
STOCK MARKET FORECASTING
a machine learning case study

Table of Contents

Background

Data Scientist working at a brokerage firm.

As the data scientist of the firm, we are given a task to create a model to help stock market forecasting .

Business Objectives

The objective of this project is to help stock investors gain margin or high advantage in the stock market

Expected Output

The output of this project is a model prediction on the direction of the market stock trend to help analyst in stock market forecasting.

Project Limitation

With the limitation of time and data we will be focusing only on a certain number of stock exchange indexes

Analytic Approach

- Machine Learning Technique

Supervised learning (clasification) to predict market direction

- Perfomance Measures

Precision, **Recall**, F1, Accuracy

Data Collection

The dataset of stock exchange market is taken from Kaggle Stock Exchange Data by Mattiuzc

Data Shape

104,229 rows

9 columns

Data Description

Dataset statistics	Count	Variable Types	Count
Number of variables	9	Categorial	2
Number of observations	104224	Numeric	7
Missing cells	0		
Missing cells(%)	0.0%		
Duplicate	0		
Total size in memory	7.2MiB		
Average record size in memory	72.0 B		

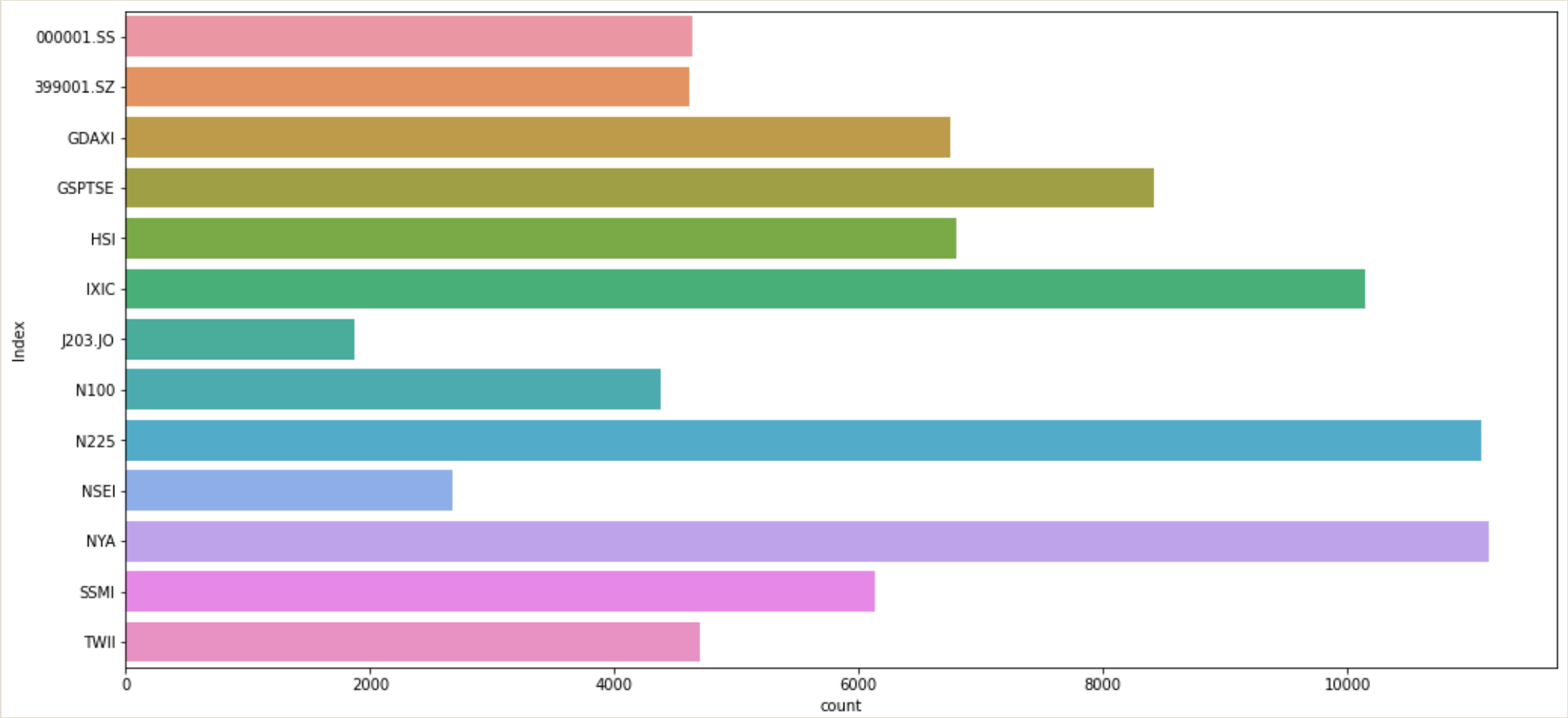
Column Explanation

No	Feature	Data Type	Description	Feature Type
1	Index	Object	Stock Indexes	Categorial
2	Date	Object	Time of Observation	Categorial
3	Open	Float64	Opening price	Numerical
4	Hight	Float64	Highest price during trading day	Numerical
5	Low	Float64	Lowest price during trading day	Numerical
6	Close	Float64	Closing price	Numerical
7	Adj Close	Float64	Closing price adjusted for dividends and stock splits	Numerical
8	Volume	Float64	Number of shares traded during trading day	Numerical
9	CloseUSD	Float64	Close price in terms of USD	Numerical

