**TOPIC – Stock market prediction**

**Business problem:**

When it comes to Data Science, there are a lot of words and phrases or jargon used that many do not know. We are here to solve all of that. Inherently, data science involves knowledge of statistics, math, and programming.

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.

**Datasets:**

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 linear regression because 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.

In linear regression, the relationships are modeled using linear predictor functions whose unknown model parameters are estimated from the data. Such models are called linear models. Most commonly, the conditional mean of the response given the values of the explanatory variables is assumed to be an affine function of those values; less commonly, the conditional median or some other quantile is used. Like all forms of regression analysis, linear regression focuses on the conditional probability distribution of the response given the values of the predictors, rather than on the joint probability distribution of all of these variables, which is the domain of multivariate analysis.

**Ethical consideration:**

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.

**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.

**Reference:**

1. [🤔Stock Market📈 Prediction with Linear Regression | Kaggle](https://www.kaggle.com/spidy20/stock-market-prediction-with-linear-regression/data)
2. [NIFTY-50 Stock Market Data (2000 - 2021) | Kaggle](https://www.kaggle.com/rohanrao/nifty50-stock-market-data?select=ASIANPAINT.csv)