|  |  |
| --- | --- |
| Summer 2021 | BUAN 6337 |
|  | Final Project |

**STOCK PREDICTION**

**&**

**PORTFOLIO DESIGN ANALYSIS**

*Under the guidance of Prof Sourav Chaterjee*

**Group 5**

Roland Cadotte

Shravika Reddy Gudepu

Wenjuan Han

Jeremiah Chacko

**Executive Summary:**

Stock market news replaced sports highlights last year for many people, and today stocks are omnipresent on social media and television. Many employers often offer retirement plans that their employees are encouraged to participate. Typically, participants follow the “set-it and forget-it” approach where they decide on their respective contribution percentage and continually deposit into a diversified portfolio tailored for a specified retirement year. Since, the number of people investing in stocks has been increasing, it is necessary to have a good stock portfolio and also having insights on the best opportunities available might influence investments plans.

In this paper, we analyze a select group of stocks, perform data analysis on one of the predictable stocks and develop price prediction models, using three methods:

* Arima
* ESM
* UCM

then create a portfolio comprised of these stocks based on performance and predictability. In addition to the stock portfolio, an ETF portfolio is created based on the industries and/or indexes the group of stocks belong to. These two portfolios are compared to two diversified portfolios, one with an aggressive composition, and one with a more conservative composition.

1. **Introduction**

A stock (also known as equity) is a security that represents the ownership of a fraction of a corporation. Investors buy them for the income these stocks generate. Value stocks may be growth or income stocks. People buy value stocks in the hope that the market will grow and that the increase in stock's price will generate huge profits. In this paper, we aim to analyze stocks that will help us find the best investment opportunities as an individual or for the investors. Utilizing historical stock market pricing data, we want to analyze if there is a benefit to invest in individual stocks, certain ETFs, or a blend of each. The analysis on few particular stocks will also help in building a stock portfolio and customizing it to get significant profits in the future.

To predict an accurate model for the future stocks based on historical data, we started with exploratory data analysis on different stocks. We generated times serious plots, white noise, trend analysis, and autocorrelation analysis to determine if the stock is not a white noise series and has a potential growth from 2005 to 2017. Among the tested stocks, we selected JNJ for our analysis as we rejected that the series is not a white noise, auto correlated and has predictable outcomes. In addition to it, we also did the empirical analysis to determine the best suitable model for prediction. We generated three models of Time Series Analysis which includes ARIMA, ESM, and UCMs models.

Based on the prediction results of the models discussed in the Empirical analysis section, a custom portfolio was created by balancing the predicted return with the low-case and high-case scenarios. These are described as Aggressive and Conservative

1. **Data Description**

We have chosen the stocks/ETFs dataset from Kaggle and whole analysis is based on this dataset. This dataset contains historical daily prices for all stocks currently trading on NASDAQ. This data set contains observations from February 2005 to November 2017 for 5 days of every week. This data set contains 7 variables for a particular day and Time Series analysis have to be performed.

**Data Structure**

The data for every stock is saved in .txt format with common fields:

* **Date** - specifies trading date
* **Open** - opening price
* **High** - maximum price during the day
* **Low** - minimum price during the day
* **Close** - close price adjusted for splits
* **Adj Close** - adjusted close price adjusted for both dividends and splits.
* **Volume** - the number of shares that changed hands during a given day.

All the data of stocks is then stored in either ETFs or stocks folder, depending on its type. Moreover, each filename is the corresponding stocks.

**Data preparation:**

On further analysis, we found that not at all the stocks start at the same date, so we need to adjust accordingly for this. We have missing data for some of the ETFs. We noticed that the SPY ETF’s data started on February 5, 2005, whereas Fidelity says the fund was incepted in 2000. However, we selected Johnson and Johnson stock and performed our preliminary analysis. We generated different Forecasting models using the same stock and applied it to different other stocks to generate a portfolio.