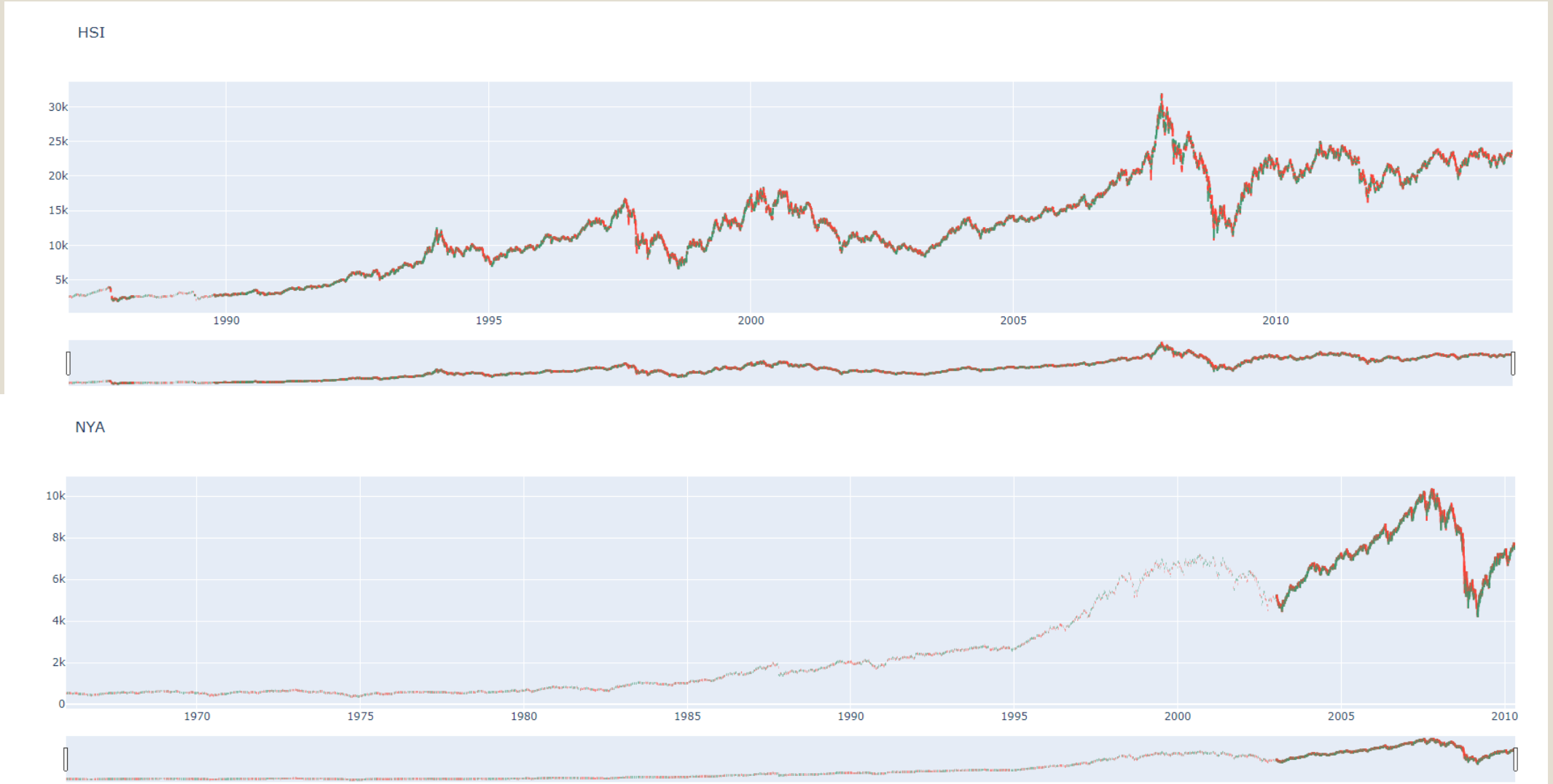
Splitting Data set

Before preprocessing we will be splitting each indexes price history based on the date of the day. The split will be a 80 %, 20% split.

