Exploration of a trading strategy system based on meta-labeling and hybrid modeling using the SigTech Platform.

Petri Nousiainen

Master's Thesis
Master of Engineering - Big Data Analytics
June 3, 2021

MASTER'S THESIS	
Arcada University of Applied Sciences	
Degree Programme:	Master of Engineering - Big Data Analytics
Identification number:	8277
Author:	Petri Nousiainen
Title:	Exploration of a trading strategy system based on meta-labeling and hybrid modeling using the SigTech Platform.
Supervisor (Arcada):	Ph.D. Magnus Westerlund
Commissioned by:	

Abstract:

The thesis aims to study a machine learning (ML) supported trading system. The methodology is based on a process that is utilizing meta-labeling, thus provides labels for a secondary model, where losses and gains are labeled outcomes. The secondary model provides a prediction based on the primary model output correctness. The secondary model predicts whether the primary model succeeds or fails at a particular prediction (a meta-prediction). A probability correctness measure of the direction prediction of the secondary model is used to size the position. They are "meta" labels because the original simple trading strategy predicts the ups and downs of the market (base predictions or labels). The metalabels predict whether those base predictions are correct or not.

The research question is finding a trading strategy process that can predict the next trading opportunity. The system constructed is a binary classification that aims to determine profitable trades, either buying or selling opportunities. The proposed system is a hybrid model setup of a primary and secondary model.

The process starts with processing raw price data given to a primary model, a trend following Donchian Channel technical analysis model. The primary model output is binary labeling used for the secondary model. The secondary model is the Random Forest machine learning algorithm that predicts if the next trade is a profit or not and with what probability. The prediction probability is used for order sizing. Finally, the rule for executing a trade is that both the primary and secondary models agree on the decision.

The result shows that the study supports the hypothesis that machine learning can improve any trading strategy. The coded trading strategy system successfully predicts the profitable trades with probabilities utilizing meta-labeling and the designed hybrid model. Furthermore, the machine learning prediction of the secondary model significantly improves the performance of the primary model trading strategy.

Keywords:	Machine learning, feature selection, metalabeling, meta-labeling, labeling, random forest, trading strategy, investment strategy, hybrid process, hybrid model, research process, trading system, bet size, trading platform, binary classification, SigTech
Number of pages:	92
Language:	English
Date of acceptance:	30.6.2021

CONTENTS

FIGURES

TABLES

FOREWORD

This basis for this research initially stemmed from the passion for developing better methods of trading strategy systems combined with machine learning (ML) algorithms.

The research question was formulated together with the supervisor, Magnus Westerlund. The research was complex, but conducting an extensive investigation has allowed answering the question that we identified.

27th May 2021 Helsinki

Petri Nousiainen

Table 1. Definitions

Definitions	
Term	Explanation
Active Portfolio Strategy	An active portfolio strategy defines an investment strategy targeting generating maximum value to a portfolio outperforming index investing by moving the capital to profitable securities according to available information and forecasting techniques. News (2021)
Asset management	Asset management is the direction of a client's portfolio by a financial services institution for growing the portfolio value while mitigating risk. Ganti (2021)
Classification	Classification, in statistics, refers to an identification problem to which of the category sets an observation belongs. Wikipedia contributors (2021a) In machine learning, classification means a predictive modeling problem where a class label predicts a given example of input data. Binary classification refers to classification tasks with two class labels. (Brownlee 2020b)
Compound annual growth rate	Compound annual growth rate (CAGR) defines a business and investment-specific term for the geometric progression ratio producing a constant rate of return over a period. Anson et al. (2010) The CAGR is the rate of return requiring an investment to grow from its starting balance to its ending balance, assuming reinvesting the profits at the end of each year during the investment's lifespan. Fernando (2021)
Conditional Probability	In probability theory, conditional probability defines a measure of the probability of an event occurring—given that another event has (by assumption, presumption, assertion, or evidence) occurred. Gut (2013)
Data dredging	Data dredging meaning also data fishing, data snooping, data butchery, and p-hacking is the exploitation of data analysis to discover patterns in data presenting statistically significant. Wasserstein & Lazar (2016)
Donchian Channel	Donchian Channels defines three lines generated by moving average computations that comprise an indicator established by upper and lower borders around a mid-range or median band. The upper band labels the highest security price over N periods, while the lower band indicates the lowest security price over N periods. The area between the upper and lower borders represents the Donchian Channel. Career futures trader Richard Donchian developed the technical indicator in 1936 to identify trends. He nicknamed later "The Father of Trend Following." (Chen 2021a)
Feature importance	Feature importance in machine learning indicates how useful the features are in predicting a target variable. Brownlee (2020a)

Table 2. Definitions

Definitions	
Term	Explanation
Feature selection	Feature selection attempts to improve the transparency and interpretability of machine learning models by ranking the importance of the features used. Man & Chan (2021b)
GPT-3	Generative Pre-trained Transformer 3 (GPT-3) defines an autoregressive language model using deep learning to produce human-like text. GPT-3 is the third-generation language prediction model in the GPT-n series created by OpenAI. (Brown et al. 2020)
Hybrid model	A hybrid is the combination of two or more different elements aimed at achieving a special objective or goal. A model is a simplified version of something complex used in analyzing and solving problems or making predictions. (Alliance 2021)
Investment Strategy	An investment strategy defines as a set of principles helping investors achieve the investor's financial and investment goals based on investment goals, risk tolerance, and future requirements for capital. Chen (2021c)
Machine learning (ML)	Machine learning (ML) defines the study of computer algorithms that improve automatically through experience and data use. Mitchell (1997)
Meta-labeling	"The goal of meta-labeling is to train a secondary model on the prediction outcomes of a primary model, where losses are labeled as "0" and gains are labeled as "1". Therefore, the secondary model does not predict the side. Instead, the secondary model predicts whether the primary model succeeds or fails at a particular prediction (a meta-prediction). The probability associated with a "1" prediction can then be used to size the position." de Prado (2020) They are called "metalabels" pioneered by Marcos Lopez de Prado (De Prado 2018). They are "meta" because Marcos assumed the original simple trading strategy is used to predict the ups and downs of the market itself- those are the base predictions or labels. The metalabels are on whether those base predictions are correct or not.
Non-farm payroll employment	Non-farm payroll employment is a compiled term for goods, construction, and manufacturing companies in the US. It is excluding farm workers, private household employees, or non-profit organization employees. The financial assets most affected by the non-farm payroll (NFP) data include the US dollar, equities, and gold. Wikipedia contributors (2021b)

Table 3. Definitions

Definitions	
Term	Explanation
Portfolio	Portfolio is a collection of assets held by an institution or a private individual. Wikipedia contributors (2021c)
Process	A Process is a series of actions or steps taken in order to achieve a particular end. Languages & Google (2021)
Random Forest	Random forests (RF) defines a combination of tree predictors, each tree depending on the random vector values selected independently and with the same distribution for all trees in the forest. Breiman (2001) The RF is a method for classification and regression for constructing a multitude of decision trees. Kam Ho (1995)
SHAP	SHAP (SHapley Additive exPlanations). The SHAP algorithm assigns each feature an importance value for a particular prediction. Lundberg & Lee (2017a) SHAP is a method to explain individual predictions and based on the game theoretically optimal Shapley Values. Molnar (2019)
The Sharpe Ratio	"The Sharpe ratio is designed to measure the expected return per unit of risk for a zero investment strategy. The difference between the returns on two investment assets represents the results of such a strategy. The Sharpe Ratio does not cover cases in which only one investment return is involved." (Sharpe 1994) It refers to the additional amount of return that an investor receives per unit of increase in risk. The developer was William F. Sharpe, 1966.
System	System is a set of principles or procedures according to which something is done; an organized scheme or method. Dictionary (2021)
Strategy	Strategy definition means maintaining a growing balance between ends, ways, and means; about identifying objectives and the resources and methods available for meeting such objectives. (Freedman 2015)
Strategy_2	$Strategy_2$ defines a general plan to achieve one or more long-term or overall goals under conditions of uncertainty. Wikipedia contributors (2021f)
Theory	A theory defines a well-substantiated explanation of an aspect of the natural world that can incorporate laws, hypotheses, and facts. (McComas 2003)
Trading	Trading, financial trading is buying and selling financial assets or instruments (such as shares, currency or bonds, or derivatives such as CFDs, futures, or options) to make a profit. (Lill 2021)

Table 4. Definitions

Definitions	
Term	Explanation
Trading Process	A trading process is an act of buying, selling, or exchanging financial instruments, e.g., stocks, bonds, or currency: Broker companies generally charge a commission per trade. (a series of actions or steps taken to achieve a particular end.)
Trading Strategy	A trading strategy defines a method of buying and selling in markets based on predefined rules used to make trading decisions. (Chen 2021b)
Trading System	A trading system defines as a set of rules that generate buy and sell trading signals without any ambiguity or any subjective elements. The primary objective of a trading system is to manage risk and increase profitability in any market environment. (Moreira 2021)
Trend Following Strategy	Trend following or trend trading is described as a trading strategy according to which an investor should purchase a security when its price trends up and sell when its trend decreases, expecting price movements to continue. (Wikipedia contributors 2021g)
Triple-Barrier Method	Meta-labeling: suppose that a model sets a trading signal side (long or short); you only need the size of the bet. Meta-labeling is a secondary ML model that can tell how much money we should risk in such a bet. (De Prado 2018 chap. 3.6) The Triple-Barrier Method is de Prado's labeling method, and it labels an observation according to the first barrier touch out of three barriers. (De Prado 2018 chap. 3.4)

1 INTRODUCTION

1.1 Introduction

Trading starts with a dream, goal, and motivation. A Dream gives value and purpose to the hard work. Achieving a Goal is a process of planning and preparation, execution, and evaluation. Motivation helps to be fascinating and passionate about what to do and sustain a strong focus over time. (Ward 2010)

"Becoming a great trader is a marathon, not a sprint!" -Abe Cohen, trader mentor. (Ward 2010)

Why is it important to think in probabilities? What does it mean? The best traders treat trading like a numbers game. Casinos make consistent profits facilitating an event that has a purely random outcome. Events with probable outcomes can produce consistent results if one can get the odds in one's favor and a large enough sample size. It is the way casinos and professional gamblers approach gambling. (Douglas 2000)

A Trader, gambler, and casino are all dealing with both known and unknown variables or features impacting the outcome of each trade or event. The known variables are the rules of the game (gambling) and the results of the market analysis (trading). The unknown variable in trading is all the other traders that may enter or leave the market at any time, contributing to the collective behavior pattern at that moment. (Douglas 2000)

(Covel 2009) summarized the sum of the parts as if building a system that gives an entry and exit, tells *how much* to bet along the way, and adjusts to your current capital and current market volatility at all times, and no more analysis is needed.(Covel 2009)

1.2 Background and Need

Successful investment strategies are specific implementations of general theories. Asset managers should focus on researching theories rather than backtesting potential trading rules. Only a theory can identify the precise cause-effect mechanism that allows generating profits against the behavior of crowds. (x implies y, and the absence of y implies the absence of x) A historical simulation of an investment strategy's performance (backtest) is not a theory. It is a simulation of a past that never happened. Theory, not historical simulations, must support strategies. Machine learning (ML) is a powerful instrument for building financial theories. (de Prado 2020)

Machine learning refers to algorithms that learn complex patterns in a high-dimensional space without being directed. ML learns without being directed, and the algorithm derives the structure from the data. (de Prado 2020) Machine learning using hybrid modeling introduced by Psichogios & Ungar (1992) and Thompson & Kramer (1994). Their hybrid models in machine learning include a knowledge-driven and a data-driven submodel. Two submodels are combined in different ways depending on the problems. (Hu Baogang 2017)

Machine learning is used in asset management for hedging, portfolio construction, detection of outliers and structural breaks, credit ratings, sentiment analysis, market making, bet sizing, securities taxonomy, risk management, detection of false investment strategies, and so on. However, maybe the most popular ML application in asset management is price prediction. (de Prado 2020)

Machine learning challenges in asset management are that modeling financial series is harder than diving cars or recognizing faces. The reason is, the signal-to-noise ratio in financial data is low because of arbitrage forces and nonstationary systems. Moreover, ML is not a substitute for economic theory but rather a powerful tool to develop better modern financial theories. (de Prado 2020)

The advantages of machine learning to academic financial research offer exciting times in the future. First, ML provides the power and flexibility needed to find dim signals in the sea of noise caused by arbitrage forces. Second, ML allows dividing the research process into two stages: 1) searching for important variables without functional form and 2) searching for a functional form that binds those variables. Third, ML offers the possibility for simulating synthetic data.(de Prado 2020)

1.3 Purpose of the Study

The purpose of this study is to find a machine learning (ML) supported trading strategy system. The fulfilling of the purpose initiates by ways of understanding the field of recent research and theories and further, leading to deeper recognizing of the possible components, processes, and methodologies—the improved understanding enabling designing the trading strategy system for testing and analyzing—finally, the concluding the implemented trading strategy performance with findings and proposes for future studies. To summarize, the purpose of the study is to increase trading strategy and machine learning-related knowledge to the level possible for implementing the found or designed trading strategy system resulting in academically valid conclusions for further study.