We considered values of closing price for our analysis. There we few missing values in the dataset for the corresponding variable. We imputated the values with the value from the previous day for the UCM model. To make the analysis easier, we merged all the closing values of each stock into a single table based on the date using ‘PROC SQL’.

1. **Exploratory Data Analysis:**

Looking at few stocks of companies like Apple, Ford, Johnson and Johnson, Microsoft, proctor and Gamble, Verizon ranging from 2005 to 2017, the stock market has increased in value with SPY etf having the highest increase comparatively. Investment in the SPY ETF at anyone of it’s peaks would have led to huge profits.

Graphical user interface, chart

Description automatically generated

**Fig 1:** Time-series plot for different companies

Looking at the summary statistics for Daily Return (day-to-day difference in close price), the Mean of AAPLs daily return is an outlier compared to other Large-Cap industry leading stocks chosen for an initial analysis.

Table

Description automatically generated

**Fig 2:** Summary statistics of Daily returns

We were expecting the correlation between SPY and AAPL to be much larger because it is presently the highest weighted asset in that fund. It may be interesting to compare performance of a Stock when it is added to certain Funds or Indexes.

A picture containing text, newspaper

Description automatically generated

**Fig 3**: Correlation between different stocks

The data Set has the observations from February 2005 to November 2017. Total number of observations recorded are 3202. The data has been recorded for 5 days in a week starting from Monday and ending on Friday. Therefore, the data has been accumulated for every week based on average.

We have chosen JNJ (Johnson and Johnson) stock for our preliminary analysis as it looked the most predictable stock and applied the predicted model generated using JNJ stock on other stocks for portfolio analysis. We performed time series analysis on JNJ stock and generated ARIMA, ESM and UCM Models.

**Time Series Analysis on JNJ stock:**

Fig 4shows the plot of all the observations accumulated to average for every week from 2005 to 2017. The plot shows that trend exists, and the mean is not Zero indicating that the series is non-stationary. The possibility of series being a white noise should be considered and rejected before estimation.

Table

Description automatically generated

Chart, line chart

Description automatically generated

**Fig 4:** Time series plot of JNJ Stock

**Autocorrelation Plot:**

Autocorrelation is the correlation of present values versus lagged values. The first plot in the below Figure enables us to see if the autocorrelation exists at multiple lags. The bars in the plot are very high and the value of bars are close to value zero. ACF plots indicates that there is strong correlation within 95% confidence interval and series is not white noise.

Graphical user interface, chart, Excel

Description automatically generated

**Fig 5:** Auto-correlation Plot for JNJ

**White Noise:**

A series is said to be white noise if it has following characteristics:

* If it varies randomly around its mean
* If it has no systematic variation
* If consists of only random variation
* If it has constant variance

Null hypothesis states that the series is a white noise, and the alternative hypothesis states that one or more autocorrelations up to lag m are not zero.

* H0: The series is white noise
* H1: The series is not white noise

Chart, bar chart, histogram

Description automatically generated

**Fig 6:** White Noise probabilities for JNJ

From the above figure, as we can see that bars are extremely high. We can reject null hypothesis that the series is a white noise. Hence from the above tests it is concluded that series is auto correlated and is not a white noise.

**Trend Analysis:**

Chart, line chart

Description automatically generated

**Fig 7:** Decomposition of JNJ series

Chart, line chart

Description automatically generated

**Fig 8:** Trend Component for JNJ

The above figure shows that for the first 8 years, the series did not have any trend but eventually, the series have a positive trend and one of the ESM models can be applied.

**Seasonality :**

A picture containing graphical user interface

Description automatically generated

**Fig 9:** Seasonal Component of JNJ Stock

The above figure indicates that there exists some kind of seasonality. Further analysis should be done in order to confirm the same.

Graphical user interface, chart, histogram

Description automatically generated

**Fig 10:** Seasonality of JNJ stock

It looks like there is no seasonality in the initial years, but the most recent years indicate that there exists some sort of seasonality with first weeks of the year having lowest values and last few weeks having highest values. But the increase in variation is high in the recent years.

