Clear Street®
Markets

Clear Street Markets Quantitative Researcher Take Home Project

We are excited to invite you to the next step in our interview process, the Quantitative Researcher Take Home Project. We thank you in advance for deciding to take part in this assessment - We hope you find it as exciting as we do!

We designed our project to be a canvas for you to showcase your best. It’s a space for you to share what you are capable of.

Before getting started, please read through this entire document. It contains all the information you need to complete and submit your output.

Your submission deadline is the End of Day on **January 7, 2024**. We welcome early submissions.

Our team can be reached at grproject@clearstreet.io. Contact us here to ask us any questions you have about the challenge and we will try our best to respond promptly. Please use the same email to submit your project.

Good luck!

Objective:

We want you to produce two calibrations **Y1 ~ f(X1...X375)** and **Y2 ~ f(X...X375)**.

Data:

We have provided 298 daily files, covering the 15-month period 20220104 - 20230331, for you to use in this project.

The column names of the files are:

time: timestamp specifying milliseconds from midnight. Start time is 35101000 (9:45am) and end time is 57600000 (4:00pm)
Q1: quality variable #1 - to be used as a quality flag for calibration of Y1
Q2: quality variable #2 - to be used as a quality flag for calibration of Y2
Y1: dependent variable #1 to be predicted
Y2: dependent variable #2 to be predicted
X1 to X375: 375 independent variables to be used as inputs to predict both Y1 and Y2

Problem Instructions:

Use 2 subsets of the 375 X variables to make 2 separate calibrations for Y1 and Y2.

If Q1 is less than 1 (make the test Q1 < 0.9999 to handle floating point precision), then the row should be excluded from the calibration of Y1.

Similarly, if Q2 is less than 1, the row should be excluded from the calibration of Y2.

Q1 and Q2 are used to indicate the quality of Y1 and Y2 respectively. And if they are less than 1, it means that Y1 or Y2 are not in a good state.

We have decided to limit the techniques used for the calibrations to either linear regression, gradient-boosted trees, or neural networks.

You can use any sub-periods of the 15-month period as in-sample or out-of-sample. You can use all the data or can sample the data.

A value of 999999 indicates a bad value so you should skip it.

You may want to skip a row if more than a certain percentage of values (50% or 70%) are bad values.

The calibration can be for the whole day (coefficient 0.15 for variable X12) or split by time of day (between 9:45 and 11:00 the coefficient for X12 is 0.12 and 0.18 between 11:00 and 14:00 and so on) - it's up to you but try to be reasonable when you split the day in periods (i.e. avoid having a different coefficient for each minute or each hour of the day).

We have provided you with real financial data for this project. Many times such data is less clean than the stylized data samples that you encounter in class.

Make your judgment of which of the 375 X variables should be included in your calibration. In practice, we generally observe that using more than 20 variables does not yield good results - many times between 5 and 10 variables would suffice..

Your calibrations will be evaluated using data starting on 20230401, dates that occur after the sample data we have provided for this project.

In addition to your calibrations, please provide a brief explanation of the:

1. feature selection method you used
2. the calibration method you used
3. data sampling (if any)
4. in-sample and out-of-sample periods (if any)
5. any other information that will help us judge your technique

We value your thought process behind your chosen approach and decisions. This is just as important as the final result of your research to us.

Please keep in mind that we need to reproduce and test your solution on our side so it should be presented with sufficient comments, explanations, and clarifications.

Sample solution:

Linear regression / sampling 50% of data / in-sample 20220103-20221031; out-of-sample 20221101-20230331 / use LASSO to select variables...

Y1 = 0.11*X33 - 0.14*X125

Y2 = (between 9:45am and 11:00am) 0.13*X34 - 0.21*X233
(between 11am and 4:00pm) 0.16*X34 - 0.25*X235

Accessing the Data:

We provide an ftp endpoint where you can download the data files in **.parquet** format for this project. Use the credentials listed below to access the site.

Hostname: ftp://st38539.ispot.cc
Username: candidate@st38539.ispot.cc
Password: udANwVgU4CcRwAKzTXZVKUCF

*** Mac users: please use FileZilla if you have trouble accessing the ftp site.