrapohl/UWF_DataScience_Capstone/**

## 1.3 Research Problem

The objective is to use my course work, current literature, and intent on future research to classify the market movement as either upward or downward. Because Random Forest has no memory I will use both machine learning, time series modeling, and quantum circuits to identify optimal lags and moving averages and introduce these variables during feature engineering.

## 1.4 Research Objectives

I intend to develop a hybrid methodology combining ARMA feature engineering with Random Forest classification, identify optimal lag structures and moving average windows through systematic time series analysis, evaluate model performance using multiple metrics, and determine feature importance for market direction prediction.

## 1.5 Purpose of Study

I will use my coursework, readings, coding, and statistical knowledge to synthesize an approach to analysis that, although not novel in academia, is new to me. I will not be using any of the tools developed in my coursework to identify, train, tune, and measure the models I build so that I may perform real-world analysis of a type I believe to be relevant to many datasets with which I've worked.

### 1.6 Scope and Limitations

The data will be from the United States Standard & Poor's S&P 500 index covering 1990-2024. The forward-looking limits will be 20 days and the predicted outcome will be binary (up/down).

### 1.7 Capstone Project Organization

This project will consist of a section covering the background, theory, recent research, and explanation of: - Random Forest models - Auto Regressive Moving Average (ARMA) models - Generalized AutoRegressive Conditional Heteroskedasticity (GARCH) models of volatility - Vector Autoregressive (VAR) models and Multivariate Time Series analysis - Hybrid models - Quantum Random Forest

I will being with a literature review, outline my methodology and dataset, state dataset statistical information, perform feature engineering, train and measure my models, review the findings, and discuss their implications.

## 2 Literature Review

### 2.1 Overview of Stock Market Prediction

Stock market prediction prior to the 1960s was based on technical or fundamental analysis, both of which are used today. Technical analysis involves analyzing charts of stock prices to look for long- and short-term cycles and patterns. Fundamental analysis is the use of company and industry data including balance sheets, contracts, and forecasts to try to determine the current and future value of a company. In the 1960s the Efficient Market Hypothesis (EMH) was the most common theory of how market pricing worked. In this, the price of a stock instantly reflected all information that could affect the price with the implication that constant changes in price are largely random and unpredictable. In the 1980s more computing power and advanced mathematical approaches identified subtle patterns within this "randomness" indicating the movements are not entirely random. Behavioral Economics showed that human and group psychology provided one mechanism by which pricing changes could violate the Efficient Market Hypothesis. The development of Autoregressive Integrated Moving Average (ARIMA) models provided the ability to forecast with more quantitative rigor. In the 2000s computing power and algorithm development advanced further leading to machine learning developments including Random Forest, Support Vector Machines, Recurrent Neural Networks, and Long- Short-term memory (LSTM) models the latter of which benefitted from both temporal memory as well as the ability to "forget" weakly interacting data points.

## 2.2 SARIMA and ARIMA Time Series Model Development

There were some foundational research projects in the 1920s that set the stage for the development of Seasonal Autoregressive Integrated Moving Average (SARIMA), a form of ARIMA in 1970 in part by Box and Jenkins (Box et al. 2015). SARIMA adds seasonality to ARIMA models and tries to find the simplest (most parsimonious) model by identifying the stationarity of data, estimate model parameter values, and checking the validity of the model. The concept of stationarity is the measure of whether a series of data have a trend or seasonality. The removal of trend and seasonality was determined to provide a more robust model(Box et al. 2015). One aspect of these time series models that limit their use is that the data must be able to be rendered stationary for the models to be valid.

## 2.3 GARCH Models of Volatility

Generalized AutoRegressive Conditional Heteroskedasticity (GARCH) is an extension to the 1982 Nobel prize winning AutoRegressive Conditional Heteroskedasticity (ARCH) system that rely on the observation that periods of high volatility tend to cluster together in time. This allows ARIMA models to capture risk over the timeframe of the model and can compensate for limitations of the heterskedasticity assumption of an ARIMA model by allowing variance to be dynamic.

## 2.4 Vector Autoregressive (VAR) and Multivariate Time Series Analysis

The addition of Vector Autoregression (VAR) models were developed in the early 1980s to capture the reality of financial markets, that they are influenced by many internal and external factors such as interest rates, unemployment, current volatility, current pricing levels, and many others. These factors have complex and dynamic influence on each other. VAR models are designed to capture current values and relationships as well as past values and their relationships. This is in contrast to the commonly 1- or 2-dimensional SARIMA models capturing linear dynamics over time.

So far I've discussed univariate time series models (ARMA, GARCH) that model a single variable over time. But financial markets don't exist in isolation - the S&P 500, VIX, interest rates, and unemployment all influence each other simultaneously. **Vector Autoregression (VAR)** models capture these dynamic relationships. VAR models also capture systemic shock propogation and the duration and power of their effect on the market.

## 2.5 Quantum Random Forest

Within the scope of my Random Forest (RF) study, traditional Random Forest uses standard compute approaches. Cloud services such as Amazon Braket provide quantum compute and

compute simulators that add quantum compute paradigms to the RF and other machine learning algorithms. One of the key capabilities in quantum RF (QRF) is the ability to move Gini impurity index calculation into a higher dimensional space, which may make the data more separable.

## 3 References

Box, George E. P., Gwilym M. Jenkins, Gregory C. Reinsel, and Greta M. Ljung. 2015. *Time Series Analysis: Forecasting and Control*. 5th ed. Hoboken, NJ: John Wiley & Sons. https://www.researchgate.net/publication/299459188_Time_Series_Analysis_Forecasting_and_Control5th_Edition_by_George_E_P_Box_Gwilym_M_Jenkins_Gregory_C_Reinsel_and_Greta_M_Ljung_2015_Published_by_John_Wiley_and_Sons_Inc_Hoboken_New_Jersey_pp_712_ISBN_.