Predictive Analytics Challenge – the dataset explained

The files:

Training set (snapshot IDs 0-36249)

- TRAINING_independent_variables.csv *(see CompetitionFiles.zip)*
- TRAINING_target.csv *(see CompetitionFiles.zip)*
- TRAINING_Interbank_Trades.csv *(see InterbankTradeFiles.zip)*
- TRAINING_mids.csv *(see MidFiles.zip)*
- HistoricalVolumesByTimeOfDay.csv *(see CompetitionFiles.zip)*

Testing set (snapshot IDs 36250-48331)

- TESTING_independent_variables.csv *(see CompetitionFiles.zip)*
- TESTING_Interbank_Trades.csv *(see InterbankTradeFiles.zip)*
- TESTING_mids.csv *(see MidFiles.zip)*

The structure of the data

The data provided is time-series data. Each row in the datasets has a unique snapshot ID. This is the equivalent to a single snapshot in time. Each snapshot contains the latest value of 30+ variables along with the Target variable that we want to predict. The data has been randomly shuffled in time and then split into a testing and training set. As such, at each snapshot ID, you should only use the state of the 30+ variables at the same corresponding ID.

Requirement

The Target variable for the Testing dataset is not provided. The challenge is to predict the Target variable for IDs 36250-48331. In other words, you need to create the withheld TESTING_target.csv file.

How the data was constructed

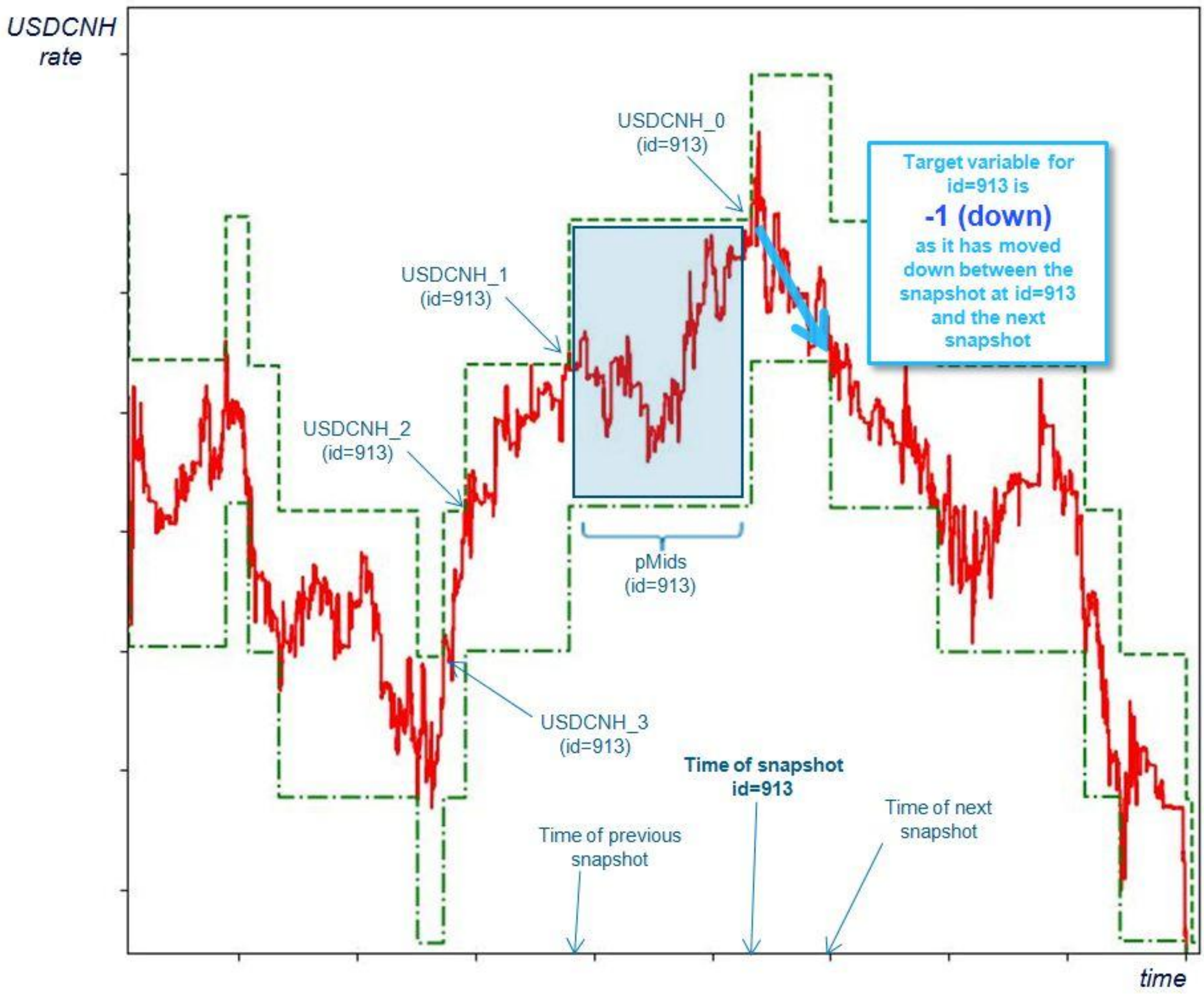
The data comes from the SCB algorithmic trading platform and relates to spot FX rates. The SCB trading platform captures and processes high frequency market data. The data was sampled every time USDCNH moved up or down by more than 7 pips. This is the equivalent to a move up or down of approximately 0.01%. At each sampling time, the state of various related data feeds was evaluated. These data feeds are presented in the data files and described below.

