Utilizing Machine Learning To Increase Returns For Retail Investors

A Project Proposal

12/9/2020

Pink Team: Kevin Fu, Minjie Xu, Ricardo Jenez

Overview

- **Intended Audience:** Robinhood executives
- **Mission/Value Proposition for Robinhood customers:** Our team's mission is to further democratize the access to financial markets using machine learning (ML) to reduce the cost and time spent by individual investors on deep financial analysis. Our tool analyzes current and historical financial data, such as stock prices, financial statements filed on the SEC and social sentiment analysis, to predict the future performance of stocks. We provide timely investment recommendations to Robinhood investors.
- **Value Proposition for Robinhood executives:**
  - Our “portfolio-recommendations-as-a-service” benefits Robinhood in 3 ways without assuming much incremental liability.
    - i. Our ML-powered portfolio attracts new investors to join the Robinhood platform.
    - ii. It serves as a mechanism for retaining existing Robinhood investors on their platform.
    - iii. It introduces an additional revenue stream for Robinhood. We can charge an annual fee based on the percentage of total returns generated for the client.
- **Goal:** Our goal is to design a research study that investigates if our machine learning algorithm can outperform the S&P 500 by 5% per year. We will follow our ML algorithm on a daily basis in our studies and monitor the risks that it takes on, as well as if the total portfolio balance is trending in the right direction. Our research consists of performing a multitude of analyses (fundamental, technical, sentiment) on liquid US stocks (100K+ in volume per day) via backtesting strategies.
- The rest of the document outlines the proposed design.

Research Question

Research Question: Can a ML-backed stock portfolio generate outperformance versus the S&P 500 index over the long-term?

Hypothesis: Based on a hybrid of fundamental, technical, and sentiment analysis, the ML portfolio will outperform the S&P 500 index (9.8% average annualized return) by more than 5% annually.

Sub-question: Can the ML-backed portfolio provide effective advisory to Robinhood users in terms of affecting their trading performances and returns?

Hypothesis: On average, Robinhood users who used our service will outperform their past performances and a control group.

Data

We expect to use already available sources of financial, stock and sentiment data:

- Data from the Robinhood trading platform.
  - Real-time market index data and stock ticker data[trades]
  - Financial and market news
  - Securities and Exchange Commission (SEC) reports (annual and quarterly financial statements, earnings reports, etc.).
  - Customer stock portfolios, current and past. Used to evaluate our tool's performance and help train our model in the later stages of this project.
- Current and historical data (stock trades, fundamental analysis) from the Center on Research of Stock Prices (CRSP/Compustat Merged Database (CCM)) to build our data model and create training and test data sets. This data source accounts for survivorship bias on returns.
- For real-time and historical stock and market sentiment analysis from Twitter and services like Stock Twits (https://stocktwits.com). Both offer excellent APIs for developers to gather specific information to train and test our system. These go back only 10-12 years, but will become more important as we move forward.

We intend to use a subset of this current and historical (time-series) data to train our ML models and a subset to test the models (we may use several schemes to select training and test sets). We think we have all the primary sources of the data we need. Risks around the use of this data is covered below.

In addition, we will collect trading data from the Robinhood users who will sign up for our pilot ML-back service with consents.

Study Design

This study consists of two phases to test the research questions and hypotheses. First, we will establish a ML-based portfolio. Second, we will launch a beta testing in Robinhood and sign up Robinhood users to cross validate the system.

1. Test our ML strategy with on our own small nest egg: Achieve 5% outperformance per annum vs. the S&P 500
   - Baseline portfolio: Initially, we will use fundamental analysis to create a baseline portfolio that contains 20 stocks in equal weight. We will then train a neural network and backtest against a subset of the data.

- Then we will invoke technical analysis to further improve the algorithm. During the step, our algorithm will auto adjust the shares of the portfolio to avoid major sell-off/market crash while buying dip/rebound.
- We will be working off of historical time-series data sets to train the models. We will gather the past 20 years of stock market data to backtest (retrospective validation) the hypotheses. This approach will allow us to validate that we can provide excess returns of 5% above the historical S&P 500 based upon historical market trends and price activity.
- Last, we will incorporate real-time market data and sentiment analysis to further tweak the adjusted portfolio to see if it can beat the SPY by at least 5% annualized return over six-month, while predicting (prospective test) future market trend and portfolio performance.

