Financial Forecasting: The Future of Investing

Jun Lee

Finance + Data Science at UIUC

01
Introduction

Introduction

- Economic indicators provide information about an economy and whether it is expanding or contracting
- Most indicators are released monthly by government departments and agencies, such as Federal Reserve Bank
- Financial Forecasting has been a popular method for investors to manage risk in their investments, and we will be exploring the method by utilizing mathematical and statistical theories

https://proteafinancial.com/financial-forecasting-101/

Background

- There are a variety of strategies investors use in order to predict stock prices
  - Some investors in the stock market utilize fundamental analysis
  - Other investors put more emphasis on technical analysis, which involves locating patterns of historical stock price movement and making predictions based on those patterns
- With the pressure of the Russo-Ukrainian War and tension between the United States and China, the economy was on a downward trend for the fiscal year of 2022—a question remains: what happens now? In other words, how can the **banks better retain stability** in the midst of incoming recession?

Proposal

- Machine learning models can give investors an accurate way to predict stock prices in the stock market
- Two models that have been becoming increasingly popular in the past couple of years to predict stock prices are:
  - LSTM neural network model
  - Autoregressive Integrated Moving Average (ARIMA) Model through the ADF Test + Granger Causality Test

< Objectives >

1.) Demonstrate that financial forecasting is a necessary step forward in market investing.
2.) Show how the ARIMA model can be used to predict stock prices and make informed investment decisions.

02
Brief Technical Overview

Visual Analytics

< Pearson Correlation Matrix >

The following diagram shows a visual representation of correlation matrix of the input data:

Dark green indicates there is a strong positive relationship with correlation closer to 1;

Dark red indicates there is a strong negative relationship with correlation closer to -1;

White indicates no relationship.

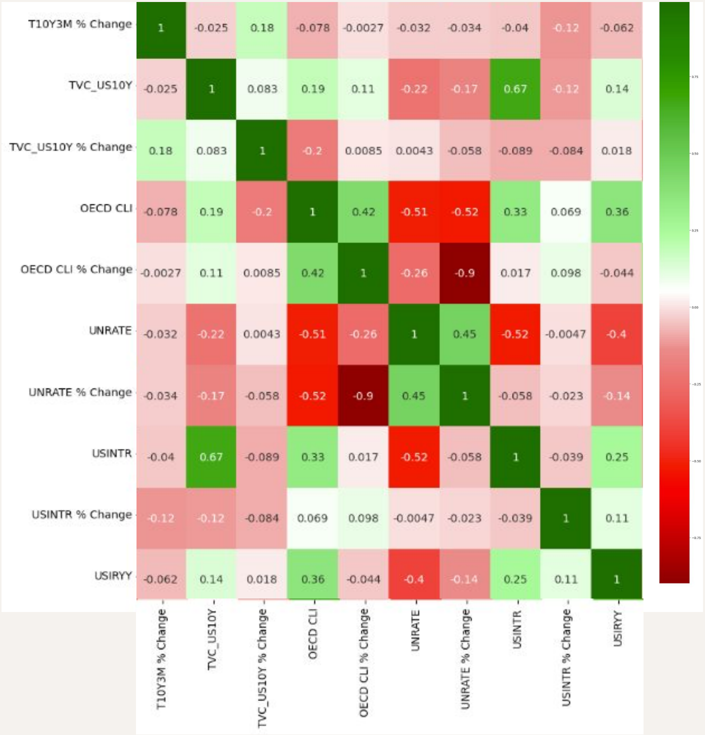


Figure: Pearson Correlation Matrix (Heatmap) condensed

Visual Analytics

< Dual Axis Plots >

From the correlation matrix, we have drawn dual axis plots, and the following chart indicates that **OECD CLI seems to lead S&P Sales % Change**, which can be helpful for accurate financial forecasting

