Abstract

Forecasting for DJI June 2016-June 2017

Dow Jones Industrial Average Forecasting

DATA MINING AND BUSINESS INTELLIGENCE PROJECT-opim 5671 SUMMER 2016

Chen Chen

Honglun Su

Satinder Sondhi

Siddharth Kajampady

**Table of Contents**

Introduction………………………………………………………………………………………..2

Objective…………………………………………………………………………………………..2

Data Preparation…………………………………………………………………………………...2

Exploratory Data Analysis………………………………………………………………………...4

Modelling Approach………………………………………………………………………………5

Pre-whitening & Cross Correlation……………………………………………………………….6

Final Model………………………………………………………………………………………10

Model Comparison……………………………………………………………………………….12

Recommendations & Business Insights………………………………………………………….12

References………………………………………………………………………………………..14

**INTRODUCTION**

DJIA is an index made up of 30 large publicly owned companies in the United State and it is the most influential stock index in the world. The DJIA index is not the average price of the related companies, but the sum of the prices divided by a factor, which changes when the stock split or dividend. The data contains Dow Jones industrial average index, crude oil price, gold price and number of Jobs from January 1987 to May 2016.

**OBJECTIVE**

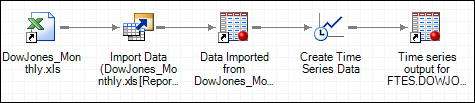
The objective is to predict the DJI closing for the next 12 months (Staring June 2016) based off from the available historical data of DJI closing, crude oil, gold, and job market by applying SAS forecasting models. Investors may benefit from this forecasting.

**DATA PREPARATION**

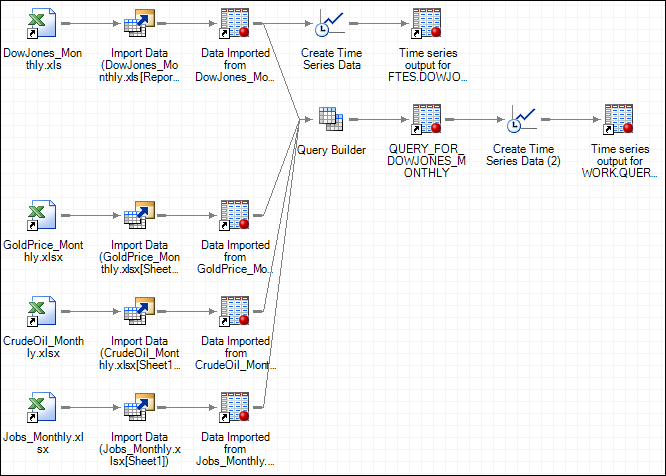
Process created using SAS Enterprise Guide:

* To import the datasets from initial CSV into SAS dataset. The excel file containing the Dow Jones monthly index values was imported into SAS. (please refer to appendix for more information regarding the dataset)
* Created a Time series data for the dependent variable ‘Dow Jones Closing value’ using the ‘Create Time Series Functionality’
* Once we realized that Gold, Crude Oil and Job Growth can be good predictors of Dow Jones index values, we imported the excel files for these 3 datasets into SAS and combined them.
* Once again, we created time series data on the combined dataset to make it available for forecasting.
* The above steps gave us the final data set with 5 columns, which includes the TimeID variable (Date), 3 independent variables (Job Growth, Crude Oil price in dollar, gold price in dollars and the dependent variable ‘Dow Jones Closing Value’

The process used to create the Dow Jones Time series data is shown below-



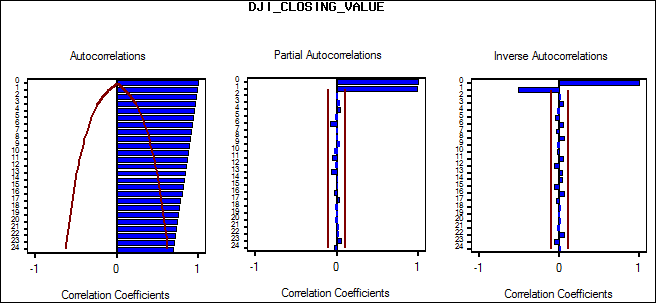
The process used to combine all independent variables and the Dow Jones monthly data is shown below-

