Data Mining Project

Introduction

Group-7

Anshul Agarwal, IWM2017008, Arya Krishnan, ICM2017501
Harsh Aryan, ITM2017003, Priyal Gupta, IHM2017004
Ritik Raj Gupta, IIM2017003

Introduction

This project deals with the time series analysis of the stock market and then using it to predict future trends.

Stock Market:- A stock or share is a financial instrument that represents the percentage of ownership in a company or corporation and represents a proportionate claim on its assets and earnings (what it generates in profits).

Every stock exchange can be represented by a Stock Index value, which is an average value calculated by combining several stocks. This value helps in representing the entire stock market and predicting the change in the market's movement over time.

Early researches used machine learning to create models using the time series data which can predict or forecast sequences or outcomes. But this approach turned out to be very difficult because of the complex features involved in this prediction.

Motivation

New investors often face the issue of when to invest in stocks. Most people consider stock markets as highly risky for investment and do not consider it suitable for trade and investments. The calculation of seasonal variance and the study of the steady flow of any index will help both current and new investors to understand and make a decision to invest in the correct stock/share market.

Problem addressed-

Recently, researchers have tried to improve the accuracy of stock market prediction using machine learning classifiers.

However, the relationship between the prediction factors and the prices is very complex and may not be implementable by individual classifiers. Hence, we will try to form an ensemble model of an SVR and SARIMA to try and model the complex relationship with their aggregate opinion.
We will independently observe their results and then accordingly weigh their opinions by their accuracy of prediction.

The techniques to be used-

Technical Indicators:-
We will try to include some technical indicators such as RSI, MA(moving average), MACD, etc. The number of Wiki mentions and google news mentions will be used as indicators for every company in the Nifty50 group. Also, the same is done for the related products of each company. The relation between the company and its products is determined by the Google knowledge graph which is then adjusted based on its result score. The product which has a greater score affects the company more.

We'll be using the following base learners:

- SVR:-
  Support Vector Regression works on the principle of Support Vector Machine. It allows the prediction of a non-linear model without changing any explanatory variables, based on the functions of the kernel like linear, sigmoid, polynomial, radial basis, etc, rather than the distributions of the dependent and independent variables underlying it.
  It follows the idea of maximal margin and is viewed as a convex optimization problem. Radial basis function is the default used kernel function, the main role of the kernel function is to convert a non-linear function to a linear function, finding a fit and mapping it to the original function. Hence, in SVR we try to fit the error in a threshold.

- SARIMA:-
  Seasonal Autoregressive Integrated Moving Average (SARIMA) model is an expansion of the ARIMA model. SARIMA is used to forecast a time series on the basis of some past values. It also supports the seasonal data which is not a part of the ARIMA model. It includes three hyperparameters- autoregression, differencing and moving average to add the seasonality feature in the model.

Dataset used-

Yahoo Stock (from March 2010 to March 2020) has been collected to be used in training.
The companies used are part of the Nifty50 group of companies as of 17 November 2019.

The dataset collected consists of 8 features (date, open, high, low, close, volume, adj close, comp).

Trends-

Fig1: Shows the closing price trend of Adani Ports for over 10 years-

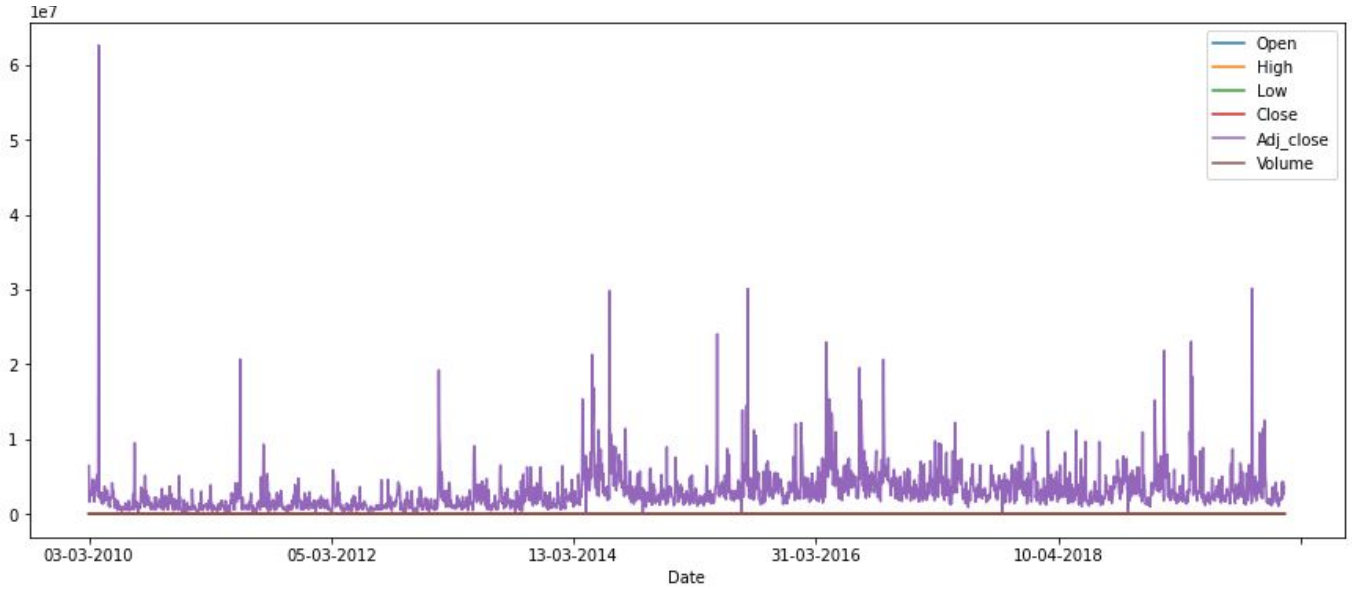
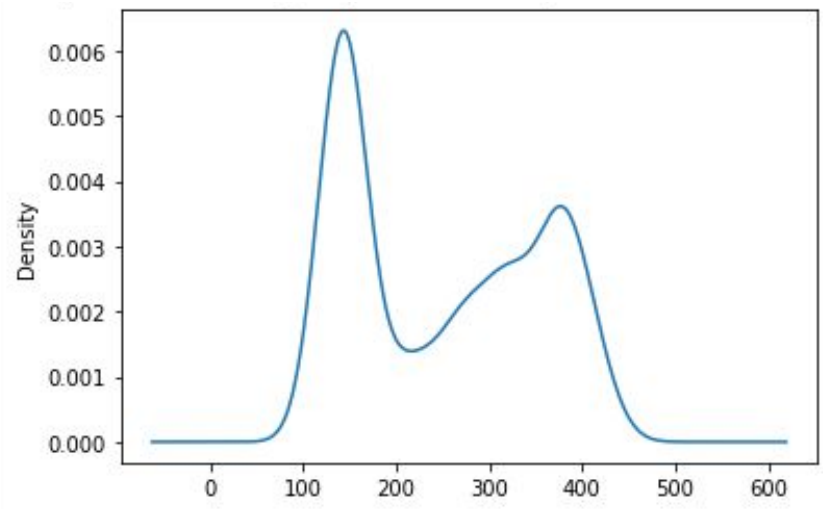


