TOPIC – Stock market prediction

Business problem:

The challenge of the stock price forecast is the most crucial component for companies and equity traders to predict future revenues. A successful and accurate prediction of future stock prices ultimately results in profit maximization.

The stock market is one of the major fields that investors are dedicated to, thus stock market price trend prediction is always a hot topic for researchers from both financial and technical domains. In this research, our objective is to build a state-of-art prediction model for price trend prediction, which focuses on short-term price trend prediction.

Background / History:

Stock market prediction is the act of trying to determine the future value of company stock or other financial instruments traded on an exchange. The successful prediction of a stock's future price could yield significant profit. The efficient-market hypothesis suggests that stock prices reflect all currently available information and any price changes that are not based on newly revealed information thus are inherently unpredictable.

Data Explanation:

The datasets are fetched from the National Stock Exchange website. For this project, we are planning to predict two random stocks and the dataset will have the following fields:

1. Date – Trade date.
2. Symbol – Tick value of the stock.
3. Prev. Close – Previous day's close price.
4. Open – Open price of the day.
5. High – The highest price in a day.
6. Low – Lowest price in a day.
7. Last – Last traded price of the day.
8. Close – Close price of the day.

Methods:

For this project, I am planning to use time series methods because it is recorded at regular time intervals, and the order of these data points is important. Therefore, any predictive model based on time series data will have time as an independent variable. The output of a model would be the predicted value or classification at a specific time.

For a new investor general research that is associated with the stock or share market is not enough to make the decision. The common trend towards the stock market among the society is highly risky for investment so most of the people are not able to make decisions based on common trends. The seasonal variance and steady flow of any index will help both existing and new investors to understand and make a decision to invest in the share market.

Analysis:

Stock and financial markets tend to be unpredictable and even illogical. Due to these characteristics, financial data should necessarily possess a rather turbulent structure which often makes it hard to find reliable patterns. Modeling turbulent structures requires machine learning algorithms capable of finding hidden structures within the data and predicting how they will affect them in the future.

Stock prices are not randomly generated values instead they can be treated as a discrete-time series model which is based on a set of well-defined numerical data items collected at successive points at regular intervals of time.

Conclusion:

As of now, to conclude we have changed the method from linear regression to time series as directed. We will split the training dataset into train and test sets and we will use the train set to fit the model and generate a prediction for each element on the test set. Finally, we will track of all observations in a list called history that is seeded with the training data and to which new observations are appended at each iteration.

Assumptions:

As mentioned early, we are going to implement various time series methods to achieve the results. It works perfectly to predict the results based on time. We are assuming here none other factors are affecting the results except the time. Apart from this assumption, there is no assumption made and the values which we are going to implement in this project are real-time values only.

Limitations:

In this project, we don't know the exact accuracy until we complete the coding part. And there are some other factors like company's revenue, products they realize, management people will directly affect the stock price which is not considered here. So, probably the accuracy of the results may be less in this approach which can't be mentioned now.

Challenges:

With the resurgence of machine learning and artificial intelligence, never has it been easier to implement predictive algorithms both new and old. With just a few lines of code, state-of-the-art models can be readily accessible at the fingertips of the budding data enthusiast, ready to conquer whatever insurmountable digital task may lay at hand. But a little bit of knowledge can be a dangerous thing. While much of machine learning can be attributed to statistics and programming what is equally important, but often skipped over in favor of instant gratification, is domain knowledge. But there are reasons for the project might fail which are listed as follows:

1. Selection Bias – This is problematic as the stock selection is not an arbitrary process, it is part of the investment decision-making process that requires a model in itself.

2. Incorrect correct application of pre-processing – Standard rinse, wash and repeat data pre-processing techniques like standardization cannot be directly applied to stock prices.
3. Look ahead bias – Frequently, observations associated with particular dates would not have been available at that date.

Future Uses:

Stock market prediction aims to determine the future movement of the stock value of a financial exchange. The accurate prediction of share price movement will lead to more profit investors can make. If the predictions came well with accuracy, we can implement it as a mobile app with a good User Interface for public use.

Recommendations:

As per websites like Kaggle and other data science websites, the recommended model for stock market prediction is the ARIMA model. A famous and widely used forecasting method for time-series prediction is the AutoRegressive Integrated Moving Average (ARIMA) model. ARIMA models are capable of capturing a suite of different standard temporal structures in time-series data.