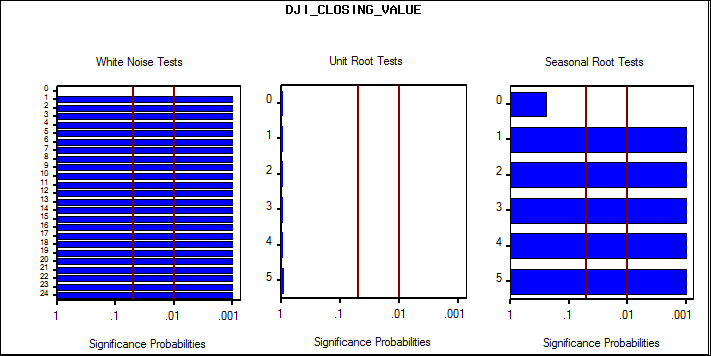


# 

# **EXPLORATORY DATA ANALYSIS**

After processing the data, the initial target time series was plotted to check the correlations, trends and seasonality. The trend and correlation analysis is given below:





From the above graph, we can observe the decaying pattern in ACF plot. Also, the white noise test shows significance at every lags, hence we can reject the null hypothesis that prediction error represents white noise. The unit root test shows very little significance, therefore we can accept the null hypothesis that time series is not stationary.

# 

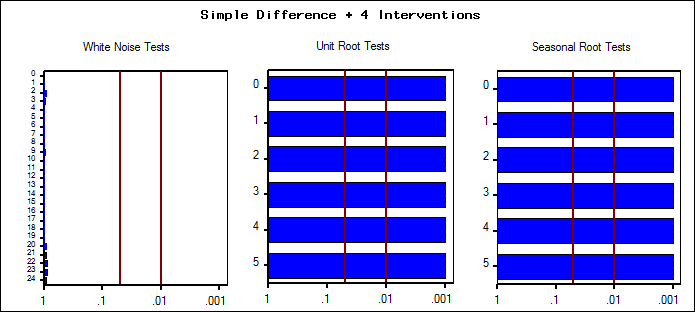
# **MODELLING APPROACH**

After checking the plots for our data sets we decided to first run the simple differencing model and improvise as we move further. After applying the Random Walk with Drift, we found out that the stationarity and trend improved but the forecast did not look promising.

In order to further improve the forecasting abilities of our model, we decided to inspect the major events that had a significant impact on the value of Dow Jones Industrial Average. After recognizing the events, we added them as different Interventions based on their significance. We came across four such events:

1. August 1998 (Point)- Russian Economic Uncertainty and Fall in Russian Ruble
2. September 2008 (Point/Exponential)- US Financial Crisis
3. March 2009 (Ramp)- Stocks Bottomed Out after the Crisis (Best Buying Opportunity in a decade).
4. July 2015 (Point/Wave)- Significant Drop in Chinese Stock Markets and Uncertain US Monetary Policy

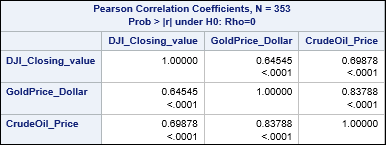
After adding the interventions, the residual series improved significantly but the quality of the forecast did not improve much.



So, we decided to add some Dynamic Regressors to our model. Based on our research and expertise we selected three input variables:

* Gold Prices- Month End Price for Gold
* Crude Oil Prices- Month End Price for WTI Crude Oil
* Job Growth- Number of Additional Jobs Added from Last Month

But before adding the Dynamic Regressors to our models, we tried to check the correlation of input variables with the output variable.



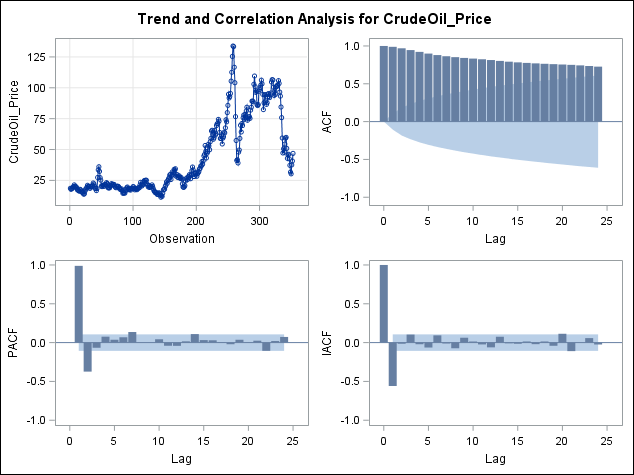
The Jobs Growth had the lowest correlation with DJIA value- around 12%, that is why we decided to not to go ahead with adding Jobs Growth as one of our input variables.

