**Introduction**

Financial indices are imaginary securities portfolios representing a particular portion of the stock market. The indices consist of the shares of leading companies in the economy or a specific sector. As a result, the stock indices provide an easy way to track the overall health of the stock market with statistical measurements of the stock prices. By looking at the statistical measures, it is easy to observe the current state of the market. Further, historical data guide investors on how the market has reacted in the past. That can help investors to identify any risk in the indices and forecast the future; therefore, it allows investors to make a better investment decision.

The stock markets tend to be unpredictable and even illogical. Due to these characteristics, finding a reliable pattern in the financial data is challenging. Fortunately, by giving the historical index prices, the machine learning model can discover hidden structures within the data and forecast how they will affect them in the future. The index prices are discrete-time series models based on numerical data; the index price is collected at successive points at regular intervals. Transforming the time series using the Seasonal Autoregressive Integrated Moving Average (SARIMA) model is a recommended algorithmic approach as it gives more reliable results from stationary data. If the Augmented Dickey-Fuller (ADF) test verifies that the series is stationary, the ARIMA model is robust and the most efficient for the index market analysis1.

This project will focus on one of the most popular financial indices, the S&P 500. It is a daily indicator of the US economy, including 500 of the largest companies traded on NYSE, Nasdaq, or Cboe. This project involves a descriptive analysis of the S&P 500 index where its attributes, such as risk-reward and forecast, are observed. In other words, this project will determine the risk-reward profile of the S&P 500 index by looking at its historical data collected from the 1960s to the recent and propose a forecast for S&P 500 index for the next five years by using the ARIMA model in the stationary data. The S&P 500 data is univariate time series containing a series with a single-time dependent variable. The index prices are the explanatory variable dependent on the Date. It will help Burnaby Financial Planning, Inc. to communicate with clients so that they can make better decisions on their investments.

**Data**

Data for this project has been collected from the 1960s until now. The overall sampling strategy and baseline study protocol of the S&P 500 index are based on the eligibility criteria of the Index Committee. It also considers sector balance in the selection of companies. Even if the economic market has a daily index, we will use the monthly index price by taking the ending balance and either adding back net withdrawals or subtracting out net deposits during the period. Then divide the result by the starting balance at the beginning of the month.

The monthly index price of the 500 selected companies has the unit of price where it changes worldwide. For instance, the index price drastically dropped at the beginning of the COVID 19. Therefore, the index price during the unique situation could differ from the overall price. The response variable is the index price, whereas the explanatory variable is the monthly Date.

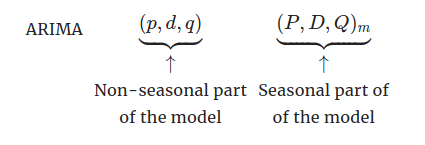
The index price will naturally fluctuate around the actual price due to the measurement error. As we use the monthly index price, it possesses a potential bias where it may not represent the daily index price. Since the parameter estimate will depend on the sample, the forecast would be different from the study population's actual value if we use the daily index price.

The excel will store the data, and we will use the R software for subsequent analysis. Since the data is time-series, we will work on an Augmented Dickey-Fuller (ADF) test to assure stationarity. Once we confirm the data quality, we build a Seasonal Autoregressive Integrated Moving Average (SARIMA) model for an actual analysis. Ultimately, we use an Akaike Information Criterion (AIC) for evaluating forecasts at the end.

**Proposed Analysis**

First of all, we must confirm whether it is stationary or not with the given univariate time series data. It is a critical initial step before actual analysis as we cannot build any model if it is not stationary. It will use the Augmented Dickey-Fuller (ADF) test, which performs the definitive null hypothesis test and returns a p-value. In our project, the null hypothesis is that the series is not stationary, while the alternative hypothesis says that the series is stationary. If the p-value is less than 0.05, we reject the null hypothesis and assume that the data is stationary. But if the p-value is greater than 0.05, we fail to reject the null hypothesis and determine the data to be non-stationary. If the data is non-stationary, we must transform the data before fitting the model. "Differencing" is a common way to transform the data to stationary, finding the differences between consecutive data terms and shifting them gradually until the data becomes stationary. But each differencing step comes at the cost of losing a row of data. If the data shows seasonality, then the differencing is done by season. For instance, since we have monthly index price data with yearly seasonality, we difference the data by a time unit of 12.

Once we transform the data to be stationary or the data is found initially to be stationary on the ADF test, we fit the Seasonal Autoregressive Integrated Moving Average (SARIMA) model. SARIMA is an extension of ARIMA that supports univariate time series data with a seasonal component. That is, it includes seasonality contribution to the risk analysis and the forecast. As it is an extension of the ARIMA model, the Autoregressive (AR), Integrated (I), and Moving Average (MA) remains in SARIMA with the addition of seasonality that adds robustness.

where m is the number of observations per year. We use the uppercase notation for the seasonal parts of the model and lowercase notation for the non-seasonal parts of the model. The additional seasonal terms are multiplied by the non-seasonal terms2. After fitting the SARIMA model, we would like to evaluate the model's fit by computing Akaike Information Criterion (AIC), whose goal is to maximize fit on out-of-sample data; the lowest AIC offers the best fit3.

The selected SARIMA model with the lowest AIC is the best representative of the risk and the forecast in the S&P 500. However, there must be uncertainty due to sampling and measuring errors. Fortunately, we can quantify the uncertainty using the Monte Carlo Dropout (MCD), computed as the average of the predictions4.

**Conclusion**

Using the given time-series S&P 500 index data, we can fit the SARIMA model to investigate the risk-reward profile and even further forecast the future index price. The ADF test should verify the data quality before the model fitting. Also, the AIC evaluates the model fitting to ensure the analysis's quality. This way, we help Burnaby Financial Planning, Inc. to communicate with clients so that they can make better decisions on their investments by looking at the risk-reward profile and forecast of the S&P 500 index price.

**Reference**

1Chauhan, N. S. (n.d.). Stock market forecasting using time series analysis. KDnuggets. Retrieved 15 July 2022, from https://www.kdnuggets.com/2020/01/stock-market-forecasting-time-series-analysis.html

28.9 Seasonal ARIMA models | Forecasting: Principles and Practice (2nd ed). Otexts.com. Retrieved 15 July 2022, from https://otexts.com/fpp2/seasonal-arima.html.

3Zajic, A. (2019). *Introduction to AIC — Akaike Information Criterion*. Medium. Retrieved 15 July 2022, from https://towardsdatascience.com/introduction-to-aic-akaike-information-criterion-9c9ba1c96ced#:~:text=In%20plain%20words%2C%20AIC%20is,lower%20AIC%20score%20is%20better.

4Milanés-Hermosilla, D., Trujillo Codorniú, R., López-Baracaldo, R., Sagaró-Zamora, R., Delisle-Rodriguez, D., Villarejo-Mayor, J. J., & Núñez-Álvarez, J. R. (2021). Monte Carlo Dropout for Uncertainty Estimation and Motor Imagery Classification. Sensors (Basel, Switzerland), 21(21), 7241. https://doi.org/10.3390/s21217241