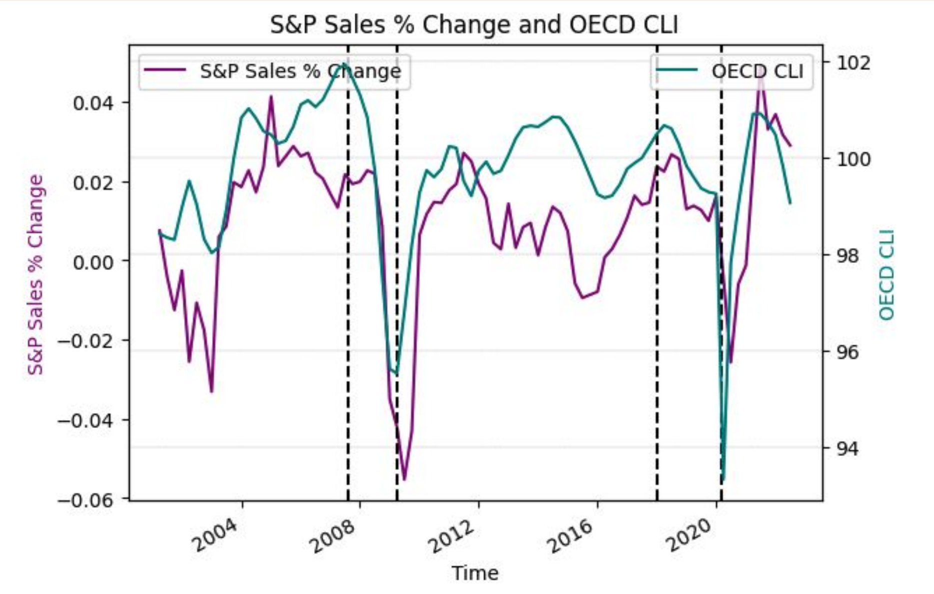


Figure: Dual Axis Plot of OECD CLI and S&P Quarterly Sales Percent Change

03
Statistical Modeling

Augmented Dickey-fuller Test

< ADF Testing >

- Test for stationarity with Augmented Dickey-Fuller test, looking for ADF Stat less than critical
- Null hypothesis = data is NOT stationary
- Alternative hypothesis = data is stationary
- Condition: p-value <= 0.05

< Key Takeaways >

• p-value of *SPX % Change Diff* = <u>0.000031</u>

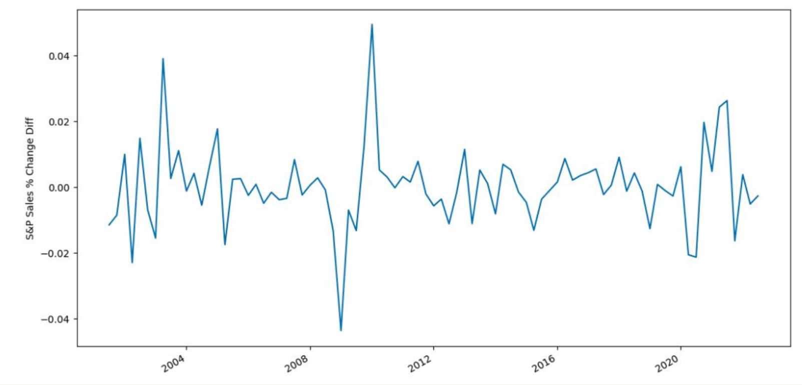
• p-value of *OECD CLI p-value* = <u>0.007479</u>

∴ Since the p-value < 0.05,

We **reject the null hypothesis** → both variables are in stationary time series

```
SPX % Change Diff
ADF Statistic: -4.924933
p-value: 0.000031
Critical Values:
        1%: -3.519
        5%: -2.900
        10%: -2.587
```

```
[OECD CLI]
ADF Statistic: -3.520373
p-value: 0.007479
Critical Values:
        1%: -3.511
        5%: -2.897
        10%: -2.585
```



<u>Figure: ADF statistics of selected variables and visualization of stationary data</u>

Granger Causality Test

```
gc_results = grangercausalitytests(df_granger[['SP Sales Pct Chg Diff', 'OECD CLI']], maxlag=5, verbose=True)

print(gc_results)
```

```
Granger Causality
number of lags (no zero) 1
ssr based F test:         F=0.8786  , p=0.3518  , df_denom=70, df_num=1
ssr based chi2 test:   chi2=0.9162  , p=0.3385  , df=1
likelihood ratio test: chi2=0.9105  , p=0.3400  , df=1
parameter F test:         F=0.8786  , p=0.3518  , df_denom=70, df_num=1

Granger Causality
number of lags (no zero) 2
ssr based F test:         F=15.2915 , p=0.0000  , df_denom=67, df_num=2
ssr based chi2 test:   chi2=32.8652 , p=0.0000  , df=2
likelihood ratio test: chi2=27.0727 , p=0.0000  , df=2
parameter F test:         F=15.2915 , p=0.0000  , df_denom=67, df_num=2

Granger Causality
number of lags (no zero) 3
ssr based F test:         F=12.1390 , p=0.0000  , df_denom=64, df_num=3
ssr based chi2 test:   chi2=40.4000 , p=0.0000  , df=3
likelihood ratio test: chi2=31.9818 , p=0.0000  , df=3
parameter F test:         F=12.1390 , p=0.0000  , df_denom=64, df_num=3
```