1.4 Research Question and Hypothesis

The domain of study leads to the following research questions aimed at understanding the elements, and the combinations of trading strategies and machine learning algorithms. By exploring ways of constructing a trading strategy system, the aim is to find an understandable and straightforward solution utilizing machine learning. Engineering requirements for scalability, repeatability, and interpretability are important design considerations. The research questions are following.

1.Exploring how to find and construct a machine learning (ML) supported trading strategy system predicting the next trading opportunity?
2.How can metalabeling enable and improve the integration of machine learning in a trading strategy?
3.What is a hybrid model that combines the elements of the trading strategy system?
4.What are the empirical results from the implemented trading strategy system?

The research hypothesis is that machine learning can improve any trading strategy.

1.5 Significance To The Field

de Prado (2020) argues that Asset managers should focus on researching theories, not backtesting trading rules. A common misunderstanding is to consider of backtesting as a research tool. De Prado (2018) Most backtests in journals are flawed due to selection bias on multiple tests. Bailey et al. (2014); Harvey et al. (2016) Backtesting is a historical simulation of how a trading strategy would have performed should the strategy have been run over a past period. It is a hypothetical experiment. The past would not repeat itself. De Prado (2018)

Marcos' first rule of backtesting - ignore at your own peril : "backtesting is not a research tool, feature importance is." De Prado (2018)

What is the point of backtesting? A backtest is a sanity check on several variables, including bet sizing, turnover, resilience on costs, and behavior under a given scenario. The purpose of backtesting is to discard bad models, not to improve them. (De Prado 2018)

A team of quantitative analysts in Deutsche Bank, led by Yin Lou, published a study of "Seven Sins of Quantitative Investing." in 2014. Luo (2014) It discusses the seven common mistakes investors tend to make when they perform backtest and build quantitative models. (De Prado 2018)
1. Survivorship bias: Using only companies currently in business ignoring bankruptcy, delisting, and acquisitions.
2. Look-ahead bias: Utilizing non-public information or information not available at the moment of the historical simulation.
3. Storytelling: Creating a story *ex-post* to justify random patterns.
4. Data mining and data snooping: Training the model on the testing set.
5. Transaction costs: The only way to be sure about the cost is to make the actual trade.
6. Outliers: building strategy on a few extreme outcomes that may never happen again
7. Shorting: Taking a short position requires finding a lender, and the costs may be unknown.
De Prado (2018) argues that most papers published in journals make these basic errors routinely. Sarfati (2015), at the CBOE Risk Management Conference, discussed other common errors, including computing performance using a non-standard method, such as ignoring hidden risks, focusing only on returns while ignoring other metrics, and ignoring the existence of stop-out limits or margin calls. (De Prado 2018)

Successful investment strategies are specific implementations of general theories. Theory must support strategies. A historical simulation of an investment strategy (backtest) is not a theory. Only a theory can identify a cause-effect mechanism that extracts profit and explains factual evidence. ML helps discover theories. (de Prado 2020)

Feature selection algorithms may improve machine learning accuracy. Man & Chan (2021a) argue that the feature selection algorithms in machine learning suffer from the "random seed" problem. Meaning that features ranked most important can be ranked to a low level in another run. Sometimes the selected features are the only desired output of an ML program. Thus, if the random seed produces different output every run, the output is not interpretable. The study compared MDA, LIME, and SHAP algorithms and introduces "instability-index" to measure the stability of the feature selection algorithms. A finding is that high feature stability does not necessarily improve the predictive performance. In applying a LIME algorithm to a trading strategy, both Sharpe ratio and cumulative return were improved out-of-sample. (Man & Chan 2021a)

Man & Chan (2021b) study feature importance in machine learning that indicates how much information the features contribute in predicting a target variable. de Prado (2020) introduced the cluster-based MDA (cMDA) method that improves predictive performance, feature stability, and model interpretability. Man & Chan (2021b) study shows that the stability and interpretability of the cMDA selected features are superior to MDA selected features.

The machine learning-supported trading strategy target is to maximize the profit of the initial capital. Thus, the strategy has to identify trading signals of the side (long or short) in financial instruments. As important as identifying good trading opportunities is determining the best trading size (size the bet) properly. (De Prado 2018 chap. 3.7) Otherwise, achieving high accuracy on small bets and low accuracy on high bets will ruin you. (De Prado 2018 chap. 3.7) We want the ML algorithm to learn to tell us what is the appropriate size of the bet, including the possibility of no bet at all (zero bets). (De Prado 2018 chap. 3.6) De Prado calls this sizing problem meta-labeling. (What is meta-labeling? Meta-labeling: suppose that a model sets a trading signal side (long or short); you only need the size of the bet. Meta-labeling is a secondary ML model that can tell how much money we should risk in such a bet. (De Prado 2018 chap. 3.6) The Triple-Barrier Method is de Prado's labeling method, and it labels an observation according to the first barrier touch out of three barriers (De Prado 2018 chap. 3.4) building a secondary ML model that learns how to use a primary exogenous model.) (De Prado 2018 chap. 3.6) De Prado claims that no book or paper had discussed this common sizing problem, meta-labeling, before de Prado. (De Prado 2018 chap. 3.6)

1.6 Limitations

The limitations in data mean that the SigTech platform provides the data including tradable rolling front futures of about 20-year historical data. The data input financial futures must be tradable for the trading strategy to avoid liquidity risk. Liquidity risk meaning that nobody wants to sell when trying to buy and vice versa.Faith (2019) The possibility for executing a trade is critical for the trading strategy when there is a trading signal.

The limitation of the parameter time refers to using only the daily closing price data. The input data ignores intraday, weekly and monthly periods.

The study limits to trend-following strategy even if there is a wide universe of trading strategies. In following a trend, the trader is trying to join large price movements lasting several months. Trend followers enter trades when prices are historical highs or lows and exit when prices reverse. Challenges in trend following strategies are that 1) they generate losing trades when trying to enter a trend start and 2) when there are no trends. Moreover, the trend-following strategy may need 3) a high amount of initial capital if there is a large distance between the entry and stop-loss prices.(Faith 2019)

The limitation in ML algorithms is that they must predict probabilities. Thus, we have chosen Random Forest ML algorithm to run the prediction in the secondary model.

The study limits out feature selection, knowing the importance of improving the predictive performance, feature stability, and model interpretability.

The study ignores transaction costs, taxes, and the risk-free interest rate on calculations.

Trading execution costs for futures contracts are meager, and the study trades only futures contracts. The assumption is that the risk-free interest rate is zero because there is no notable yield, currently, if an investor holds capital on the bank account. Moreover, the study ignores a trader's mental and psychological development Steve Ward (2021) focusing on a mechanical or technological trading system.

1.7 Ethical Considerations

Zicari et al. (2021) introduces a valuable process for auditing, or before implementing, an AI system considering ethical, social, technical, and legal risks and pitfalls. The study states that assessing trustworthy artificial intelligence (AI) applications based on machine learning (ML) and/or deep learning (DL) is difficult. The sophisticated AI applications raise new ethical and legal questions that significantly impact society. (Zicari et al. 2021)

Brundage et al. (2020) argues, and on Artificial Intelligence (2020) confirms that "techniques and best practices have not yet established for auditing AI systems." This study is not developing an AI system but utilizing a machine learning algorithm. The seven ethical requirements established by the EU High-Level Experts Guidelines for Trustworthy AI (HLEG) apply for ethical assessing the machine learning-based trading system. (on Artificial Intelligence 2020)

Table 5. The seven ethical requirements for trustworthy AI.

The seven ethical requirements for trustworthy AI		
Nr	Requirement	Brief Description
1.	Human agency and oversight:	The trading strategy systems have human interaction only in interpreting the performance.
2.	Technical Robustness and safety:	The trading strategy system is reproducible and accurate. Further, the system implements in the SigTech platform that is secure and safe.
3.	Privacy and data governance:	The trading strategy system data provided by the SigTech platform is reliable with high integrity and high quality.
4.	Transparency:	The trading strategy system utilizes only historical price data: public information. The primary model is the Donchian channel technical analysis model publicly available, and the secondary model uses statistical systems. The secondary model random forest transparency validates by the features importance table.
5.	Diversity, non-discrimination, and fairness:	The trading strategy system is not possible to use for cheating or front running.
6.	Societal and environmental well-being:	The trading strategy system is sustainable and environmentally friendly.
7.	Accountability:	The primary model Donchian channel parameters optimized by maximizing the cumulative percent return of the output with S&P500 E-mini front futures contract closing prices of 917 days testing period. The model testing implements with the same data. The secondary model random forest uses statistical systems.

2 RELATED

2.1 Financial Theory

The study examines a trading system that observes the price movement of any preferred financial instrument. Therefore, the literature of asset pricing theory or model (such as CAPM and APT) valuations and return predictability is not a concern for this examination. However, it observes the price movement of the security.

2.1.1 Information Efficiency

An ideal market is where the investors can choose securities under the assumption that security prices at any time "fully reflect" all available information. A market where prices always "fully reflect" available information is "efficient." (Fama 2021)

2.1.2 Efficient Markets Hypothesis

The efficient markets hypothesis discusses the adjustment of security prices with three information subsets. First, the *weak-form* tests the information set of the historical prices. Second, the *semi-strong form* tests whether prices efficiently adjust to other information publicly available (e.g., announcements of annual earnings, stock splits, and so on.). Third, the *strong form* tests whether given investors or groups have monopolistic access to any information relevant for price formation. (Fama 2021)

2.2 Literature Review

Although technical analysis is not the focus of this study, a brief review presents the evolution elements in technical analysis, trend following, and trading strategy. Moreover, the overview continues with fixed fraction asset allocation and futures contract introductions.

2.2.1 Technical Analysis

Technical analysis and fundamental analysis are the primary methods for analyzing securities and assess investment decisions. Fundamental analysis determines a company's fair value by analyzing financial statements. At the same time, technical analysis assumes that the security price reflects all publicly available information and focuses on analyzing price movements. However, academics have been skeptical for technical analysis, perhaps for three reasons. First, there is no theoretical basis for it. Second, the stock price random walk model rules out any profitability from technical trading. Third, earlier empirical findings, such as Cowles 3rd (1933) and Fama & Blume (1966) are mixed and inconclusive. Zhu & Zhou (2007) This study is utilizing the technical analysis indicators presented in the (Qiu & Song 2016) paper.

