# Predicting Stock Movement of Hang Seng's Components

Presentation by

Lakshmi Prabha S.,

NYCDSA boot camp student.

## Stock market index:

- **A stock market index (or just an 'index') is a number that measures the relative value of a group of <u>stocks</u>.**

- **As the stocks in this group change value, the index also changes value.**

- **If an index goes up by 1% then that means the total value of the securities which make up the index have gone up by 1% in value.**

# Hong Kong's Hang Seng Index and Components

0. **AAC Technologies Holdings Inc. (2018.HK)**

1. **AIA Group Limited (1299.HK)**

2. **BOC Hong Kong (Holdings) Limited (2388.HK)**

3. **China Life Insurance Company Limited (2628.HK)**

4. **China Mengniu Dairy Company Limited (2319.HK)**

5. **China Mobile Limited (0941.HK)**

6. **China Petroleum & Chemical Corporation(0386.HK)**

7. **China Resources Land Limited (1109.HK)**

8. **China Resources Power Holdings Company Limited (0836.HK)**

9. **CITIC Limited (0267.HK)**

10. **CK Infrastructure Holdings Limited (1038.HK)**

11. **CLP Holdings Limited (0002.HK)**

12. **CNOOC Limited (0883.HK)**

13. **Galaxy Entertainment Group Limited (0027.HK)**

14. **Hang Lung Properties Limited (0101.HK)**

15. **Henderson Land Development Company Limited (0012.HK)**
16. **Hengan International Group Company Limited (1044.HK)**
17. **Industrial and Commercial Bank of China Limited (1398.HK)**
18. **Lenovo Group Limited (0992.HK)**
19. **New World Development Company Limited (0017.HK)**
20. **Ping An Insurance (Group) Company of China, Ltd. (2318.HK)**
21. **Power Assets Holdings Limited (0006.HK)**
22. **Sands China Ltd. (1928.HK)**
23. **Sino Land Company Limited (0083.HK)**
24. **Sun Hung Kai Properties Limited (0016.HK)**
25. **Tencent Holdings Limited (0700.HK)**
26. **The Bank of East Asia, Limited (0023.HK)**
27. **The Hong Kong and China Gas Company Limited (0003.HK)**
28. **WH Group Limited (0288.HK)**
29. **Wharf Real Estate Investment Company Limited (1997.HK)**

# Literature Review

| Year pub. | Index | Exp Year | Market moves | Techniques | Predictors | Mean acc. % | Out-performer |
|---|---|---|---|---|---|---|---|
| 2009 | Hang seng index | 2001-2006 | Smooth – very few shocks | Supervised ML techniques | OHLC for major indices, currency | 84 | SVM |
| 2014 | Indian BSE (4 comp) | 2003-2012 | With Shocks | Supervised ML and DL | Technical Indicators | 90 | RFC |

# Literature Review

| Year pub | Index | Exp Year | Market moves | Techniques | Predictors | Mean acc. % | Out-performer |
|---|---|---|---|---|---|---|---|
| 2016 | Thailand's SET50 index | 2009–2014 | With Shocks | ANN + GA | Technical Indicators | 62-68 | ANN + GA |
| 2016 | Norway Index | 2011–2015 | Very few Shocks | Sup. ML and DL | One Technical Indicator | 60 – 62 | Many |

## Questions:

- **How to train each model to extract its best for Hang Seng's components?**

# (or)

- **In what way a model has to be trained to make it perform well for Hang Seng's components.**

- **Which stocks will influence the movement of a particular stock component?**

- **Predict major components of Hang Seng index**

- **Stock prediction was made into a two-class prediction problem.**

- **This was done by breaking the problem down to whether or not a stock would rise.**

- **Experimental Period: 11$^{th}$ march 2013 to 8$^{th}$ March 2018**

# Performance Evaluation Metric

- **Mean Accuracy of all the components**

- **Accuracy is the percentage of correctly classified samples.**

- **Mean ROC-AUC Score of all the components**

- **ROC-AUC score is the area under the curve plotted between true positives and false positives.**