Figure: Results of Granger Causality Test with identified Autocorrelation lag of 4 periods (months)

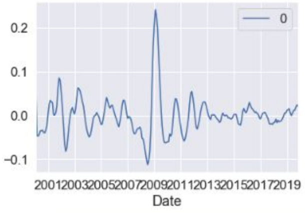
Granger causality test is an investigative method to find causality between two variables that are in time series (time-series variable takes time as a factor).

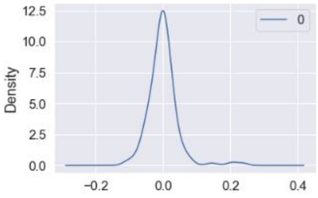
< Key Takeaways >

The p-values and F-statistics on Granger Causality Test results indicate that there is a probable chance that OECD CLI indicator leads S&P Index

Autocorrelation using ARIMA model

```
                             ARIMA Model Results
==============================================================================
Dep. Variable:             D.OECD CLI   No. Observations:                  238
Model:                 ARIMA(1, 1, 0)   Log Likelihood                 398.849
Method:                       css-mle   S.D. of innovations              0.045
Date:                Wed, 05 Feb 2020   AIC                           -791.699
Time:                        21:22:54   BIC                           -781.282
Sample:                    02-29-2000   HQIC                          -787.501
                         - 11-30-2019
====================================================================================
                       coef    std err          z      P>|z|      [0.025      0.975]
------------------------------------------------------------------------------------
const                0.0036      0.074      0.049      0.961      -0.141       0.148
ar.L1.D.OECD CLI     0.9642      0.016     61.679      0.000       0.934       0.995
                                    Roots
=============================================================================
                  Real          Imaginary           Modulus         Frequency
-----------------------------------------------------------------------------
AR.1            1.0372           +0.0000j            1.0372            0.0000
-----------------------------------------------------------------------------
```





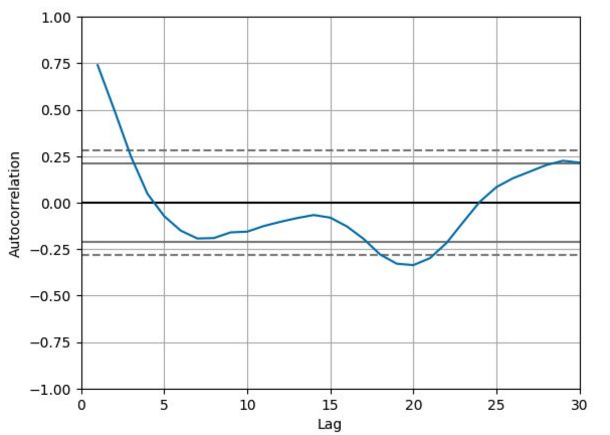


Figure 1: ARIMA Model Output
Figure 2: Visualization of Autocorrelation with lagging periods as X label

< Key Takeaways >

The autocorrelation plot below indicates we can play with lagging the OECD CLI up to 4-5 periods to get a decent fit.

Predicting and Back-testing yields

< Analyzing Predicted Values>

The forecasted CLI numbers suggest that it will gradually rise but will remain below 100 for at least the next three months. Below 100 indicates that economic activity is falling short of the long-term trend