**PREWHITENING & CROSS-CORRELATION**

**Prewhitening:** When we try building a model with dynamic regressors, we need to perform certain filtering operations on the input series. This filtering is implemented using prewhitening.

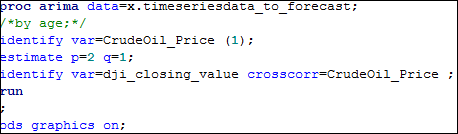
As input series are themselves auto correlated, the direct cross-correlation function between the input and response series gives a misleading indication of the relation between the input and response series. To proceed with prewhitening, we need to first looks at the plots of the input regressors and then use appropriate filters to the series.

The preliminary plots of the input regressor ‘CrudeOil\_Price are shown below-

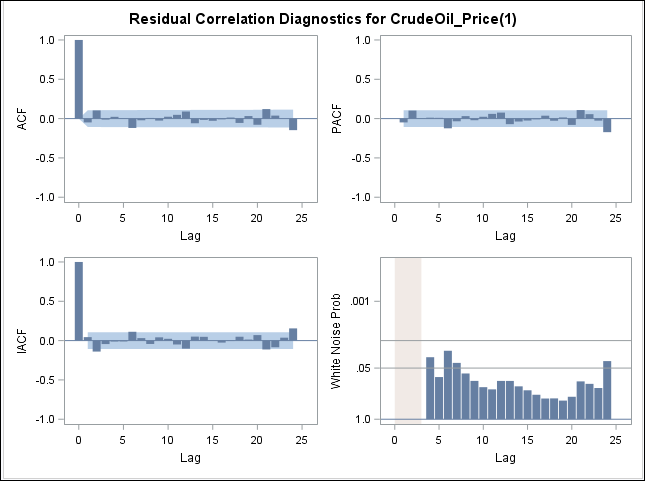


From the above plot, we can observe that there are significant spikes in ACF, PACF and IACF plots. This suggests that the series is auto-correlated with itself.

Prewhitening helps to eliminate these auto correlations and reduce the residuals to white noise. The intention of prewhitening is to make the input time series stationary. This has been implemented using the PROC ARIMA procedure. The ARIMA procedure used to prewhiten the ‘CrudeOil\_Price’ variable is as follows-

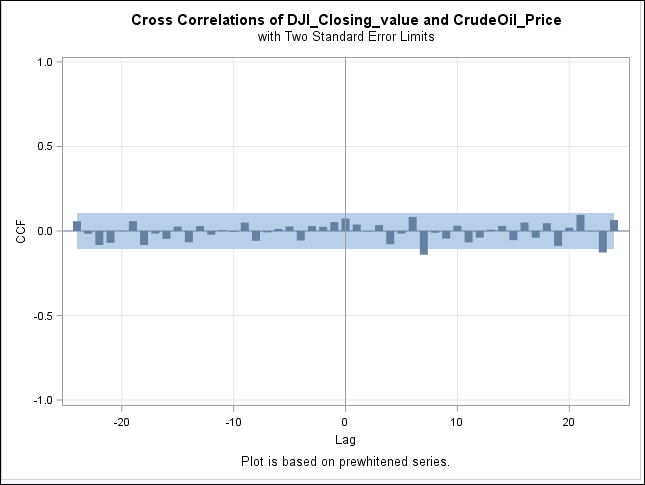


Once the variable ‘CrudeOil\_Price was filtered with AR=2 and MA=1 and a first order differencing, the residual diagnostic plots are as given below-



The lags for ACF, PACF and IACF are all insignificant with one or two lags which may be considered spurious. Also the white noise test suggests that the residual series is white noise. Thus, we have removed the trend and succeeded in making the input series stationary.