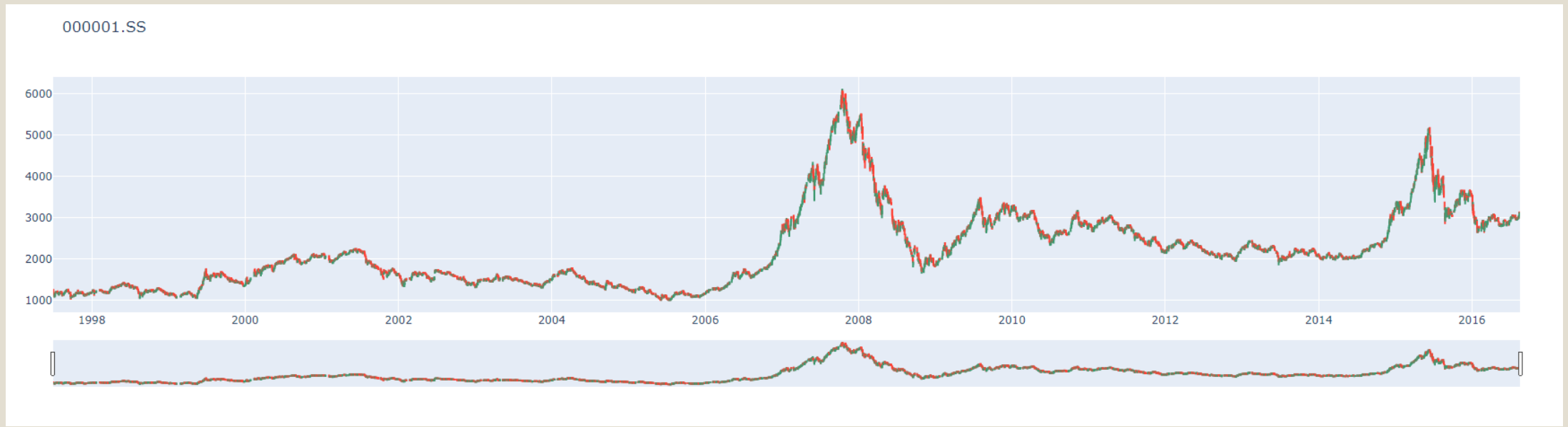
Index's record distribution



Opening and closing price based on their respective date

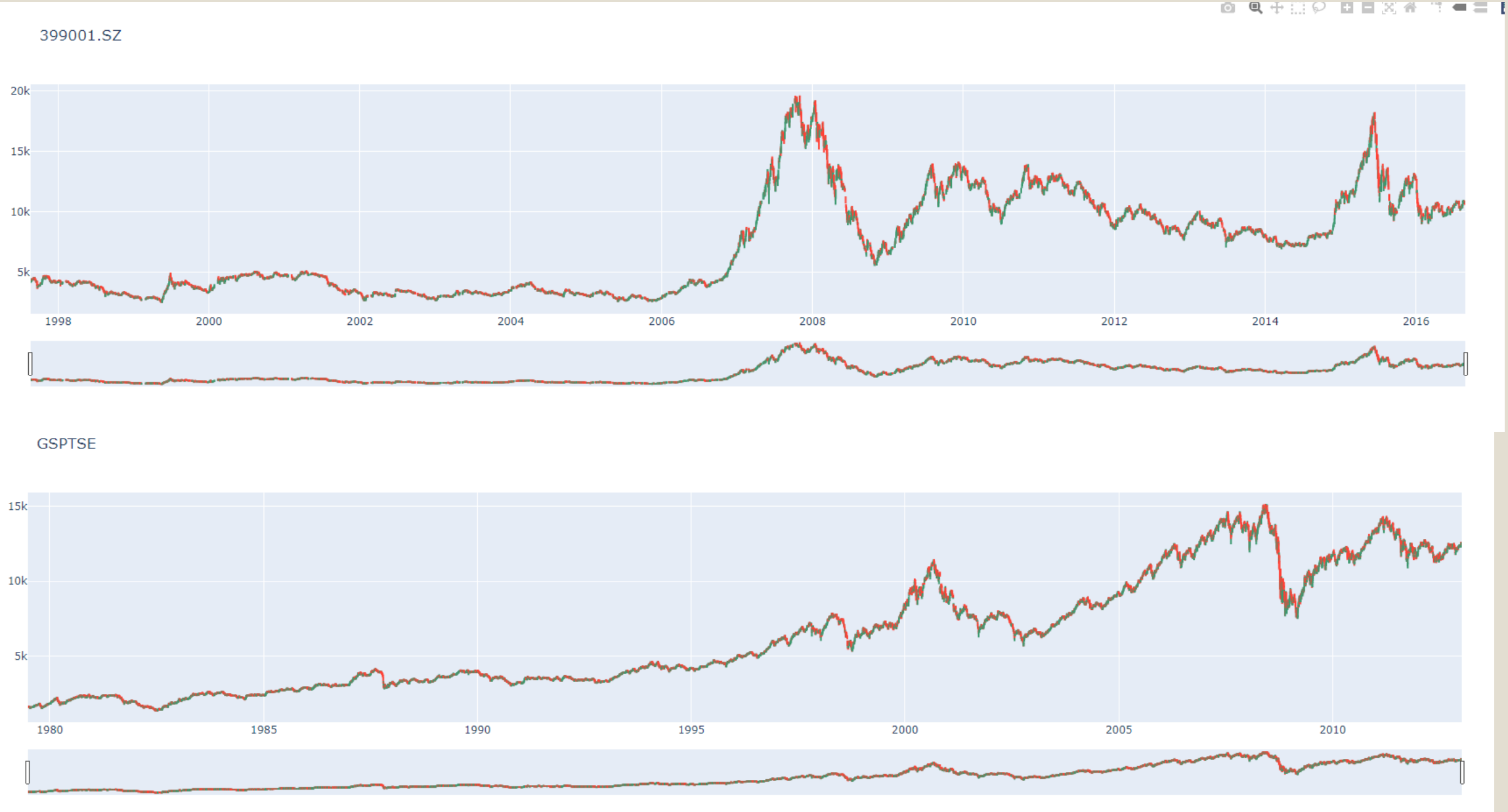


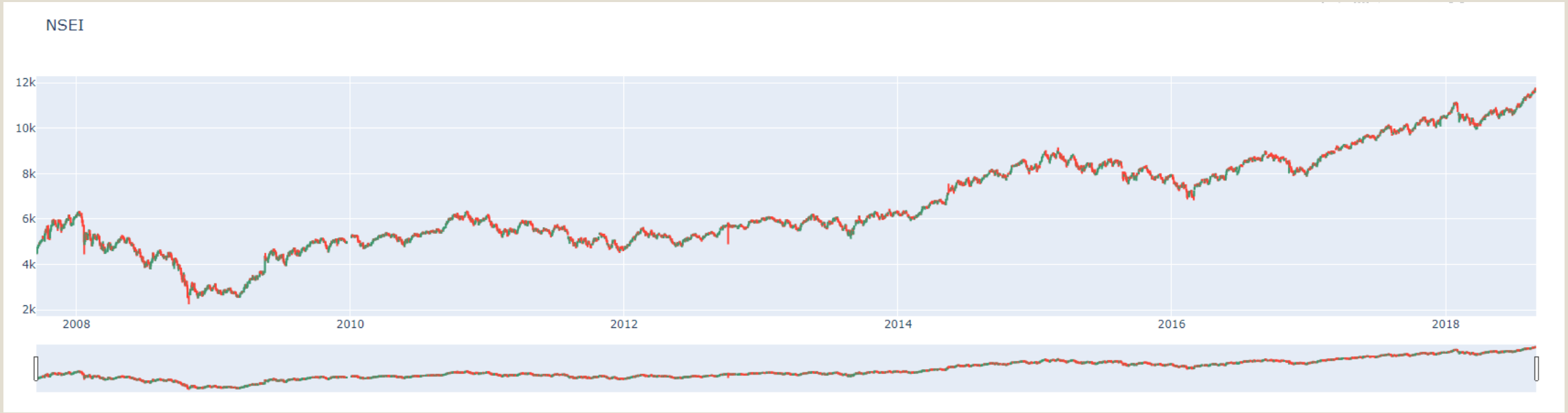

IXIC
5000
4000
3000
2000
1000
0
1975
1980
1985
1990
1995
2000
2005
2010


000001.SS
6000
5000
4000
3000
2000
1000
1998
2000
2002
2004
2006
2008
2010
2012
2014
2016


N225
40k
30k
20k
10k
0
1970
1975
1980
1985
1990
1995
2000
2005
2010


N100
1000
800
600
400
2000
2002
2004
2006
2008
2010
2012
2014
2016


399001.SZ
20k
15k
10k
5k
1998
2000
2002
2004
2006
2008
2010
2012
2014
2016
GSPTSE
15k
10k
5k
1980
1985
1990
1995
2000
2005
2010