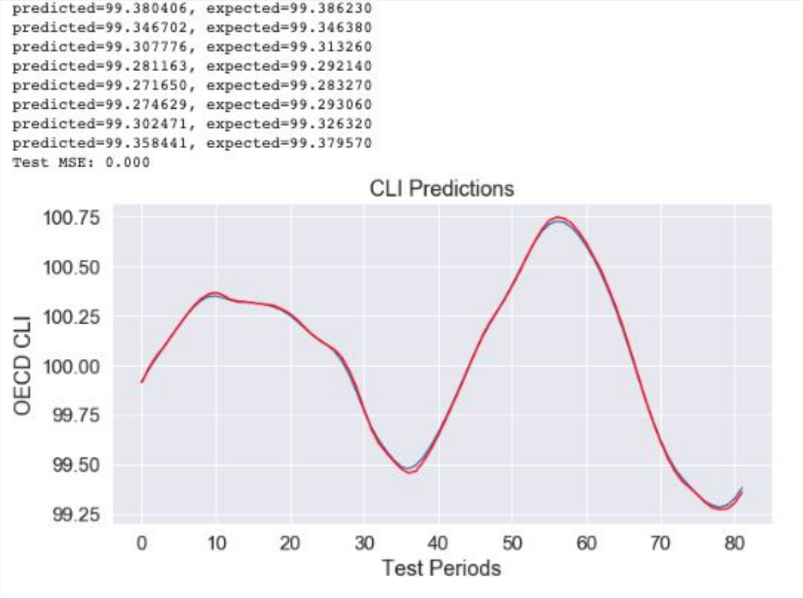


Figure: Visualization of CLI indicator predictions; red line = predicted values, blue line = original historical values

Results

The prediction model proves to have around 57% accuracy in average. It is evidently proved that there is a clear linear relationship between OECD Composite Leading Indicator and S&P Quarterly Sales movement

```
# Printout relevant metrics
from sklearn.metrics import mean_squared_error, r2_score, mean_absolute_error
print("Model Coefficients:", model.coef_)
print("Mean Absolute Error:", mean_absolute_error(y_tst, model_prediction))
print("Coefficient of Determination:", r2_score(y_tst, model_prediction))

print('Conclusively, Out of {} Price Moves, {:.2f}% Accuracy to predict the price accurately'.format(y_tst.shape[0],
```

```
Model Coefficients: [[0.00890028]]
Mean Absolute Error: 0.00970225924715269
Coefficient of Determination: 0.5687882569057696
Conclusively, Out of 17 Price Moves, 56.88% Accuracy to predict the price accurately
```

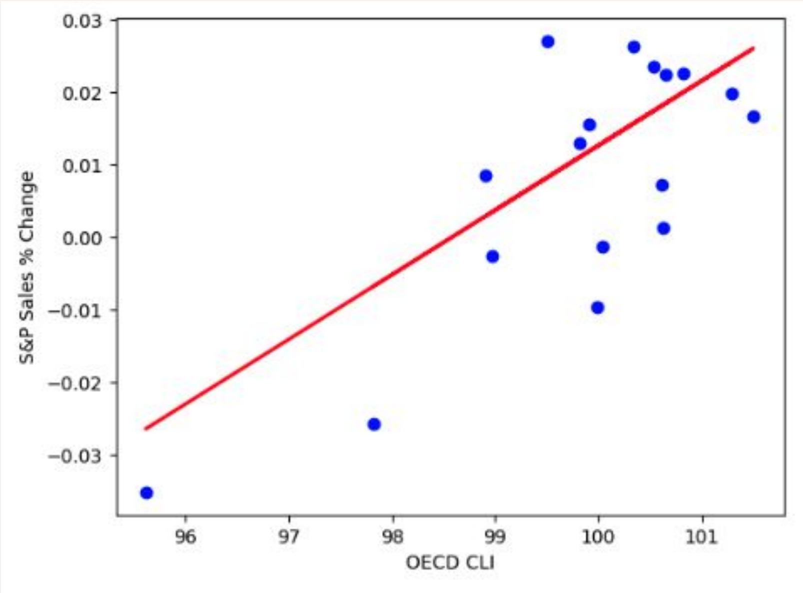


Figure 1: The scatter plot of predicted S&P values using OECD CLI indicator
Figure 2: Model Coefficient, Mean Absolute Error, and Accuracy of the model and predicted values

04

Discussion & Recommendations

Discussion

< Limitations >

- Integrating complex modeling schemes is computationally inefficient; thus, **increased overhead costs**
- ARIMA model maps linear relationships **only** and therefore cannot handle nonlinear nature of stock market exchange

< Potential Impact >

- Human tendencies skew investment decisions
- ARIMA model/financial forecasting gives non-emotion based predictions → **aids decision-making process**
- Financial modeling provides additional confidence to investors

Recommendations

< Integrate Quantitative Financial Modeling >

- Proven method of maximizing profit + minimizing loss
- Enables learning based off past market trends
- Provides greater financial opportunity for banks

< Investigating Models + Their Design >

- ARIMA model is a promising method of financial forecasting
    - However, the model is limited to linear relationships
- Conducting research to tune the model for different financial domains
- Implementing a non-linear model could yield more accurate results given the non-linear nature of the economy
    - i.e., Long-short-term Memory model, Gated Recurrent Unit model