| Capstone Project Proposal |  |
| --- | --- |

*Mohammed Tayor*

**Business Goals**

| **Project Overview and Goal**  What is the industry problem you are trying to solve? Why use ML/AI in solving this task? Be as specific as you can when describing how ML/AI can provide value. For example, if you’re labeling images, how will this help the business? | Financial markets have consistently attracted speculators, investors, and institutions across the globe. Especially with the covid 19 crisis and many people stuck at home, a significant wave has entered the markets, including Millenials. However, as professional people in finance like to call them, these retail investors, with the hope of getting rich quick or simply investing into their retirement accounts, often are hit with a cold truth, predicting the stock market is no easy task. The so-called smart money or professional and institutional investors hire the brightest PhDs in maths and finance, have the most advanced infrastructure servers with nanoseconds latency order execution, and employ state-of-the-art artificial intelligence trading systems. It is next to impossible to beat smart money on this zero-sum game by Average Joe. |
| --- | --- |
| **Business Case**  Why is this an important problem to solve? Make a case for building this product in terms of its impact on recurring revenue, market share, customer happiness and/or other drivers of business success. | The goal is to give an Alpha to normal users and investors using AI that will be a real edge towards profitability.  For both traders and investors, a tool based on AI that can help them in their decisions of buying and selling against the rest of the market to achieve profitable predictions of the future price action of stocks can be of vital help. |
| **Application of ML/AI**  What precise task will you use ML/AI to accomplish? What business outcome or objective will you achieve? | Stock data is primarily time series of price and volume. It is the backbone of what appears on a lot of charts on brokers like CME or Robinhood. So, for executing profitable trades, it is required to be right on the price direction, whether it is going up or down in the future. This can be translated in ML as time series forecasting.  Furthermore, the collective mood of the crowd represents a significant predictor of future price action. This can be achieved through sentiment analysis, a niche in NLP (natural language processing).  The objective would be to combine both time series forecasting and sentiment analysis in a model to suggest profitable setups to normal people in the stock market. |

**Success Metrics**

| **Success Metrics**  What business metrics will you apply to determine the success of your product? Good metrics are clearly defined and easily measurable. Specify how you will establish a baseline value to provide a point of comparison. | The model accuracy is the first success metric. The more future setups the model can achieve consistently, the better the product's success. An 80% live accuracy hit would be phenomenal as with sound risk management (setting stop losses and profit target and position sizing), professional traders can achieve profitability with even 51%. As we will consider a 50% hit rate as a random chance outcome with the same chance of coin flip. |
| --- | --- |

**Data**

| **Data Acquisition**  Where will you source your data from? What is the cost to acquire these data? Are there any personally identifying information (PII) or data sensitivity issues you will need to overcome? Will data become available on an ongoing basis, or will you acquire a large batch of data that will need to be refreshed? | We need historical data of both stock prices and volumes across supported exchanges and brokers. We will need access to their APIs for live trading in our terminal. Most large exchanges offer such services are different rates. Some are free and require just an account creation in TD Ameritrade or Robinhood, for example. And for public opinions, we will use Twitter API, which has a free version, to analyze historical tweets of profiles related to finance to train our model and implement it live to filter through the noise of millions of tweets per day and to find correlations with the movement of prices. |
| --- | --- |
| **Data Source**  Consider the size and source of your data; what biases are built into the data and how might the data be improved? | Price and volume data time series are never complete or clean. Exchanges can go offline, historical orders can be missed to be included in the data, and so on. This means we will have to deal with missing data and outliers etc.  For tweets, it is unstructured data, people say something meaning the complete opposite either because of sarcasm, making fun of something, or even bad intentions. So the data will need a lot of filtering in this regard and not fall into any one-side bias. Also, structuring tweets into one format will make the tasks much easier. |
| **Choice of Data Labels**  What labels did you decide to add to your data? And why did you decide on these labels versus any other option? | For price and volume time series, on a low timeframe, for example, the 15-minute chart, the label for each 1-minute data point (open, close, high, low, volume) would be the percent the stock will move in the next 1 hour either positive to the upside or negative to the downside.  For sentiment analysis, we will rate a tweet that is related to our stock on a scale from 1 to 5, where 1 is very negative to 5 very positive. So we can quantify the actual response from that tweet. A different approach would be to only use two labels of positive and negative, but that would ignore more data that can add more predicting power to our price action prediction.  The initial filtering of tweets would be by hashtags and keywords, but another ML model would be developed to select the most profiles correlating historically with predicting the price action. So if we have a universe of 100 stocks, for example, each profile would have 100 binary labels corresponding to each stock, 0 if not to use their tweets for that particular stock, and 1 if their tweets have a good track record (or bad, to use the inverse predictions, most importantly there is a correlation whether positive or negative). Hence another label of the multiclass variable between 0 and 100 (as for the example of 100 stocks) would be added to each tweet. The 100th class would be assigned if a tweet is to be ignored and wouldn't appear to the user nor be used in the models. |