"Technical analysis is anathema to the academic world. We love to pick on it." - Malkiel (1981), A Random Walk Down Wall Street.

2.2.2 Trend Following

The founder of Wall Street Journal, Charles Dow, is considered the father of technical analysis since his Dow Theory was the first technical method to predict future movements of security prices. Dow suggested that prices tend to move in trends, and utilizing trends was a key to profitable investing. Pring (2002) Moreover, Kirkpatrick Charles & Dahlquist Julie (2007) discusses that the main objective of technical analysis is to identify trends in security prices. A key element of trend following strategies, channel breakout, has been studied in financial literature since the 1960s. (Swart 2016)

2.2.3 Trading Strategies

Numerous trend-following trading strategies identify trends by differing entry and exit signals for investors with various investment appetites. Swart (2016) A traditional trend-following moving average crossover strategy was examined by Dahlquist (2005) resulting in a simple moving average cross over strategy outperforming, on average, a contrarian strategy, neither outperforming a simple buy and holding strategy. The strategy provides an entry signal for the investor to go long (buy) or go short (sell) once a shorter period moving average of n days penetrates a longer period moving average of (n+t) days. Swart (2016) However, Jegadeesh & Titman (1993) studied trading strategies based on relative strength, an investor buying past winners and selling past losers, showing results of abnormal returns. Moreover, Wilcox & Crittenden (2005) performed a comprehensive study of a trend following strategy over 24 000 securities during 22 years. The strategy examines a long-only technique of buying a stock at a new all-time high and exiting a ten-day average true range trailing stop. Swart (2016) The results show that trend following on stocks does offer a positive mathematical expectancy. Wilcox & Crittenden (2005)

2.2.4 Fixed Fraction Asset Allocation

The machine learning-supported trading strategy target is to maximize the profit of the initial capital. Thus, the strategy has to identify trading signals of the side (long or short) in financial instruments. As important as identifying good trading opportunities is to size the bet properly. (De Prado 2018 chap. 3.7) Otherwise, achieving high accuracy on small bets and low accuracy on high bets will ruin you. (De Prado 2018 chap. 3.7) We want the ML algorithm to learn to tell us what is the appropriate size of the bet, including the possibility of no bet at all (zero bets). (De Prado 2018 chap. 3.6) De Prado calls this sizing problem meta-labeling, building a secondary ML model that learns how to utilize a primary exogenous model. (De Prado 2018 chap. 3.6) De Prado (2018) claims that no book or paper had discussed this common sizing problem, meta-labeling, before de Prado. (De Prado 2018 chap. 3.6)

Miner (2008) discusses that position sizing is crucial to successful long-term trading. Big positions may result in big losses that are difficult to cover. A 50% drawdown requires 100% gain to get back to even. Successful traders have relatively close stops. The best professional traders rarely have a greater than 50% win record. Therefore, it is critical to minimize the losses and preserve capital. The standard for the maximum capital exposure (risk), a dollar amount, per trade is three percent and six percent for all open trades. The capital exposure percents calculate from the funds available for trading. Most traders also have a maximum acceptable loss of ten percent per month. If the closed trades have resulted in a 10% drawdown, stop trading for the month's balance. (Miner 2008)

However, (Kelly 1956) introduces a scientific gambling method Kelly criterion (or Kelly strategy or Kelly bet) formula for bet sizing. The formula conducts almost certainly to higher wealth than any other strategy in the long run (number of bets proceeds to infinity). The Kelly criterion conducts by maximizing the expected value of the logarithm of wealth, identical to maximizing the expected geometric growth rate. The Kelly criterion computes the needed bet size percentage by multiplying the percent probability to win by two, then subtracting one hundred percent. Therefore, the Kelly bet size formula shows that with a 70% probability of winning, the optimal bet size is 40% of the total risk capital. (Kelly 1956) In summary, the Kelly criterion is a method to find the optimal bidding fraction for achieving the maximal asset growth rate.Wu & Chung (2018)

The Kelly criterion for maximizing the long-term growth rate of capital. Kelly (1956)

$$f* = p - \frac{q}{b}.$$

f* = is the fraction of current capital to bet (how much to bet, in fraction)
b = net fractional odds received on the bet (betting $10, winning $4 plus bet, b = 0.4)
p = probability of a win
q = 1 - p is the probability of a loss.
The formula simplifies for even money bets (b = 1)

$$f* = p - q.$$

Since, q = 1 - p, it simplifies the formula to

$$f* = 2p - 1.$$

Kelly (1956) is further investigated by Gehm (1983), and Vince (1990) introducing the *"Optimal f"* technique for future traders that aims identifying the optimal fixed fraction of the portfolio to bet on any outcome. Further, Anderson & Faff* (2004) found the *"Optimal f"* technique functional for speculative future traders in money management while applying the turtle trading system. The *"Optimal f"* method is in operation in the online trading scenarios having accurate capital amounts and observed losses. (Swart 2016)

This study examines the fixed fraction asset allocation for the primary model and the secondary (ML) model prediction of profit or not. However, the secondary (ML) model prediction of probability applies the flexible position fraction method. In the flexible position fraction method, the (ML) probability prediction result determines the bet size for the trade. The added value of the (ML) probability prediction of the next trade is that it allows increasing the position risk, which may impact increasing expected return. The higher the position risk, the higher the expected return of the trade. In the behavior finance research, Haigh & List (2005) shows a trader changing position size according to the trading performance, increasing position risk while profiting and decreasing position size while losing; an expected outcome under utility theory.

2.2.5 Futures Contracts

A futures contract is a derivative instrument, a function of an underlying asset, a contract between a buyer and a seller of an asset for an agreed price today (the current futures price) with delivery and payment materializing on the delivery date in the future. The risk allocation feature of the futures contract allows the trading parties to secure rates or prices for future transactions in advance. Transaction costs for futures contracts are low, which allows taking exposure to the underlying (commodity, index, currency, and so on) for a fraction of the cost of a direct trade to the underlying. The low transaction costs and the ability to sell futures short (short-sell) easily are the elements of liquid financial instruments for testing active trading systems and strategies. (Swart 2016)

2.3 Body of the Review

Metalabeling is a potent tool to have because of four advantages. First, meta labeling allows transforming a fundamental (founded on economic theory) model into an ML model. Second, overfitting is limited because ML will not decide the side of the bet, only the size. Third, meta labeling enables sophisticated strategy structures; for instance, one primary model including feature set for long positions, and another primary model including features for short positions. Fourth, meta labeling ML models can deliver more robust and reliable outcomes than standard labeling models. It is essential to develop an ML algorithm solely focused on getting the critical decision of bet sizing right because achieving high accuracy on small bets and low accuracy on large bets will ruin one. The section describes the blog Meta Labeling (A Toy Example) Thames (2020), originally based on (De Prado 2018)

2.3.1 Trading Strategy Process - Predicting Conditional Probability of Next Trade's Profit

For computing the size of the bet, implement a trading strategy process that predicts the conditional probability of profit of the next trade. (Chan 2020a) Based on that, Dr. Ernest P Chan has developed a Tail Reaper strategy, and the process overview is here. The unconditional probability of profit has no dependency and no forecasting use, for example, winning ratio counting, winning and losing trade, from historical data. The conditional probability of profit is conditioned on a specific environment and time of the expected trade. In a long volatility strategy, the conditional probability of profit would be high in crises and low in a calm market. How to calculate this probability? Machine learning is the only way to compute this probability. 1) Prepare a spreadsheet with returns of historical trade data. 2) Label y; we only care about whether the trades are profitable or not. Label the trades profitable 1, otherwise 0, and ignore the magnitude of returns. 3) Input x, add features like VIX, 1-day SPY return, change in the interest rate on the previous day. It is possible to add as many features as one like, valid or not. It is not a problem for ML because it will drop useless features via the feature selection process. These feature values must be familiar before each trade's entry time. Otherwise, there will be look-ahead bias, and the system is disabled for live trading. 4) Choose a random forest model for prediction because the nonlinear co-dependences between predictors need to be discovered and utilized. Moreover, too many relationships are to be captured for a linear model. An example of a relationship is that if VIX is lower than 15, the 1-day SPY return is of no use for predicting the probability of profit of the next trade. Moreover, if VIX is higher than 15, the 1-day SPY return is beneficial. 5) Train random forest model, ready for live trading. 6) Input all the latest data and other features into a new spreadsheet and make live predictions. 7) Download the prediction outcome of random forest whether the trade is profitable and with what conditional probability. Use the probability to size the trade. 8) The sizing of a trade decision may go like following: a) if the probability of profit is higher than 0.6, buy 20Keur of AAPL, b) if the probability is between 0.51 and 0.6, buy only 5Keur, c) if the probability is under 0.51, do not buy at all. 9) Prediction repeating will take only about 1 second or less, and repeating can perform as frequently as possible. The API method interacts trading system seamlessly. 10) Training frequency of the prediction model random forest need not be more than once a quarter. (Chan 2020a)

11) For improving your prediction model by feature selection, they utilize a predictive algorithm SHAP. (Lundberg & Lee 2017b) for feature selection (Man & Chan 2020). The SHAP will remove all the features "below the average" and retrain the model with the remaining features for improving your underlying simple (primary) trading strategy. In the feature selection study, they examined MDA, LIME, and SHAP algorithms. (Man & Chan 2020) To reduce variability in feature importance ranking and to improve interpretability; they found that LIME is preferred to SHAP and certainly preferred to MDA Man & Chan (2021a). However, they applied a hierarchical clustering methodology, pioneered by de Prado (2020), prior to MDA feature selection to the same data sets with a method called cMDA, and it shows surprising results. Additionally, they applied the cMDA method to the Tail Reaper strategy with public data of the S&P500 monthly returns. The model accuracy improves from 0.517 to 0.583 when using cMDA instead of MDA, and the AUC score improves from 0.716 to 0.779.Man & Chan (2021b)

12) Critics to Tail Reaper's strategy is that it is very hard to predict the market's movement (for example, if the coronavirus hits the US president, what is its impact on the market return) because of the low signal-to-noise ratio. The strategy should find out how to filter out the noise by telling machine learning prediction what is favorable or unfavorable to your strategy and with what probability. Anyway, Dr. Chan discusses the Kalman filter (KF) in his book. (Chan 2013) The optimal filtering theory of Kalman filters shows a better prediction of the price moves as lag is reduced while computing with SP 500 mini-futures contract. (Benhamou 2016) 13) Advantage of this model is that fellow traders can not arbitrage away your trading alpha when you are using predict numbers like company earnings surprise, credit rating change, or the US nonfarm Chan (2019) payroll surprise. 14) Random forest model was chosen because it is simple but not too simple, and deep learning algorithms are too complicated. Complicated deep learning algorithms like LSTM can use time series dependence of the features and labels, but they run a serious risk of data snooping due to the large number of parameters to fit. Moreover, the GPT-3 deep learning algorithm has more than 175 billion parameters to fit, too much for

1000 historical trades. 15) Tail Reaper strategy is a tail hedge strategy, a crisis alpha, and it was working successfully by creating a 64 percent YTD return as of June 2020. Moreover, the ML model told not to enter any trade (due to the low conditional probability of profit) from Nov 2019 to Jan 2020. Suddenly, 1st Feb 2020, the ML program signaled to expect a crisis. It captured a 12 percent return later that month. However, the Tail Reaper system trades only the E-mini SP 500 index futures.Chan (2020b) Past performance is not a guarantee of future results.Chan (2020a) The whole process description of Tail Reaper strategy paragraph was based on the blog of Dr. Ernest P Chan, 6th Aug 2020.(Chan 2020a)

To summarise the process of predicting the conditional probability of profit for your next trade with the Tail Reaper strategy. For the model random forest, your input x features include all the possible financial instruments in the market with the historical return data. You input y label includes the "metalabeled" market return (SP 500 index) profit true (1) or otherwise false (0). The random forest model outcome is improved by utilizing the SHAP algorithm for feature selection and for reducing noise, which was possibly used Kalman filter. After train and prediction with random forest, the model will tell you the result if your trade is profitable and with what probability. The probability is used for sizing the bet of the trade. The improvement of the ML model accuracy is performed by studying feature selection. The feature importance table shows the ranking of each feature used in the ML model. The cluster-based cMDA method used in the Tail Reaper strategy with public data of the S&P500 monthly returns shows the model accuracy improving from 0.517 to 0.583 when using cMDA instead of MDA, and the AUC score improving from 0.716 to 0.779. Finally, to test trading strategy online, connect the trading system to the market via API and run the model with live predictions trading with S&P500 mini-future contract.