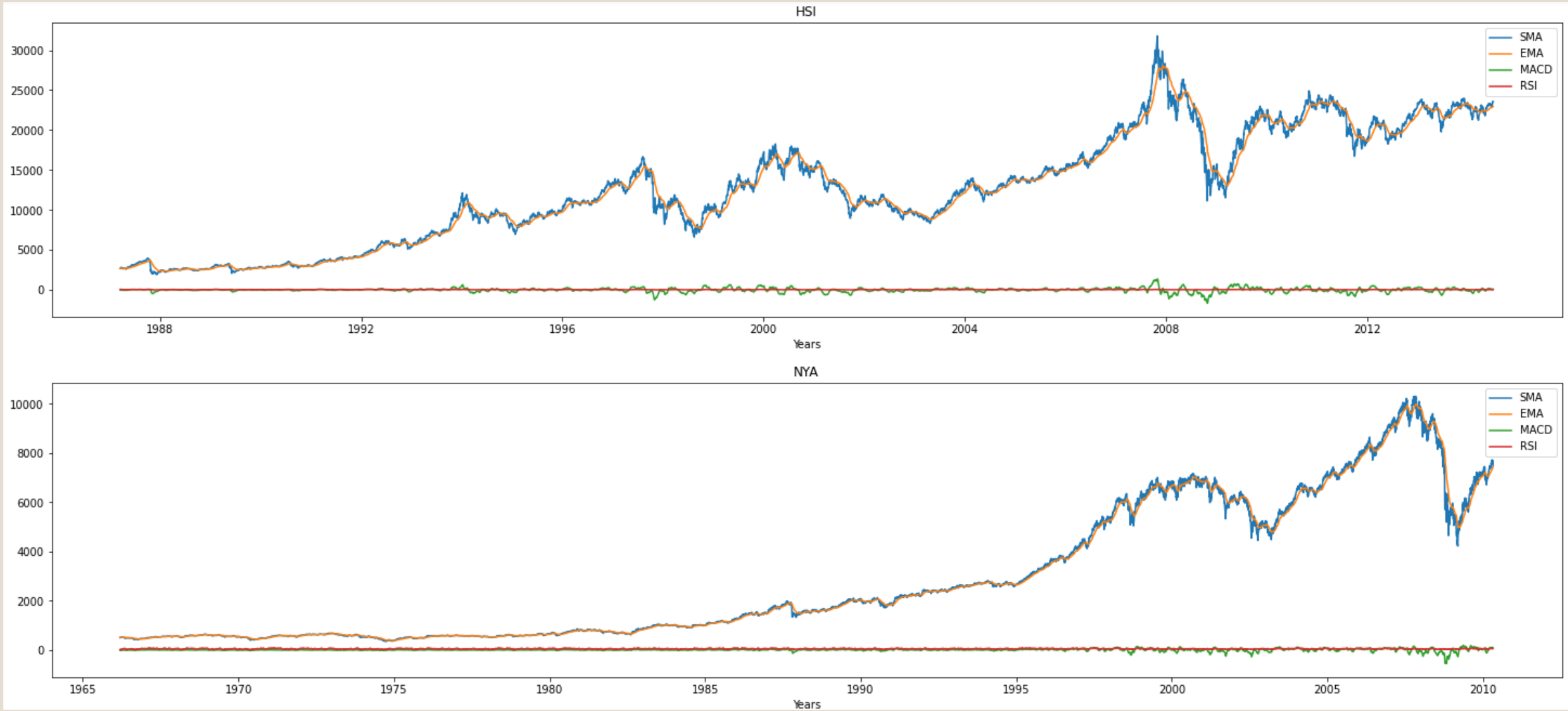

NSEI
12k
10k
8k
6k
4k
2k
2008
2010
2012
2014
2016
2018

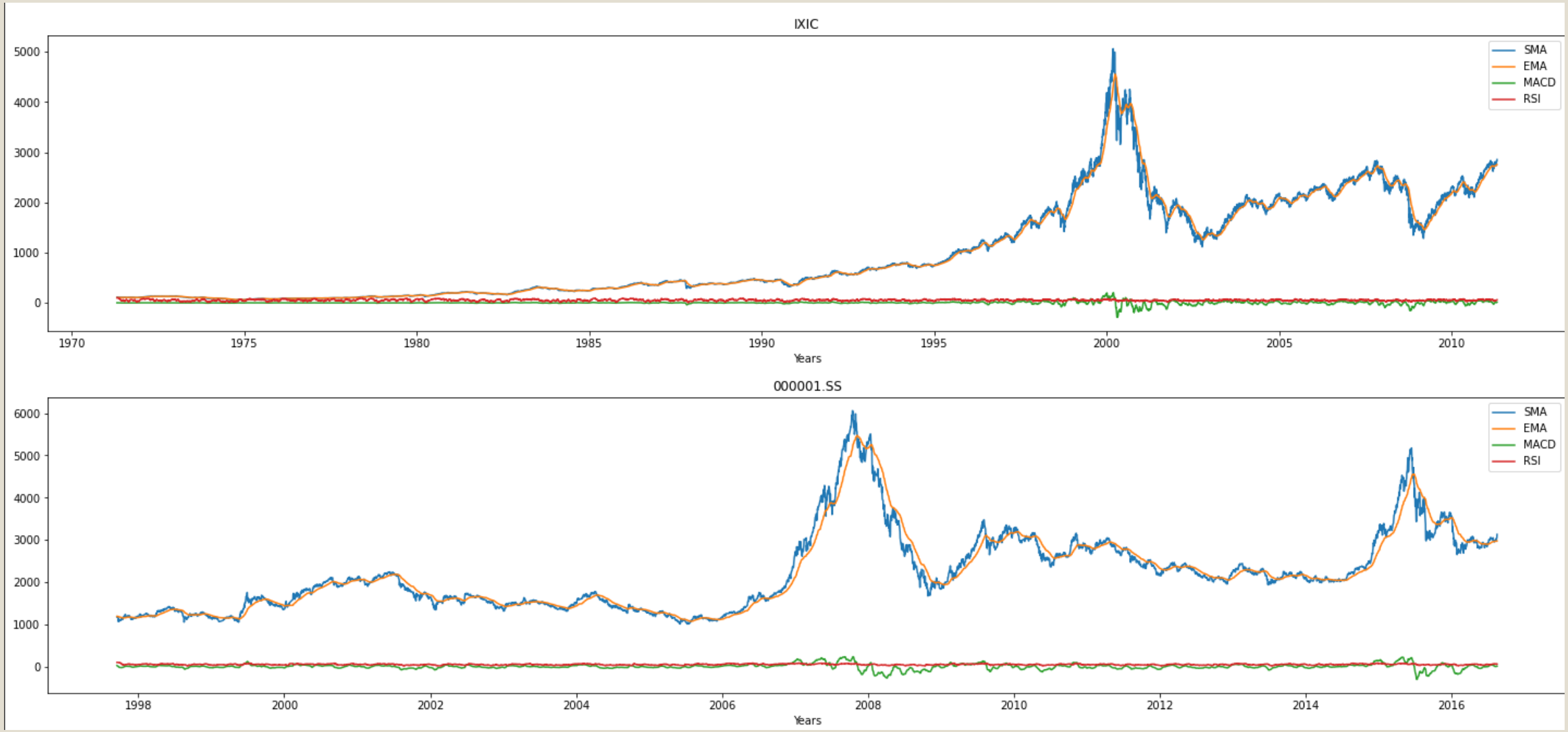

GDAXI
10k
8k
6k
4k
2k
1990
1995
2000
2005
2010


SSMI
10k
8k
6k
4k
2k
1995
2000
2005
2010
2015


TWII
10k
8k
