1. Started with S&P500
2. Ran PCA to identify a list of 50 ‘similar’ stocks
3. Fired a technical analysis that can produce 5 features (MA,EWMA,BB, Stoch, Oscillator) on 50 stocks on a daily basis. Using QUANDL data
4. Additionally we will harvest 10 fundamental factors

For each trading day a table like below will be created

|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Stock 1 | TF 1 |  |  |  | TF 5 | FF 1 |  |  |  |  |  |  |  | FF 10 |
|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| Stock 50 |  |  |  |  |  |  |  |  |  |  |  |  |  |  |

In theory the dimensionality of the dataframe = 15 Year \* 50 Stocks \* 15 indicators \* 252 trading days

5 . How do we choose the 10 Fundamental Factors (SIDEBAR QUESTION) BACKTESTING

1. Application of ML techniques to split above dataset into train test. Is it on a daily basis
2. Manual train test split
3. Add a column for returns ; and classify on a binary basis at stock level; or portfolio level. i.e. did the entire portfolio make money; or did the stock make money

|  |  |  |
| --- | --- | --- |
| Stock 1 | Actual Return (3 month) | Predicted Return (3 month) |
|  |  |  |
|  |  |  |
|  |  |  |
| Stock 50 |  |  |
| Portfolio |  |  |
|  |  |  |
|  |  |  |

1. From the above table we will predict top 10 stocks (90 day return)