

Predicting stock price from a time series analysis entails quality data, meticulous model selection, and rigorous assessment of model performance.

### **Description of Data**

Yahoo and Google provide information on General Electric stock dating back to 1962. The Adjusted Closing Price (ACP) includes corporate actions, distributions, and equity value for a specific stock and is often the de facto metric for conducting time series analysis (investopedia.com). General Electric's ACP is the response variable in the time series analysis. Prior research recommends including Real Economic Output, Interest Rate, Real GDP, Inflation, CD Rates, and CPI. The Federal Reserve Bank of St. Louis (FRED) is the primary data repository for accessing these predictor variables. In addition, there are more than fifty additional predictor variables within FRED for use in the initial assessment. Since data from the time series variables is released at different time intervals, aggregating the data quarterly allows the analysis to capture the most variables. It is expected that 100 unique observations are sufficient for a time series that includes testing and training sets. Automated variable selection and erudite intuition will be used for variable reduction. The data is currently housed in a CSV file in Microsoft Excel and will be imported into R via Quantmod, which simplifies quarterly aggregation of the variables.

### **Modeling Methods**

Meticulous modeling entails multiple modeling techniques along with iterations at the individual model level. When modeling economic and financial data, time varying volatility is more common than constant volatility. Autoregressive and moving average (ARIMA) models use past history to model conditional expectations of a process. Generalized autoregressive conditional heteroskedasticity (GARCH) models address generalized conditional heteroskedasticity that is in economic data. It is common for the explanatory variable to be expressed as the proportion gained or the log ( $r_t/r_{t-1}$ ) (Shumway, 2006) (Vose, 2008). When auto regressive errors have a time series structure and the error

terms are not independent, ARIMAX modeling is used to accept exogenous variables. This inclusion of regression on autoregressive errors extends to vector auto regressive models and GARCH models as well.

Support Vector Machines (SVMs) and Random Forests are often used for predicting stock prices based on their ability to overcome the noisy nature of the data involved in stock price prediction. Random Forests filter available parameters, while the SVM is applied against the most effective predictors. The R language provides a number of packages for developing predictive models based on SVMs and Random Forests. The `e1071` package contains functions and objects for SVMs, and the `randomForest` package contains functions for building Random Forests. The modeling methods include, ARIMA, GARCH, ARIMAX, SVM, and Random Forest, but more models may be included if further research allows.

### **Assessing Model Performance**

Model performance will be assessed using time series cross validation in a test and validation regimen. First, the model is fitted to sample data. The sample data is the minimum length needed to fit the model at the earliest observations in time. The next observation is forecasted from the fitted model and the error is collected. This process is repeated in a forward chain. At each iteration, the training data increases in size by one observation until the length of the time series is exhausted. Root mean square error (RMSE) is calculated using the collected error terms. The RMSE is minimized during the parameter estimation process and determines the width of the confidence intervals for predictions. The RMSE is preferred over mean squared error (MSE), because the metric is the same units as the data and is representative of the size of the typical error. The best value of RMSE is dependent on the variable's measurement units and the degree of forecasting accuracy.

The time series analysis will have a solid predictive foundation based on accurate data, different models, and multiple assessment metrics.

Hyndman, Rob. "Hyndsight." *Hyndsight*. N.p., 4 Oct. 2010. Web. 2 Nov. 2013.

<<http://robjhyndman.com/hyndsight/crossvalidation/>>.

Hyndman, Rob J, and George Athanasopoulos. "Forecasting: Principles and Practice." Otexts. N.p., n.d.

Web. 10 Oct. 2013. <https://www.otexts.org/fpp/>

Levine, Helisse, and Jonathan B. Justice. *Handbook of local government fiscal health*. Burlington, MA: Jones & Bartlett Learning, 2013. Print.

Shumway, Robert H., and David S. Stoffer. Time series analysis and its applications with R examples. 2nd [updated] ed. New York: Springer, 2006. Print.

Vose, David, and David Vose. *Risk analysis: a quantitative guide*. 2nd ed. Chichester: Wiley, 2000. Print.