TO_DO_modelling

Nicholas Mitchell

March 4, 2016

Contents

1 DONE Clean up and improve data

1. **DONE** Remove duplicate (integer) columns from sentiment analysis results
2. **DONE** Add some more macroeconomical data:
   (a) **DONE** oil prices -> one month forward month -> both ICE Brent and WTI Crude

(b) **DONE** natural gas -> NYMEX Naturla Gas Futures; 1-month forwards

(c) **DONE** copper cash price [additionally added bid/ask spread as vola proxy]

(d) **DONE** Volatility from VIX (S&P500)

(e) **DONE** Asian markets: Nikkei 225, Shanghei SE Composite, ETF for Emerging Markets ("EFA")

(f) **DONE** Currencies:

i. **DONE** USD_{EUR}

ii. **DONE** USD-GBP

iii. **DONE** USD-JPY

iv. **DONE** USD-AUS

v. **DONE** USD-CAD

3. **DONE** Remove duplicate (integer) columns from sentiment analysis results

4. **DONE** Extra search terms from Twitter?

5. **DONE** Merge all data into one data table

6. **DONE** Remove variables of high-covariance/correlation

(a) **CANCELLED** function: nearZeroVar() {caret} - see: "Building Predictive Models in R Using the caret Package.pdf"

(b) Used the findCorrelated() function from caret

(c) Also removed the Dow_{SPDR} data, as it had zero variance over many time periods within the time-series

2 DONE Research alternative methods for imputing data

1. What are the underlying assumptions of our model?

2. Are we altering or breaking our assumptions by the way we impute data?

3. Can we use several methods and compare the output of the model?

(a) If we do this, we need to be careful about the interpretation. One model may return brilliant results, but without robustness, i.e. the data does not correspond to the assumptions used in the model, or the data way 'fixed' to produce better results

4. Current Imputation Methods *<2015-12-09 Wed>*

Method	Weekends	is.monday	was.weekend.pos	# Lags	test.ID	Notes
LOCF	remove all	used	not usec (yet)	1:5	$ds_{\text{locfL}[1}$:5]	First value fo Nikkei225: taken from preceding Fri
Spline	remove all	used	not usec (yet)	1:5	$ds_{\text{splineL}[1}$:5]	
Predictive Models	remove all	used	not usec (yet)	1:5	NOT YET DONE	Using a separ model for the to model itse replacing NA
kNN	remove all	used	not usec (yet)	1:5	$ds_{\text{caretL}[1}$:5]	In {caret}: PreProcess
bagImpute	remove all	used	not usec (yet)	1:5	$ds_{\text{caretL}[1}$:5]	In {caret} PreProcess

3 DONE Run using parallel processing

1. **DONE** doMC() for Mac
2. **DONE** doParallel for Windows

4 STARTED Define a simple model that can be used to optimise the parameters

1. **DONE** glmboost() -> use Niko's examples for a quick start
2. **STARTED** Work through the 'mboost$_{\text{tutorial}}$'
3. **DONE** Ensure that the script works on both windowns and Mac -> doParallel() for Windows
4. **TODO** Use different loss functions:
   (a) NOTE: this can be done within caret by specifiying custom models - see: `http://topepo.github.io/caret/custom_models.html`
      i. least squares
      ii. least absolute deviation
      iii. Huber –> can be done within mboost, using family =
      iv. Quantile
      v. Possibly not available (meant for classfication):

A. Exponential loss -> e.g. only possible for binary classification, tree depth of two?
B. Binomial deviance
C. Multinominal deviance

5 DONE Create simple loop to perform cross validation

1. Use fixed window size - between 25 and sixty days Find some sources for this claim. Where is it used? Why is it better than increasing window size?

6 STARTED Implement model in optimisation functions

1. Note: the glmboost() function doesn't work in caret for optimisation This function is therefore used to test the water, manually and get some results to act as 'control' for optimsation within {caret}
2. **STARTED** define a testGrid of all parameters:
   (a) n-iterations???
   (b) tree.depth
   (c) shrinkage
   (d) Model specific:
      i. number of observations in node after split: "n.minobsinnode"
      ii. sample-size for stochastic boosting

7 STARTED Significance of weekend sentiment

1. **STARTED** Can we use another dummy variable to model the market movements on mondays?
2. **STARTED** Aggregate the weekend SA scores to DV: 'was.weekend.positive' -> see notes in book
3. Plot the weekend sentiment against movements in markets on Monda

8 STARTED Plot results:

1. **STARTED** loss function, using cvrisk()
2. Plot showing how the approximation becomes better over more iterations (reduced variance if stochastic GB?)
   (a) Graded colouring for more and more iterations (see: `https://youtu.be/IXZKgIsZRm0?t=17m1s`)
   (b) This might also work using the alpha = I(1/10) in ggplot, somehow!
3. Interesting data exploation plots of most significant predictors (e.g. Peter Prettenhof, sci-kit-learn)
4. ROC -> sensitivity vs. specificity
   (a) This can be done easily within caret, but it only works for classification problems. If we predict for example if the Dow moves upwards or downwards, (using the binomial 'family'), then using the ROC curve/as the metric within caret's optimisation works well. Not forgetting it can't be done for glmboost()!
   (b) This can be done within the model extension predicting the direction of movement and the magnitude separately. There we have a classification problem using the binomial model!

9 TODO Compare several models

1. **DONE** base model: glmboost()
2. **STARTED** gbm (not within the mboost package -> mboost literature refers to it as a black-box method')
3. gamboost()
4. blackboost()
5. lm()
6. AdaBoost() -> minimises the exponential loss, c.p. our loos function, by default the mean squared error
7. C.5.0()
8. Ensemble methods?
9. Which involve stochastic gradient boosting? Make sure it is involved!

(a) AdaBoost covers this - subsampling of data for each base learner. See: `http://scikit-learn.org/dev/modules/ensemble.html#subsampling`

(b) This may well be exactly the same as 'out of bag' selection in mboost or the equivalent using {gbm}!

10 STARTED Compare GoogleTrends data to our search terms, another way to show correlation

1. Obtain some google trends data for same time frame Done using {gtrendR}
2. Plot the data on its own to see the correlation This shows the relative frequency with which each search term is 'googled'
3. Plot the Google data alongside the twitter data This is to see what the relationship is and also if there are large discrepnacies in particular areas
4. Further work ideas using both data sets We could weight the tweets with the relative frequencies coming from Google

11 TODO Try a few more ensemble methods for comparison in Weka/RWeka