PORTFOLIO PREDICTION EXERCISE

The following is a brief guide to the toy data exercise that we ask all parties interested in working with us to submit. The problem is designed to suit all backgrounds i.e. it can be done in one or two evening sittings.

Jacob Ayres-Thomson, Head of Data Sciences

MARKET PREDICTION EXERCISE

This data exercise is designed to give us both a chance of a small scale example of "working together". We consider this to be a good chance for us to explore how each other works in practice and the thinking behind what one does when dealing with real data. We consider the conventional interview process woefully unsuitable for our purpose here. We also believe it is unsuitable for yours.

We encourage you to be experimental and not become disheartened when your attempts struggle to derive strong information. Our typical daily prediction model has about 10% correlation between forecast and actual outcome and we utilise additional methods to bolster this to over 20% - but that is not part of this model development process.

You will not be assessed on the performance of your work so much as the thinking and intelligence behind it as well as the effort. You may, for example, find that certain portfolios are not predictable. We therefore encourage you to document all of your attempts and to retain and submit them.

The data attached is real data that we have created. It covers the daily investment returns of 20 investment portfolios that we have statistically constructed. The challenge is to create a way to predict the return of those 20 portfolios in the out-of-sample period. You can do as much or as little research as you wish.

OBJECTIVE

The objective is to build an automated method that forecasts the returns in the out-of-sample period as accurately as possible for as many of the twenty portfolios as possible.

> ***"I have not failed. I've just found 10,000 ways that won't work" – T. A. Edison***

DATASET

This is data that we have created and represents a real financial market opportunity on proprietary information. The dataset is relatively small.

The data is split into two samples IS and OS. IS represents the "in-sample" data and is 3 years long. OS represents the "out –of-sample" data and is one year long.

The **reference** column is meaningless for your purposes and can be ignored.

There are 20 different **Portfolio Code'**s, each one represents a financial security, statistically constructed by us to have unique properties making them easier to predict than conventional stocks. The daily return is the target which we seek to learn and then predict. The data covers data created from US banking stocks.

There are 24 simple features already derived for you, namely X1 to X24. These are simple technical indicators typically used by traders. From the perspective of the learning challenge, it doesn't really matter how X1 to X24 are derived. They may, or may not, contain information useful to the task. Also, what is informative in predicting one portfolio may or may not be informative in predicting another portfolio. There is therefore a hanging question as to whether the portfolio model should be learn't at the individual portfolio level or across this whole bucket of portfolios (that all come from the same sector) or some combination thereof.

The date column contains the time ordering of the data, this will be useful if you wish to either construct an online algorithmic method or, for example, use the previous outcomes to form your own time series indicators on the past data for any given point, though neither of these are expected.

ACTIVE DATA SNOOPING ("look-ahead") BIAS

Forecasting financial market data is perhaps the toughest data to predict of any field for a number of reasons, firstly it is very noisy and secondly, it does not maintain static properties like those we observe for large bodies in the physical universe. In fact, when dealing with daily data, there are few known reasons that any patterns need exist or maintain themselves at all.

For this reason, and to enable you to progress you work through iterative feedback loops, we provide you with the data you are trying to forecast, the "out-of-sample" returns. We encourage you to exploit look-ahead bias to see what is working out-of-sample and progressively iterate modelling attempts.

The only condition that we ask, is that parameter selection and model tuning is always conducted based on analysis of present and past data only (i.e. preferably automatically if online) and also, that any model learning is only conducted based on observation of past data i.e. the X features and possibly past return time series etc. We therefore ask that the rationale for any manually selected parameters is based upon non-hindsight data and the graphs/data summaries etc are presented.

Where parameters are set manually, they must be determined based solely on analysis of in-sample data and we will wish to see this evidence included in the submission. Where they are set systematically by algorithms/models, they can be based on the most present information i.e. 120 days into the out-of-sample period they will be able to view everything up to that point but nothing beyond that time period.

What we are getting at here is that the 'method' for model creation can be hindsight optimised by snooping but the setting of models parameters itself cannot. We are pushing you to find the method that creates the best models, not to find the best models themselves. Some examples to explain this thus:

TACKLING THE CHALLENGE

QUICK SOLUTION

Train the models on the in-sample data only, ignoring the time series optionality of online learning. Use purely the X features as your learning data and train a model for each portfolio individually (using the same method). Might want to investigate whether it is possible to learn forecasting information at the global portfolio level or only locally or both.

ADVANCED SOLUTION

Construct a time series of past 40 days returns at all points from the data. Consider using technical trading indicators to construct new X features or parsing the historic time series of outcomes through something like a deep learning network to see if additional information can be sourced in addition to the 24 pre-selected X features etc in order to expand the X feature dataset.

