1. **Abstract :**

We prepare have imported dataset from yahoo finance and features data from alpha 101 and few calculations using artificial intelligence techniques to predict stock direction. Here the direction of stock and our prediction signal i.e. alpha should match to give good prediction.. This paper can help in predicting the stock market share like google direction to help in building trading strategy. With the help of this algorithm user will be able to perform sell and buy correctly.

1. **Summary of problem statement, data and findings**

Predicting trends in stock market prices has been an area of interest for data scientist & researchers for many years due to dynamic nature. Volatility in stock market makes the task of prediction of stock direction is challenging.

Market risk, strongly correlated with forecasting errors, needs to be minimized to ensure minimal risk in investment. This paper is the outcome of experiments with a different approach. We just try to predict whether prices will increase or decrease. The problem is posed as a classification problem, where the class labels may be -1, +1, indicating an increase or a decrease in the price of a stock with respect to n days back.

For this purpose, the potential of Random Forests and KNN. Random Forests use an ensemble of Decision Trees to improve the accuracy of classification. Technical indicators such as Alpha 101, rolling average is used as features to train the model. The algorithms are shown to outperform the algorithms used in the existing literature.

For this we have selected few of alpha signal from word quant company.

Data selected for Apple for five years, sample dataset is shown below

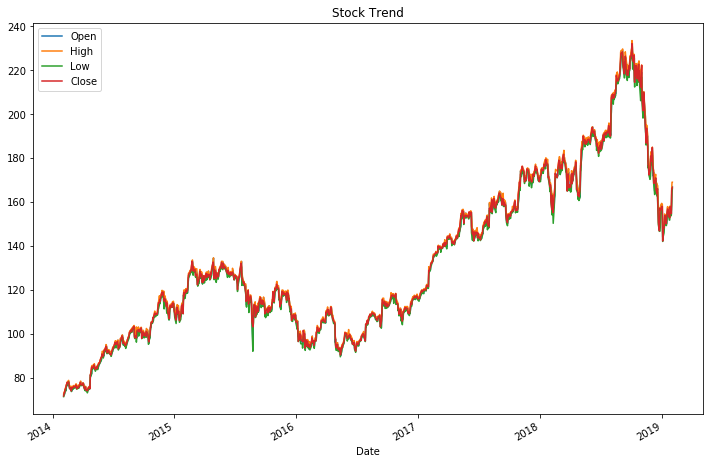
**Table 1: Apple stock price (Raw data)**

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| Date | Open | High | Low | Close | Adj Close | Volume |
| 2/3/2014 | 71.80 | 72.53 | 71.33 | 71.65 | 60.98 | 100366000 |
| 2/4/2014 | 72.26 | 72.78 | 71.82 | 72.68 | 61.86 | 94170300 |
| 2/5/2014 | 72.37 | 73.61 | 72.32 | 73.23 | 62.32 | 82086200 |
| 2/6/2014 | 72.87 | 73.36 | 72.54 | 73.22 | 65.02 | 64441300 |
| 2/7/2014 | 74.48 | 74.70 | 73.91 | 74.24 | 65.93 | 92570100 |
| 2/10/2014 | 74.09 | 76.00 | 74.00 | 75.57 | 67.11 | 86389800 |

**Table 2: Basic Statistics of stock price**

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  | Open | High | Low | Close | Volume |
| count | 1259.0 | 1259.0 | 1259.0 | 1259.0 | 1259 |
| mean | 132.5 | 133.7 | 131.4 | 132.6 | 41839490 |
| std | 36.9 | 37.2 | 36.5 | 36.8 | 21504980 |
| min | 71.8 | 72.5 | 71.3 | 71.6 | 11475900 |
| 25% | 105.5 | 106.5 | 104.8 | 105.7 | 26518500 |
| 50% | 121.1 | 122.2 | 120.3 | 121.3 | 36379100 |
| 75% | 159.0 | 160.1 | 157.6 | 158.7 | 51141350 |
| max | 230.8 | 233.5 | 229.8 | 232.1 | 189977900 |

**Figure 1: Trend of 5 year stock price**



Highest value of stock is 232

Lowest value of stock is 72

1. Overview of the final process

In our experiments, the time series data acquired is first exponentially smoothed. Then the technical indicators are extracted. Technical indicators provide insights to the expected stock price behaviour in future. These technical indicators are then used as features to train the clas­sifiers. The indicators used in the current work will be discussed in this section.