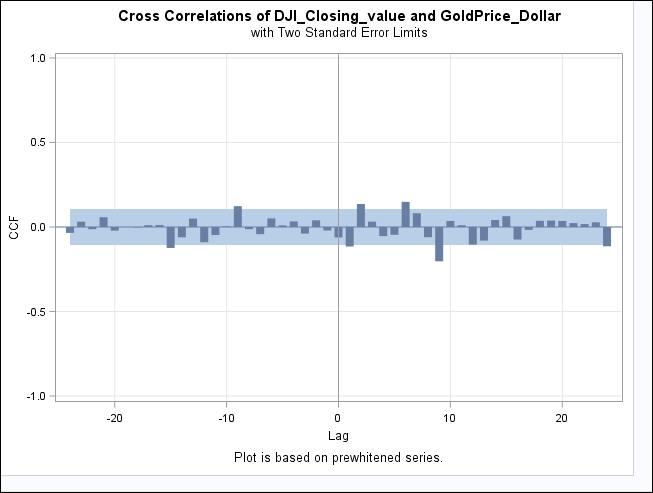
The cross-correlation plot for ‘CrudeOil\_Price and the response series is given below-



There are few significant lags in the cross-correlation plot. The significant spike at lag 7 suggests that the present value of response series ‘DJI\_closing\_value’ is influenced by the 7th lag of the variable ‘CrudeOil\_Price’.

Based on the cross- correlation plot, a suitable transfer function has been chosen and is given by 1$ (1)/(1).

Similary, the other input regressor ‘GoldPrice\_Dollar’ has been prewhitened using suitable filters and based on the cross correlation plot, an appropriate transfer function was determined. The CCF plot is shown below-



Once the transfer functions for both the input regressors are obtained, the next step is to build an ARIMAX model to forecast the response variable.

The PROC ARIMA procedure used to build the model is given below-

**proc** **arima** data=x.final\_timeseries\_data ;

identify var=dji\_closing\_value crosscorr= (goldprice\_dollar (**1**) CrudeOil\_Price (**1**));

estimate p=**2** q=**0** input=(**1** $ (**1**)/(**2**) goldprice\_dollar **1** $ (**1**)/(**1**) CrudeOil\_Price) noconstant plot ;

forecast id = date LEAD = **12** interval = month out=x.output printall;

**run**;

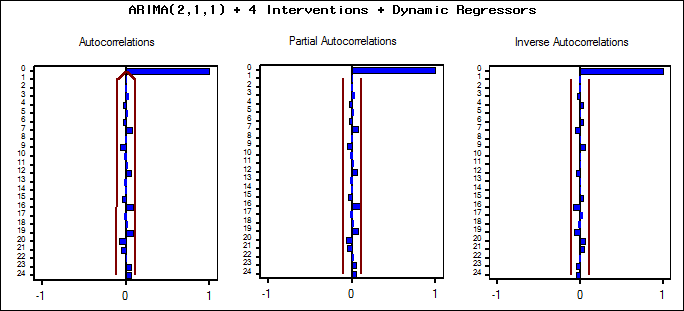
This code forecasts the Dow Jones Index values for the next 12 months.

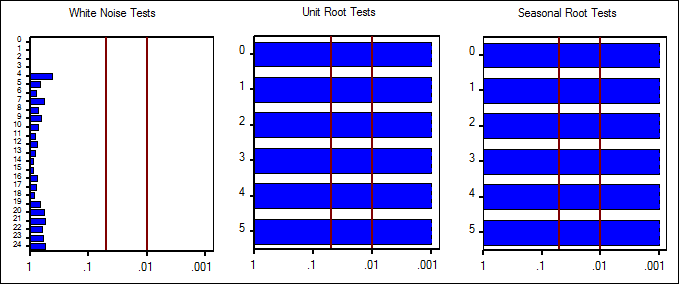
The forecasts provided by the ARIMAX model developed in SAS Base was not giving excellent results. The 95% confidence intervals had a wider interval and forecasted results were not good as the model we built earlier using a seasonal Difference and the 4 interventions.

**FINAL MODEL**

After obtaining the transfer function from the CCF plots, we decided to run them in TSFS. The final model that we decided was ARIMA (2,1,1) with 4 Interventions and Dynamic Regressors.

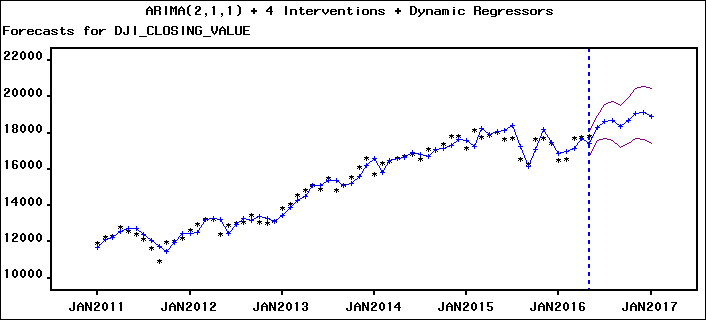
The diagnostic plots and the forecast results looked promising and was by far the best model we got. The below shows the results of Ljung Box and Box-Jenkins test of the new model.



