Proposal for Capstone Project - Global X Copper Miners ETF Price Prediction

Domain Background

Since 2021, inflation in the United States of America has increased dramatically. Indeed, from 1.2% in 2020, the US inflation rate rose to 6.8% in 2021, its highest since 1982. In these uncertain times, how does one who has invested into the stock market protect their investments from the negative impacts of inflation? One does by investing in assets that performs well during inflationary periods.

Historically, copper is one of the best performing assets during inflationary periods. As a result, predicting changes in the value of copper has obvious financial benefit including knowing when to buy or sell.

Problem Statement

The problem, that I will be attempting to solve with this project, is to predict the next day's price of copper based on historical data. By comparing a price predicted by my model with an actual market price, I will be able to determine the copper prediction accuracy of my model. Based on thus accuracy, I should be able to make more informed decisions regarding my investments in copper.

Solution Statement

In order to achieve my goal, I will start by downloading historical stock prices for a specific copper ETF. The data will be obtained from Yahoo Finance! After analyzing and adding new statistical information to my data, it will be divided into three datasets: a training dataset; a test dataset; a validation dataset.

At this point, I will used the training dataset to train three machine learning models (a Linear Regression model; a Random Forest model; a Long Short-Term Memory model) in predicting copper prices. Finally, I will compare the prediction accuracy of my model in order to choose the best performing one.

Datasets and Inputs

The historical data used for this project will be the one for the 'Global X Copper Miners' ETF, an ETF that invests in companies involved in the mining of copper. The data will be downloaded from Yahoo Finance!

In its raw form, the dataset is a collection of daily trading records. Each record is identified by a trading date ('Date'), the highest price reached during the day ('High'), the lowest price reached during the day ('Low'), the opening price for the day ('Open'), the closing price for the day ('Close'), the number of shares traded during the day ('Volume') and the closing price adjusted for stock splits and dividends ('Adj Close').

Benchmark Model

The benchmark that I will use for this project is based on the naive method of forecasting. This method dictates that one use a previous period to forecast for the next period. In my case, I will be using the previous adjusted closing price of the stock to predict its next day closing price.

Evaluation Metrics

In order to evaluate the accuracy of my models, I will be using to two evaluation metrics. The first metric that I will use will be the Root Mean Square Error (RMSE). The Root Mean Square Error is one of the most commonly used measures for evaluating the quality of predictions. It shows how far predictions fall from measured true values using Euclidean distance.

The second metric that I will use is the Coefficient of Determination (R^2). The Coefficient of Determination tells one how well the data fits the model. It can take any values from 0 to 1, the closer to 1 the better the model fit.

Project Design

For this project, I will be using AWS Sagemaker. Most of the project will be done in a Jupyter Notebook. Within this Notebook, I will be downloading the historical data into my instance's environment. I will then proceed by analyzing the data and enhancing it via featuring engineering. Once satisfied with the state of the data, I will divide it into three distinct datasets. I will use the first dataset to train three distinct models (a Linear Regression model; a Random Forest model; a Long Short-Term Memory model). My second dataset will be used to test my models. Based on my test results, I will deploy my best performing model to an AWS endpoint. Finally, I will implement a lambda function that will allow me to use my deployed model in real life.