6k
4k
1998
2000
2002
2004
2006
2008
2010
2012
2014
2016

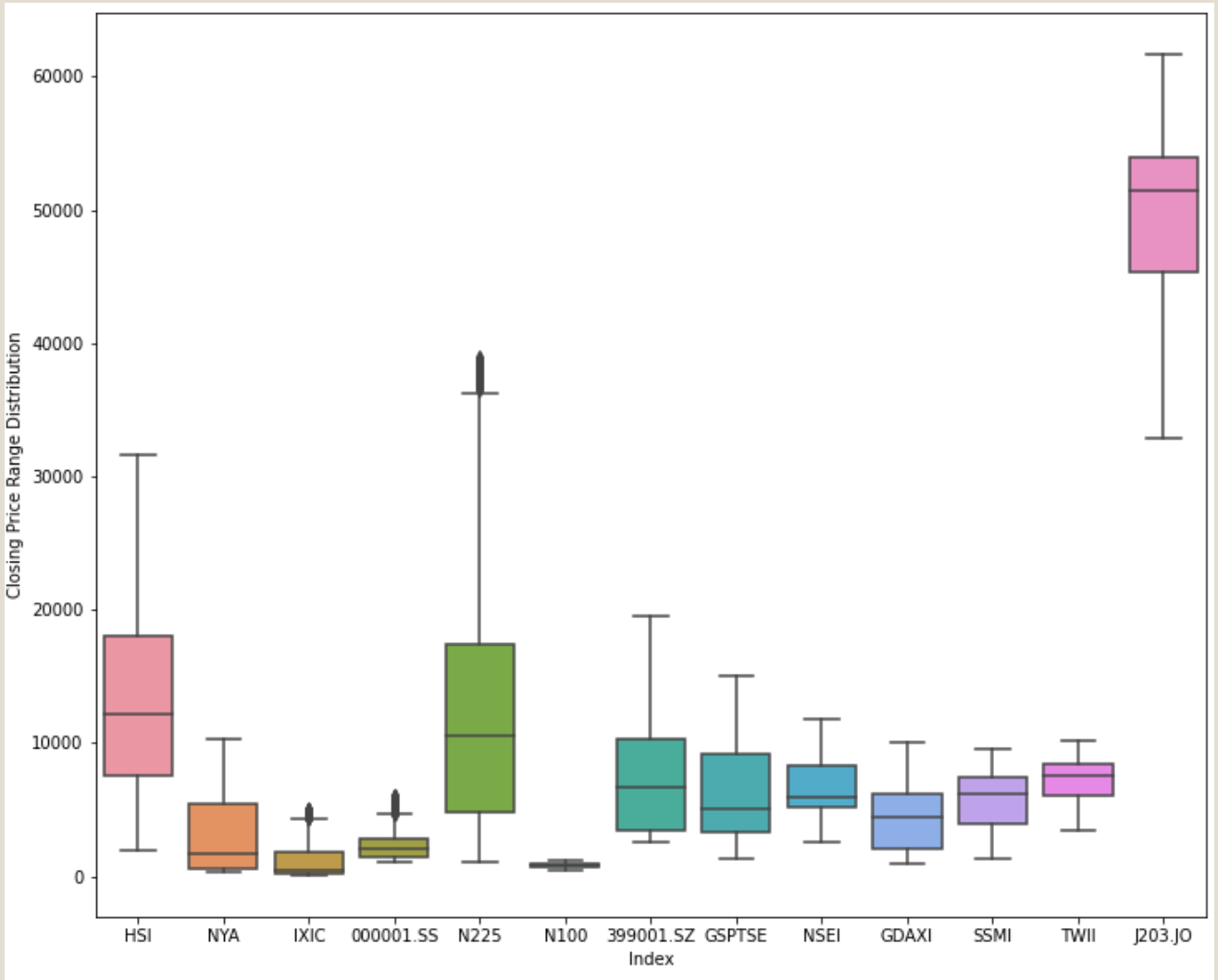
There are patterns occuring in the graph throught out the year we will try to generalize this pattern by smoothing the graph using moving averages.


