# Capstone Project Proposal

# Stock Index Predictor

## Domain Background:

Investment firms, hedge funds and even individuals have been using financial models to better understand stock market behavior and make profitable investments and trades. A wealth of information is available in the form of historical stock prices and company performance data, suitable for machine learning algorithms to process.

In this project I propose to study various machine learning models for predicting future prices and direction of Stock Index Nifty50 of NSE, India. NSE is the leading stock exchange in India and the fourth largest in the world by equity trading volume in 2015, according to World Federation of Exchanges (WEF). It is the largest stock exchange in India in terms of total and average daily turnover for equity shares every year since 1995, based on annual reports of SEBI. (https://www.nseindia.com/global/content/about_us/about_us.htm)

This prediction can help Investors, Traders and others to take informed decision about their trading strategy and may benefit financially.

We will be using daily stock price data which is open, high, low and close price of Nifty50 along with various technical indicators like Bollinger bands, Moving averages and others as input for predicting future price and direction of Nifty50.

## Problem Statement/Task:

In this Project of Stock Index Predictor, I will build a prediction/forecasting framework that takes daily trading data of Nifty50 over a certain date range as input, and outputs projected estimates for given query dates. This would be ideal for someone looking for short term trading ideas for a horizon of few days to weeks and may not be suitable for years of or long term price forecasting. Note that the inputs will contain multiple metrics, such as opening price (Open), highest price the stock traded at (High), lowest price the stock traded at (low), how many stocks were traded (Volume) and closing price of the stock (Adjusted Close) along with different statistical measures like Bollinger bands, moving averages etc. We will only predict the closing price and direction of the trend. That is whether the market is expected to close up or down from today's closing price.

## Datasets and Inputs:

Datasets that will be used for our project is freely available to download from NSE India official website. We will be using daily trading data from 04$^{th}$ Jan 2010 to 29$^{th}$ Dec 2017 for training and testing of our model.

https://www.nseindia.com/products/content/equities/indices/historical_index_data.htm

The data in its raw format from downloaded csv file looks as follows:

| Date | Open | High | Low | Close | Shares Traded | Turnover (Rs. Cr) |
|---|---|---|---|---|---|---|
| 04-Jan-10 | 5200.9 | 5238.45 | 5167.1 | 5232.2 | 148652424 | 6531.61 |
| 05-Jan-10 | 5277.15 | 5288.35 | 5242.4 | 5277.9 | 240844424 | 7969.62 |

Using this raw data we will calculate various technical indicators like Moving Averages, Bollinger Bands and others for input to the model.

We will also use few fundamental indicators like P/E ratio, P/B value for Nifty50 downloaded from NSE India website.

https://www.nseindia.com/products/content/equities/indices/historical_pepb.htm

**Solution Statement:**

I propose to use various machine learning algorithms for forecasting Stock Index price and direction. I will be studying different Regression algorithms like Linear Regression, MLP and LSTM for predicting future prices. For predicting the direction of market, I will use various Classification algorithms like Logistic Regression, SVM, and Random Forest.
The idea is to study different algorithms to see if we can predict prices and direction of Stock Index with greater accuracy and suggest the best model out of it.

**Benchmark:**

For Regression we will consider Linear Regression as our benchmark model for comparing regression learning algorithms.
For classification task, we will consider Logistic Regression as our benchmark model.

**Evaluation Metrics:**

The key metrics that we will use for regression models will be R2 score which is Coefficient of determination and RMS error.
For classification we will use Accuracy Score.
Finally depending on the accuracy of these models, final model will be suggested for Stock Index forecasting.

**Setup:**

Software's and libraries that will be used for this project are:
Python, Numpy, Pandas, Scipy, Matplotlib
Scikit Learn, Keras, Tensorflow
Jupyter Notebook

## Analysis

**Data Exploration:**

The data will be in csv file. This raw data will consist of Open price, High price, Low price, Close price, Volume and Turnover
**Open Price**: This is the first price at which trade took place when the stock market opens
**High Price**: This is the highest price that a particular stock has traded for the day
**Low Price**: This is the lowest price that a particular stock has traded for the day
**Close Price**: This is the price at which the last trade for the day took place
**Volume:** This gives the total number of shares traded on a particular day
**Turnover**: This is the actual value of the shares in crores (INR) that got traded in a particular day.
1 crore is equal to 10 million
These features along with few added features will be used as input to our model for training.
These other technical features will be moving averages, Bollinger bands and others.

**Moving average**: this is a simple moving average (rolling mean) of close prices taken over a defined period called window. For our training purpose we will be experimenting with various window sizes from 3 day to 21 day
**Bollinger band**: This is the standard deviation of closing prices forming a band across the simple moving average with 2 standard deviations above and below. This tries to gauge the volatility of stock price.
We will use some fundamental indicators like P/E ratio (price to earnings ratio) and P/B value (Price to book value).
We will also be experimenting with various other technical and fundamental indicators that may be helpful in forecasting the future price of Nifty50.

We will be checking data for any inconsistencies, outliers, Nan or missing values etc. We will also be checking these input features for any skewness.
We will be using normalization, scaling, forward fill and backward fill techniques to deal with data abnormalities and inconsistencies.
We will calculate and report various statistics for the data like min, max, mean, standard deviation etc.

**Exploratory Visualization:**

We will be doing initial exploratory visualization on the data. This is to figure out any correlation between the features for deciding which ones to be added or removed from inputs to the model. Also we will be plotting the data to check linearity. We will also plot different features to check for skewness and if any found will use log transformation on that particular skewed feature.

**Algorithms and Techniques:**

We will be using various algorithms in our project as stated above.

First we will preprocess and visualize the data for initial exploration and remove any inconsistencies. We will use various techniques like scaling, log transformation, normalization, feature selection, dimensionality reduction to get the data ready. We will prepare this data into features and labels for inputs to the model as Numpy arrays.

Next we will split our data into 80 % as training and 20 % as testing set for model evaluation and validation and to avoid overfitting. With this training data we will train our different models and finally evaluate these models using testing set. We will use evaluation metrics for deciding on the best model.

**Deliverables:**

All necessary project code
Readme file
Data
Project report as .PDF file