Strathclyde Business School, finTech MSc

Becoming an effective technology analyst - fall 2018

Olivier Bauthéac

01/11/2018

As part of the fall 2018 iteration of the 'becoming an effective technoloy analyst' class of the Strathclyde Business School finTech MSc program coursework, below are the instructions for your data-science finance assignment. Examples solutions in both the R and Python programming languages will be provided in due time.

Full stack data-science finance (small) project

Preprocessing (ELT)

Extract

Minimum required

In an excel woorkbook, query Bloomberg for historical (bdh) as well as contemporaneous (bdp) data for a market index as well as a broad cross-section of U.S. stocks. Historical data should be retrieved from October 1st 2016 to today at the daily frequency on individual ticker specific sheets (one sheet per name). All names' contemporaneous data, on the other hand, should sit on a single sheet. The Bloomberg ticker for the market index is 'RAY Index' while those for the corporation names are listed below:

BBG stock tickers				
ADM US Equity	CIVI US Equity	GBX US Equity	LIND US Equity	SERV US Equity
AE US Equity	CLGX US Equity	GDI US Equity	LZB US Equity	SGA US Equity
AGCO US Equity	CLR US Equity	GHC US Equity	MAN US Equity	SITE US Equity
AJRD US Equity	COMM US Equity	GME US Equity	MEI US Equity	SMP US Equity
ALG US Equity	CRL US Equity	GOLF US Equity	MLR US Equity	SPXC US Equity
AMD US Equity	CTB US Equity	GPN US Equity	MRC US Equity	STRT US Equity
AMOT US Equity	CTLT US Equity	GTLS US Equity	MTD US Equity	SUPN US Equity
ASGN US Equity	CTXS US Equity	HFC US Equity	MTZ US Equity	TAST US Equity
ATRO US Equity	DHI US Equity	HOFT US Equity	NC US Equity	TMO US Equity
AVT US Equity	DKS US Equity	HPE US Equity	NGVT US Equity	TNET US Equity
AWI US Equity	EBIX US Equity	HURC US Equity	NHC US Equity	TPB US Equity
BBBY US Equity	EEFT US Equity	HWKN US Equity	NUE US Equity	UBNT US Equity
BFAM US Equity	ELF US Equity	HY US Equity	OSIS US Equity	UFPI US Equity
BID US Equity	ELVT US Equity	IAC US Equity	OSK US Equity	UFS US Equity
BIG US Equity	EML US Equity	IART US Equity	PFGC US Equity	USAK US Equity
BKNG US Equity	ENTG US Equity	IBP US Equity	PGTI US Equity	VLGEA US Equity
BLD US Equity	ERI US Equity	IDTI US Equity	PKI US Equity	VLO US Equity
BSET US Equity	ETH US Equity	INT US Equity	PLPC US Equity	VRSK US Equity
BWA US Equity	FICO US Equity	IOSP US Equity	PRAH US Equity	WBC US Equity
BYD US Equity	FISV US Equity	ITRI US Equity	PSX US Equity	WERN US Equity
CAL US Equity	FL US Equity	JLL US Equity	RBC US Equity	WGO US Equity
CBRE US Equity	FLR US Equity	KHC US Equity	RS US Equity	WRK US Equity
CENTA US Equity	FLT US Equity	KSU US Equity	RXN US Equity	XPO US Equity

BBG stock tickers				
CHEF US Equity	FTV US Equity	LGND US Equity	SCL US Equity	ZBRA US Equity

The historical time series should include the following market & book data fields:

Field	Bloomberg symbol
close price	PX_LAST
book value per share	BOOK_VAL_PER_SH
earnings per share	TRAIL_12M_EPS
dividend per share	TRAIL_12M_DVD_PER_SH
debt	SHORT_AND_LONG_TERM_DEBT
equity	TOTAL_EQUITY
current assets	BS_CUR_ASSET_REPORT
current liabilities	BS_CUR_LIAB
sales	SALES_REV_TURN

Contemporaneous data on the other hand should include the number of shares outstanding, number of directors on the board, number of women on the board, number of board meetings per year, long company name and company description. Explore Bloomberg to find the corresponding field symbols.

Going further

- Using VBA, make your workbook updatable. Ammend your workbook so that it retrieves up to date data in one clic. I.e. if in the future you open the workbook you created today, the workbook should be able to retrieve up to date data.
  - Hint 1. Update doesn't necessarily mean adding most recent values to an existing time series. Requerying the whole data up to the most recent date would work as well.
  - Hint 2. Inspect the BQL syntax in Bloomberg formula cells, ammend accordingly.
- Using VBA, make your workbook flexible. Ammend your workbook so that it can retrieve data for any set of stocks/indexes & market/book fields at various frequencies (year, month, week, day), from and to any date. The user should only have to list the tickers/fields and set the parameters on one sheet.
  - Hint 1. Object oriented programming could help; excel table objects in particular.
  - Hint 2. Create an 'update' sheet with tickers list, parameters (frequency, start and end dates) and fields. This sheet could also be used to host the contemporaneous dataset.
- Using VBA, make your workbook fully portable. If you open your workbook without a live Bloomberg connection you'll notice you loose the contemporaneous dataset; try to fix that problem somehow.
  - Hint 1. VBA events could help.

