# Capstone Proposal

## Domain Background:

A stock market is the aggregation of buyers and sellers of stocks, when people talks stocks, they are usually talking about companies listed on major stock exchanges like the New York Stock Exchange (NYSE) , the Nasdaq, the Bombay Stock Exchange (BSE) and many others where stockbrokers and traders can buy and sell shares of stock which represents ownership/share claims on businesses.
When we buy a stock of a company and if that company does good then it's more likely that stock price goes up and we gain the same percentage of profit in the investment of that stock and vice versa .

Stock market prediction has been one of the challenges for researchers and financial investors, over the years they have been trying to better understand stock price behavior and make profitable investments and trades. However, predicting stock market movements is extremely challenging because according to Zhong and Enke (2017), stock markets are affected by many highly interrelated factors that include economic, political, psychological, and company-specific variables.

Stock market price prediction is a tricky thing. Several theories regarding stock markets have been conceptualized over the years. One is Efficient Market Hypothesis (EMH) and another one is Random Walk Theory.

**Efficient Market Hypothesis (EMH):** It states that at any point of time, the market price of a stock incorporates all information about that stock. Hence it is not possible to outperform the stock market. Efficient market Hypothesis exists in three forms:

- Weak EMH: Only the past data is considered
- Semi-Strong EMH: All public information is utilized
- Strong EMH: Publicly and privately available information is used

**Random walk theory:** Random walk theory assumes that it is impossible to predict stock prices as stock prices don't depend on past stock. It also considers that stock price has great fluctuations, so it is infeasible to predict future stock prices.

Many widely accepted empirical studies show that financial markets are to some extent predictable (Chong et al. 2017). Criticism of EMH has given rise to an increasing number of studies that question the validity of EMH and introduce new and successful approaches that combine technical analysis indicators and chart patterns with methodologies from econometrics, statistics, data mining, and artificial intelligence (Arévalo et al. 2017).

I have been doing stock trading for last couple of years and analyzing stock market behavior fascinates me a lot. Being an active stock trader motivates me to utilize all my technical and functional skills to build the stock market predictor, so that I can utilize this predictor for my stock trading. I believe it will be a perfect project to evaluate my final model with real time stock data which can help me to optimize this predictor over period.

## Problem statement:

The Market researchers and financial investors have been facing the problem of not being able to predict stock market price movement, hence they are struggling to make profitable investments and trades.

In this project, I will be building a stock price predictor that takes daily trading data for a list of ticker symbols (e.g. GOOG, AAPL) over a last 5 years span as input, and outputs the predicted stock prices for each of those stocks on the given dates.

The predictor should be able to predict stock price value until 28 days in future (1 day, 7 days, 28 days), and the predicted stock value for 7 days or 28 days from the date of prediction should be within +/- 5% of actual value, on average.

## Datasets and Inputs:

For this project, I will use historical stock data of top 10 stocks holdings of ARK Next Generation Internet ETF (ARKW) ETF from Yahoo! Finance for last 5 years (**Jul 12, 2015 - Jul 12, 2020**).
I have chosen this very popular ETF because I have invested on this ETF and a lot of top Market analysts have been suggesting buying this, so I am curious to know more hidden insights about this ETF.

The ARK Next Generation Internet ETF is an actively-managed fund, therefore it does not track a particular index. ARKW aims to identify companies that will profit from developments in cloud computing, artificial intelligence (AI), financial technology, and similar innovations. Its largest holdings are Tesla Inc. (TSLA), the electric car company; Square Inc. (SQ), the mobile payments company; and Roku Inc. (ROKU), the digital streaming equipment maker.

**The list of top 10 stocks of ARKW ETF that will use in this project.**

1- Tesla Inc : (**TSLA** )
2- Square Inc A : (**SQ**)
3- Roku Inc Class A : (**ROKU**)
4- 2U Inc : (**TWOU**)
5- Zillow Group Inc C : (**Z**)
6- Splunk Inc : (**SPLK**)
7- LendingTree Inc : (**TREE**)
8- Pinterest Inc : (**PINS**)
9- Xilinx Inc : (**XLNX**)
10- Facebook Inc A : (**FB**)

The historical data for last 5 years (**Jul 12, 2015 - Jul 12, 2020**) fetch from Yahoo finance has seven columns such as Date, Open (opening price ), High (highest price the stock traded at), Low (lowest price the stock traded at), Close, Adjusted Close (closing price adjusted for stock splits and dividends) and Volume (how many stocks were traded).

I will split data into 80:20 ratio for training/cross validation and testing, four years data (**Jul 12, 2015 - Jul 12, 2019)** will use as training data and last 1 year (**Jul 12, 2019 - Jul 12, 2020)** data will use for testing and performance evaluation of my model.

## Solution Statement:

As this is a regression problem where we have to predict stock adjusted price, so I will be using these three regression algorithms such as XGBoost (eXtreme Gradient Boosting), Amazon SageMaker linear learner algorithm, Amazon SageMaker DeepAR forecasting algorithm, and will make three separate models for them and will choose the one which will have best performance among these three algorithms.
As per existing study Amazon SageMaker DeepAR forecasting algorithm is very well suited for financial time series forecasting, so I will consider this algorithm as my first preference for this stock prediction, however I will use it after comparing it's performance against other two algorithms.

First step will be to get data from Yahoo Finance and perform the require preprocessing into the data, then feed this input data to the model, and the model will return the forecast stock Adjusted Close value in the chosen day range(7 days,14 days,21 day, 28).

## Benchmark Model:

The benchmark model will be Amazon SageMaker linear learner algorithm as this model is well known for regression problems, I will use the same data and features to my benchmark model and main DeepAR model, and will compare DeepAR model against this benchmark model based on the evaluation metrics to decide the best performing model.

