Udacity AWS Machine Learning Engineer Nanodegree Capstone project

Domain Background

Student briefly details background information of the domain from which the project is proposed. Historical information relevant to the project should be included. It should be clear how or why a problem in the domain can or should be solved. Related academic research should be appropriately cited. A discussion of the student's personal motivation for investigating a particular problem in the domain is encouraged but not required.

In sensitive markets, like Taiwan, regulators post a list of stocks daily that are deemed to have high execution risk that could lead to financial contagion due to their volatility (fast/significant price fluctuations) or because a particular account is trading a large majority of the daily volume of the security. So, Regulators require buys/sells of these securities to be pre-funded/delivered. The purpose of this project would be to use historical/current data as inputs to train a model to predict which securities will be flagged the next day at the close of trading on the current day in order to improve fund operations by being able to move money/securities around 12 hours in advance if portfolio managers plan on making a trade in one of these securities on market open the next day before prices move out of favor.

In Taiwan the regulators are not completely specific as to the rules used, hence the need for an ML model that's accessible through an API using attributes such as price fluctuation, PE ratio, price to book ratio, trading volume, and turnover as factors to retrieve their predictions.

Problem Statement

Student clearly describes the problem that is to be solved. The problem is well defined and has at least one relevant potential solution. Additionally, the problem is quantifiable, measurable, and replicable.

- The TWSE issues a daily "watch list" that identifies stocks where the price fluctuation, PE ratio, price to book ratio, trading volume, or turnover rate is too high. The problem is to predict on a daily basis, at the close of trading (T), which stocks will be placed on the watch list for the next day (T+1).
- The risks of a false positive are that if the front office has an open order for the stock the following day, that a same-day (T+0) FX is executed for the trade when it doesn't need to be. This may subsequently need to be reversed because the portfolios do not want to accept extra foreign-exchange rate risk. The larger the FX, the greater the risk, and the larger the TWD/USD exchange rate pair volatility, the greater the risk.
- The risks of a false negative are that middle office must initiate a T+0 FX with the custodian but this will often not be settled until mid afternoon, then cash needs subsequently moved to the broker cash account. Often this will mean front office has effectively lost most of the equity trading day and if the order was marked to trade on the open we will have missed the benchmark by almost a full day. The higher the market value of the order, the larger the risk. The higher the volatility of the underlying stock, the larger

the risk. The most concerning scenario is from same-day orders generated by portfolio managers (most of the orders are coming from US, for next day).

- Qualitatively, the middle office has stated that the risks of a false negative are expected to be greater than the risks of a false positive, because in general the market opportunity cost is greater than the FX volatility. This is because in general, equity market volatility is greater than FX market volatility, and because typically the reason that a stock is placed on the watch list is due to market events which present a critical window of time in which to trade the stock.

Solution Statement

Student clearly describes a solution to the problem. The solution is applicable to the project domain and appropriate for the dataset(s) or input(s) given. Additionally, the solution is quantifiable, measurable, and replicable.

- The solution should be a binary classification model. The model must be able to perform batch predictions on a list of stocks within the investible universe of stocks listed in Taiwan.

Datasets and Inputs

The dataset(s) and/or input(s) to be used in the project are thoroughly described. Information such as how the dataset or input is (was) obtained, and the characteristics of the dataset or input, should be included. It should be clear how the dataset(s) or input(s) will be used in the project and whether their use is appropriate given the context of the problem.