1. **Empirical Analysis**

With the goal of creating a portfolio that outperforms traditionally constructed retirement portfolios, the prediction and analysis of our models were evaluated on long-term projections. The dataset spans over 12 years from 2005 to 2017. First 8 years of the dataset were used for modeling purposes, and the last 4 years for forecasts and accuracy comparisons. Our modeling and analysis process criteria are the following:

* Run model on data and forecast 4-year price prediction
* Record predicted price and predicted return compared to price at the start of the forecast
* Using confidence intervals, record lower limit as the low case potential and upper limit as the high case potential
* Analyze model fit statistics for each stock to select the forecast for each respective stock
* Build custom portfolio and weightings based on the confidence in each stock’s model and potential range of outcomes

The following table XX below shows the summary of the models for each stock that was modeled.

The below sections will contain a detailed analysis into the modeling process for the JNJ stock, but this process was similarly applied to the other stocks mentioned into the summary table in the appendix.

**4.1 ARIMA**

Through our exploratory data analysis, it was discovered that the time series data for the JNJ stock is not stationary. The below graphic shows a comparison of the autocorrelation analysis for the original JNJ data and the first-differenced JNJ data. The ACF plot exhibits high values of autocorrelation which signifies a non-stationary data set and suggests an ARIMA model is more appropriate. Comparatively, the first-differenced data in the ARIMA model removes the autocorrelation after 1 lag. Analyzing the differenced JNJ data against time shows that the differenced data has a mean value around 0 and increases in variability around Observation 1000 and the later Observations.

This corresponds with our exploratory data analysis, that revealed increased volatility during the 2008 financial crisis and then a rapid increase in stock valuations from 2014 to 2018.

Chart

Description automatically generated

Graphical user interface, chart

Description automatically generated

**Fig 11:** Autocorrelation analysis for original JNJ stock price(left) and first-differenced data (right).

Looking at the ACF plot and the IACF plot, there is strong correlation with the first lag of the differenced data. Meanwhile, the PACF plot shows no correlation throughout the displayed lags. This suggests that the differenced data is more suitable for a Moving Average of 1 lag. An ARMA (0,1,1) model was performed, and the resulting autocorrelation check is below. The hypothesis test for the autocorrelation white noise test shows that the p-values varying around 0.05 depending on the lag value. This suggests that the variation still in the model may not all be attributable to white noise and that there may be a more adequate model.

Graphical user interface, chart

Description automatically generated

Table

Description automatically generated

**Fig. 12:** Residual autocorrelation analysis for the JNJ ARIMA(0,1,1) model.

Because the test for correlation amongst residuals for the ARIMA(0,1,1) model did not successfully reject the null hypothesis for all lag values, an ARIMA(1,1,1) model was ran. The results of this model are displayed below and have similar residual correlation results as the MA(1) model. Based off of these similar results, the Goodness of Fit Statistics for each model is used to determine which ARIMA model to trust.

Table

Description automatically generated

**Fig. 13:** Residual autocorrelation analysis for the JNJ ARIMA (1,1,1) model.

The table below shows the comparison of the AIC and SBC values for each model. ARIMA (0,1,1) model has the lowest fit statistic models so that is the model we will base our stock selection on in terms of ARIMA models. The table also displays the predicted 4-year value, as well as the lower and upper limits. As you can see, each prediction yields similar results and will result in similar recommendations in terms of how confident you are to invest in JNJ for your retirement portfolio.



Chart

Description automatically generated

**Fig. 14:** JNJ ARIMA Model Prediction Results.

The results of the ARIMA models for the other stocks are included in the Appendix. While performing these models, it was noted that the prediction errors for the two tech stocks, AAPL and MSFT, were the largest. Those two stocks saw rapid increases in value from 2014 to 2018 that the model was unable to capture. MRK and PG were the only two stocks that had an actual return greater than 20% and a percent error in the predicted price that was less than 10%.



