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

Description of Data

Information on General Electric stock is available from 1962 to current daily values from websites including Yahoo and Google. The Adjusted Closing Price (ACP) includes corporate actions, distributions, and equity value for a specific stock and is often the de facto metric when conducting a historical time series (investopedia.com). General Electric’s ACP is the response variable in the time series analysis. Within stock prediction analysis, there are many predictor variables available and prior research findings coupled with a thorough exploratory variable analysis are the joint rudders steering the time series through the sea of potential variables. Prior research recommended 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 the predictor variables. In addition to the six variables mentioned earlier, there are more than fifty additional predictor variables that will be initially assessed found within FRED. Data from the time series variables are released at different time intervals and aggregating data at the quarterly level allows the analysis to capture the most variables. 100 unique observations is sufficient for a time series and the data will be divided into testing and training sets. Some variables will require data transformations based on the integer value compared to the predictor variable. Automated variable selection will be used along with erudite intuition for variable reduction. The data is currently housed in a CSV file in Microsoft Excel and will be imported into R via Quantmod. An advantage of utilizing Quantmod entails its ability to aggregate the variables at the quarterly time interval.

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 (ARMA) models are used to model conditional expectations of a process given the past, but ARMA models have conditional variance. GARCH models allow one to model heteroskedastic distributions of financial markets and express the variance as well as the explanatory variable as the proportion gained or the log (rt/rt-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. Support Vector Machines (SVM) 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 Forest is used to filter available parameters and the SVM is applied against the most effective predictors. The R language provides a number of packages that can be used 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 ARMA, ARIMA, GARCH, ARIMAX, SVM, and Random Forest, but more models may be added upon further research.

Assessing Model Performance

Precision and Recall are often used to determine the effectiveness of classification models, as this information is readily available through a confusion matrix. Also, the F-statistic, which can be generated from the Precision and Recall measurements, is used to evaluate models. In addition, ROC curves or a measurement of the Area Under the ROC curve provides an additional method of evaluation. When comparing classification verses regression models, a rolling window approach will be used comparing the testing to the training data. (Testing to Training, dataset)

Extra Stuff

Typically, the Random Forest is used to filter the available parameters--given their noisy nature--and the SVM is applied against the most effective predictors. Precision and Recall are often used to determine the effectiveness of the models, as this information is readily available through a confusion matrix. Also, the F-statistic, which can be generated from the Precision and Recall

In AR modeling, the dependent variable is rt and the explanatory variables are some number of previous observations of the variable; rt-1, …, rt-k. Within MA modeling, the dependent variable is rt and the explanatory variables are some weighting of the q previous terms from the mean. ARMA models assume variance of the error terms, and when violated generalized conditional heteroskedasticity comprises model integrity.

Research shows that stock return data has heavy tails and outlier prone probability distributions (Ruppert 2010). ARMA and GARCH models require stationarity, and differecing is often used. In terms of stocks, this can be considered the daily change or rt – rt-1 and this is often enough to make a time series stationary.

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 regression with autoregressive errors is referred to as ARIMAX where X represents eXogenous variables. One must use caution when using ARIMAX models as they tend to reduce the in sample error. Arimax modeling often reduces the in sample error and easily over fits the model (Levine, 2013).

The ARIMAX and GARCH models don’t allow for interaction between forecast and predictor variables, they only allow for the predictor variables to influence the forecast variable. In cases where two way interactions occur the Vector Autoregressive (VAR) can be used. The VAR model is a generalization of the univariate AR model except now the model forecasts a vector of time series. The first n lags of each time series in the system of equations is used as regression predictors for each forecast variable. There is one forecast variable for each time series. (Hyndman, 2013) VAR models can be generalized into VARMA and VARIMAX models.

Neural Networks embed previous lags into the input nodes and train a feed forward network. (Hyndman, 2013) MA terms can be added using recursive neural network but this functionality is not readily available in R at this time. Research shows that stock returns have heavy tails or outlier prone probability distributions (Ruppert 2010).

Support Vector Machines (SVM) and Random Forests are often used for predicting stock prices. They are particularly useful, because the noisy nature of the data involved in stock price prediction often reduces the efficacy of machine learning algorithms such as Artificial Neural Networks. The R language provides a number of packages that can be used for developing predicitve models based on SVMs and Random Forests. The e1071 packages contains functions and objects for SVMs, and the randomForest package contains functions for building Random Forests. Typically, the Random Forest is used to filter the available parameters--given their noisy nature--and the SVM is applied against the most effective predictors. Precision and Recall are often used to determine the effectiveness of the models, as this information is readily available through a confusion matrix. Also, the F-statistic, which can be generated from the Precision and Recall measurements, is used to evaluate models. In addition, ROC curves or a measurement of the Area Under the ROC curve provide an additional method of evaluation.

Assessing Model Performance

Conclusion

References