HSI
SMA
EMA
MACD
RSI
30000
25000
20000
15000
10000
5000
0
1988
1992
1996
2000
2004
2008
2012
Years
NYA
SMA
EMA
MACD
RSI
10000
8000
6000
4000
2000
0
1965
1970
1975
1980
1985
1990
1995
2000
2005
2010
Years


IXIC
SMA
EMA
MACD
RSI
5000
4000
3000
2000
1000
0
1970
1975
1980
1985
1990
1995
2000
2005
2010
Years
000001.SS
SMA
EMA
MACD
RSI
6000
5000
4000
3000
2000
1000
0
1998
2000
2002
2004
2006
2008
2010
2012
2014
2016
Years

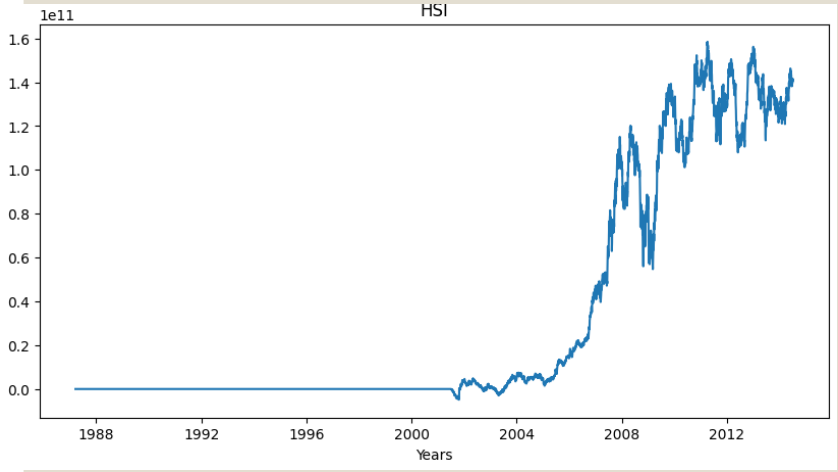

TWII
10000
8000
6000
4000
2000
0
1998
2000
2002
2004
2006
2008
2010
2012
2014
2016
Years
SMA
EMA
MACD
RSI
J203.JO
60000
50000
40000
30000
20000
10000
0
2012
2013
2014
2015
2016
2017
2018
2019
Years
SMA
EMA
MACD
RSI

