The **transform\_data** function takes a pandas dataframe **df** as input and performs a series of transformations on it.

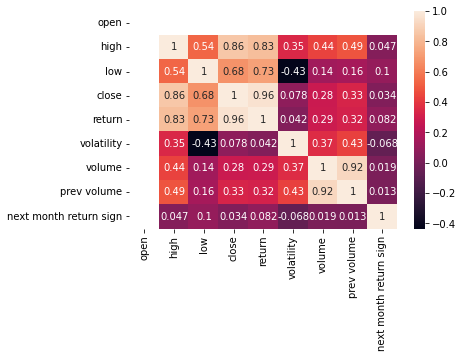
1. The missing values in the dataframe are filled in with the mean / mode value of the respective column .
2. A new column ‘diff\_high\_low’ is added to the dataframe, which represents the difference between ‘high’ and ‘low’ features
3. A new column ‘diff\_open\_close’ is added to the dataframe, which represents the difference between ‘open’ and ‘close’ features
4. A new column ‘ratio\_currvol\_prevvol’ is added to the dataframe, which represents the ratio of ‘volume’ vs ‘prev volume’
5. A new column ‘ratio\_diffhighclose\_difflowclsoe’ is added to the dataframe, which represents the (df['high']-df['close'])/(df['low']-df['close'])
6. Unnecessary columns(high,low,open,close,volume,prevvolume) are dropped from the dataframe. AnsThe function returns the scaled dataframe.

The purpose of these transformations is to prepare the data for use in a machine learning model, by filling in missing values, mapping categorical data, dropping irrelevant columns, adding a new relevant feature, and scaling the data to ensure that all features have similar ranges of values.

The generate\_model(x) function :

1. Trains the model on Training data using Random Forest Classifier and return the model

Why df['diff\_high\_low'] = df['high'] - df['low'] and other features?

* 
* As you can see ‘next month return is not getting affected by all the features already present in the dataframe. So , we need new features.
* As the difference between month high and month low is a good metric to see and trade , I chose it to be my feature. Same goes for (open , close) . Also volume in which a particular stock is being traded also affects return. If volume traded increases significantly from month to month then return will be highly affected by it. So I chose this ratio to be my feature. Again the difference between (high ,close) and (low,close) also affects return but their difference did not add much accuray to my result . So after randomly trying division , we got something good and hence it was chosen.

Model finally trains on X = ['ratio\_diffhighclose\_difflowclsoe','return','volatility','diff\_high\_low', 'diff\_open\_close', 'ratio\_currvol\_prevvol']

y=[‘next month return sign’]

I got an accuracy of 0.65 from this model.

How do final features affect the ‘next month return sign’ ?