The target variable at each sampling time is equivalent to the next 7 pip move in USDCNH – is it up or down?

The data has been randomly shuffled and split into Training and Testing sets. Models should be built against the Training set.

An explanation of the variables

The figure below is provided as an aid to understanding the variable descriptions that follow. Note, the diagram is just for illustration and does not represent the actual data provided.



In TRAINING_target.csv file

Target

The target variable is the future movement of the USDCNH foreign exchange spot rate (hereon just called USDCNH). A value of +1 indicates that the future movement is up, -1 indicates that the future movement is down.

In TRAINING_independent_variables.csv and TESTING_independent_variables.csv files

id

The snapshot ID. The dataset has been randomly shuffled and split – as a result, consecutive IDs are not consecutive in time. You should not try to encode history or memory into your solution by using information from proceeding IDs. For example, use only the values of the variables at id=10 to predict the target variable at id=10.

USDCNH_0, USDCNH _1, USDCNH _2, USDCNH _3

The latest price of USDCNH and 3 historical prices

In the independent variables dataset, the column titled 'USDCNH_0' is the price of USDCNH at the time of the snapshot. The columns 'USDCNH_1', 'USDCNH_2', and 'USDCNH_3' are the prices of USDCNH at the previous three snapshots ('USDCNH_3' being the price three snapshots previously). As such (USDCNH_0 - USDCNH_1) tells you how much the price has moved up in the time since the previous snapshot.

Again, it should be noted that for (eg) id=10, USDCNH_1 is not equal to USDCNH_0 from id=9 since the IDs are not in time order.

USDAUD_0, USDAUD _1, USDAUD_2, USDAUD_3,

USDCAD_0, USDCAD _1, USDCAD_2, USDCAD_3,

USDEUR_0, USDEUR_1, USDEUR_2, USDEUR_3,

USDGBP_0, USDGBP _1, USDGBP_2, USDGBP_3,

USDNZD_0, USDNZD _1, USDNZD_2, USDNZD_3,

Latest and historical prices for 5 currencies (AUD, CAD, EUR, GBP and NZD) that may be related to CNH

Similarly to above, the columns titles 'USDAUD...', 'USDCAD...', 'USDEUR...', 'USDGBP...', and 'USDNZD...' are the prices of 5 other currencies (against the US dollar) that may be useful. The prices are given with the same definition as for USDCNH, ie, the amount of that currency that is equivalent to 1 USD. Again, for each currency, the prices at the snapshot (..._0) and the previous 3 snapshots are given (..._1, ..._2, ..._3).

vol_1, vol_3, vol_5, vol_10

Latest degree of Orderbook imbalance observed in interbank market

These columns detail the market mid price corresponding to different trade sizes. Column 'vol_1' is for a trade of size 1000000 USD of USDCNH, vol_3 for 3000000, vol_5 for 5000000 and vol_10 for 10000000. These variables could be used to determine the imbalance in the orderbook. The number of interested buyers and sellers in the market will not (in general) be balanced and may influence the future price movement and, as such, orderbook imbalance is a common attribute to consider.