2. Beta test with Robinhood users: Treatment vs. Control
   - To test the effectiveness and validity of the ML-backed portfolio with Robinhood users, our phase two will propose a quasi-experiment study by inviting Robinhood users to beta test our ML-backed service from a pool of 20,000 Robinhood users. We will seek the help from Robinhood management team to randomly select 10,000 free users and 10,000 gold users (who paid $5 monthly subscription fee), and send out the test invitations. The one month signup will be the first come first server and first 5000 free users and 5000 gold users will get free access to our ML-backed portfolio and advisory service (the intervention) for 6-month, and become our treatment group. The rest of 10,000 users will be the control group.
   - After 6 months, we will send a short follow-up survey to the treatment group to ask 1) their trading behavior and/or performance changes if any, 2) their expected and actual returns with/without the service, 3) their willingness to use and/or pay for the ML-backed service in future, 4) their willingness to recommend it to their friends or promote it to the public, and 4) other feedback (e.g. risk control, user experience, suggestion for improvement).
   - Furthermore, all 20,000 users' trading behavior and performance will be analyzed and compared in the following groups: 1) before vs. after, 2) treatment vs. control, and 3) gold vs. free.
   - In addition, each customer's portfolio will be analyzed by our machine learning systems and we will try to predict as our customers execute stock trades in their portfolio whether or not they are getting excess returns commensurate with the risk they may be adding. This will train our systems to help our customers build model portfolios adjusted for risk and return on investment in future versions.

Sample

For model build and testing phase:

- Out of the universe of 6100 or so stocks listed on the NYSE and Nasdaq, we are selecting the 1000 with ample liquidity (> $20-80M of trade volume) to put in our training data set for our ML model.

    - The ML will then look at different stock portfolio sizes (5, 10, 20 stocks - small, medium, large) at any given time that it and a trading pattern of (< 5, < 20, <100 trades per year) believes will generate the most alpha for the next 12 months. With these different size considerations we can better match our customers propensity to different size portfolios and different trading patterns.
    - Roughly 60% of robinhood customers hold mid-cap ($2B+) or larger companies. Therefore, in a further effort to manage risk, we can filter out small cap companies, leaving the larger, more established companies within our ML algorithm's universe.
    - This portfolio must be easily purchasable by Robinhood's retail customers.

For cross-validation phase:

- After we have successfully tested our ML algorithm's strategy with our small, internal nest egg, we will invite 20,000 Robinhood users (10,000 free users and 10,000 gold users) to our ML-backed beta testing, and the first 10,000 users (5000 free users and 5000 gold users) who signed up will get the 6-month free access to the system.

Variables and/or Intervention

For this longitudinal case-control study, the ultimate outcome variable is annual return of our ML-backed portfolio. The initial predictor variables will be generated from:

- Fundamental analysis variables: Return on Assets (over different periods > one year), Long Term Free Cash Flow(over different periods> 1 year), Gross Margin etc. from SEC/CRSP will be used to train the base models. We plan to use the bible in this area 'Quantitative Value' by Gray & Carlisle (2012) to explore additional variables.
- Technical analysis (stock price, moving average, etc.) [additional predictor variable]
  For technical analysis we will look at chart patterns (patterns traced out by stock trades over different intervals - seconds, minutes, hours, days, weeks, months, etc.), moving averages of stock prices, reversals, candlestick patterns, etc. We will gather this data as we said from Robinhood's agreements with exchanges.
- Sentiment analysis (mentions) [additional predictor variable]
  For sentiment analysis we will look mostly at mentions and whether those are positive or negative. The data will be from Stocktwits founded in 2009, and so we only have sentiment data going back 11 years and not of large volume so we must manage the impact of this data on our models in the early stages of training vs the fundamental and technical analysis. In general we will weigh the sentiment analysis from Stocktwits less than the other two data sources for our ML algorithm, fundamental data from the SEC, and historical stock price data directly from the exchanges.

We recognize that as we look at historical data we need to manage the timing of when the data appears to the model so that it mimics arrival patterns in the real world (such as when financial analysis gets released - so we don't have perfect knowledge at every step of the training).

Statistical Methods

We will apply a batch of mainstream machine learning models used in trading to train our ML-backed system. We will first use models of supervised learning and their relevance in the financial market to train the baseline portfolio (e.g. linear model, support vector machine, random forest, artificial/deep neural network). Then we apply unsupervised learning models (e.g. K-means clustering) and reinforcement learning models to solve common financial market prediction problems and validate the portfolio. After that we will predict trend using classification to backtest our strategies (e.g. support vector classifier). In addition, we will also use natural language processing for sentiment analysis to adjust the portfolio.

Potential Risks

Hindsight Bias. We rely significantly on analyzing past data to predict future data points. There is no guarantee that our ML algorithm will successfully predict future stock prices solely based on past data points.

- **Response:** We can utilize a Monte Carlo scenario analysis, in which we run thousands/millions of scenarios given a limited historical data set. For example, we can pretend that it is 2008, and only feed our machine learning algorithm historical prices and SEC filings from 2008 and before. We will withhold 2009 data and see how the algorithm performs. We will be able to repeat this process for any given time period. This way, our ML algorithm will learn from millions of possible future scenarios.