ADVICE

Be reasonable with yourself, if you decide, for example, to train an algorithm using genetic algorithms then consider approaching it via the quick solution below rather than re-training the model every single day in the out-of-sample period.

VAGUENESS

This exercise can be taken forward in many different ways and approaches, more so than a conventional Kaggle and yet, the dataset is pretty small. This problem is constructed that way on purpose. In data science, there are nearly always infinite approaches possible to tackle any problem. Making decisions and moving ahead is critical and producing a terrible model is better than indecisively sitting on the fence, indeed engaging in the very act of modelling and learning from real data is the fastest route to expertise improvement.

DEADLINE

There is none, take as long or as short as you wish. We are likely to be hiring on a continual basis where we find the right people. If you are in the market now actively looking for a new position then we would encourage you to get this back to us within no more than 1-2 weeks.

QUESTIONS

Please send any questions you have regarding the exercise to the following addresses:

- jacob.ayres-thomson@justretirement.com
- edward.conway@justretirement.com

SUBMISSION

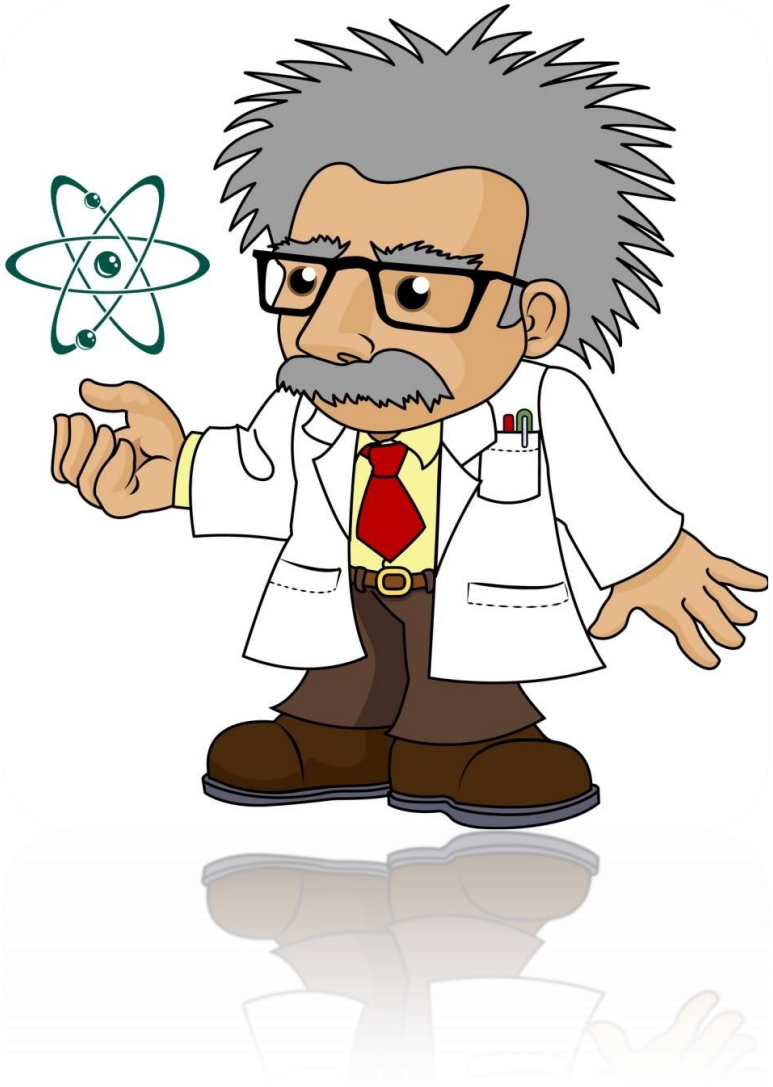
We require the following to be submitted:

- Brief explanation of the final model used as well as the other models and attempts you made.
- Any relevant graphs or data summaries used in determining parameters, settings or algo/model choice or structure (where these were selected by yourself).
- Amount of time spent on the exercise.
- Explain current working/study situation[1]: part-time, full-time, on career break, etc.
- Predictions for out-of-sample days for your best chosen model(s).
- Code for the final model (s) used.

Please email submissions to the email addresses given above.

If you have any problems sending them through please drop Jake a text on 07958 255 763.

[1] This is so we can fairly consider the amount of time available to you to conduct this exercise.



AFTER SUBMISSION

We will invite all submissions of a minimum standard in to discuss their work and for a first or second stage interview with us. At that stage, there will also be some other cognitive probing into your motivations and interests.

Good luck and most of all, enjoy the process!