2.3.2 The Process of a Strategy in the SigTech Platform

The process of a strategy in the SigTech platform is based on the queue of methods. These methods are run in order of time and (if times match) priority. Each process can make any decision to manipulate the contents of a strategy. This is the novelty compared to the common back-testing methodology of using large time-series data arrays. The key

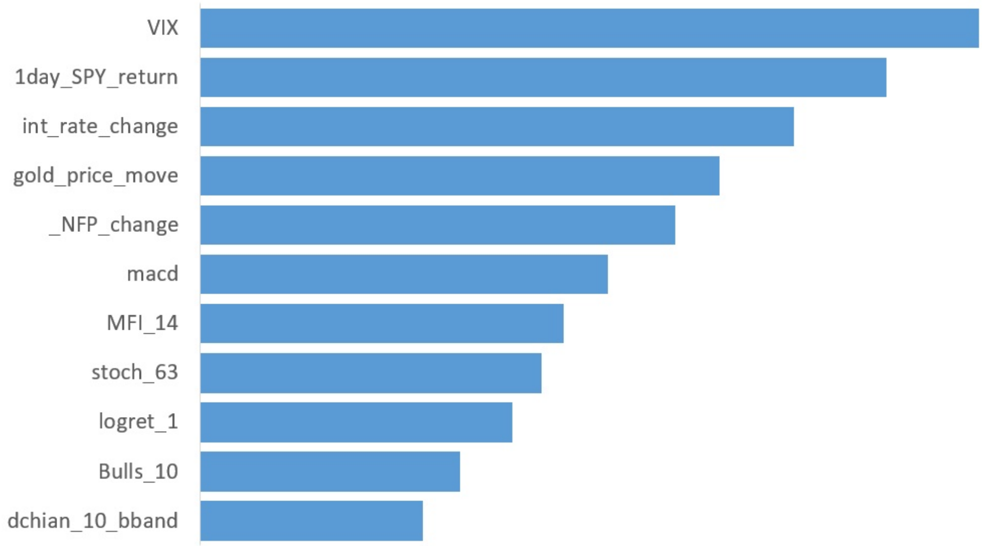

Figure 1. Features Importance Table of The Meta Labeling Process - Tail Reaper Trading Strategy "

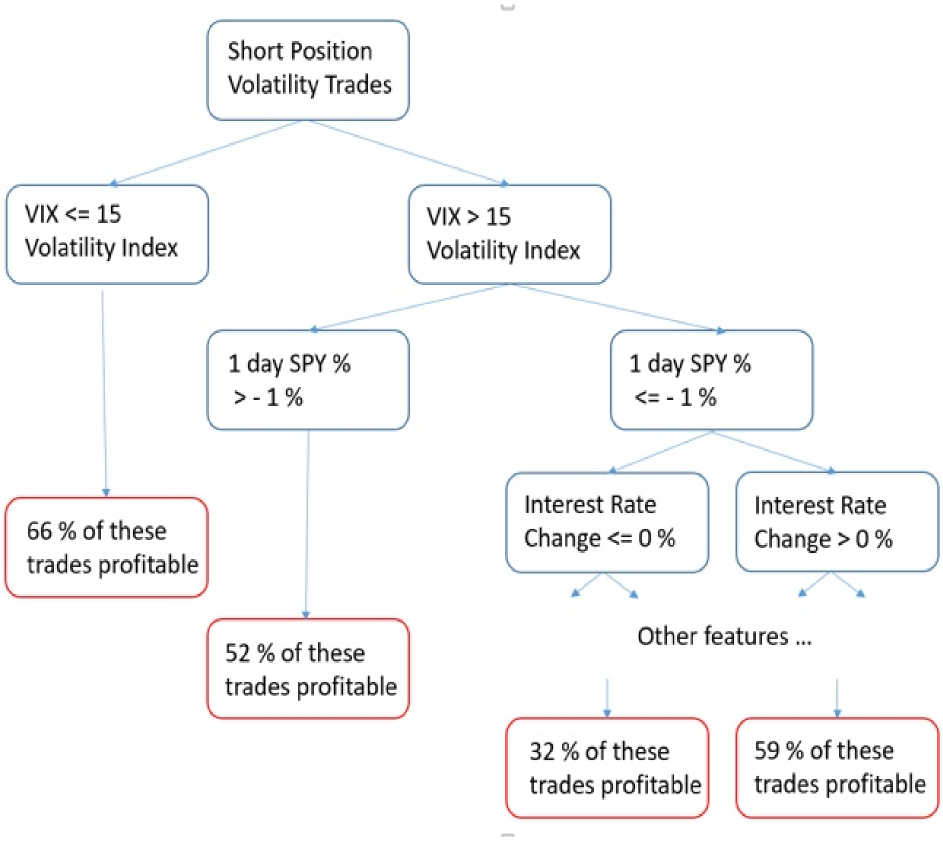

Figure 2. Classification Tree Example by predictnow.ai - Tail Reaper Trading Strategy "

advantages of the process queue are the following. First, better requirement scaling - in the time-series approach, it becomes increasingly costly when every element (instrument/signal) of a strategy has information for every point of interest. Second, detailed order and transaction modeling - the process approach allows intricate order logic to be correctly modeled. Third, greater flexibility - the process approach allows a modular strategy construction where strategies can inherit characteristics and hold positions in sub-strategies. Since each process can perform any code, it allows greater complexity than a simpler way.[1]

2.3.3 Another Machine Learning Supported Process

The section is based on the blog of Thames (2020), "A Toy Example" investigates Meta Labeling (chap. 3 De Prado 2018 p. 50) that improves model and strategy performance metrics by filtering out false positives. The blog shows the process and the metrics by utilizing MNIST handwritten digit classification. Generally, when you have a strategy model that predicts the side of the bet (long or short), you need to learn the size of that bet, including no bet at all (zero sizes). That problem is called meta labeling because we want to build a secondary ML model that learns to use a primary exogenous model. In other words, the primary ML model is trained for binary prediction to take a bet or pass. When the predicted label is 1, we use the probability of the secondary model prediction to size the bet, where the side (sign) of the position is set by the primary model. The binary classification problem is a trade-off between false positives and false negatives. Generally, increasing the true positive rate tends to increase its false positives. The receiver operating characteristic (ROC) curve measures the cost of increasing the true positive rate by accepting a higher false-positive rate. Meta labeling will increase the F1 score by filtering out the false positives. The secondary model task is to determine whether the positive of the primary (exogenous) model is true or false. Moreover, the secondary model will determine if we should act or pass the opportunity presented. The model process of the MNIST example is the following. First, a primary model (binary classification) is trained with high recall (split 90/10 train/test) to see when we are over-fitting. Second, we determine the threshold level for high recall of the primary model by utilizing the ROC curve. A high recall means that the primary model captures the majority of positive samples even if there is a large number of false positives. The meta-model will correct it by reducing the number of false positives and thus improving the metrics. Third, the features from the primary model are concatenated with the predictions from the primary model into a new feature set for the secondary model. Meta labels are targets in the secondary model. Fit

[1] https://www.sigtech.com/ The section is based on the SigTech strategy framework description.

the secondary model. Meta labels are true (1) if the primary model prediction matches the actual value, else false (0). Here, if an observation is a true positive or true negative, then label it as 1 (the model is correct), else 0 (the model is incorrect). Because it is categorical, this example adds One Hot Encoding. Fourth, we combine the prediction from the secondary and primary models. Only when both are true, the final prediction is true. If the primary model predicts a three (3) and the secondary model gives a high probability of the primary model result to be correct, is the final prediction a 3, else not a 3. The section is based on the Hudson and Thames blog "A Toy Example".(Thames 2020)

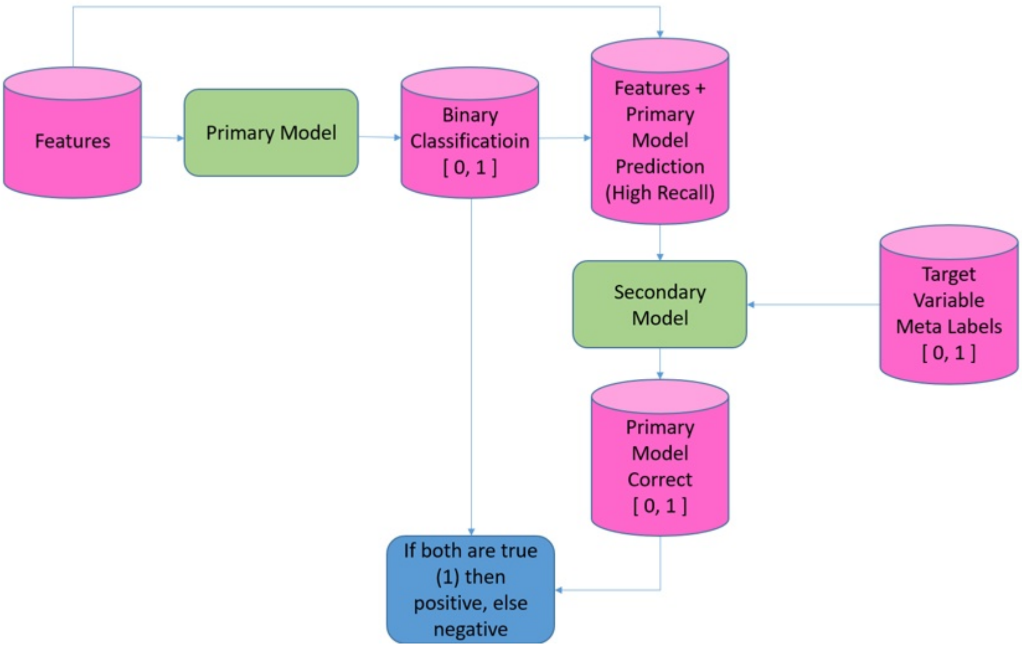

Figure 3. Process of Meta Labeling - Hudson and Thames "A Toy Example"

Table 6. Base Model Metrics

Base Model Metrics				
	Precision	Recall	f1 Score	Support
1.False	0.95	0.94	0.94	892
2.True	0.94	0.96	0.95	1010
3.Micro avg	0.95	0.95	0.95	1902
4.Macro avg	0.95	0.95	0.95	1902
5.Weighted avg	0.95	0.95	0.95	1902

The figure 4 shows the receiver operating characteristic (ROC) curve that measures the cost of increasing the true positive rate by accepting a higher false-positive rate.