Heightened competition. Substantial and increasingly intense competition within the financial industry may harm our business. Many of our competitors (Morgan Stanley, Goldman Sachs, JP Morgan) have substantially greater financial, technological, operational and marketing resources than we/Robinhood do/does.

- **Response:** Data is king. We have first-mover advantage in terms of providing a democratized ML-backed solution for the everyday, retail investor. This means our machine learning algorithm will have a longer track record and more experience in terms of working with real-time data vs. our newer competitors who enter the market later.

Reliance on Accurate and Timely Data from Robinhood & Third Party Providers. We depend on order flow information from Robinhood, as well as accurate and up-to-date market information from the exchanges (NYSE/NASDAQ, CME). We have limited control over the third party firms, and may not be able to ensure the accuracy of the third-party information on our platform.

- **Response:** There is little we can do within this study to control for this potential issue. However, we can look for fail-safes (backup data sources) in case one data provider goes offline. We can also seek partnerships or regular internal checks to ensure that the data we receive is accurate.

Increased Market Efficiency over time. If other market participants catch on and utilize similar ML strategies like us, our ML systems may not be able to deliver outsized returns over the long term.

- **Response:** We do not believe in a complete efficiency of markets; there will be some inefficiencies, especially if there are human participants in the market.

Automation Bias. Investors, including our team, may overly rely on the ML algorithm to make correct decisions, even though we may question it from time to time due to our natural human instincts (e.g. selling fear, buying greed).

- **Response:** We can place limits on how much of one's total net worth can be allocated to our strategy, especially in its initial stages where a track record with actual, real-time data sets are still unavailable.

Privacy. Unauthorized disclosure, destruction or modification of data, through cybersecurity breaches, viruses, etc. could expose us to liability. Protracted and costly litigation could damage our reputation.

- **Response:** Data access should be restricted to monitors and employees within Robinhood. Wherever possible, pseudonymization techniques should be employed to prevent identifying customers. In addition, we believe that, if every customer felt that they could improve their market returns and reduce the risks in a difficult market by utilizing our ML-backed service, they would be willing to forgo some of their concerns to adopt our system.

Regulatory risks. Any money management operation is subject to complex and evolving regulations and oversight. Increased regulatory requirements could harm Robinhood and our ML-backed portfolio.

- **Response:** This is an unavoidable risk. In all industries, especially the financial industry there will be, from time to time, regulations we have to adhere to.

Deliverables

The goal of this research is to deliver a working prototype of the full system. This first step of demonstrating that we can deliver excess returns is critical. As we have highlighted, once this occurs there are follow on deliverables. The phases of our project are:

- Phase 1: Data gathering/collection/cleansing - 3 months
    - From Twitter/Stocktwits/CRSP/SEC.gov/any public data source(s)
    - Develop model for the training sets and test sets from the data
- Phase 2: Experimental phase/testing - 3 months
    - Program our ML algorithm and train our models on our training sets and validate on our test sets
    - Run models on a unrestricted set of historical data to see if it does better than the market (backtests)
- Phase 3: Model validation - 6 months
    - If we do not reach the hurdle of outperformance by 5% per annum, repeat Phase 1-2
    - We can annualize our returns if there are time constraints
- Phase 4: Present our findings to the product team and senior leadership at Robinhood

  - Write up the final report with insights and recommendations.
  - Document lessons learned
  - Project plan and budget for next project to get to a production system
  - Financial outline of the expected benefits to RobinHood and our customers.

We fully expect that we will be successful in our endeavor and deliver a product that will increase our customer base and will prompt our individual investors to trade with greater security and far less risk.

Statement of Contribution

- Ricardo:
  - Worked on the concept with all three members, ideating and determining who was the client and the approach.
  - Did research on value investment.
  - Led write-up on the Data Section.
  - Contribution to the Overview, led by Kevin.
  - Added information to the Study Design led by Minjie
  - Contributions to the Risk section led by Kevin
  - Did work on the deliverables section with team members.
- Kevin:
  - Worked on the concept with all three members, ideating and determining who was the client and the approach.
  - Researched academic journals to see if using ML in terms of stock price prediction has been well documented especially in terms of the growth and value stocks
  - Led write-up for the Overview and Potential Risks sections
  - Contribution to the Research Question and Study Design sections
- Minjie:
  - Led write-up for the research question, study design, sample, and statistical methods sections
  - Contribution to the overview, data and variable sections led by Ricardo and Kevin
  - Researched academic journals related to ML application on stock

---

Reference:

- Gray, W. R., & Carlisle, T. E. (2012). Quantitative Value: A Practitioner's Guide to Automating Intelligent Investment and Eliminating Behavioral Errors. John Wiley & Sons.