Exponential smoothing grants larger weights to the recent observations and exponentially decreases weights of the past observations. The exponentially smoothed statistic of a series Y can be recursively calculated as:

**Figure 2: Framework of supervised learning in the current work**

Calculation of alphas, there are several alphas and we have selected Alpha 101 for our price predictions.

In this section we describe some general features of our 101 formulaic alphas. The alphas

are proprietary to WorldQuant LLC and are used here with its express permission. We provide as many details as we possibly can within the constraints imposed by the proprietary nature of the alphas. The formulaic expressions

Formula 1: **Alpha#101**: ((close - open) / ((high - low) + .001))

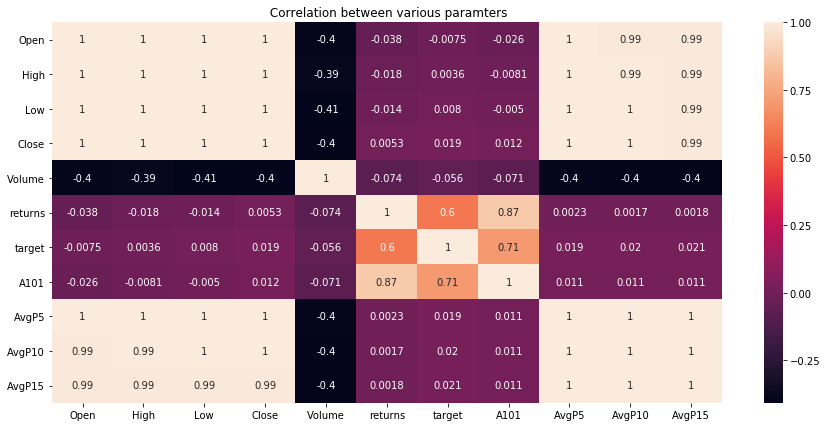
The target label or Y of the machine learning model is created based on this formula. When the value of target is +1, it indicates that there is a positive shift in the price, -1 indicates that there is a negative shift, giving us an idea of the direction of the prices for the respective stock. The target values are assigned as labels to the feature matrix.

Formula 2: **target** = sign(Today close price - Yesterday close price)

1. Solution overview

For this we have pair plot to compare the overall signal correlation amongst various parameters. This pair plot shows that all parameters are highly correlated.

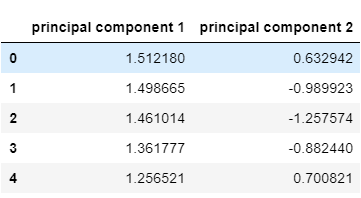
**Figure 3: Heatmap of stock parameters**

****

**Principal component analysis (PCA):**

Apply PCA on all the feature for dimension reduction: Linear dimensionality reduction using Singular Value Decomposition of the data to project it to a lower dimensional space. PCA is applied to reduce the dimensions of the data set and the reason for this is as per Linear Algebra if the prediction be showcased on 2 dimensional then can be easily shifted to N-Dimensions by doing PCA we are calculating eigen values for eigen vectors. Every pair of the eigen vector are perpendicular to each other and hence are termed to be 0 because the cos 90 is 0

**Table 3: Principal component**



This is final data frame which is going to be pass into model, feature extraction

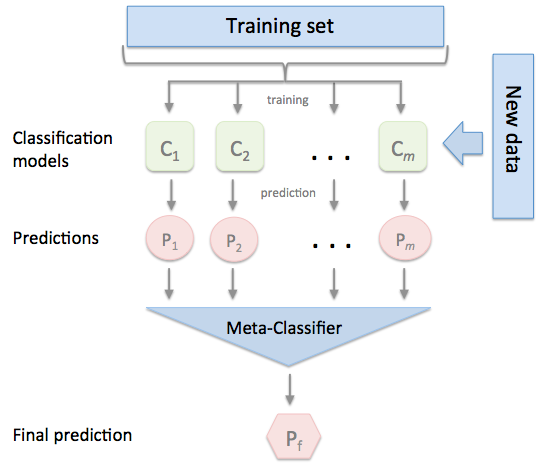
**Table 4: Final Data frame for model development**