**Model**

| **Model Building**  How will you resource building the model that you need? Will you outsource model training and/or hosting to an external platform, or will you build the model using an in-house team, and why? | We will build the model internally using an in-house team. Once historical data is acquired and API for live data from exchanges and Twitter feed, we can start training our models on cloud providers like Google Cloud Platform or Amazon Web Services.  The intellectual property of the models is the core added business value, and the Quant finance world is a very high competitive world where all firms hide their intellectual property, and it is rare to find advanced models or state-of-the-art in the hands of the public or a competitor since then they can reproduce your edge to eventually where your models won’t work anymore. |
| --- | --- |
| **Evaluating Results**  Which model performance metrics are appropriate to measure the success of your model? What level of performance is required? | We will target an 80% success rate (accuracy )of predictions, including risk management (where the model suggests for each trade setup, the entry price, the exit, and the stop loss). |

**Minimum Viable Product (MVP)**

| **Design**  What does your minimum viable product look like? Include sketches of your product. | Please find bigger sketches in the pdf sketches.pdf |
| --- | --- |
| **Use Cases**  What persona are you designing for? Can you describe the major epic-level use cases your product addresses? How will users access this product? | The primary target will be enthusiast traders and investors without much knowledge of technical analysis, quantitative trading, or even fundamental trading.  An introductory course will be developed in the documentation for the initialization of these customers.  The product will initially be a responsive website that works on all devices. Afterward, we will develop our own mobile app to add push notifications directly to users’ phones and be more customized to phones. |
| **Roll-out**  How will this be adopted? What does the go-to-market plan look like? | We will do surveys first for idea validation, most wanted features to prioritize development, most traded instruments, etc.  In development, we will follow an agile development philosophy where every 15 days, the team should ship new features.  The product will start in a closed alpha release with a few dozen test users to gather feedback from them and fix any bugs or add any enhancements.  Then we will have a beta release where it is open to the public, and we will scale our server's API according to demand.  After the final launch, marketing will be scaled to acquire more users from Ad platforms like Twitter but also from sponsored ads in financial Youtube influencer channels. |

**Post-MVP-Deployment**

| **Designing for Longevity**  How might you improve your product in the long-term? How might real-world data be different from the training data? How will your product learn from new data? How might you employ A/B testing to improve your product? | In our case, we don’t have much difference between training data and live data as they are homogeneous and won’t represent any abrupt discontinuity between the two.  All models will be monitored and updated regularly as the new dataset is acquired from APIs and users' feedback.  We can use A/B testing to get get more feedback from users about some nuances between two versions of models or the website like when we are in the process of developing a new feature, changing esthetic style, etc. |
| --- | --- |
| **Monitor Bias**  How do you plan to monitor or mitigate unwanted bias in your model? | First, we will use cross-validation techniques to mitigate the risk of overfitting on the past data where the model would be absolutely accurate in the past but have very little predicting power for the future. Then we will keep monitoring the live predictions and the trades taken by users as another validation data for the model to measure its accuracy on live data if at some point it starts falling, then we can assume a regime change in the market and a need for new training, adding more data or tweaking of some parameters. |