Text-based Crude Oil Price Forecasting

Central University of Finance and Economics

Fan Jia

Nine 12th , 2019

□ Introduction:

Forecasting Methods: Time Series, Econometrics

$$\hat{Y} = f(X) + T$$

T: Sentiment Score, Topics, Word Embedding, Sentence Embedding

Forecasting Model: Long-Short Term Memory(LSTM)

□ Variables:

Target : y_t ——WTI spot price

1. $\boldsymbol{y_t}$: WTI spot price → F(y_t);

2. $\boldsymbol{T_t}$: Text Variable → F(T_t) ;

3. $\boldsymbol{M_t}$: Marketing Variables, including futures, DIJA, USDA and google→ F(M_t) ;

4. $\boldsymbol{S_t}$: Sentence Embedding Variables

S_{1t}—SIF Embedding Variables ,

S_{2t}—Power Mean Embedding Variables ,

□ Model:

$$\boldsymbol{y_t = F(y_t) + F(T_t) + F(M_t) + S_t}$$

1. $\boldsymbol{y_t = F(y_t)}$ --ARIMA，LSTM
2. $\boldsymbol{y_t = F(y_t) + F(M_t)}$--Market
3. $\boldsymbol{y_t = F(y_t) + F(T_t)}$--Text
4. $\boldsymbol{y_t = F(y_t) + F(T_t) + F(M_t)}$--cross impact analysis
5. $\boldsymbol{y_t = F(y_t) + F(T_t) + F(M_t) + S_t}$--Sentence Embedding

□ Result:

Table 1: Performance of different models(RMSE)

Proportion of Testing Datasets		**20%**	**30%**	**40%**
Time Series Model	ARIMA	0.0144	0.0128	0.0121
	Auto Arima	0.0141	0.0131	0.0124
	ETS	0.0137	0.0127	0.0121
LSTM	WTI	0.0143	0.013	0.0123
	WTI + M	0.0084	0.0082	0.0087
	WTI + T	**0.0153**	**0.0144**	**0.0138**
	WTI + M + T	0.009	0.0094	0.0094
	WTI + M + T + S_1	0.0079	0.0079	0.0083
	WTI + M + T + S_2	**0.0072**	**0.0077**	**0.0076**

□ Update:

Table 2: items updated

Target: y_t	**Before**	**Update**	**Notes**
Numeric Variables	$M = \{M_t, M_{t-1},\dots M_{t-p}\}$	$M = \{M_{t-1},\dots M_{t-p}\}$	more practical
Corpus	yearUnited', 'U.K.', 'kkkkkkk'	token	cost a long time
Topic	LDA	to ensure converge	
Sentiment Score	TextBlob	TextBlob(to be comparable)	few methods available
Word Embedding	trained by our full text	Google (d = 300)	larger corpus, more precise
Sentence Embedding	S= $\{S_t$, d = 80}	S= $\{S_{t-1}$, d = 300}	more practical
lag	VAR	Granger Causal Text	may lag more

□ Appendix:

1. 20%: ARIMA(5,2,0)

	drift	s.e.
Coefficients:	-7e-04	5e-04

sigma^2 estimated as 0.0002284
log likelihood=2352.3
AIC=-4700.61 AICc=-4700.59 BIC=-4691.12

2. 30%: ARIMA(2,1,1)

	ar1	ar2	ar3	ar4	ar5
Coefficients:	-0.8791	-0.7066	-0.5423	-0.3168	-0.1201
s.e.	0.0296	0.0384	0.0407	0.0384	0.0296

sigma^2 estimated as 0.000228: log likelihood=3136.73
AIC=-6261.46 AICc=-6261.38 BIC=-6231.28

3. 40%: ARIMA(0,1,0)

	ar1	ar2	ma1
Coefficients:	0.9159	0.0570	-0.9598
s.e.	0.0428	0.0321	0.0293

sigma^2 estimated as 0.0002092: log likelihood=2787.74
AIC=-5567.48 AICc=-5567.44 BIC=-5547.89