Table 7. Base Model Metrics

Base Model Metrics	
Confusion Matrix	
700 11	192 999
Accuracy	0.8933

Table 8. Meta Label Metrics

Meta Label Metrics				
	Precision	Recall	f1 Score	Support
1.False	0.95	0.96	0.95	892
2.True	0.96	0.95	0.96	1010
3.Micro avg	0.96	0.96	0.96	1902
4.Macro avg	0.96	0.96	0.96	1902
5.Weighted avg	0.96	0.96	0.96	1902

Table 9. Meta Label Metrics

Meta Label Metrics	
Confusion Matrix	
857 47	35 963
Accuracy	0.9569

2.3.4 Analysis

The state-of-the-art literature for finding a machine learning supported trading strategy is written by Marcos López de Prado (De Prado 2018) and (de Prado 2020). The main point of his literature is the importance of feature selection in a trading process instead of back-testing. "Backtesting is not a research tool; feature selection is.". De Prado argues that he was the first-ever discussing in the literature the process for finding the right trading size of the bet by the process of metalabeling. When your model knows the side (long or short) of your bet, your research problem is to find the probability of your bet that helps you to size your bet. To summarize, the financial literature of de Prado is quite mathematical, which is challenging, but they are also practical because they include example codes.

The key event for my literature review study was the webinar by Dr. Chan for the Society of Technical Analysts. He explained the machine learning prediction process of predictnow.ai SaaS that supports trading strategy development. Moreover, Dr. Ernest

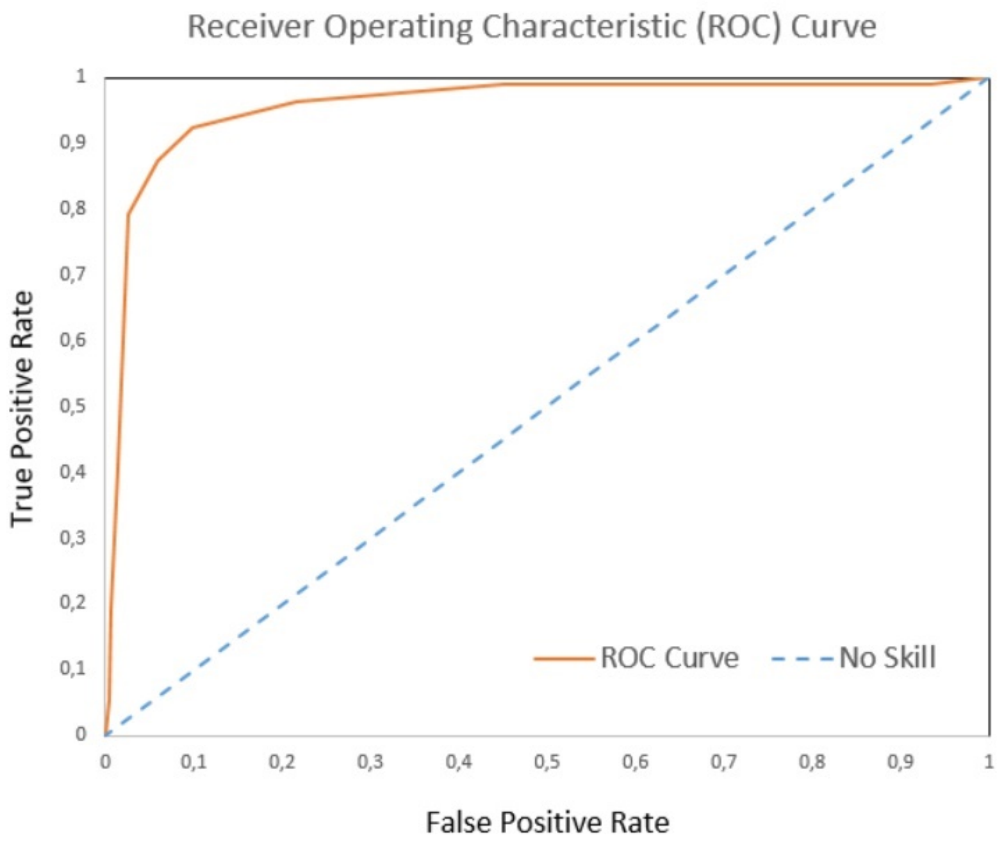

Figure 4. The Receiver Operating Characteristic (ROC) Curve Example.

Chan explained the concept of the conditional probability of your next trade based on de Prados literature. That concept and metalabeling are utilized in his Tail Reaper strategy showed in his Quantitative Trading blog. Both Dr. Chan and Professor de Prado highlight the importance of feature selection. The newest literature (Man & Chan 2020) compares LIME, SHAP, and MDA methods in feature selection. The finding is that LIME and SHAP are performing quite similar. However, Dr. Chan is using the SHAP method in his Tail Reaper strategy process for measuring the feature selection importance. In summary, the literature of Dr. Chan is giving practical tools by having blog writing, strategy process descriptions, and feature selection research paper. Additionally, it was possible to listen to his webinar that was a key moment in the literature review.

2.4 Summary

The key findings are the state-of-the-art process of finding a machine learning-supported trading strategy. The role of ML is that it can improve any trading strategy. Asset managers should focus their effort on building theories, and ML is the critical tool in that

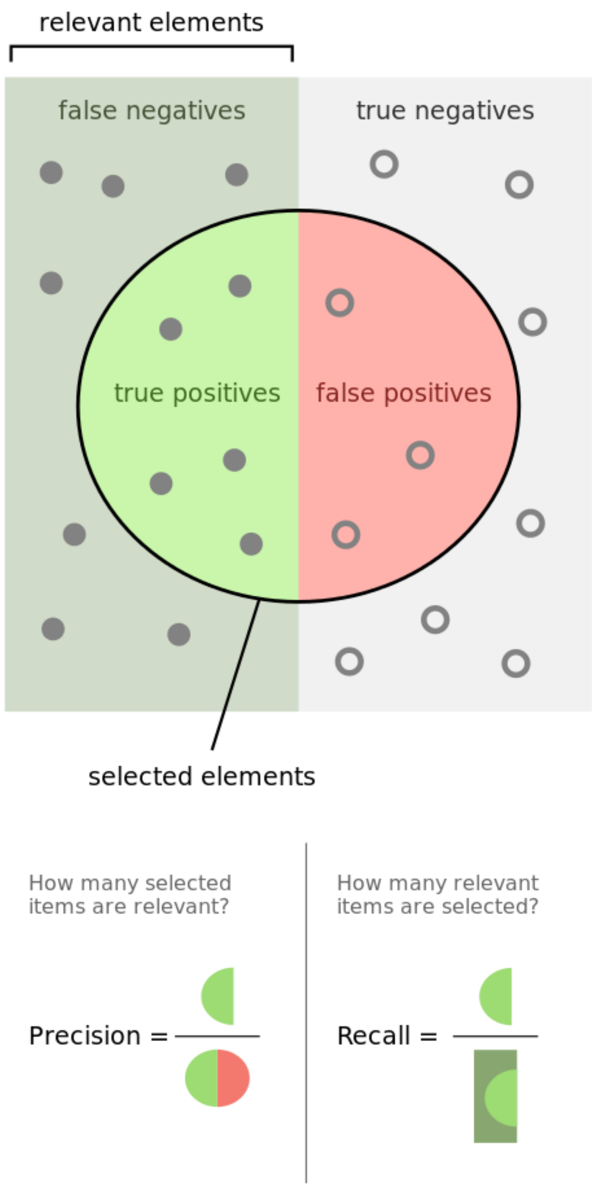

Figure 5. The Precision Recall Table of The Meta Labeling Process - "A Toy Example" Wikipedia contributors (2021d)

work. The state-of-the-art strategy process includes feature selection, metalabeling, predicting if the next trade is profitable or not, and the conditional probability of the next trade. That probability is sizing the bet of the next trade. De Prado claims that nobody had discussed bet sizing in the literature before (De Prado 2018), and he stopped the misery. Bet sizing of the trade is at least as necessary as the side of the trade (De Prado 2018). In summary, ML is the critical tool to build financial theories or portfolio strategy processes. We live in fascinating times, and ML technology will transform how everyone invests for generations - Professor de Prado.

3 METHODOLOGY

This methodology section aims to study the methodology of the research process of finding a machine learning (ML) supported trading strategy. The study presents two methodology processes for finding a trading system.

The research question is finding a trading strategy process to a binary classification problem to predict whether the next trade is profit or not and the probability for that outcome. The predicted probability sizes the bet of the next trade.

3.1 Setting

de Prado (2020) argues that the process to discover new financial theories is following. First, we apply ML tools to uncover hidden features involved in a complex phenomenon. Thus, ML tools identify the ingredients of the theory without informing the exact equation that binds them together. Second, we formulate a theory that connects the ingredients using a structural statement, a system of equations, that hypothesizes a cause-effect mechanism. Third, the theory has a broader range of testable implications predicted by ML tools than in the first step. Therefore, a successful theory will predict events out-of-sample. The theory will explain not only positives (x causes y) but also negatives (the absence of y is due to the absence of x).(de Prado 2020)

"The most insightful use of ML in finance is for discovering theories" de Prado (2020)

Our theory hypothesizes that we combine the parts of the hybrid model by meta-labeling. (meta-labeling implies hybrid model combination, the absence of hybrid model combination is due to the absence of meta-labeling)

The trading system hierarchy includes a theory that can pin down the cause-effect mechanism to extract profits from the market. A trading system is a build implementation of testing the theory. A trading system includes a trading strategy that is a buying and selling method in the market. A trading strategy consists of a trading process that performs buying and selling financial instruments according to the strategy signals. A trading process initiates by trading signals that trigger to buy or sell a financial instrument based on a pre-determined set of criteria.

Table 10. The Setting of the Trading System

The Trading System Hierarchy:	
Process	Description
1.Theory	A theory can pin down the cause-effect mechanism that allows extracting profits from the collective wisdom of the crowd - a testable theory explains factual evidence as well as counterfactual cases (x implies y, and absence of y implies the absence of x). (de Prado 2020)
2.System	A trading system is a practical implementation of testing the theory. A trading system includes a trading strategy with trading processes (acts of buying and selling) that executes according to the trading signals. (a collection of principles or methods according to which something is done; an organized scheme or method.)
3.Strategy	A trading strategy is the method of buying and selling in markets basing on predefined rules used to make trading decisions. (a group of actions designed to achieve a long-term or overall aim.)
4.Process	A trading process includes buying and selling financial instruments according to the strategy. (a series of actions or steps taken to achieve a particular end.)
5.Signal and Sign	Trading signals are triggers to buy or sell a security based on a pre-determined set of criteria. Traders can create trading signals using various criteria, from simple ones, such as earnings reports and volume surges, to more complex signals derived using existing signals. (a gesture, action, or sound used to transfer information or instructions, typically by prearrangement between the parties concerned.)

