Data engineering final

Marcus Rapacioli

Goal

The goal of this project was to improve on my work in homework 5.

I decided to do this in 3 ways:

1. Establish a connection to **mongodb**
2. Use **more data**
3. Add **macro-economic data** to the regression model

Establish a connection to MongoDB

MongoDB is a noSQL database.

- The major benefit for developers is the **query speed.**
- MongoDB is cloud-based and so offers a **scalable solution.**

Add more data

I used **360 records** to train the model for the final, vs 100 records in hmwk 5. They are spread over 6 different days and so should offer strong representative sample.

More data = better predictions (in theory).

Add macro-indicators

I was aware that inflation was a key determinant of exchange rates.

It is difficult to find inflation data broken down into intervals that are small than a month, so I decided to use treasury bond yields as a proxy. **We know that treasury prices are highly dependent on the inflation rate.**

Most of my data collection had taken place overnight, however, when the treasury market was closed. This ruled out being able to use data at the minute level.

However I had collected data on 6 different days, which gave me 6 different prices for treasury data.

I did this for 5 year, 10 year and 30 year bonds.

One of my closing dates was Dec 4th (A Sunday). So I had no closing price available. I decided to use the Monday's opening price instead.

Some things I noticed

- Installing pycaret was quite fiddly on M1 MacBooks. I managed to do it by creating a new anaconda environment specifically for pycaret.
- Interestingly, after I established the MongoDB connection the code to longer to run. I think this may have been due to the time it takes to write to the cloud vs locally. (Code was taking 2 seconds).
    - In order to deal with this I eliminated the sleep period and reduced the counter by half
- When I was preparing the data I had to deal with Best practice #3. I had misspelled ‘volatility’ for example.
- With more data the new vol and fd classifications were mostly in the second band *(there were a number of exceptions to this so I know that it wasn’t a coding error)*

Results

Currency pair	Profit loss (hmwk5)	Profit loss (final)
EURUSD	0.018	**0.042**
GBPUSD	-0.499	**-0.148**
USDAUD	**-0.002**	-0.034
USDCAD	-0.032	**-0.022**
USDCHF	**0.081**	0.060
USDHKD	-0.021	**0.003**
USDNZD	**0.143**	-0.456
USDSGD	-0.051	**-0.030**