Old Mission Quantitative Researcher Project

General Description

Your task will be to build two models for predicting intraday bid/ask spreads for a universe of equity symbols. Having a robust predictive model for market spreads is useful for signal research and strategy construction. This project is meant to evaluate your ability to do quantitative research applicable to trading.

Manifest

The project folder should contain the following:

- spreads_project.pdf – This document
- spreads_data_train.par – A Pandas dataframe of the training data stored as a parquet file
- data_descriptions.csv – A csv with descriptions and categories for the dataframe columns

Data

The data consists of intraday bid/ask spreads for a universe of equity securities. The data is sampled at 10 minute intervals, and the variable you are trying to predict, fut_spread, is the spread from the subsequent interval. The columns in the Quote category are derived bid/ask prices and sizes from the current interval. The Trade category contains aggregated trade information over the current interval. The Static category contains characteristic information for the given symbol derived from aggregated historical data.

Tasks

The first task is to do some EDA on the data and note any interesting findings, and any cleaning or adjustments that need to be made to the data.

The first model should predict the fut_spread variable only using columns in the Time and Static categories. The goal of this model to have a robust heuristic for what the spread of a symbol should be given the current time of day, and the historical characteristics of the symbol.

The second model can use the remaining columns in the Trade and Quote categories. This model should better predictive performance with its access to current trade and quote information.

Deliverables

The write up of your findings can take the form of a separate document, or as part of a formatted jupyter notebook. Include the code written for the analysis and any code needed to reproduce and run the models.

Evaluation

The project will be evaluated on the following criteria

- The methodological rigor of the analysis
- The ability to find novel and useful insights from the data
- The predictive performance and robustness of the models produced on a withheld out of sample dataset

Tips

- The timestamps are in UTC, the US equity markets trade in EST
- The fut_spread is in price space, a common normalization is to divide by the price to represent the spread as a percent

Example Code

Python 3.8.6

Pandas 1.3.2

```
>>> import pandas as pd

>>> pd.__version__

>>> df = pd.read_parquet("spreads_data_train.par")

>>> df.shape

(3406042, 28)
```