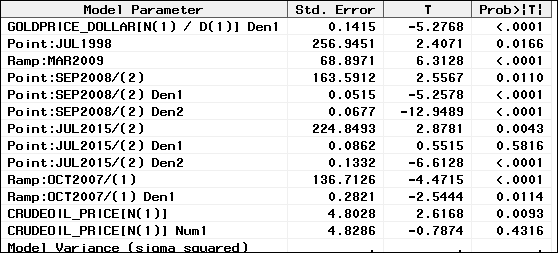


The above tests clearly show that there are no significant spikes at any lags in ACF & PACF plots. Also, the unit root/stationary tests suggested that the residual series is white noise, also the series is stationary.

This model provided a better forecast plot as given below-



The forecast plot passes a face validity test and also the parameter estimates are significant as shown below-



**MODEL COMPARISON**

The model accuracy statistics were calculated for both models and the results are shown below-

|  |  |  |
| --- | --- | --- |
| **Fit Statistics** | **Model without Regressors** | **Model with Regressors** |
| Mean Square Error | 131085.6 | 113464.5 |
| Root Mean Square Error | 362.05741 | 336.84491 |
| Mean Absolute Percent Error | 3.1718 | 3.06272 |
| Mean Absolute Error | 262.32384 | 246.31363 |
| R-Square | 0.994 | 0.994 |
| Akaike Information Criterion | 4140.3 | 4111.7 |
| Schwarz Bayesian Information Criterion | 4171.1 | 4185 |

After adding the dynamic regressors Crude and Gold to the model, there is a significant decrease in the RMSE value and also MAPE has dropped down. The SBC value has slightly increased, since it takes the model variables into account. Overall, after adding the dynamic regressors, there has been a significant improvement in the model fit statistics.

# 

# **RECOMMENDATIONS AND BUSINESS INSIGHTS**

Accurately predicting the movement of a stock market index is a billion-dollar problem and many renowned Economists and Data Scientists are trying to solve that. We tried our best to predict the future movement of DJIA but still aren’t confident enough to pitch it to institutional investors. There can be various other input variables that can go into the model like the Fed Funds Rate, US Treasury Yield, 10-Year US Bond Yield, Housing Starts, Debt-to-GDP ratio etc. and we will wish to add few more input variables before we can be absolutely confident about our model’s predicting capabilities.

With that being said, the significance of this model will also depend upon an investor’s horizon and risk taking abilities. For a short-term trader, a model with 97% accuracy can still result in huge losses, but an investor with a longer horizon might realize the upward-linear trend of the next 12-month forecast and start investing today. Even a risk-averse investor can utilize our model and exploit Dollar Cost Average method. It means that based on our forecast he can invest a portion of his money today and then keep on investing as there are smaller corrections in the markets.

To conclude, I will say that it’s not the 97% accuracy of the model that matters, but the 3% inaccuracy that can result in huge losses.

As a great investor once said,

“Markets are like Weather, if you don’t like it just wait for a day and it will change…;)”

High accurate sales forecasting model, Investors would be able to

* Accurately estimate the future movements of Dow Jones Industrial Average
* Guarantee sufficient product will be manufactured or ordered to service customers on a timely basis resulting in better service
* Better inventory management to avoid overstock and production if any
* Accurately predict revenue and profit. Explore possibilities to increase both revenue and profits.
* Have a better control on the internal operations
* Sales can be managed on a thoughtful basis
* Schedule promotions or discontinue slow moving products

**REFERENCES**

**Data reference:**

Crude Oil: <https://www.eia.gov/dnav/pet/hist/LeafHandler.ashx?n=PET&s=RWTC&f=M>

DJIA: <http://us.spindices.com/indices/equity/dow-jones-industrial-average>

Gold: <http://www.gold.org/research/download-the-gold-price-since-1978>

Job: <http://www.bls.gov/data/>

**Others**

<https://onlinecourses.science.psu.edu/stat510/node/75>

<http://support.sas.com/documentation/cdl/en/etsug/63939/HTML/default/viewer.htm#etsug_arima_sect033.htm>