Predicting Oil Prices with Sentiment Data

David Hirschey

Spring 2016

Introduction

This report demonstrates a predictive model of crude oil prices through a profitable trading strategy. Behind the trading strategy is sentiment data, which aims to quantify human emotion. In this case, sentiment data are numbers on a (-1, 1) interval recorded across time for emotions like fear, optimism, or uncertainty. The model also includes similar data for economic ideas like demand, supply, and price expectation. To calculate a sentiment score, computer driven algorithms evaluate the frequency and relevance of key words in news articles and social media posts. These data are collected by MarketPsych, who follow a growing group of 8,000 edited news sources and 2,000 individuals on social networks.

Model

Behavioral Economics

The behavior of traders motivated by news and social media may be predictable through news and social media sentiment data. For example, widespread news of violence in the Middle East might scare oil traders into expecting rising prices, even though the violence would have little eventual effect on the supply and demand for oil. Recognizing this as emotional behavior, a knowledgeable trader would sell oil at the excessive price. This model similarly assumes negative or positive relationships between sentiments and prices. The directional associations between individual sentiments and oil prices in this model are chosen after individually testing sentiments with the model and observing the correlation. These are displayed in Figure 1 of the Appendix.

Interpreting Sentiment Data

Sentiment data can be difficult to interpret because the data may vary wildly form day to day. To extract signal form noise, this model uses exponential smoothers, popularized by Brown, Holt, and Winters in the 1940s. A simple exponential smoothing algorithm produces a number representing the recent average value of a data sequence by weighting observations with a smoothing factor. As the smoothing factor approaches one, the smoother matches the original data.

$S_0 = X_0$

$S_t = \alpha * X_t + (1 - \alpha) * S_{t-1}, t > 0$

(α is the smoothing factor, $0 < \alpha < 1$, S is the smoothing curve, and X is the data series)
Comparing two smoothing curves of the same data but different smoothing factors gives a sense of momentum in the data, as a positive movement in the data would have greater influence on the smoother with smaller smoothing factor. The result of comparison is a binary series that tells which smoothing curve has higher value than the other.

Creating a Trading Strategy

This model predicts an intraday increase in oil price when the smoothing curve with higher smoothing factor exceeds the secondary curve, given a positive association between oil prices and the sentiment modeled. Negative association would imply an intraday decrease in price. The higher valued smoothing parameter is selected to minimize the one step ahead prediction error of the smoothing curve to the sentiment score. The second smoothing parameter is selected to maximize the correlation between the resulting binary series and the daily oil price change. After creating comparison smoothing models for each sentiment, the model takes a mixed long and short position balanced according to the average prediction over all individual

sentiments. If 60% of sentiment signals expect a price increase, the model invests 60% long and 40% short. The oil prices used are the closing prices of CME Group's West Texas Intermediate Crude oil futures. Transaction costs of trading are assumed zero.

Empirical Analysis

Separating the Training and Testing Sets

Validating a model that chooses optimal parameters across time requires a training period and testing period. Data throughout the training period are used to optimize parameters, in this case the smoothing parameters. Throughout the testing period, the model creates next day predictions using only past data and the parameter values picked during the training period. When measuring the model's performance, only the testing period results are credible, because accuracy in the training period was deliberately optimized.

Model Performance

Optimizing parameters over 2011 through 2014 and testing in 2015 delivers compounded returns exceeding the baseline oil price by a 46% margin over the testing period. Equating a majority long position to predicting a price increase over the next day, the model correctly predicts next day closing prices 193 of 252 trading days, a 76% success rate. If the model had no predictive value, the trading strategy would perform no better than a coin flip. Random chance would almost never deliver better results, corresponding to a p-value near 2.2×10^{-16}. These results suggest very strong evidence for the model's merit. An equity curve demonstrating the growth of one dollar continually reinvested in the strategy is shown in Figure 2 of the Appendix.

Discrepancy in the model's accuracy between long and short positions would indicate bias in the model methods. However, the model accuracy remains consistent across majority long or majority short positions.

Statistical Inference

The 2015 testing set performance makes a powerful case for model. Over the 2011 to 2014 training period, oil prices remained stable without much trend, but declined sharply near the end of 2014 and into 2015. This dramatic difference, likely caused by production surges from the Organization of Petroleum Exporting Countries, could be reasonably expected to tarnish the testing set results. A model excessively tailored to the training set could mishandle the testing set data if assumed relationships that held in the past changed along with the paradigm shift. Fortunately, sentiment based data and modeling appear robust to these effects. During a year when the price of oil tumbled, this model continued to profit.