3.1.1 Tool - Platform SigTech

The process of a strategy in the SigTech platform is based on the queue of methods. These methods are run in order of time and (if times match) priority. Each process can make any decision to manipulate the contents of a strategy. This is the novelty compared to the common back-testing methodology of using large time-series data arrays. The key advantages of the process queue are the following. First, better requirement scaling - in the time-series approach, it becomes increasingly costly when every element (instrument/signal) of a strategy has information for every point of interest. Second, detailed order and transaction modeling - the process approach allows intricate order logic to be correctly modeled. Third, greater flexibility - the process approach allows a modular strategy construction where strategies can inherit characteristics and hold positions in sub-strategies. Since each process can perform any code, it allows greater complexity than a simpler way.[2]

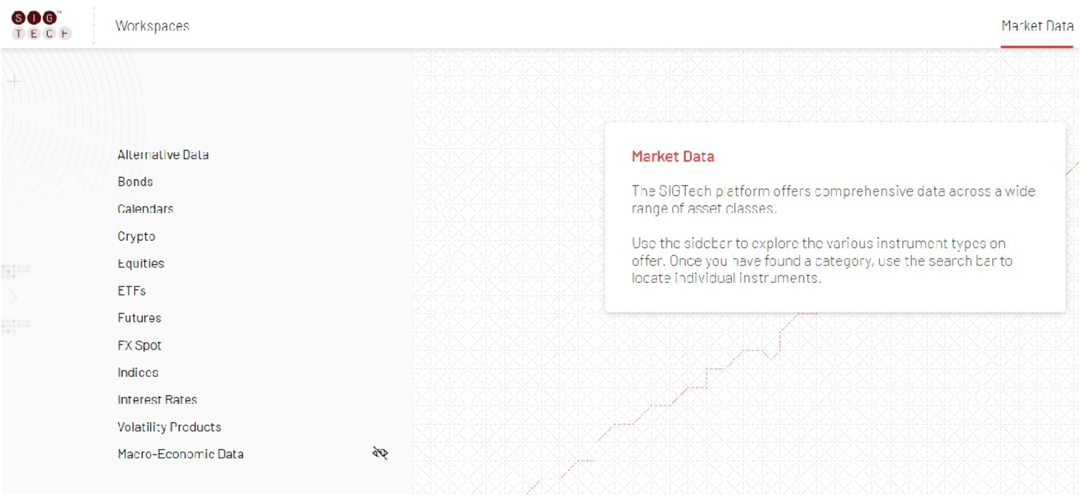

Figure 6. SigTech platform - the data browser view for searching data

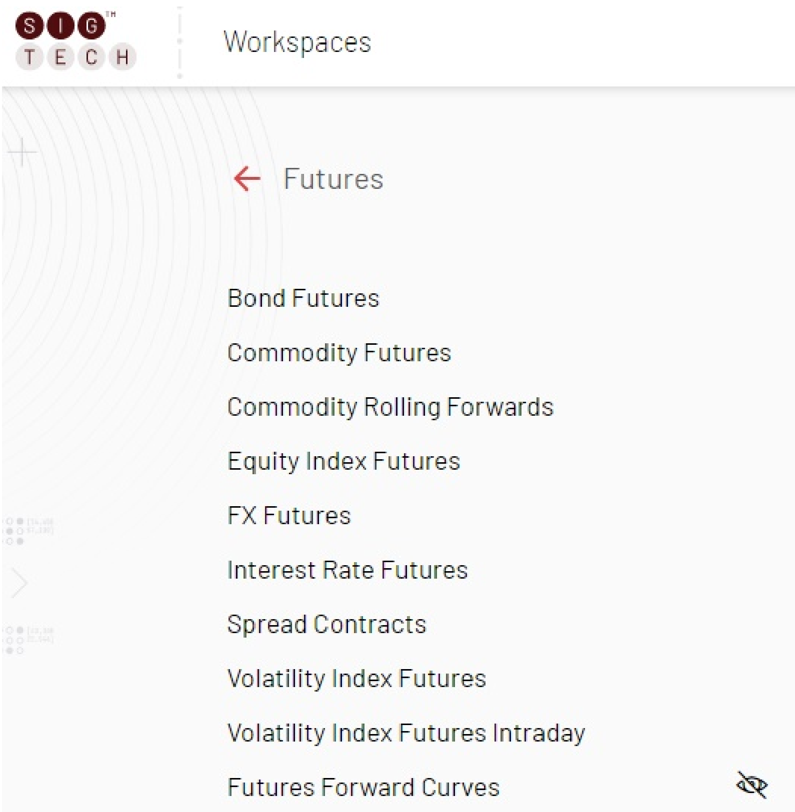

Figure 7. SigTech platform - the list of data types

[2] https://www.sigtech.com/ The section is based on the SigTech strategy framework description.

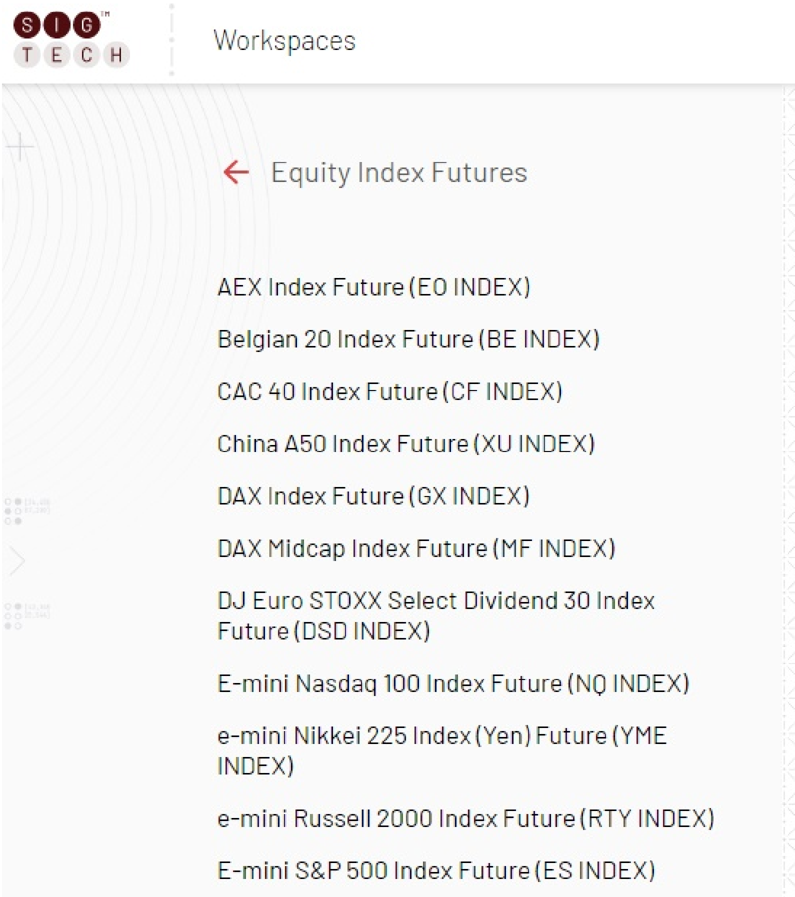

Figure 8. SigTech platform - the data instrument examples of equity index futures

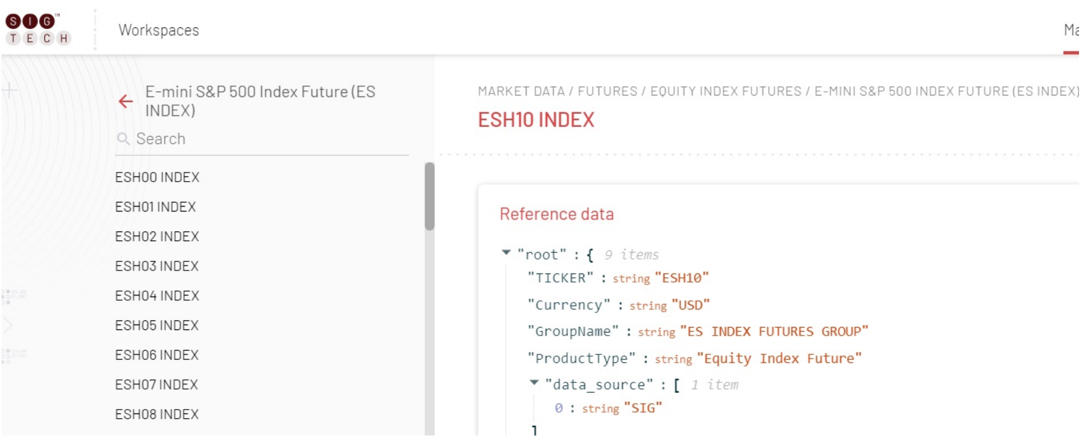

Figure 9. SigTech platform - the list of data instruments and the reference data

3.2 Data and Features

The input feature is a data set of 17 rolling front futures. They include three fixed income futures, six fx currency futures, five commodity futures, and three equity index futures. The feature set is downloaded to a dataframe and calculated 1-day historical returns.

Table 11. The Input Feature Futures

The Input Feature Futures	
Input Feature	Clarify
1. TU	US 2YR NOTE Future (TU COMDTY)
2. TY	US 10YR NOTE Future (TY COMDTY)
3. US	US LONG BOND Future (US COMDTY)
4. BP	British Pound Future (BP CURNCY)
5. EC	British Pound Future (BP CURNCY)
6. CD	British Pound Future (BP CURNCY)
7. JY	Japanese Yen Future (JY CURNCY)
8. AD	Australian Dollar Future (AD CURNCY)
9. NV	New Zealand Dollar Future (NV CURNCY)
10. CT	Cotton #2 Future (CT COMDTY)
11. LC	Live Cattle Future (LC COMDTY)
12. BO	Soybean Oil Future (BO COMDTY)
13. QS	Gas Oil Future (QS COMDTY)
14. HO	NY Harbor ULSD Future (HO COMDTY) Heating Oil
15. ES	E-mini S&P 500 Index Future (ES INDEX)
16. NQ	E-mini Nasdaq 100 Index Future (NQ INDEX)
17. NX	Nikkei 225 Index Future (NX INDEX)

In addition, there are 14 technical analysis models included in to set of features. The target label is the rolling E-mini SP500 future (ES). The data period is from 2000-01-04 to 2021-04-21, including a total of 5557 close price data points.

Table 12. The Descriptive Statistics of Equity Index Futures Features

The Descriptive Statistics of Equity Index Futures Features			
Measure	ES	NQ	NX
1. count	5557.000000	5557.000000	5557.000000
2. mean	1461.191635	1083.138077	930.940285
3. std	773.164419	864.525605	317.293349
4. min	545.729547	237.907109	468.882487
5. 25%	903.015262	476.612086	652.810856
6. 50%	1106.140790	698.863374	903.815800
7. 75%	1883.613873	1380.553533	1158.853535
8. max	4115.285679	4474.308308	1992.329449

The input feature table includes input data in the SigTech platform of 48 features (17 futures, 14 technical analysis features, 17 1-day returns) and a target feature of E-mini SP500 rolling front future from 2000-01-04 to 2021-04-21 including totally 5557 close prices.

Table 13. The Input Feature Table

The Input Feature Table		
Input Feature	1-Day Return of Feature	Technical Analysis Indicator with default value
1. TU	TU_1D	Momentum (default 10)
2. TY	TY_1D	ROC Rate Of Change (default 14)
3. US	US_1D	Disparity (default 5)
4. BP	BP_1D	Disparity (default 10)
5. EC	EC_1D	Price Oscillator
6. CD	CD_1D	RSI Relative Strength Index (default 14)
7. JY	JY_1D	Moving Average (default 5)
8. AD	AD_1D	BIAS (default 6)
9. NV	NV_1D	PSY (default 12)
10. CT	CT_1D	ASY (default 5)
11. LC	LC_1D	ASY (default 4)
12. BO	BO_1D	ASY (default 3)
13. QS	QS_1D	ASY (default 2)
14. HO	HO_1D	ASY (default 1)
15. ES	ES_1D	
16. NQ	NQ_1D	
17. NX	NX_1D	

The figure 10 shows the relative development 17 rolling input feature futures during the time 2000-01-04 to 2021-04-21.

The figure 11 shows the development of the E-mini SP500 rolling front future target variable during the time 2000-01-04 to 2021-04-21.

The figure 12 shows the target variable E-mini SP500 rolling front future 2-days, 20-days and 60-days volatilities from 2000-01-04 to 2021-04-21 including totally 5557 close prices.