# The ROC Curve



Logistic Regression ROC Curve
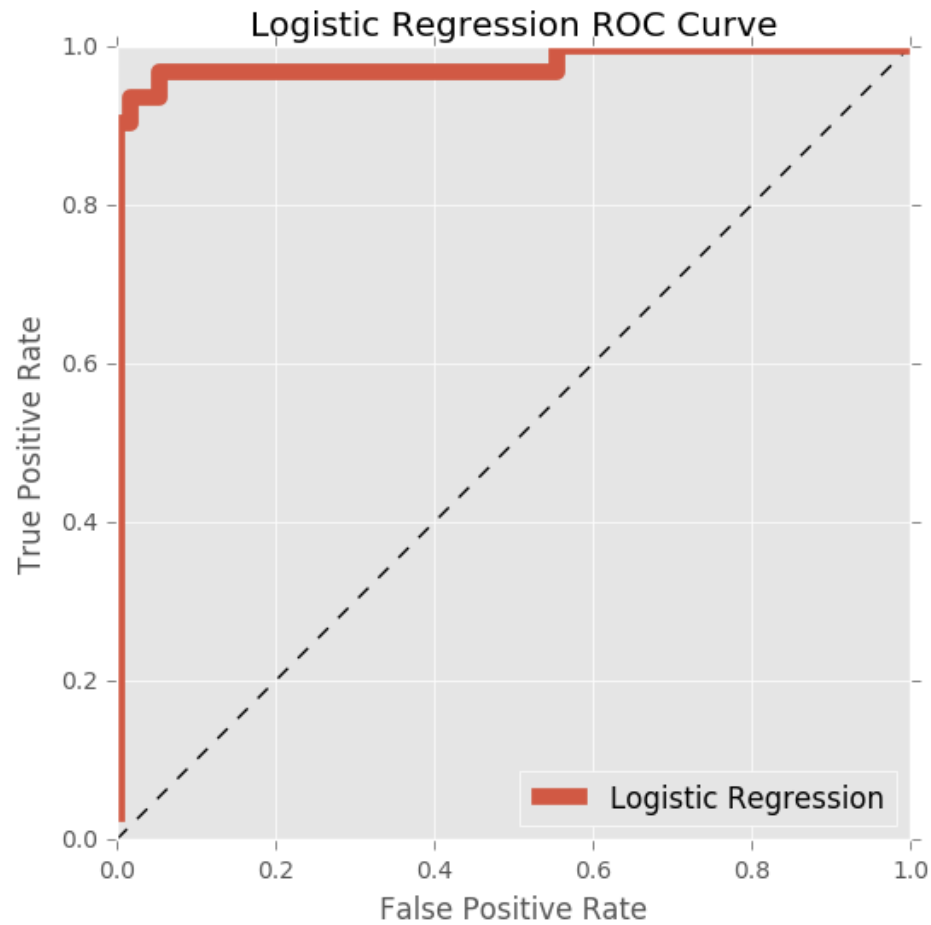
# Experimental Setup I - Considering Raw Prices

## Predictors:

- **Open, High, Low, Close values and Volume traded for**
  - **each of the 30 components**
  - **Hang-seng Index**
  - **Major world stock indices**

- **Span: 1 day (previous) - Predictors: 290**

- **Span: 5 days – Predictors:1040**

## Missing values imputed with 0

## Target

**=[1 if next day's close > current day's close, else -1]**

**Train and test sets: using train_test_split**

# Preliminary Tests

- **set the time line for the prediction, meaning how many days of data that were used to predict forward.**
  - Span considered: 1 day, 5 days, 10 days

# Experimental Setup II –
## Considering Technical Indicator:

Percentage Change for OHLC values
(old-new/old) *100

**Predictors:**

- **Volume traded and % change of OHLC, for**
  - **each of the 30 components**
  - **Hang-seng Index**
  - **Major indices**
  - **Currency**

- **Span: 1 day (previous) - Predictors: 310**

- **Span: 5 days – Predictors: 1060**

- **Span: 10 days - Predictors: 1810**

**Missing vales imputed with 'ffill' and 0**

**Target = [1 if %Change in Close > 0.5, else 0]**

**Train and test sets: using train_test_split**

# Preliminary Tests

- **set the time line for the prediction, meaning how many days of data that were used to predict forward.**

  - **Span considered: 1 day, 5 days, 10 days**

- **set the threshold for the two classes, meaning how much a stock needs to increase in value before it is considered a positive.**

  - **Thresholds considered: 0.5%, 0.1% and 0%**

# Experimental Setup III – Considering Technical Indicator:

## Percentage Change for OHLC values
## +
## Train set: 11th March 2013 – 26th May 2016
## Test Set: 27th May 2016 – 7th March 2018

## Setup III

| Classifier | Mean Accuracy % (1-) | Mean ROC_AUC (1-) |
|---|---|---|
| Logistic Regression | 59.18 | 0.53 |
| SVC | 66.35 | 0.53 |
| Random Forest | 63.59 | 0.5 |
| Gradient Boosting | 67.01 | 0.52 |