|  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Date | Open | High | Low | Close | Volume | returns | target | A101 | AvgP5 | AvgP10 | AvgP15 |
| 1/11/2019 | 152.4 | 154.0 | 151.3 | 152.8 | 27023200.0 | 0.0 | -1 | 0.150 | 154.339 | 155.387 | 158.401 |
| 1/10/2019 | 152.4 | 153.8 | 151.1 | 152.9 | 35780700.0 | 0.0 | -1 | 0.190 | 153.842 | 154.769 | 157.493 |
| 1/9/2019 | 152.1 | 154.0 | 150.7 | 153.0 | 45099100.0 | 0.0 | 1 | 0.281 | 153.381 | 154.355 | 156.596 |
| 1/8/2019 | 151.4 | 153.5 | 150.0 | 152.5 | 41025300.0 | 0.0 | 1 | 0.319 | 152.928 | 154.016 | 155.690 |
| 1/7/2019 | 150.5 | 152.2 | 148.8 | 151.2 | 54777800.0 | 0.0 | 1 | 0.200 | 152.487 | 153.647 | 154.921 |

1. Model evaluation
   1. stacking **algorithms**

**Stacking** is an ensemble learning technique to combine multiple classification models via a meta-classifier. The individual classification models are trained based on the complete training set; then, the meta-classifier is fitted based on the outputs -- meta-features -- of the individual classification models in the ensemble. The meta-classifier can either be trained on the predicted class labels or probabilities from the ensemble.

**Figure 4: Stacking Classifier**

Output is shown below.

3-fold cross validation:

Accuracy: 0.83 (+/- 0.03) [KNN]

Accuracy: 0.81 (+/- 0.04) [Random Forest]

Accuracy: 0.83 (+/- 0.03) [Support vector]

Accuracy: 0.81 (+/- 0.03) [Stacking Classifier]

* 1. K- Nearest Neighbors:

K nearest neighbors is a simple algorithm that stores all available cases and classifies new cases based on a similarity measure (e.g., distance functions). KNN has been used in statistical estimation and pattern recognition already in the beginning of 1970’s as a non-parametric technique.

A case is classified by a majority vote of its neighbors, with the case being assigned to the class most common amongst its K nearest neighbors measured by a distance function. If K = 1, then the case is simply assigned to the class of its nearest neighbor

In our case KNN & Support vector are giving the best accuracy.

K Nearest Neighbours (NN = 20)

Accuracy Score: 85.8288770053476%

Recall : 86

Precision: .86

Equation 1: Distance calculation

* 1. Random forest classifier:

Random forests or random decision forests are an [ensemble learning](https://en.wikipedia.org/wiki/Ensemble_learning) method for [classification](https://en.wikipedia.org/wiki/Statistical_classification), [regression](https://en.wikipedia.org/wiki/Regression_analysis) and other tasks that operates by constructing a multitude of [decision trees](https://en.wikipedia.org/wiki/Decision_tree_learning) at training time and outputting the class that is the [mode](https://en.wikipedia.org/wiki/Mode_(statistics)) of the classes (classification) or mean prediction (regression) of the individual trees. Random decision forests correct for decision trees habit of [overfitting](https://en.wikipedia.org/wiki/Overfitting) to their [training set](https://en.wikipedia.org/wiki/Test_set).

Random forest classifier creates a set of decision trees from randomly selected subset of training set. It then aggregates the votes from different decision trees to decide the final class of the test object.



RF Accuracy Score: 86.08%

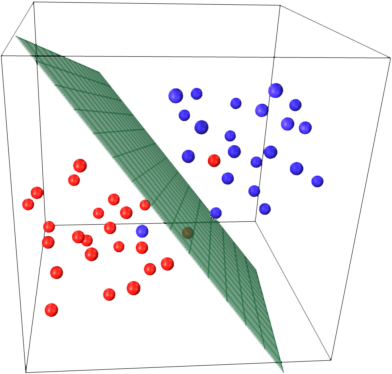
Recall : 86%

Precision: 86%

Figure 5: Random forest classifier

* 1. Support **Vector Machine:**

Support Vector Machine (SVM) is a discriminative classifier formally defined by a separating hyperplane. In other words, given labeled training data (supervised learning), the algorithm outputs an optimal hyperplane which categorizes new examples. In two dimensional space this hyperplane is a line dividing a plane in two parts where in each class lay in either side



