Stock market predictions using time series modeling

Prashank Kadam
Yogesh Nizzer

Abstract

Stock markets have always been one of the most important applications of Machine Learning where we could calculate the fluctuations in a company's stock price based on various factors like capital expenditure, Revenue, Investments, Intraday stock fluctuations, etc. As a part of this project, we will be performing careful feature selection and analysis of the New York Stock Exchange dataset [1] to predict the stock prices of any of the listed companies based on their previous stock and their annual SEC 10k fillings [2] for a period of four years.

Introduction

There have been many implementations of stock market prediction models. Most of them have been using very rudimentary techniques to make these predictions. Very early models, which are still being used by many finance firms are based on the heuristics that the companies have formulated by themselves and do not involve any formal machine learning. Large subset of stock prediction models involve basic regression techniques like linear regression and only recently have researchers started to implement more advance machine learning algorithms in these models. But there have always been limitation on the predictive accuracy of these models due to data constraints or model robustness. As a part of this project we intend to find the best possible technique for predicting stock prices, we will first use more widely used regression techniques like linear and polynomial regressions and then move forward to more advanced algorithms based on moving average and neural networks to predict the stock prices. Finally we show a comparison of various techniques and which of them serves our purpose the best. We also intend to perform some advanced experiments using different supervised learning techniques on different features of the dataset and combining the results to make our predictions.

Proposed Project

Techniques:

Linear regression – This is the most basic form of regression that uses the relationship between the variables to find the best fit line that can be used to make predictions about the data. Although this method is widely used in various predictive stock models, it is not very accurate as the stock prices are not guaranteed to vary linearly. This algorithm will serve as a base for comparing the rest of our algorithms.

Polynomial regression – A polynomial regression uses the relationship between various input variables to find the best fitting curve for the given degree of the polynomial that can be used to make predictions. Appropriate regularization terms will also be included in the regression equation if necessary. We do not expect the dataset to contain many outliers, hence polynomial regression would be a good approximation technique for making our predictions.

Autoregressive Integrated Moving Model (ARIMA) [3] – It is a class of statistical models used for analyzing and forecasting time series data. It explicitly caters to a suite of standard structures in time series data, and as such provides a simple yet powerful method for making skillful time series forecast. A linear regression model is constructed including the specified number and type of terms, and the data is prepared by a degree of differencing in order to make it stationary, i.e. to remove trend and seasonal structures that negatively affect the regression model. This technique could prove to be very useful in out stock prediction application.

Recurrent Neural Network (RNN) [4] – It is a type of neural network that contains loops allowing information to be stored inside the neural network. This configuration of neural networks uses reasoning from previous experiences to predict the upcoming events. It works great with time series data and since our stock prediction data is completed a time series dataset, this neural network configuration should perform very well for our model.

Long Short Term Memory Network (LSTM) Networks [4] – These are modified versions of RNNs which consists of a gated memory unit which allows information to be stored for a longer time within the network. The gate has Sigmoid and tanh activators which decide the values to let through and assign weightage respectively. The vanishing gradient problem of RNNs is resolved using this method. It works great with time series data and hence is expected to perform very well with our choice of the dataset.

Dataset:

The dataset that we will be using consists the New York Stock Exchange listings from 2012-2016 and consists of following files:

prices.csv: raw, as-is daily prices. Most of data spans from 2010 to the end 2016, for companies new on stock market date range is shorter. There have been approx. 140 stock splits in that time, this set doesn't account for that.
securities.csv: general description of each company with division on sectors
fundamentals.csv: metrics extracted from annual SEC 10K fillings (2012-2016), should be enough to derive most of popular fundamental indicators.

References:

[1] https://finance.yahoo.com/quote/CSV/history/

[2] https://www.nasdaq.com/market-activity/stocks/csv/financials

[3] *Lui et. al.*, Online ARIMA Algorithms for Time Series Prediction (Thirtieth AAAI Conference on Artificial Intelligence (AAAI-16))

[4] *Alex Sherstinsky*, Fundamentals of Recurrent Neural Network (RNN) and Long Short-Term Memory (LSTM) Network (arXiv:1808.03314v7 **[cs.LG]**)