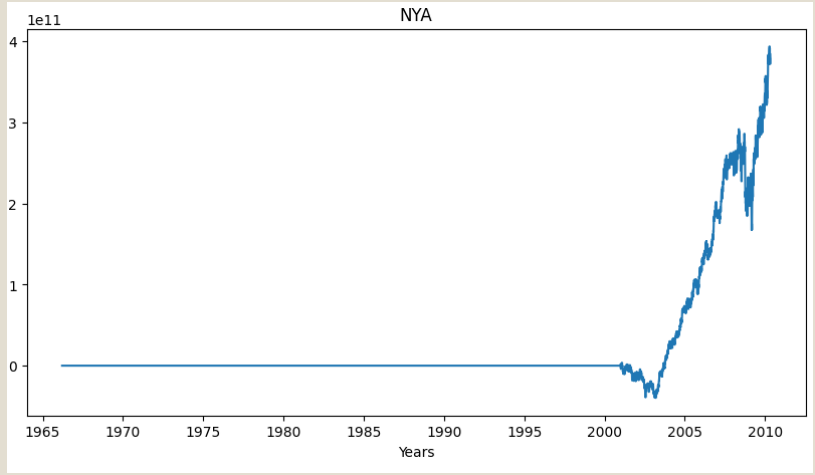
Closing price distribution

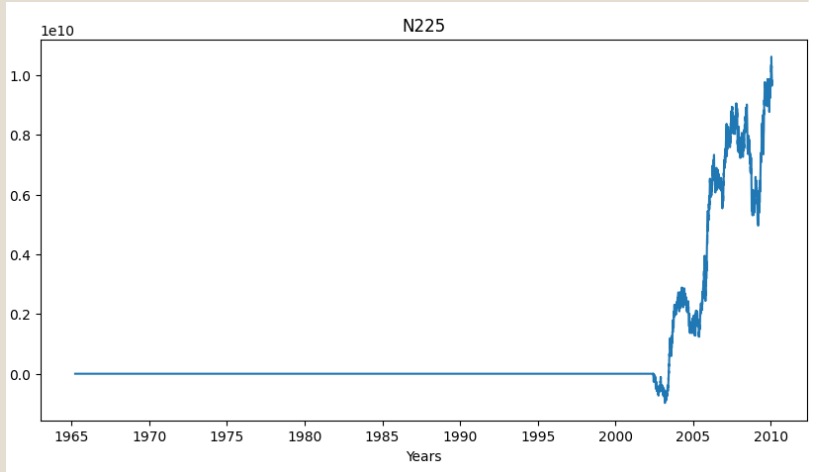


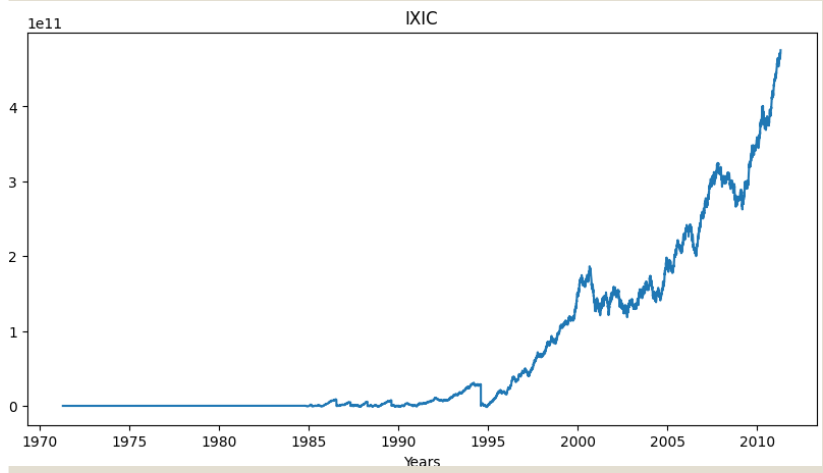
- J203.JO shows a different price range than the other indexes