**Fig. 15:** ARIMA Model Prediction Results.

**4.2 ESM**

**Exponential Smoothing Model:**

The prediction or forecasts using exponential smoothing model is generated by ESM procedure with optimized smoothing weights for time series data of closing stock of JNJ. The procedure can forecast JNJ stock data as its observations are equally spaced by a specific time interval.

ESM has eight models. Trend and seasonality of the series determine the model that can be used

to forecast.

* + Models for time series with trend:
    - simple exponential smoothing
    - double (Brown) exponential smoothing
    - linear (Holt) exponential smoothing
    - damped-trend exponential smoothing
  + Models for time series with seasonality:
    - seasonal exponential smoothing
  + Models for time series with trend and seasonality:
* Winters additive exponential smoothing
* Winters multiplicative exponential smoothing

From our exploratory analysis we observed that the series have a positive trend and seasonality after a certain period. Because both exist, the best model to forecast is the Winters additive exponential smoothing Model.

**Winters additive exponential smoothing Model:**

Pre-Processing of the data for the Winters ESM model included accumulating the stock prices by weekly averages. The seasonality component was then set to 52 because our data exploration identified a trend in the JNJ stock price that showed a gradual increase towards the end of the year. The more recent years show a shift in stock prices for the entire year, but the seasonal component was still present.

Chart

Description automatically generatedChart, histogram

Description automatically generated

**Fig. 16**:Residual distribution and its variation for the JNJ ESM Model

Approximately, the errors appear to have constant variance and normal distribution which is a positive sign to consider ESM Model.

Chart

Description automatically generated

**Fig. 16**:Residual autocorrelation analysis for the JNJ ESM model.

Looking at the ACF plot, the bars are within 95% confidence interval and looks like they are not auto correlated. However, when we observe the plot for the white noise hypothesis test, the probability is small enough to reject the null hypothesis up until lag of 50. This suggests that there is variation in the model that may not all be attributable to white noise, and thus, this model can be improved.

Chart, line chart

Description automatically generated

**Fig. 17**: JNJ ESM Prediction Results.

The above figure shows the data predicted for four years and the confidence level of JNJ stock. The table below summarizes the prediction results for the additional stocks that we analyzed. Stocks that experienced rapid changes in price from 2014-2018 were also the stocks that experienced the largest prediction errors. Part of this error could be the result of not correctly identifying the seasonal time period or misidentifying a seasonal component altogether.



**Fig 18:** ESM Prediction Results

**4.3 UCM**

UCMs are also known as structural time series models. These models decompose time series into below components:

* trend
* season
* cycle
* irregular
* regressors

UCM model can accommodate and extrapolate more features of change as a function of time. Each of the above component captures some important feature of the series dynamics and have their own models. In addition to it, each component has its own source of error. The coefficients for trend, season, and cycle are dynamic.

General form:

Change in Stock value = Trend + Season + Cycle + Regressors

**Table

Description automatically generated**

**Fig 19:** Final estimates of the values for the parameter to be estimated

The above table is generated when we run UCM inclusive of all components. The table shows that the variances for the slope component is highly insignificant. This suggests that a deterministic trend model may be more appropriate. The estimate of the damping factor is 0.97, which is close to 1. We can construct another model by holding the values of the slope disturbance variances fixed at zero.

Graphical user interface, application, table

Description automatically generated

**Fig.20**:Estimation and forecast summary of the data

Table

Description automatically generated

**Fig. 21**:Fixed parameters and starting values for the parameter to be estimated

The above table shows that Slope variances disturbances is fixed at 0 as discussed in the previous section

Table

Description automatically generated

**Fig. 22**:Maximum likelihood estimates of the free parameters

The above estimates show that, the disturbance variances of all the components in the table are significant except for error variance in both cycles.