## Evaluation Metrics:

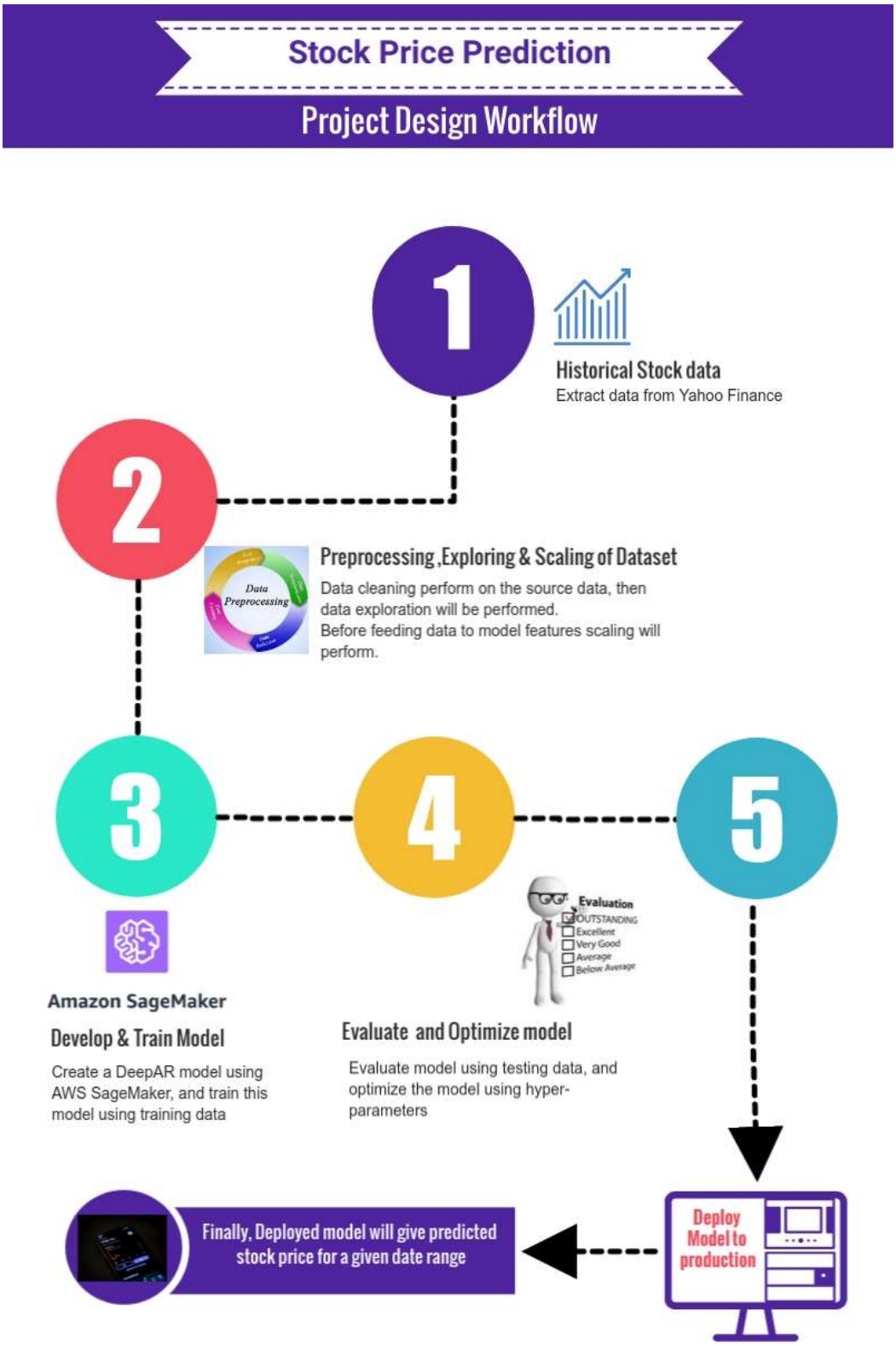
I will use Root Mean Square Error (**RMSE**) as the evaluation metric which is very popular evaluation metrics for regression problems, it is a frequently used measure of the differences between values (sample or population values) predicted by a model or an estimator and the values observed.

**RMSE formula:**

$$\sqrt{\frac{\sum_{i=1}^{n}(Predicted - Actual)^2}{N}}$$

I will compare RMSE value for Amazon SageMaker DeepAR forecasting algorithm against my benchmark model Amazon SageMaker linear learner algorithm, also with XGBoost (eXtreme Gradient Boosting), In this comparison whichever model will have lower RMSE value that will have the best accuracy among them, hence I will choose the model with lower RMSE as my best model.

**Project Design:**



**Reference:**


- https://www.researchgate.net/publication/328832048_Quantitative_Analysis_of_Stock_Market_Prediction_for_Accurate_Investment_Decisions_in_Future
- Zhong, Xiao, and David Enke. 2017. Forecasting daily stock market return using dimensionality reduction. Expert Systems with Applications 67: 126–39. [CrossRef]
- https://en.wikipedia.org/wiki/Stock_market
- https://www.investopedia.com/articles/07/stock-exchange-history.asp
- Chong, Eunsuk, Chulwoo Han, and Frank C. Park. 2017. Deep learning networks for stock market analysis and prediction: Methodology, data representations, and case studies. Expert Systems with Applications 83: 187–205. [CrossRef]
- Arévalo, Rubén, Jorge García, Francisco Guijarro, and Alfred Peris. 2017. A dynamic trading rule based on filtered flag pattern recognition for stock market price forecasting. Expert Systems with Applications 81: 177–92. [CrossRef]