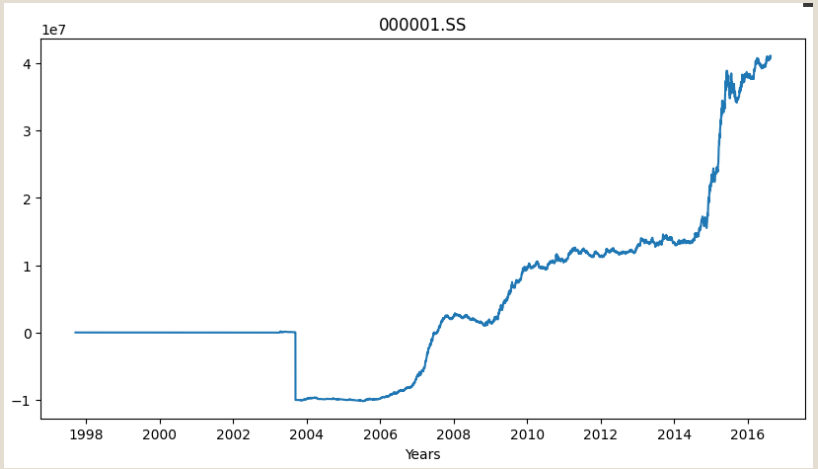
On Balance Volume of each indexes

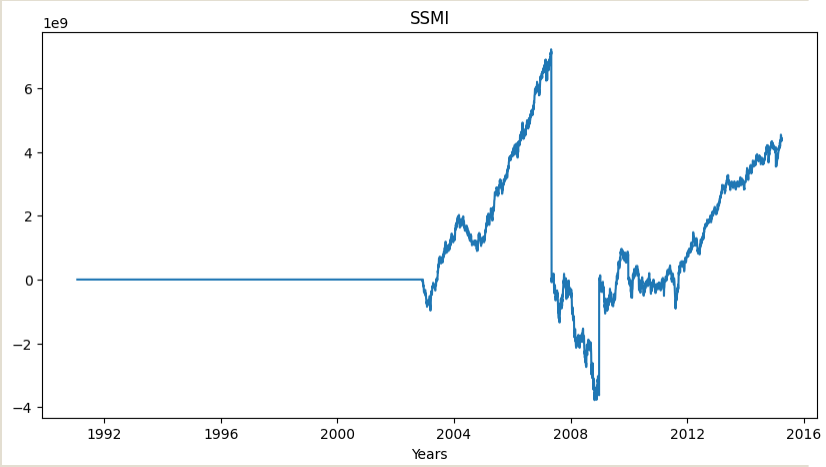












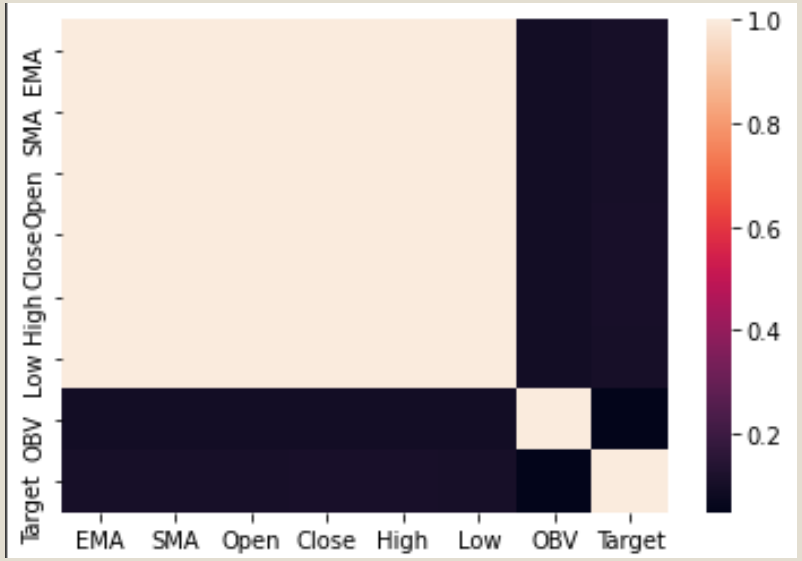
Data Preprocessing

- Feature engineering
- Feature selection
- Encoding

Feature engineering

- SMA (Simple Moving Averages) $= \frac{(A_1 + A_2 + \ldots\ldots + A_n)}{n}$
- EMA (Exponential Moving Averages) $= \text{Price Today} \times \left(\frac{\text{Smoothing}}{1 + \text{Days}}\right) + \text{EMA Yesterday} \left(1 - \left(\frac{\text{Smoothing}}{1 + \text{Days}}\right)\right)$
- CDMA (Convergence Divergence Moving Averages) = 12-EMA – 26-EMA
- RSI (Relative Strength Index) $= 100 - \left[\frac{100}{1 + \frac{Ave\ Gain}{Ave\ Loss}}\right]$
- OBV (On Balance Volume)

Feature Selection & Encoding



- Since SMA have a high linear correlation with price columns we will use the SMA and drop the other three since SMA generalize the trend of the graph better.

- Next step we hot encode the indexes column

Modeling Baseline

Baseline model using DummyClassifier

Precision scores:
0.5011150551901 06

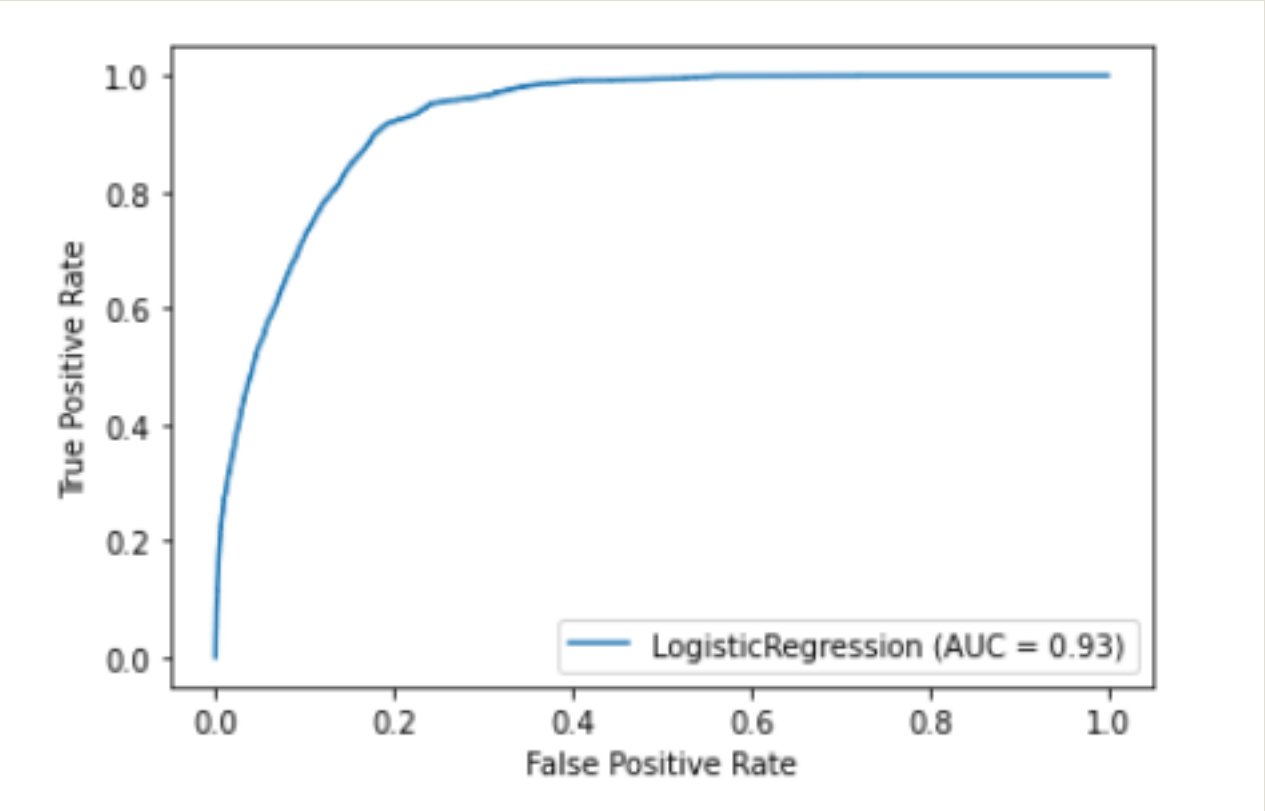
Recall scores:
0.5011149737323 81

F1 scores:
0.5011050082000 984

Model Selection

	Model	Precision	Recall	F1	ROC AUC
0	LogisticRegression	0.776489	0.780509	0.778000	0.928984
1	LGBM	0.820723	0.750410	0.778058	0.937964
2	CatBoost	0.810967	0.749652	0.774454	0.936750
3	Decision Tree	0.745111	0.746629	0.745793	0.746629
4	XGBoost	0.826467	0.744556	0.775617	0.937831
5	Random Forest	0.782515	0.740514	0.758434	0.929760

ROC AUC Curve



- Since we want to improve on recall we have to move the threshold to get a better score

Model Evaluation on unseen data

	precision	recall	f1-score	support
0	0.82	0.98	0.89	14391
1	0.89	0.48	0.62	5758
accuracy			0.83	20149
macro avg	0.86	0.73	0.76	20149
weighted avg	0.84	0.83	0.82	20149

Conclusion

- Logistic regression still prove to be one of the best model to be able to predict the stock market direction trend
- The model could help analyst predict the market stock price with promising result in given more time and resource.