- The dataset is split between two exchanges, the Taiwan Stock exchange and the Taipei stock exchange. To obtain the full universe we must combine both exchanges.
- Data can be obtained on Taipei Stock Exchange website and the Taiwan Stock Exchange website
- Due to time constraints, we will train on the Taiwan Stock Exchange (TWSE) listed stocks.
- Data for these stocks have been downloaded from the website at the following links:
    - "ratios" : 'https://www.twse.com.tw/en/page/trading/exchange/BWIBBU_d.html',
    - "short_sales" : 'https://www.twse.com.tw/en/page/trading/exchange/TWTASU.html',
    - "summary_px_vol" : 'https://www.twse.com.tw/en/page/trading/exchange/MI_INDEX.html#subtitle7',
    - "full_delivery": 'https://www.twse.com.tw/en/page/trading/exchange/TWT85U.html'
- Zipped files for each of these datasets from 1/1/2022 through 10/31/2022 are included in this repository as follows:
    - udacity-aws-machine-learning-nanodegree-capstone/taiwan-pe-pb-ratios-website-raw-01012022-10312022.zip
    - udacity-aws-machine-learning-nanodegree-capstone/taiwan-shortsale-data-website-raw-01012022-10312022.zip
    - udacity-aws-machine-learning-nanodegree-capstone/taiwan-summary-website-raw-01012022-10312022.zip

    - udacity-aws-machine-learning-nanodegree-capstone/taiwan-watchlist-data-website-raw-01012022-10312022.zip

- Additional information surrounding the TWSE's irregularity rules are listed from the website here. These "rules" guided the data we collected and will inform the feature engineering. Most of the restrictions listed refer to Article 4 detailed in the attached, this simply refers to irregularity without further expanding on what that means. https://twse-regulation.twse.com.tw/m/en/LawContent.aspx?FID=FL007225

- Article 4: At the close of trading each day, the TWSE will analyze the trading of exchange-listed securities (excluding foreign bonds, government bonds, and straight corporate bonds). Upon discovery of any of the following circumstances, the TWSE will announce related trading information (such as the degree of upward or downward movement in prices, trading volume, turnover rate, degree of concentration, price-to-earnings ratio, price to book ratio, long/short ratio, premium/discount percentage, sales quantity of borrowed securities, and day trading percentage).

#	Irregularity Rules	Attributes	Calculation
1	An irregularity in the cumulative percentage of increase or decrease in the closing price during the most recent period.	closing price	cumulative % chg
2	An irregularity in the percentage of increase or decrease in the closing price between the initial and final business days of the most recent period.	closing price	% chg. Need clarity on defn of 'most recent period'. Does this mean 'prior business day to today'?
3	An irregularity in the cumulative percentage of increase or decrease in the closing price during the most recent period, combined with an unusually large increase in the intraday volume of trade relative to the daily average in the most recent period.	closing price, intraday trade volume, average daily trade volume	cumulative % chg ratio of intraday trade volume to average daily trade volume
4	An irregularity in the cumulative percentage of increase or decrease in the closing price during the most recent period, combined with an unusually high intraday turnover rate.	closing price intraday turnover	cumulative % chg

#	Irregularity Rules	Attributes	Calculation
5	An irregularity in the cumulative percentage of increase or decrease in the closing price during the most recent period, combined with intraday consigned trading of the given security at a securities firm in which confirmed purchases or sales account for an unusually high percentage of the intraday volume of trade in the given security.	closing price, firm-level consigned trading volume, intraday trade volume.	cumulative % chg. What does 'consigned trading' mean ratio of firm-level consigned trade volume to intraday trade volume
6	An irregular price-to-earnings rate or price-to-book ratio, or an usually high intraday turnover, combined with any of the following three circumstances: a relatively high price-to-book ratio for stocks of the given industry; an intraday value for confirmed purchases or sales of the given security at any single securities firm that accounts for an unusually high proportion of the total intraday value of confirmed trades in the given security; or an intraday value for confirmed purchases or sales of the given security by any single investor that accounts for an unusually high proportion of the total intraday value of confirmed trades in the given security.	price-to-earnings rate, price-to-book ratio, intraday turnover, industry price-to-book ratio, (total, total per investor, total per firm) intraday value of confirmed trades in given security,	ratio of. POSSIBLY NOT AVAILABLE
7	An irregularity in the cumulative percentage of increase or decrease in the closing price and a significant increase in the long/short ratio during the most recent period.	closing price long/short ratio	cumulative % chg, chg in
8	An irregularity in the percentage of premium or discount calculated from the closing price of Taiwan Depository Receipts and the closing price, on the exchange market of their home country, of the shares they represent.	premium or discount of TDRs, closing price	ratio of
9	A significant increase in the daily volume of trading for a given day or several recent days relative to the daily average volume of trade for the most recent period.	daily volume of trade, daily avg trade volume	ratio of. What's the difference between 'daily volume of trade' and 'intraday volume'?