Fig2: This is a KDE plot of the closing price of Adani Ports over 10 years-



We see that the data is not Gaussian (normally distributed).

Fig3: Shows the closing price trend of Zee Entertainment for over 10 years-

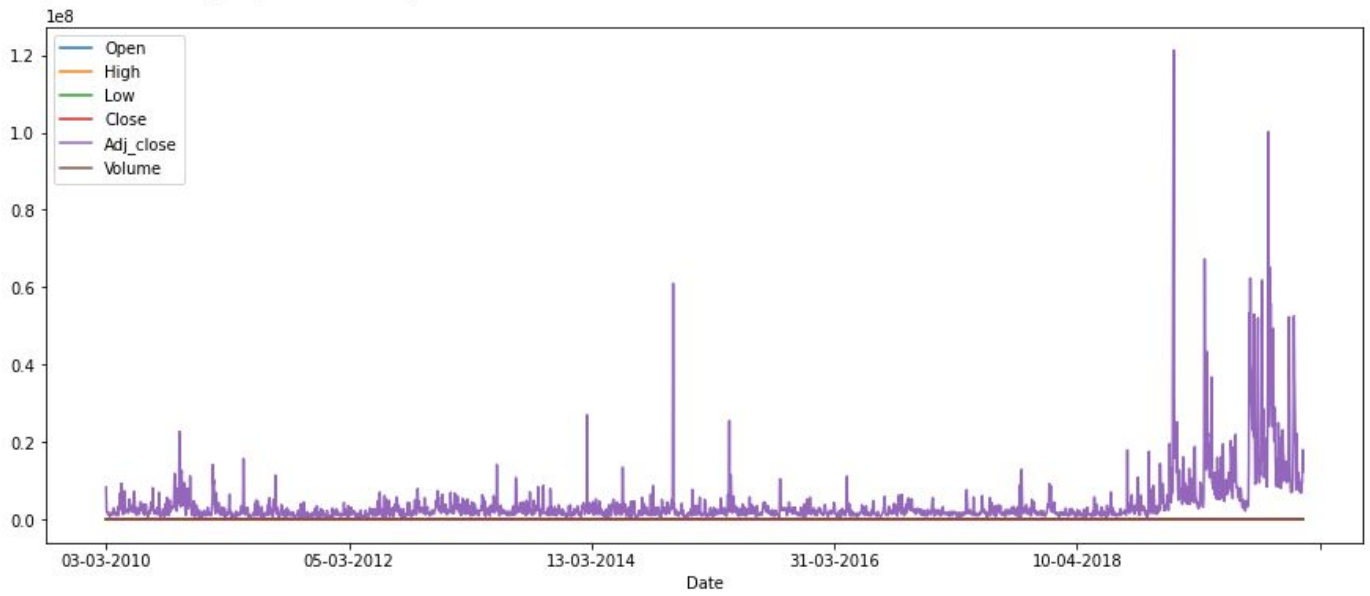
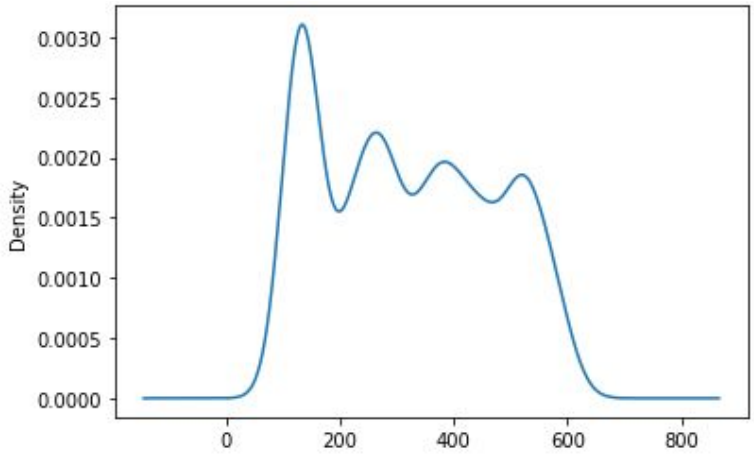


Fig5: This is a KDE plot of the closing price of Zee Entertainment over 10 years-



Similarly, we need to analyze the trend of the company to which we fit our model and make appropriate transformations to the data prior to modeling.