You now have a fully portable, customizable Bloomberg financial data extraction tool and now it's time to use it.

Load

Using R or Python (example solutions will be provided for both programming languages), load the workbook data in memory. Organise the data in two dataframes, one for the historical times series, the other for static (contemporaneous) data. The time series dataframe should have a two-level row index including tickers & dates while columns should host the corresponding time series; the dataframe should broadly look like this:

```
##                ticker       Date  PX_LAST BOOK_VAL_PER_SH TRAIL_12M_EPS
##     1:      RAY Index 2016-10-04 1273.897        476.0300       58.4400
##     2:      RAY Index 2016-10-05 1279.598        476.0500       58.4400
```

```
##     3:      RAY Index 2016-10-06 1279.430        476.0500       58.4300
##     4:      RAY Index 2016-10-07 1274.602        476.0700       58.4800
##     5:      RAY Index 2016-10-10 1281.312        476.0800       58.4700
##    ---
## 62692: USAK US Equity 2018-10-17   17.830          8.4359        0.7559
## 62693: USAK US Equity 2018-10-18   17.310          8.4359        0.7559
## 62694: USAK US Equity 2018-10-19   17.410          8.4359        0.7559
## 62695: USAK US Equity 2018-10-22   18.240          8.4359        0.7559
## 62696: USAK US Equity 2018-10-23   17.310          8.4359        0.7559
##        TRAIL_12M_DVD_PER_SH SHORT_AND_LONG_TERM_DEBT TOTAL_EQUITY
##     1:                   NA                 566.360      504.630
##     2:                   NA                 566.370      504.650
##     3:                   NA                 566.340      504.640
##     4:                   NA                 566.360      504.660
##     5:                   NA                 566.350      504.660
##    ---
## 62692:                    0                  88.958       70.125
## 62693:                    0                  88.958       70.125
## 62694:                    0                  88.958       70.125
## 62695:                    0                  88.958       70.125
## 62696:                    0                  88.958       70.125
##        BS_CUR_ASSET_REPORT BS_CUR_LIAB SALES_REV_TURN
##     1:             339.210     234.730             NA
##     2:             339.340     234.820             NA
##     3:             339.340     234.820             NA
##     4:             339.580     234.980             NA
##     5:             339.610     235.000             NA
##    ---
## 62692:              78.798      71.077        135.381
## 62693:              78.798      71.077        135.381
## 62694:              78.798      71.077        135.381
## 62695:              78.798      71.077        135.381
## 62696:              78.798      71.077        135.381
```

The static dataset on the other hand should be row-indexed by tickers and have columns hosting the corresponding static data fields. For static data, only numeric fields should be loaded with long company name and description fields left to the excel workbook for reference.

Transform

Market betas

Minimum required

Using the most recent samples in the time series data, calculate the individual 1-year market betas for the stocks. Show calculations and comment. Comments should include a detailled discussion on what market betas are, what they represent for stocks as well as details about the corresponding model. Plot your results as a histogram and comment. Hint: there are 252 trading days in a year.

Going further

Using all the time series samples, calculate the individual rolling 1-year market betas for the stocks. Randomly select five stocks and display their corresponding rolling beta time series on the same lineplot.

Features interactions

- Using the most recent samples in the time series dataset, for each name construct a set of feature interactions that include the following popular financial ratios: price to book, price to earnings, dividend yield and gearing. Show calculations and discuss these concepts from a corporate finance standpoint.
- Explore this new dataset. Hint: use visualization tools.

Modeling

Minimum required

Cluster analysis (unsupervised learning)

Hierarchical clustering

After normalizing the ratios dataset above to zero means and unit variances, apply hierachical clustering and draw the corresponding dendogram. What seems to be the optimal number of clusters for this dataset? Explain.

K-means

- Implement a two-cluster k-means analysis on this dataset. Explore the resulting cluster characteristics: calculate the cluster specific means for each ratio. Comment on the results and propose labels for the two classes. Hint: how would Warren Buffett most likely answer this?
- Label individual names accordingly in a new 'classes' dataframe.

Going further

Classification (supervised learning)

- Create a betas dataset that subsets the most recent (last sample date) samples from the rolling maket betas dataset above. Merge the classes, ratios, static and betas datasets together.
- Implement a classification analysis on the resulting dataset where the target is name's class as attributed above. Use various classifiers including logistic regression, k-nearest-neighbours, support vector machines, decision tree, random forest and neural network (multi-layer perceptrons). Use 75-25% for training-test sets split and 5-fold cross-validation.
- For each model:
  - Show training and test set confusion matrices and calculate corresponding precision & recall indicators; comment. Your comments should include a discussion on precision and recall.
  - Explain what the model does and how. Discuss model paramaters and how they contribute to model fine-tunning.
  - Find model optimal parameters using gridsearch and run model accordingly. Show corresponding learning curves.