Two types of goodness-of-fit statistics are generated after a model is fit to the series (as shown below. The first type is the likelihood-based goodness-of-fit statistics, which include the full likelihood of the data. The second type of statistics is based on the raw residuals.

Table

Description automatically generatedTable

Description automatically generated

**Fig. 23**:Residual based and likelihood based Fit statistics.

Table

Description automatically generated

**Fig. 24**:Significant Analysis of Components(Final State)

The above table provides the significance of the components of the model at the end of estimation span.

In order to consider the model to be perfect fit, we have to analyze if the residuals form a white noise. If yes, it indicates no amount of significant information is left out.

**Chart

Description automatically generated**

**Fig. 25**: Residual Diagnostics for the Airline Series Using a BSM

The residual histogram and the Q-Q plot show there is a small deviation from the normality of the residual distribution. . The sample correlation plots, the autocorrelation function (ACF) and the partial autocorrelation function (PACF), also do not show any significant violations of the whiteness of the residuals.

Chart, bar chart

Description automatically generatedChart

Description automatically generated

**Fig. 26**:Residual test for white noise

However, the residuals do not have constant variation in the recent years and also based on p-test, we fail to accept that the residuals form a white noise within 95% confidence interval. Therefore, on the whole some of the adequate data is being missed out.

**Component Plots:**

Component plots show the filtered and smoothed estimates, respectively, of the cycle component in the model. The smoothed estimate appears smoother compared to the filtered estimate. This is always true because the filtered estimate of a component at time t is based on the observations prior to time t—that is, it uses measurements from the first observation up to the .t 1/th observation. On the other hand, the corresponding smoothed estimate uses all the available observations—that is, all the measurements from the first observation to the last.

Chart

Description automatically generated

**Fig. 27**:Smoothed Irregular components

Chart

Description automatically generatedChart

Description automatically generated

**Fig. 28**: Smoothed level and Filtered level component

The below model shows the forecast of JNJ stock for 4 years based on the data before the year 2014.

Chart

Description automatically generated

**Fig. 29**: Forecast of JNJ

Chart

Description automatically generated Chart

Description automatically generated

**Fig. 30**:Smoothed Trend and Sum of smoothed Trend and Cycles for JNJ

In terms of percent error, the UCM models were accurate for stocks that did not experience large increases in value, and inaccurate in stocks that saw large increases in value. This is summarized in the table below. As you can see, the UCM predicted results over-estimated the performance of the poor performing stocks and under-estimated the performance of the good-performing stocks. The predicted returns were more conservative in general, but these results may lead an investor into poor-investments.



**Table 1:** UCM Prediction Results

1. **Portfolio Analysis**

To compare the customized portfolios, two diversified portfolios were created using the weightings of two different Fidelity Retirement Portfolios. These portfolios have a 30-year difference in target retirement date and are recreated by assigning different weights on specific ETFs to reflect the holdings of each respective portfolio. These are described below in Table **XX** and will be referred to as Aggressive and Conservative.

|  |  |  |  |
| --- | --- | --- | --- |
| **Diversified Portfolio Composition** | | | |
| **ETF Information** | | **Porfolio Weighting** | |
| **Description** | **Ticker** | **Aggressive** | **Conservative** |
| Tech | VGT | 12% | 9% |
| Fin Serv | XLF | 9% | 7% |
| Healthcare | XLV | 6% | 5% |
| Industrial | XLI | 6% | 5% |
| Cons Cyclical | XLP | 5% | 5% |
| Real Estate | VNQ | 5% | 4% |
| Energy | XLE | 5% | 4% |
| Utilities | XLU | 2% | 1% |
| Non-US | EEM | 35% | 25% |
| CorpBonds | LQD | 10% | 18% |
| US Bonds | AGG | 5% | 17% |
| **Domestic Stocks** | | 50% | 40% |
| **Foreign Stocks** | | 35% | 25% |
| **Bonds** | | 15% | 35% |

**Table 2:** Diversified portfolio composition by ETF

Based on the prediction results of the models discussed in the Empirical analysis section, a custom portfolio was created by balancing the predicted return with the low-case and high-case scenarios. Table **XX** below shows the these values averaged for each stock.

