Big Data in Finance: In-Class Presentation Assignment

Prof. Tarun Ramadorai

2021-2022 Academic Year

We have done work in class on predicting default, as well as predicting equity returns. We have also built up a suite of tools, both reduced-form and structural, over this time. Finally, we have gain (or will soon gain) domain knowledge in finance, discussing issues on the analysis of creditworthiness, as well as in asset management. *The final assignment asks that you apply these newly acquired skills on a group project that you find personally interesting. You will need to present this project to an audience of your peers in-class at the end of the term.*

The ground rules are that:

- You pick a topic within one of the areas of finance that we investigated in the course (credit or asset management).
- You undertake actual data analysis (see suggestions below) and draw conclusions from this analysis.
- You provide justification for your chosen project that tells us why it is interesting from a commercial perspective.

Projects that have been successfully completed in the past include (note: you do not need to limit yourselves to these, as these are merely examples):

- Analysis of stock, bond, or commodity return data, using machine learning techniques that lead to a useful trading strategy.
- Credit risk modelling using machine learning applied to the Lending Club (see below) or other datasets.
- Analysis of new data in areas such as marketing of financial products, modelling real estate prices, or modelling mutual fund flows.

As noted, you should feel free to deviate from the suggestions here, and experiment with topics that conform to the rules set above. You are welcome to use any machine learning method that you like including those that we may not have covered in class. That said, you will be evaluated based on the conceptual frameworks that we develop in the class. The paper that you finally come up will gain points for originality and differentiation from your peers (conditional on being correct and accurate, of course!).

Your output should be in the form of an in-class presentation of up to 10 slides max (you will likely have between 7 and 10 minutes to present it at the very maximum). The

presentation should mainly focus on visuals and tables that help to outline and market your analysis. You should also append any code you used to generate your solution when you submit your slides for assessment.

Marks will be assigned in the following fractions: 25% for clearly and correctly explaining the predictive techniques that you are using and their specific application in this context. 30% for the appropriate application of these techniques on the actual dataset (please provide your code). 25% for clearly explaining the underlying finance and economic mechanisms that underpin your results. Finally, 20% marks will be assigned for the originality and creativity of the strategy that you formulate.

1 Some curated datasets

We have placed two broad groups of datasets on the Hub that you can download if you wish to use them, but recall that you are welcome to use any other datasets that you like (more on this below in the next Section).

- **Lending Club Loan Data for Credit Analysis:** LendingClub is a peer-to-peer lending company. The company historically made datasets available which documented the performance and characteristics of issued loans. We have put together a data file that contains a selection of these loans, randomly reducing the number of loans that were fully repaid so as reduce the size of the dataset to facilitate easy analysis. If you are having trouble working with the file because of its size, you should feel free to remove observations as you see fit. The dataset contains one line per loan issued. The first column (loan_status) will tell you whether the loan was fully repaid or charged-off. Note that in the event of a charge-off, some amount may still be recovered (see field variable recoveries). The remaining columns will provide additional characteristics about the loan or borrower. A detailed description of the variables is also provided in the file Lending_Club_Dictionary.csv.
- **Stock Returns and Characteristics Data for Asset Management Analysis:** This dataset comprises multiple files, listed below.
  - Returns_Data.csv is the main data file, which contains monthly stock returns and other price and liquidity related variables for 500 U.S. stocks. In addition to stock-level returns, the file contains: VOL, the aggregate trading volume on the stock in each month expressed in hundreds of shares; and PRC, the price of the stock. Moreover, the variables VWRETD and EWRETD provide the value-weighted and equal-weighted cum-dividend index returns on the S&P500. These data are from CRSP.
  - Stock_Characteristics_Data.csv contains stock characteristics published by firms every quarter (from their SEC 10-Q filings). Be careful how you merge the return and characteristic datasets, as characteristics data are only available to investors *after* the month of release. These data are from Compustat/Capital IQ.
  - Time_Series_Data.csv contains time series data from Goyal and Welch's 2007 paper "A Comprehensive Look at the Empirical Performance of Equity Premium Prediction". The data were downloaded from Ivo Welch's website (https://www.ivo-welch.info/professional/goyal-welch/)

– StockDataDictionary.xlsx contains a data dictionary explaining the variables.

2 Some other useful references and finding datasets

Business datasets and journal/magazine resources available at Imperial (including many finance datasets):
`https://www.imperial.ac.uk/admin-services/library/subject-support/business/databases-a-z/`

Kaggle Finance datasets and competitions: `https://www.kaggle.com/tags/finance`

Varian, Hal R. 2014. "Big Data: New Tricks for Econometrics." *Journal of Economic Perspectives*
`https://www.aeaweb.org/articles?id=10.1257/jep.28.2.3`

Mullainathan, Sendhil, and Jann Spiess. 2017. "Machine Learning: An Applied Econometric Approach." *Journal of Economic Perspectives*
`https://www.aeaweb.org/articles?id=10.1257/jep.31.2.87`