**Stock Price Prediction**

**1. Introduction & Literature Review:**

Stock price prediction is one of the significant and most difficult tasks in the world of finance. Most traders rely on technical, fundamental & quantitative analysis for making predictions or for generating price signals. In the recent years with the growing trends in Machine learning, financial Industries also started using ML algorithms for stock price predictions and for generating trading signals.

In this project ‘AAPL’ stock price prediction is performed for predicting its next day’s closing price. By predicting the stock price, a trader can leverage this information by going long or short position. **The most fundamentally distinguishing feature in time series analysis is that the observations are dependent or correlated**. Data from Dec 1980 of AAPL stock is considered for building the model. Time series analysis requires different methods and commonly used statistical methods based on random samples may not be applicable. But over the time, machine learning methods are increasingly seen as an alternative to this traditional methods and have more accuracy but, they are computationally more expensive in comparison to statistical methods.

The main challenge of this project is using the large open available stock data and creating time series features in pyspark and using MLlib for building predictive models.

**2. Business Problem:**

The aim of our project is to construct and compare machine learning models for stock price prediction of Apple for different time horizons. By precisely predicting the stock prices we can anticipate the expected movement and take a position in the market. The continuing aim is to develop a trading strategy to see the profitability of the classification/Regression model.

**3. Data set:**

Source: Apple daily stock price data for over 35 years is pulled from Quandl website. The main reason for using Quandl as data source is to get the values of adjusted prices. The yahoo data only provides the Adjusted Close prices, where as Quandl provides Adjusted prices of open, high, low, close and volume. The difference in adjusted and actual close price is when a company announces a dividend, or stock splits.

**4. Target Variable:**

This is a supervised machine learning approach and the target variable or independent variable is adjusted closing price of Apple stock (please assume unless mentioned I am referring to adjusted prices). The features are selected such that the correlation, time dependent nature of the series will be taken in to account. The models considered are linear regression, Random Forests, Gradient Boosted trees and logistic regression (classification for trading). The reason for limiting only to this simple regression models is due to the easy implementation of them in the Spark ML Lib library.

**5. Feature Engineering:**

Since this is a time series problem, to capture the seasonality, trends and volatility in the data – different features like Moving averages, Average price in last few weeks, day of the week, day of the month etc., are created. For example Moving average feature captures the trend of the stock price over time. Since stock prices heavily depends on time (Eg: week start/Week end/ etc.,) features like day of week and day of month captures this seasonality.

**6. Model Development:**

For predicting next days adjusted stock price, different regression models available in the ML lib library (Linear regression, Random forest regressor, Gradient boost regressor, Decision tree regressor) are used. Initial results shows that Linear regression is performing better than other models. Also, features like daily (open – close) price, day of week & last few days moving average turned out to be the most predictive features for the model.

One problem with Regression model is that it gives the same low RMSE values for both positive and negative stock movement. Eg: Actual stock price could be $100, but predicting $99 or $101 gives the same RMSE values which results in incorrect buy/sell in trading strategy. Since regression might not be a better use for our problem, a classification model (Logistic regression) is built to predict whether next days closing price increases/ decreases

**7. Model Evaluation:**

RMSE (root-mean-square error): It is simply the square root the second sample moment of the differences between predicted and actual values or simply, quadratic mean of these differences.

**RMSE** value for linear regression model is 5.78 which looks very good, but as explained above this might not help for building a good trading strategy.

**Recall** for Logistic regression Model: 0.77 This shows that model was able to capture 77% of the times total rise in stock prices accurately.

**8. Model improvement:**

There are many ways of improving the model performance and next steps could be finding the right amount of training data and second, hyper parameter tuning. Instead of training over the full data, we can alter the amount of data used for training and evaluate the model performance. Since stock prices distribution changes over time, using too old data might not help for making current predictions.

**9. Conclusion:**

For predicting stock prices, simpler model like Linear regression performed better in comparison to Random forests, Gradient boost regressor. As expected, the machine learning has capability to explain the behaviour of time series and hence we can use ML for making trading strategies more powerful.