Table

Description automatically generated

**Table 3:**  Summarized model predictions

As stated earlier, the tech stocks (AAPL and MSFT) had the highest predicted return, as well as, the highest best-case scenario. Based off these predictions, it is hard to argue investing in any other companies over them, but predictions of tech stocks prior to the dotcom crash most likely showed similar positive predictions. In order to protect our portfolio from investing too heavily in one industry, we will cap the percentage of the portfolio arbitrarily at 30% per industry. The customized portfolio with each stock’s weighting is summarized below in Table **XX**. All of the analyzed stocks were included, except for Ford, due to the relatively low predicted return, and the relatively high predicted downside. AAPL, MSFT, JNJ, and PG received the highist weightings due to the predicted return and relative low poor-case scenario.

|  |  |  |
| --- | --- | --- |
| **Custom Portfolio Composition** | | |
| **Stock Information** | | **Portfolio Weighting** |
| **Description** | **Ticker** | **Stock** |
| Tech | AAPL | 20% |
| Pharmaceutical | JNJ | 20% |
| Cons Cyclical | F | 0% |
| Pharmaceutical | MRK | 10% |
| Tech | MSFT | 10% |
| Cons Cyclical | PG | 15% |
| Communication | VZ | 5% |
| Energy | XOM | 5% |
| CorpBonds | LQD | 10% |
| US Bonds | AGG | 5% |
| **Domestic Stocks** | | 85% |
| **Foreign Stocks** | | 0% |
| **Bonds** | | 15% |

**Table 4:** Customized portfolio

The diversified portfolios and the custom portfolio respective weightings were applied and the increase in daily security price was analyzed through time. Table **XX** shows the annualized expected returns for each portfolio and Figure **XX** shows the time series plot of the return on investment starting at February 2005.

Table

Description automatically generated

**Table 6**:Annualized Portfolio Performance

Chart, line chart

Description automatically generated

**Figure 31:** Portfolio Return on Investment

The customized portfolio significantly outperforms the diversified retirement portfolio in terms of return on investment. The figure above is summarizing the scenario of investing a lump-sum in 2005. Even though 4 years may be considered a long time, it is small in terms of a retirement account that needs to be a sufficient source of income for 20-30 years. The annualized standard deviation is still less than the aggressive retirement portfolio, but 2013 and 2015 illustrate the risks of being heavily invested in a small concentration of stocks. The decrease in value was not recovered until 1-2 years later, but the custom portfolio still outperformed the more diversified portfolios.

1. **Conclusion**

With a large database of stock and ETF time-series data, we utilized our domain knowledge to focus on ETFs that can be used to reflect common retirement portfolios, as well as select and evaluate the predictability of several large-cap stocks. The previous sections go into detail about the exploratory data analysis, data pre-processing, and the modeling processes that we followed.

While the UCM and ECM methods produced some of the most accurate predictions, the large variance in their low and high-case predictions for the better performing stocks made those models to be inadequate. Looking at the results in hindsight, we can use the UCM and ECM methods for slow, steady growing laggard stocks, but then that brings extra uncertainty in classifying a stock. Variations of the type of ARIMA models provided the best predictions, with first-order differencing and MA (1) proving to be adequate for most of the stocks.

Overall, our model predictions resulted in the creation of a portfolio that outperformed the common retirement portfolios for a duration of 2005 to 2018.

1. **Appendix**



**Fig A.1:** ARIMA Model Results Summary

Graphical user interface

Description automatically generated

**Fig A.2:** ESM Model Results Summary

A picture containing graphical user interface

Description automatically generated

**Fig A.3:** UCM Model Results Summary