Support Vector Machine RBF

Accuracy Score: 88%

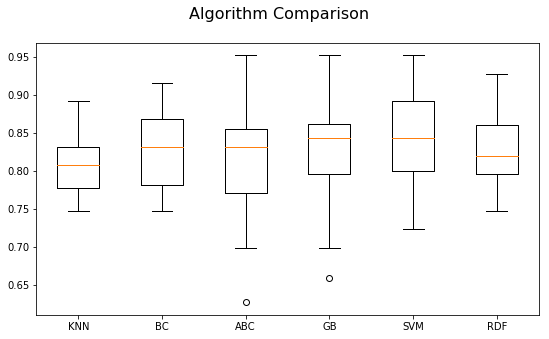
Recall: 88

Precision: 88%

Figure 6: Support vector machine

Model Comparison :

**Figure 7: Algorithm comparison using Cross validation**

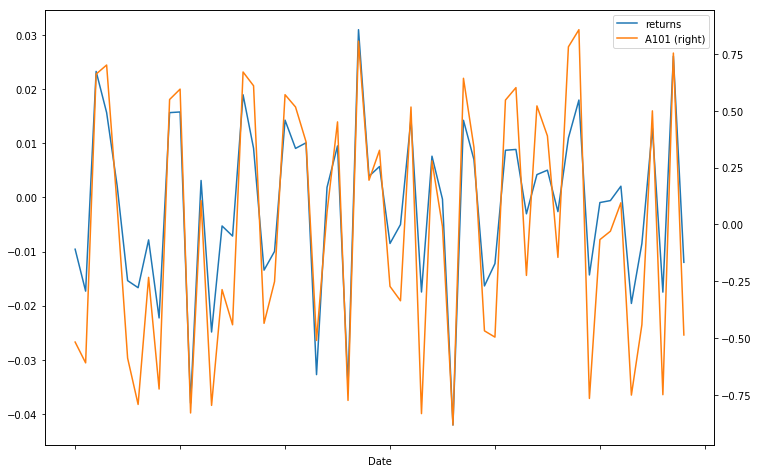


|  |  |  |  |
| --- | --- | --- | --- |
| Model | Accuracy | Precision | Recall |
| KNN | 85% | 86% | 86% |
| SVM | 88% | 88% | 88% |
| RF | 86% | 86% | 86% |

1. Comparison to benchmark
2. Visualization(s)

This trend shows that alpha 101 and close price return are following each other in almost all the occasion during market going high or low.

Here Return is plotted on primary axis and prediction signal A101 is plotted in the secondary axis in the figure.

**Figure 8: Returns vs Alpha 101**

1. Limitations

As the world quant have 101 alpha available, but for academic purpose we have used only one alpha to calculate the signal prediction till end.

1. Closing Reflections

We emphasize that the 101 alphas we present here are not “toy” alphas but real-life trading

alphas used in production. In fact, 80 of these alphas are in production as of this writing.24 To

our knowledge, this is the first time such a large number of real-life explicit formulaic alphas

appear in the literature. This should come as no surprise: naturally, quant trading is highly

proprietary and secretive. Our goal here is to provide a glimpse into the complex world of

modern and ever-evolving quantitative trading and help demystify it, to any degree possible.

Technological advances nowadays allow automation of alpha mining. Quantitative trading

alphas are by far the most numerous of available trading signals that can be turned into trading

strategies/portfolios. There are myriad permutations of individual stock holdings in a (dollarneutral)

portfolio of, e.g., 2,000 most liquid U.S. stocks that can result in a positive return on

high- and mid-frequency time horizons. In addition, many of these alphas are ephemeral and

their universe is very fluid. It takes quantitatively sophisticated, technologically well-endowed

and ever-adapting trading operations to mine hundreds of thousands, millions and even billions

of alphas and combine them into a unified “mega-alpha”, which is then traded with an added

bonus of sizeable savings on execution costs due to automatic internal crossing of trades.