## Setup II

| Classifier | Mean Accuracy % (1-) | Mean ROC_AUC (1-) |
|---|---|---|
| Logistic Regression | 62.55 | 0.56 |
| SVC | 68.84 | 0.58 |
| Random Forest | 66.52 | 0.56 |
| Gradient Boosting | 69.94 | 0.59 |

## Setup I

| Classifier | Mean Accuracy % (1-) | Mean Accuracy % (5-) |
|---|---|---|
| Logistic Regression | 62.26 | 63.2 |
| SVC | 63.3 | 63.63 |
| Random Forest | 63.28 | 63.42 |
| Gradient Boosting | 63.31 | 62.7 |
| Simple ANN | 62.06 | - |

/23/2018

# Preliminary Tests-Best Setup

| Classifier | Stock price Threshold | Span | Mean Accuracy % Exp 3 |
|---|---|---|---|
| Gradient Boosting | (0.5) | (1-) | 67.009655 |
| Logistic Regression | (0.5) | (1-) | 59.181034 |
| Random Forest | (0.5) | (5-) (1-) | 63.980000 63.59 |
| SVM | (0.5) | (5-) (1-) | 66.494828 66.35 |

**Setup III**

**Selected  Stock price threshold: 0.5%**

**Selected Span: (1-), 1 day**

# Attempts to Improve Accuracy

- **Train set: 11th March 2013 – 6th March 2017**
  **Test Set: 7th March 2017 – 7th March 2018 (smooth)**

| Classifier | Mean Accuracy % (1-) Smooth Test | Mean Accuracy % (1-) Shock Test |
|---|---|---|
| Logistic Regression | 60.40 | 59.18 |
| SVC | **67.71** | 66.35 |
| Random Forest | 64.99 | 63.59 |
| Gradient Boosting | **67.44** | 67.01 |

# Feature Engineering

- I
  - Fit the entire data with **Random Forest Classifier**
  - Extract 20 best features
  - Fit the reduced model with LR, SVC, RFC and GBC
- II
  - Fit the entire data with **Gradient Boosting Classifier**
  - Extract 20 best features
  - Fit the reduced model with LR, SVC, RFC and GBC

# Results - Feature Engineering

| Classifier | Mean Accuracy % (1-) Without Feature Eng | Mean Accuracy % (1-) With Feature Eng – RFC | Mean Accuracy % (1-) With Feature Eng – GBC |
|---|---|---|---|
| Logistic Regression | 60.40 | 66.64 | 66.52 |
| SVC | 67.71 | 67.71 | 67.91 |
| Random Forest | 64.99 | 65.13 | 66.01 |
| Gradient Boosting | 67.44 | 66.15 | 67.24 |

- **As SVC outperformed all other models, our journey continues with the other three models to extract their best.**

# One Model - Feature Engineering

- **Fit the entire data with GBC**

- **Extract 10, 20, 50, 100, 150, 200 and 250 best features**

- **Fit the reduced models with GBC.**

- **Each component has specific requirements.**

# Hyper Parameter Tuning using Spark

# Mean ROC-AUC Scores

- **1 day span and 0.5% threshold**

| Classifier | Python (before tuning) | Pyspark (before tuning) | Pyspark Tuned |
|---|---|---|---|
| **Logistic Regression** | **0.56** | **0.574** | **0.61** |
| **Random Forest** | **0.56** | **0.6296** | **-** |

# Bullish and Bearish Stocks - Clustering Techniques

- **Bullish – upward trend**

- **Bearish – downward trend**

- **Normal – mixture of positive and negative values**

- **n_clusters = 3**

- **kMeans Algorithm**

- **But the clusters are not properly grouped around 0.**

- **May be the reason for the best stock price threshold to be 0.5% in preliminary tests.**

## Observations: The results might indicate that

- **different machine learning algorithms will perform differently on different stock markets.**

- **the model which gets trained in smooth period and predicts well in shock period may be considered for future predictions.**

- **the stock movement of a stock (say A) is not only dependent on its own OHLC values, but also on other stocks.**

- **it would be better for a trader to watch all the prices (previous day), which are influencing the stock price at hand, before trading.**

## Future Directions:

- **Use other technical indicators mentioned in the literature and analyze the possibility of improving the accuracy.**

- **Extend this study to other major indices.**

## References:

- https://www.hindawi.com/journals/cin/2016/3045254/

- https://www.duo.uio.no/handle/10852/51275

- https://www.sciencedirect.com/science/article/pii/S0957417414004473

- http://ccsenet.org/journal/index.php/mas/article/view/4586

# THANK YOU!!!