#	Irregularity Rules	Attributes	Calculation
10	A significantly high cumulative turnover rate for the most recent period.	turnover rate	what is 'cumulative turnover'?
11	An irregularity in the difference between the closing prices on the initial and final business days of the most recent period.	closing price	change in closing price
12	The trading volume of the sales of borrowed securities accounting for a significantly high percentage of the total volume of trade of the most recent period.	trading volume of the sales of borrowed securities, total volume of trade	ratio of
13	The day trading volume accounting for a significantly high percentage of the total volume of trade of the most recent period.	intraday trade volume, total trade volume	does 'day trading volume' mean 'intraday' trading volume? → 'day trading' volume. does 'total volume of trade' mean for a specific stock, or for all stocks on the exchange?
14	Other trading irregularities as determined by resolution of the Surveillance Operations Oversight Committee.	?	?

- If there is no closing price for a security on a given day to serve as the basis for calculation of any irregularity listed under the preceding paragraph, the price determined pursuant to Article 58-3 of the TWSE Operating Rules shall be used instead.
- When calculating the price fluctuation limit of a security, if the formula includes factors such as the given underlying security or underlying index, the provisions of Article 2, paragraph 2 shall apply mutatis mutandis to the calculation of the cumulative percentage of increase or decrease in the price within a certain period.
- When the trading volume of a security is below 1,000 units, the provisions of Article 2, paragraph 4 shall apply mutatis mutandis to the numerical standards for the unit of the trading (or order) volume.
- Numerical standards for the irregularities listed under each of the subparagraphs of paragraph 1, and any exceptions to those conditions, shall be separately adopted by the TWSE.

Benchmark Model

A benchmark model is provided that relates to the domain, problem statement, and intended solution. Ideally, the student's benchmark model provides context for existing methods or known information in the domain and problem given, which can then be objectively compared to the student's solution. The benchmark model is clearly defined and measurable.

- No existing model exists by which to benchmark.
- The ML engineer intends to begin with a linear model, and one or more decision tree-based models (random forest, XGBoost), as well as more as they have time.

Evaluation Metrics

Student proposes at least one evaluation metric that can be used to quantify the performance of both the benchmark model and the solution model presented. The evaluation metric(s) proposed are appropriate given the context of the data, the problem statement, and the intended solution.

- The loss function should be based on log-loss.
- A baseline desired recall per the front office guidance would be a recall of 90-99% desired. This is not a hard and fast requirement, and more discussion is needed with the client on this topic.
- Because of the need to minimize the risk of false negatives, the model should try to maximize recall, or the 'sensitivity'. This is calculated as (TP / (TP + FN)), where TP stands for 'True Positives' and FN stands for 'False Negatives'. Refer to this guide for more on binary classification metrics
- Additionally we will examine the F1 score, so as to also consider the precision, or the accuracy associated with positive predictions.
- Other metrics can be referred to as desired (such as ROC-AUC).

Project Design

Student summarizes a theoretical workflow for approaching a solution given the problem. A discussion is made as to what strategies may be employed, what analysis of the data might be required, or which algorithms will be considered. The workflow and discussion provided align with the qualities of the project. Small visualizations, pseudocode, or diagrams are encouraged but not required.

- Overall strategy will be to perform EDA on the data first, within the AWS SageMaker notebook environment.
- Secondly, scikit-learn and/or other ML framework libraries, including SageMaker container images, will be utilized to train various models described previously.
- Lastly, various evaluation metrics will be observed and further exploratory opportunities will be and documented.