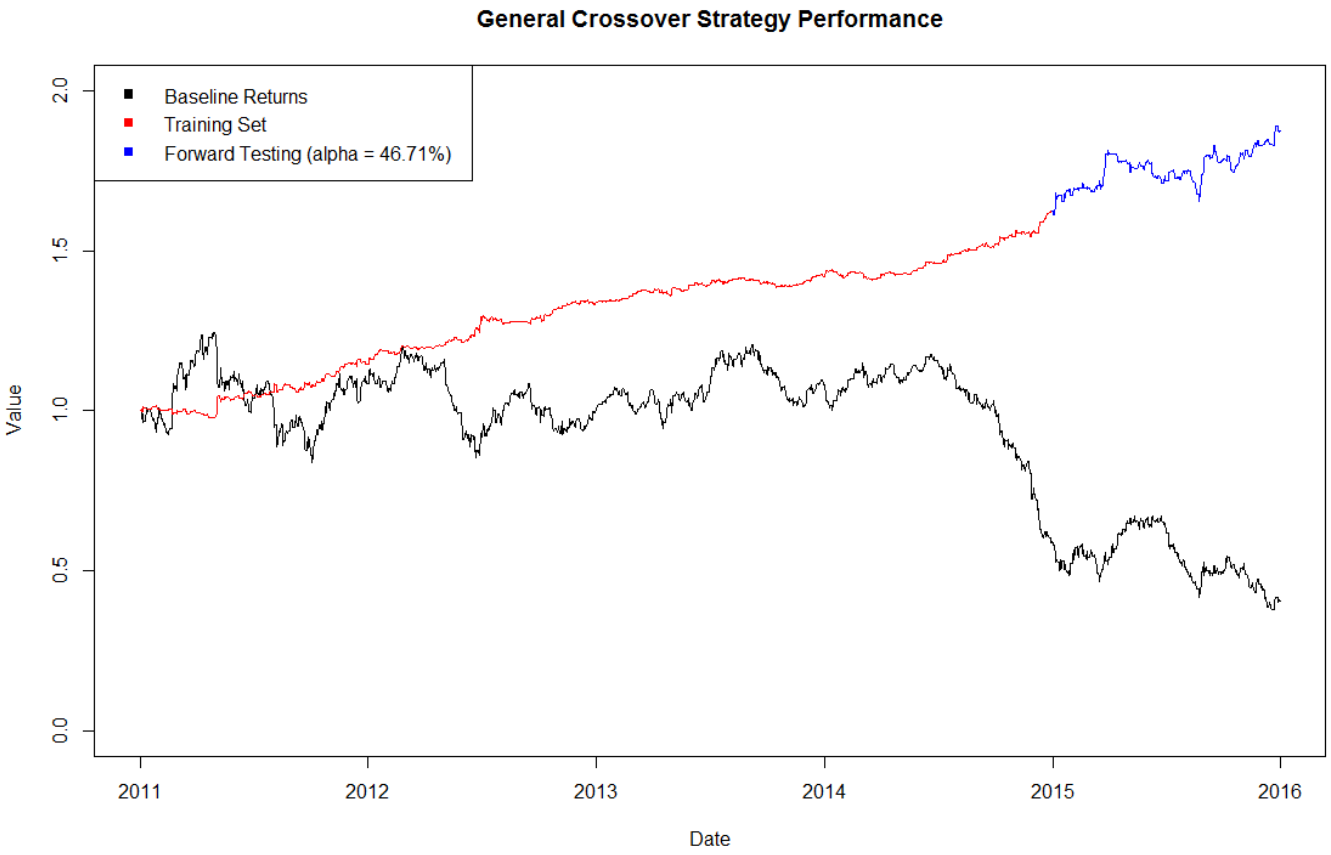
Appendix

Figure 1

Sentiment Name	Empirically Assumed Correlation
Buzz	Negative
Relative Buzz	Positive
Sentiment	Negative
Optimism	Positive
Fear	Positive
Joy	Positive
Trust	Positive
Violence	Negative
Conflict	Negative
Gloom	Positive
Stress	Positive
Time Urgency	Positive
Uncertainty	Positive
Emotion vs Fact	Positive
Long Short	Negative
Long Short Forecast	Positive
Price Direction	Negative
Price Forecast	Positive
Volatility	Negative
Consumption Volume	Positive
Production Volume	Positive
Regulatory Issues	Positive
Supply vs Demand	Positive
Supply vs Demand Forecast	Positive
New Exploration	Positive
Safety Accident	Negative

The associations shown above were selected by individually testing sentiments in the model scheme.

Figure 2



The colored line above shows compounded returns of the trading strategy given a unit starting value. The black line shows the price of CME Group West Texas Intermediate Futures contracts at closing. Alpha refers to the excess return over the baseline return for the 2015 testing period.

Notes

The R file "David Hirschey MP Model R Code.R" requires the RPostgreSQL package to request data from the MarketPsych server, though the server details have been omitted. The remaining code defines background functions ultimately called by the final plotting function, giving a one-line execution after all functions have been defined.