USDCNY_0, USDCNY_1, USDCNX, USDCNX_1

The difference between USDCNH compared against USDCNY and USDCNX

CNH has two related currencies: CNY which is only traded in China and the non-deliverable CNX which is a proxy for CNY but available for trading outside of China. The column 'USDCNY_0' is the latest price of USDCNY and 'USDCNY_1' was the last price of USDCNY at the previous snapshot. The column 'USDCNX_0' is the latest price of USDCNX and 'USDCNX_1' was the last price of USDCNX at the previous snapshot.

hourOfDay, isOpen

These two variables describe the hour of day of the snapshot (0 to 23) and whether the local China market is open at the point of the snapshot.

marketState

A computed indicator as to market liquidity and volatility

An indicator as to the state of the market in terms of volatility and liquidity. The variable can take 5 values: A, B, C, D and E.

covarianceMatrix

Each entry is a 6x6 covariance matrix for the currencies {AUD, CAD, CNH, EUR, GBP, NZD} (in that order). The diagonals of the matrix give the currency volatilities. For example, the CNH volatility at the time of the snapshot of CNH will be given by the square root of the value at the third column of the third row. The off-diagonal elements give the covariance between currencies. The USD is assumed to have zero volatility and zero correlation with other currencies. Therefore, if you wish to determine the correlation between (for example) USDCNH and USDEUR, you should use the covariance at the third row, fourth column (or fourth row, third column) and the volatility of CNH and EUR.

The covariance matrix has been constructed using the recent history at the time of the snapshot. Therefore, there is no guarantee that the correlations will hold in the future.

In TRAINING_mids.csv and TESTING_mids.csv files

The file contains all of USDCNH mid prices for the period between the previous snapshot and the current snapshot. The prices maintain their time ordering. These prices may contain information around the market dynamics immediately prior to the snapshot.

Note: These files may not open fully in Excel as they have more than 1 million rows. If you need to use Excel you will need to open the files in (for example) Notepad and then split the original csv file into smaller csv files for loading into Excel.

In TRAINING_Interbank_Trades.csv and TESTING_Interbank_Trades.csv files

The files contain the trades dealt on the interbank markets. There may be zero or multiple entries for a single id and relate to trades done in the market between the previous snapshot and the current snapshot. A trade marked as a BUY is defined as one where the counterparty initiating the trade is buying USDCNH. A SELL is the opposite case. For all trades with the same ID the time ordering of those trades is maintained.

In HistoricalVolumesByTimeOfDay.csv file

A historical distribution of interbank USDCNH buys and sells as a function of hour of day. A ‘buy’ is defined as such when the action is initiated by a market participant wanting to buy, and the corresponding situation when wishing to sell – hence the number of buys does not match the number of sells.

Helpful Information and Hints

1. **Currencies are referred to by a 3-letter code. USD is the name for the US dollar.**
2. **USDXXX is the amount of currency XXX corresponding to 1 USD**
3. **All prices are mid prices (ie, the average of the bid and offer)**
4. **We have provided the price for USDCNH and other currencies (relative to the USD) at the point of the snapshot and for the 3 snapshots immediately before. It may help to construct new variables that represent the change in price over the proceeding time periods (ie, returns) or even higher-level derivatives.**
5. ***[Reminder]* Do not encode history into your models. Only use the variable states at the time of the current snapshot to predict that snapshot's target variable.**
6. **The prices of the currencies have been provided relative to the USD. The prices may be correlated during the periods when the USD is strengthening or weakening.**
7. **Not all variables will be equally useful.**
8. **Certain times of the day may be easier to predict/model than others.**
9. **The covariance matrix may help identify correlations between currencies and/or help find times when the USD is strengthening or weakening.**
10. **This is not a standard Machine-Learning dataset. Do not expect to be able to predict the target variable with 90% success. In fact, a success rate of 51% is not a bad result.**