Implementation Plan:

We can use a library called “nsepy” to extract the historical data for the Indian stock companies. Then will create a few visualizations to show per day close price of a stock which we are selecting for analysis. Then, we need to check if a series is stationary or not because time series analysis only works with stationary data. If we fail to reject the null hypothesis, we can say that the series is non-stationary. This means that the series can be linear. If both mean and standard deviation are flat lines(constant mean and constant variance), the series becomes stationary. Then, we are going to create an ARIMA model and will train it with the closing price of the stock on the train data.

Ethical Assessment:

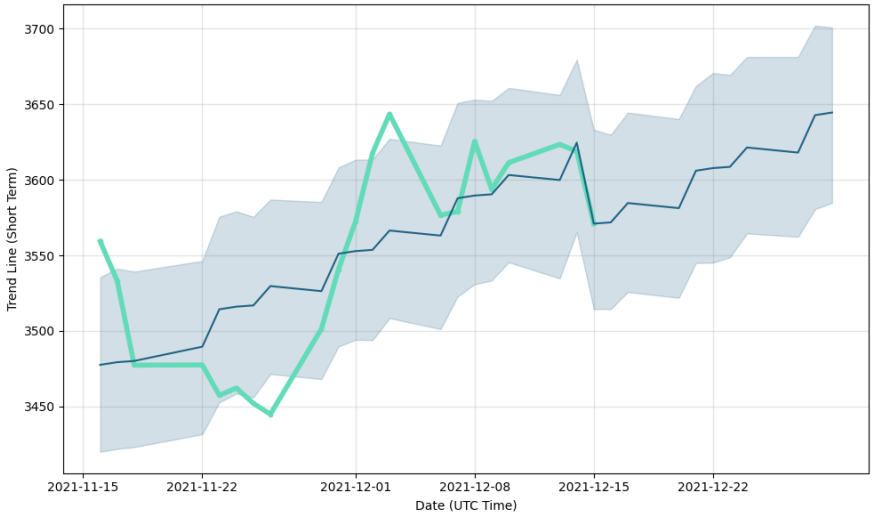
In this project, we are going to use the data which is available for public use from websites like the national stock exchange and money control. So, there are no ethical issues in handling data. But the ethical issue might raise when we release the results of this project because it's totally experimental and we don’t know exactly how the model will behave for each stock. So, there are some potential threats that the result may mislead the investors. Thus, to avoid this kind of issue, we need to test the model thoroughly with different stocks.

Reference:

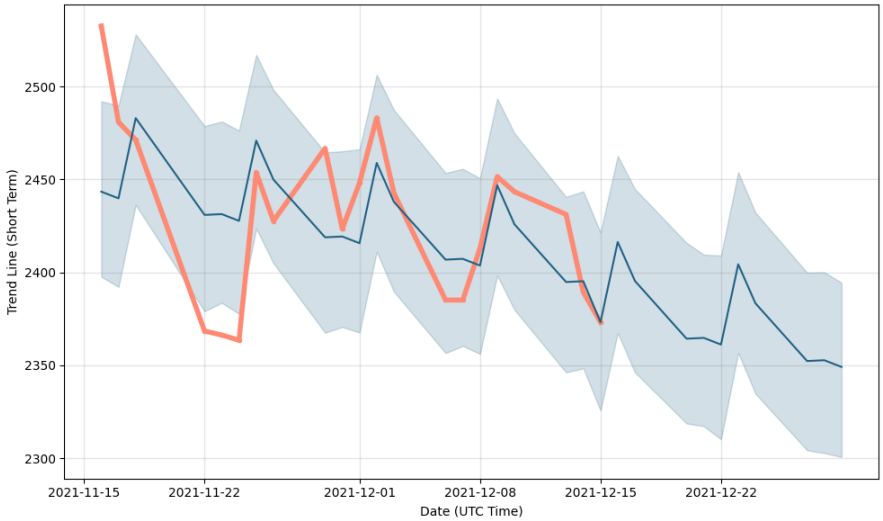
1. Stock Market Prediction with Linear Regression | Kaggle
2. Time-Series Forecasting: Predicting Stock Prices Using An ARIMA Model | by Serafeim Loukas | Towards Data Science
3. Stock Price Prediction and Stock Price Forecasting using Stacked LSTM (analyticsvidhya.com)

Illustrations:

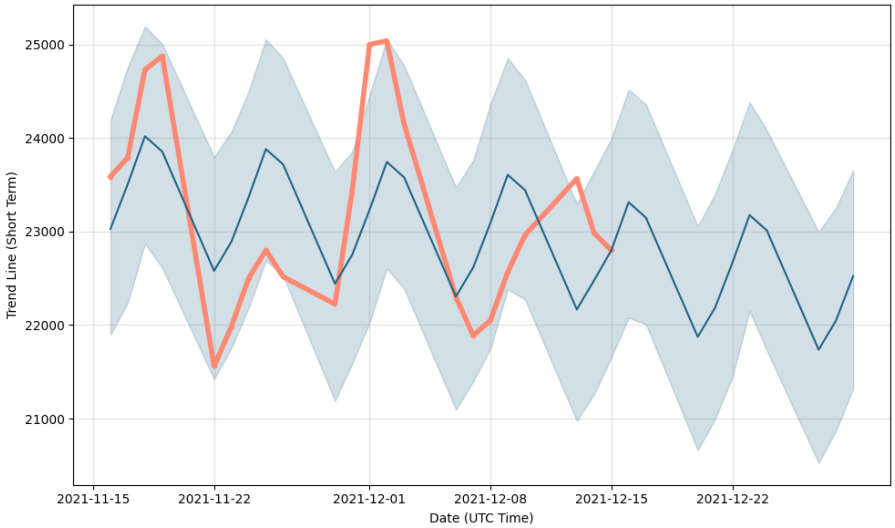
1. TCS Stock prediction:



2. RELIANCE Stock prediction:



3. ITC Stock prediction:



Appendix:

1. ARIMA model - ARIMA is an acronym that stands for AutoRegressive Integrated Moving Average. It is a generalization of the simpler AutoRegressive Moving Average and adds the notion of integration. This acronym is descriptive, capturing the key aspects of the model itself. Briefly, they are:
    a. AR: Autoregression. A model that uses the dependent relationship between an observation and some number of lagged observations.
    b. I: Integrated. The use of differencing of raw observations (e.g. subtracting an observation from observation at the previous time step) to make the time series stationary.
    c. MA: Moving Average. A model that uses the dependency between an observation and a residual error from a moving average model is applied to lagged observations.
2. NSEpy - NSEpy is a library to extract historical and real-time data from NSE's website. This Library aims to keep the API very simple. Python is a great tool for data analysis along with the scipy stack and the main objective of NSEpy is to provide analysis-ready data series for use with scipy stack. NSEpy can seamlessly integrate with the Technical Analysis Library. This library would serve as a basic building block for automatic/semi-automatic algorithm trading systems or backtesting systems for Indian markets.
3. Stock Market - The stock market broadly refers to the collection of exchanges and other venues where the buying, selling, and issuance of shares of publicly held companies take place. Such financial activities are conducted through institutionalized formal exchanges or via over-the-counter marketplaces that operate under a defined set of regulations.
4. NSE - is the leading stock exchange of India, located in Mumbai, Maharashtra. It is under the ownership of some leading financial institutions, banks, and insurance companies. NSE was established in 1992 as the first dematerialized electronic exchange in the country. NSE was the first exchange in the country to provide a modern, fully automated screen-based electronic trading system that offered easy trading facilities to investors spread across the length and breadth of the country.
5. Seaborn - Seaborn is a Python data visualization library based on matplotlib. It provides a high-level interface for drawing attractive and informative statistical graphics.
6. Time series method - Time series analysis is a specific way of analyzing a sequence of data points collected over an interval of time. In time series analysis, analysts record data points at consistent intervals over a set period rather than just recording the data points intermittently or randomly. However, this type of analysis is not merely the act of collecting data over time. What sets time series data apart from other data is that the analysis can show how variables change over time.
7. Linear regression - In statistics, linear regression is a linear approach for modeling the relationship between a scalar response and one or more explanatory variables. The case of one explanatory variable is called simple linear regression; for more than one, the process is called multiple linear regression. This term is distinct from multivariate linear regression, where multiple correlated dependent variables are predicted, rather than a single scalar variable.

Questions:

1. What are the time-series methods that have been used to predict stock prices?
2. How accurate is the model?
3. Will it predict the stock prices other than NSE?
4. How to scale it to predict the stock price for any given time?
5. How to integrate it will other applications like the web, android, iOS, and other platforms?
6. How to include fundamental analysis in this prediction?
7. What is the drawback of using linear regression for this project?
8. What is the ARIMA model?
9. Is there any other library like “nsepy” to get the stock prices directly?
10. What kind of visualizations are required for this project and what's the plan to implement it?