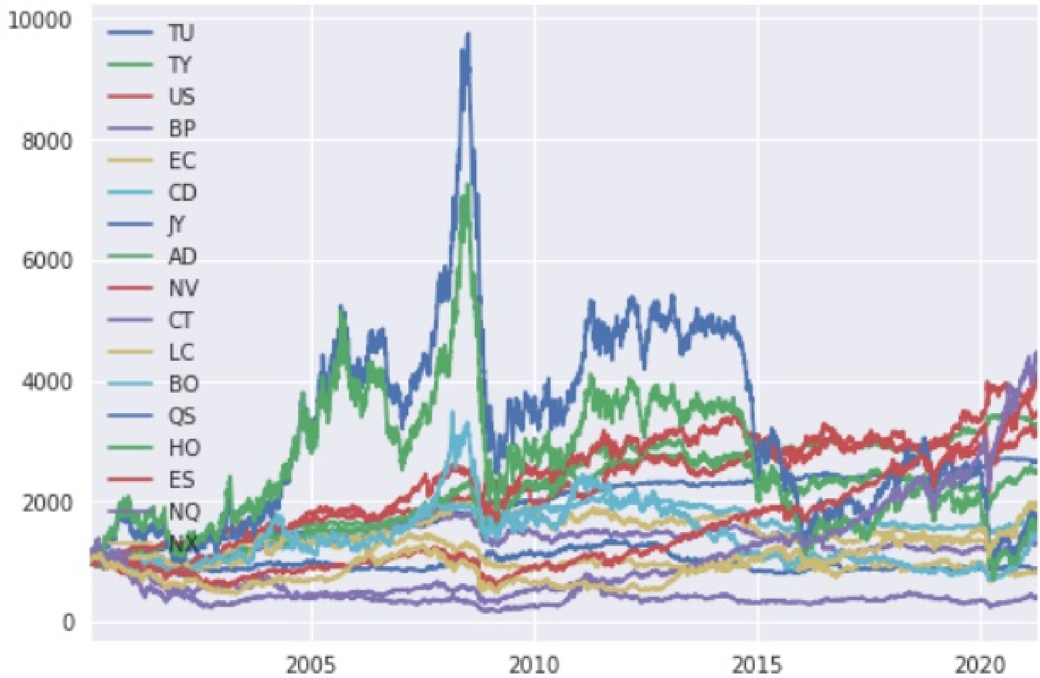

Figure 10. SigTech platform - Data feature set 17 futures rolling front from 2000-01-04 to 2021-04-21 including totally 5557 close prices.

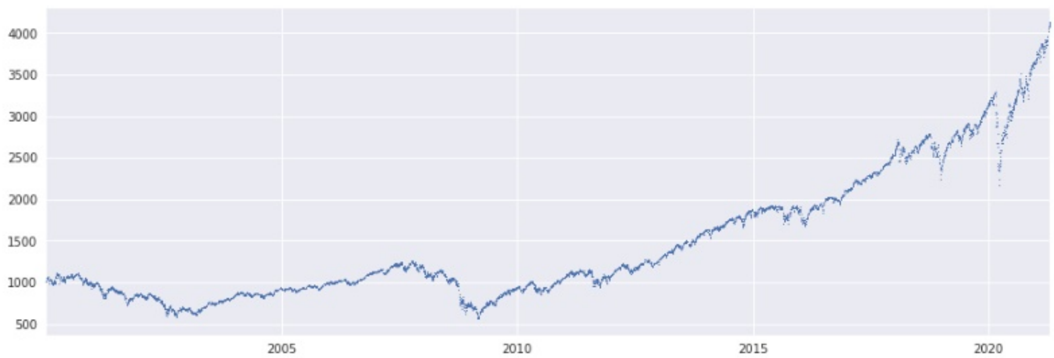

Figure 11. SigTech platform - E-mini SP500 rolling front future from 2000-01-04 to 2021-04-21 including totally 5557 close prices.

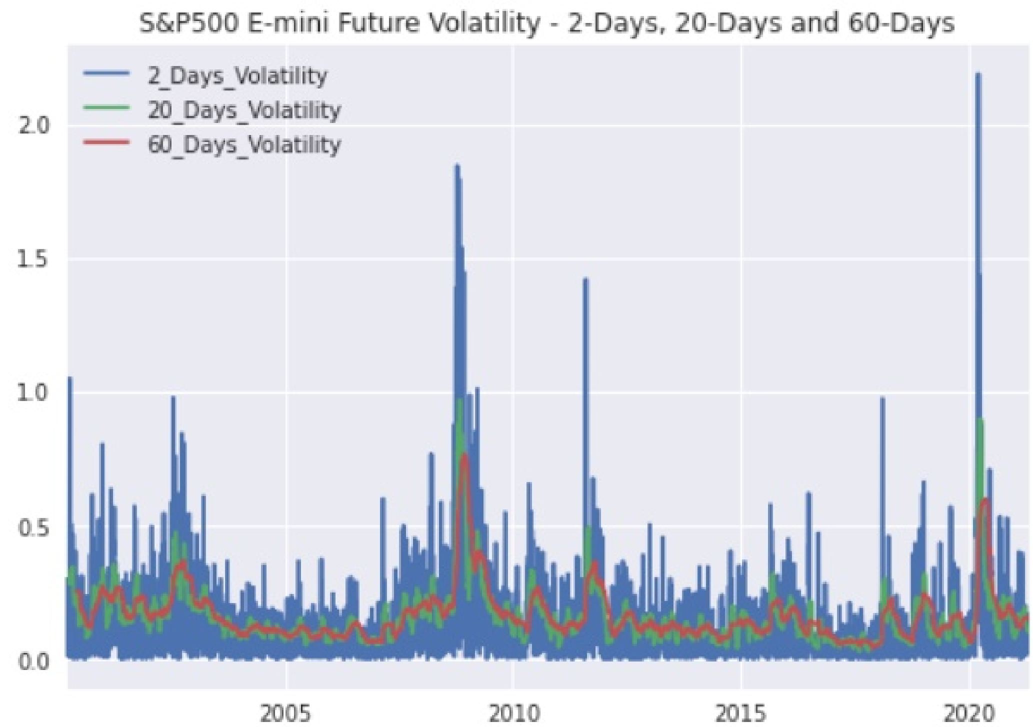

Figure 12. E-mini SP500 rolling front future Volatilities from 2000-01-04 to 2021-04-21 including totally 5557 close prices.

3.3 Model

3.3.1 Strategy Process 1 - Tail Reaper Trading Strategy

The first process example is predicting the conditional probability of profit for the next trade with the Tail Reaper strategy. For the model random forest, the input x features, including all the possible financial instruments in the market with the historical return data. The input y label includes the "metalabeled" market return (SP 500 index) profit true (1) or otherwise false (0). The random forest model outcome is improved by utilizing the SHAP algorithm for feature selection and reducing noise, which the Kalman filter could use. After train and prediction with random forest, the model will tell if the trade is profitable and with what probability—the probability of sizing the bet of the trade. Finally, one can connect the system to the market via API and run the model with live predictions trading with SP500 mini-future contract. The section bases on the blog of Dr. Chan on August 06, 2020. (Chan 2020a)

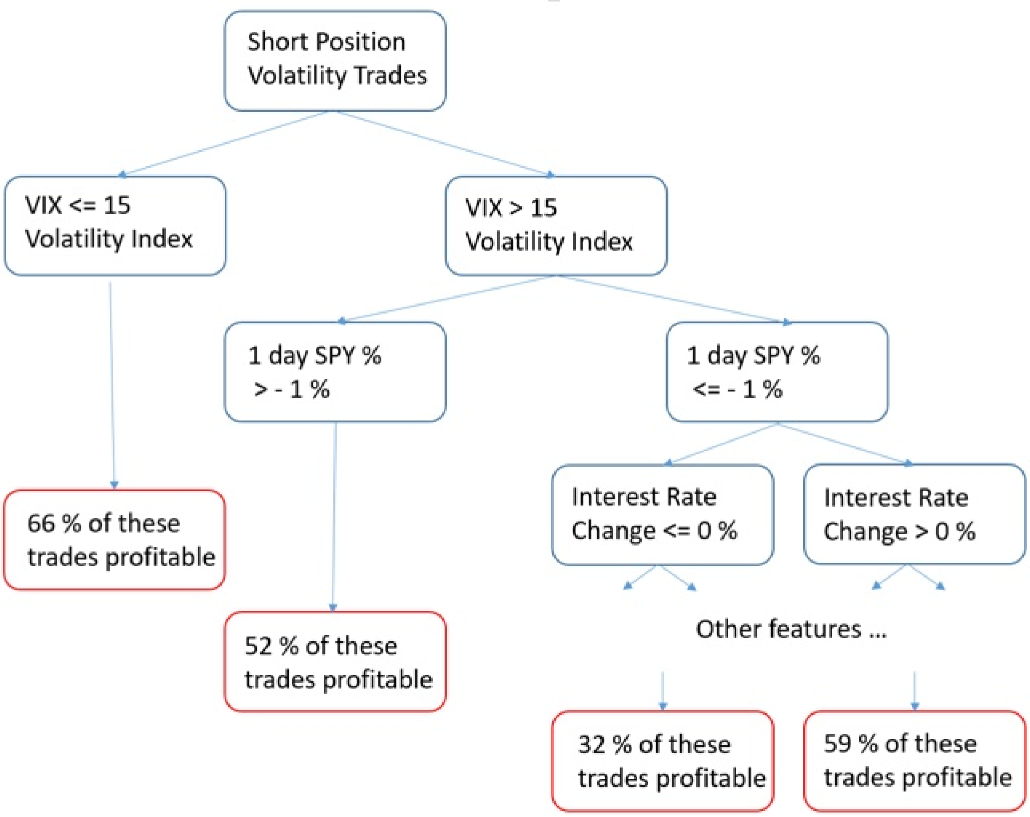

Figure 13. Classification tree of the predictnow.ai

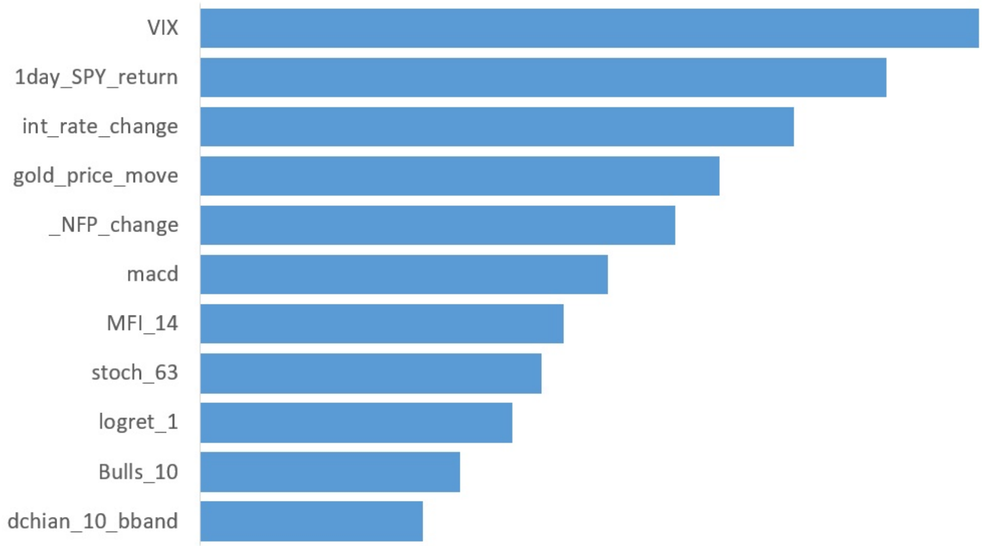

Figure 14. Feature selection of Tail Reaper strategy

3.3.2 Strategy Process 2 - MNIST Example

The second strategy process description bases on the blog of Hudson and Thames, "A Toy Example" Thames (2020) investigating Meta Labeling (chap. 3 De Prado 2018 p. 50) that improves model and strategy performance metrics by filtering out false positives. The blog shows the process and the metrics by utilizing MNIST handwritten digit classification predicting the side of the bet (long or short) needed to learn the size of that bet, including no bet at all (zero sizes). That problem is meta labeling because we want to build a secondary ML model that learns to use a primary exogenous model. The primary ML model train for binary prediction. While the label is true (1), the probability of the secondary model prediction sizes the bet, The primary model sets the side (sign) of the position. The binary classification problem is a trade-off between false positives and false negatives. Increasing the true positive rate tends to increase its false positives, the ROC curve measures that cost. Meta labeling increases the F1 score by filtering out the false positives. The secondary model determines whether the positive of the primary (exogenous) model is true (1) or false (0), deciding to act or pass the opportunity presented. The model process of the MNIST example follows. First, a primary model (binary classification) trains with high recall (split 90/10 train/test) to see over-fitting. Second, determining the threshold level for high recall of the primary model by utilizing the ROC curve, meaning the primary model captures a majority of positive samples even if there is a large number of false positives. The meta-model corrects it by reducing false positives, causing improving metrics. Third, the features from the primary model concatenate with the predictions from the primary model into a new set of features for the secondary model targeting meta-labels in the secondary model. Fit the secondary model. Meta labels are true (1) if the primary model prediction matches the actual value, else false (0). If an observation is a true positive or true negative, label it as 1 (the model is true), else 0 (the model is false). Because it is categorical, this example adds One Hot Encoding. Fourth, combining the prediction from the secondary and primary models. Only when both show true, the final prediction is true. If the primary model predicts a three (3) and the secondary model provides a high probability of the primary model resulting correct, is the final prediction a 3, else not a 3. The section is basing on the Hudson and Thames blog "A Toy Example." The process structure describes in the figure. (Thames 2020)

