Final Project

MTH 9899 Baruch College
DATA SCIENCE II: Machine Learning

May 22, 2019

- For the project, work in the assigned groups that have been sent to the TAs. If you haven't picked so far, pick now.
- You can, and are encouraged, to use sklearn and PyTorch

Goals The goal of this project is to apply the various ML algorithms you have learned in this class to a real-world financial dataset. The hope is that you will apply the algorithms you've been taught in new and exciting ways to better predict future market returns. The expectation is that you will implement and test multiple ideas from the semester, and combine them to produce superior results to what any single algorithm can do. As part of this, do your best to:

- Use cross-validation to fit and choose models
- Tune hyperparameters the best you can
- Use best practices with regards to data cleaning

As part of this project, you will submit your code at the beginning of the presentation period and your code will be tested on a hold-out sample to find out how well it performs.

Data The dataset you will be given consists of approximately 400k rows of data with 7 features. Due to the proprietary nature of the dataset things like times, symbols etc have been replaced:

Rows

- Each row represents an observation. It is your discretion whether to keep them all, consider a subset, or maybe even enlarge it.
- **YOU SHOULD NOT USE y, PAST, PRESENT, OR FUTURE, IN YOUR PREDICTIONS.**
- Some amount of features are random noise.

Columns

- Date: The date is an int - it does correspond to real dates in the same order

- Time: Can be ignored. All of the points occur at the same time and the future return is measured roughly 1 day in the future.
- fut_ret: This is the dependent variable
- sec_id: This is an identifier for the security that will not change over time. Note that there may be missing days for some securities.
- vol: This is a proxy for the stock's volatility
- 'X*': There are 7 cols with names that start with 'X' - those will be the features that you might consider using for your models

We will be using out-of-sample R2 as final benchmark. Note that we will be applying reasonable "cleaning" procedures to y to make the measure meaningful. Since there might be slightly different ways to calculate, we will use sklearn r2_score implementation as the only benchmark. For details, please see
http://scikit-learn.org/stable/modules/generated/sklearn.metrics.r2_score.html

Code We are preserving a hold-out test dataset that has about 25% of the data that you won't see. Prior to the final presentation, we will send you the out-of-sample data without y values and you must sbumit back to us a frame of date, sec_id, prediction.

Presentation Each group will give a 10 minute presentation to the class during the final exam period, 6PM, Wednesday May 22th. Every member of the group is expected to speak for a significant fraction of the presentation.

Paper Please submit a 10 to 15 page paper. The paper should describe the algorithms you tested and the results you observed. You should make sure to mention how you dealt with things like CV, data cleaning, etc. Please present your conclusions as to which techniques worked best and added the most value.

Grading Your grade will be based on the combination of the paper, the overall presentation, and your portion of the presentation. In addition, students should email us with their observation on which student in their group contributed more or less than their peers.