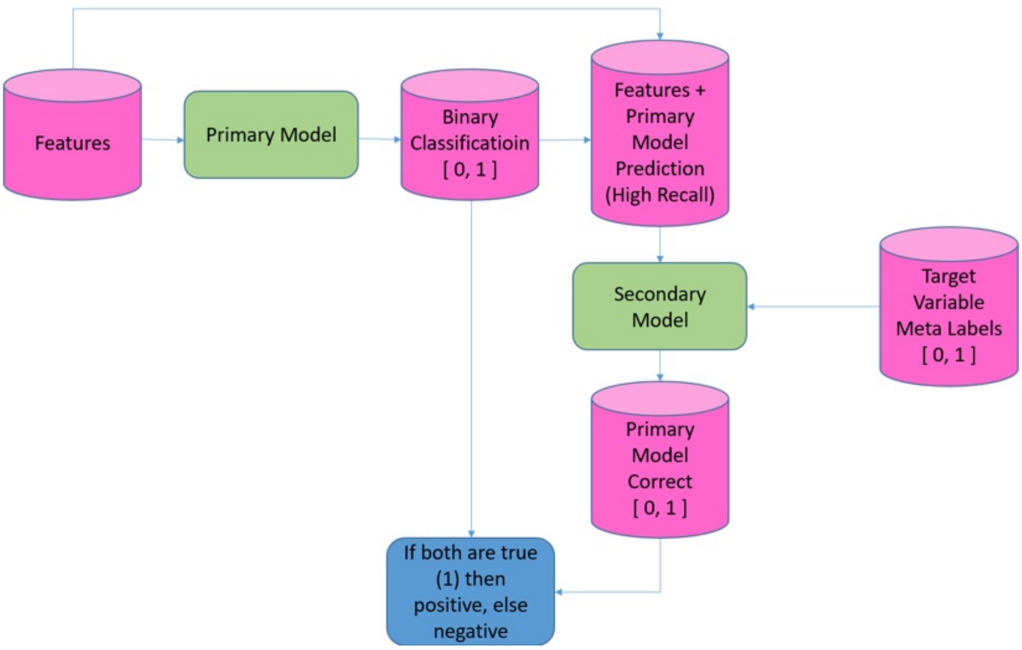

Figure 15. The process structure of MNIST example

3.3.3 The Coded Trading System Based on The Tail-Reaper and MNIST Processes

The following trading strategy system was designed for coding at the SigTech platform based on the Tail Reaper and the MNIST examples.

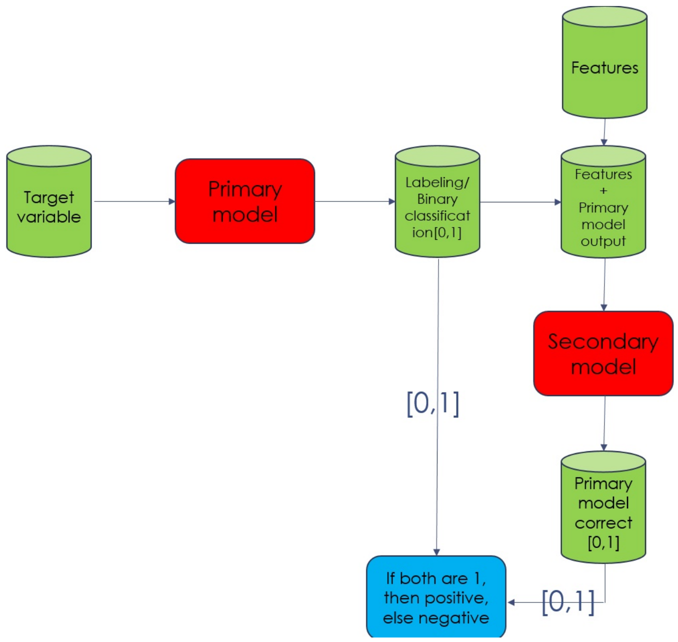

Figure 16. The Trading Strategy System.

The trading strategy system is following. The primary model is inputting a target variable and the output labels by binary classification of profitable (1) and non-profitable (0). The combination of the features and labeled output of the primary model are the inputs to the secondary model. The secondary machine learning model predicts the labeled output of the primary model if the output is correct (1) or not (0). The AND logic gate method validates the output of the primary model and the prediction of the secondary model. If and only if both models are resulting in a valid (1), the trade is executed, else not.

The trading strategy system table shows the faces and the description of the coded trading strategy system structure.

The structure of the trading strategy system is following. The set of features (48) include the rolling front future (17) raw data, one-day historical returns of each futures contract (17), and the technical analysis indicators (14) using the close price of the target variable (S&P500 E-mini front future) inputting the primary model. The primary model input

Table 14. The Trading Strategy System Structure

The Trading Strategy System Structure	
Face	Description
1.	**Features (48):** 17 rolling front futures + 1-day historical returns + 14 technical analysis indicators
2.	**Primary / Base Model (technical analysis model):** INPUT: S&P500 E-mini front future (target feature, 5557 close days, 2000-01-04 to 2021-04-21) MODEL: Donchian Channel (channel 50, exit 10) OUTPUT: daily percent returns (positive and negative)
3.	**Labeling / Binary Classification [0,1]:** Binary classification of Donchian Channel output (1=profit, 0=else) (Metalabels[0,1])
4.	**Secondary Model (machine learning prediction):** INPUT: features + labeled output of the primary model (= target variable: Metalabels[0,1]) MODEL: Random Forest OUTPUT: Prediction (profit=1 or not=0) and probability e.g. (0.35, 0.65) of tomorrow's close price
5.	**Rule** => If both Primary and Secondary model perform true (1): Then positive (do the trade), else negative (no trade)

or target feature has 5557 closing price days from 2000-01-04 to 2021-04-21. The primary model is a Donchian channel (DC) technical analysis model optimized for 50 days channel with ten days mean exit outputting daily percent returns. The primary model parameter optimizing maximizes the cumulative daily percent returns. The primary DC model output labels by binary classification to meta-labels profitable (1) and else (0). The secondary model basing random forest machine learning algorithm inputs features as x and meta-labeled output of the primary model as label y target variable. The secondary model predicts profit (1) or not (0) and the probability of tomorrow's close price. The final validation is the rule that both models must agree valid (1) to execute the trade, else no-trade entering.

3.4 Measurement Instruments

In finance, a *return* defines a profit on an investment. A return covers any change in the value of the investment and/or cash flows (or other investments) investor receiving from investment (e.g., interest payments, coupons, cash dividends). Moreover, a return measures either in absolute terms (e.g., dollars) or as a percentage of the amount invested, called the holding period return. A loss contrary to a profit describes as a negative return, assuming the amount invested is greater than zero. (Wikipedia contributors 2021e)

The formula of return.
Calculating the return or the holding period return R over a single period of any length of time is:

$$R = \frac{V_f - V_i}{V_i} \tag{1}$$

where:
V_f = final value, including dividends and interest
V_i = initial value

Compound annual growth rate (CAGR) refers to a business and investment-specific term for the geometric progression ratio producing a constant rate of return over a period. Anson et al. (2010) Moreover, the CAGR is the rate of return requiring an investment to grow from its beginning value (BV) to its ending value (EV), assuming reinvesting the profits at the end of each year during the investment's lifespan. Fernando (2021)

An advantage of the CAGR is that it is an accurate way to calculate and determine returns for anything that can increase or decrease in value over time. Moreover, investors can compare the CAGR of two alternative investments to assess how the asset performed against other assets in a peer group or against a market index. However, the CAGR is ignoring the investment risk. (Fernando 2021)

The formula of CAGR.

$$CAGR = \frac{(EV)^{\frac{1}{n}}}{(BV)} - 1 \tag{2}$$

where:
EV = Ending value
BV = Beginning value
n = Number of years

The Sharpe ratio (also reward-to-variability ratio) measures the financial performance of an asset investment compared to a risk-free asset after adjusting for its risk. The definition of the Sharpe ratio is the difference between the returns of the investment, and the risk-free return (risk-free interest rate), divided by the standard deviation of the investment (i.e., asset volatility). Thus, it refers to the additional return that an investor receives per unit of increase in risk. The developer was William F. Sharpe, 1966. (Sharpe 1966)

The formula of Sharpe Ratio.

$$S = \frac{(R_p - R_f)}{\sigma_p} \tag{3}$$

R_p = the daily return of the trading strategy

R_f = the risk-free rate, being an average of about 2% over the 2015 – 2021 period, (10 Year Treasury Rate 1.63% for March 23 2021, high 3.24% and low 0.52%). The assumption is that the risk-free rate is zero because the return is very small when leaving capital on the savings account.

σ_p = the standard deviation of returns of the trading strategy

3.5 Procedure

3.5.1 Technical Analysis Indicators as Features

The input features for the secondary model include technical analysis indicators,where C_t is the closing price of the target variable.

MA_n is the moving average of the closing price value in the last n days. Up_t is the upward price change of the target variable closing price at time t and Dw_t is the downward price change of the target variable at time t. PSY_n is the ratio of the number of rising periods over the n day period. Variable A is number of rising days in the last n days. SY_t represents the return of the target variable at time t, $SY_t = (lnC_t - lnC_{t-1})x100$. ASY_n is the average return in the last n days.

3.5.2 Donchian Channel Technical Analysis Model

Donchian channels (DC), a trend-following strategy, was developed by Richard Donchian in 1936. Donchian channel technical analysis model selected to be the primary model in the trading strategy process because it is a trend-following strategy. Trend following

Table 15. The Input Features - Technical Analysis Indicators

The Input Features - Technical Analysis Indicators	
Name of Feature	Formula
Momentum	$C_t - C_{t-10}$
ROC (Rate of Change)	$C_t / C_{(t-n)} x100$
Disparity in 5 days	$C_t / MA_5 x100$
Disparity in 10 days	$C_t / MA_{10} x100$
OSCP (price oscillator)	$MA_5 - MA_{10} / MA_5$
RSI (relative strength index)	$100 - 100/(1 + \frac{\sum_{i=0}^{n-1} xUp_{t-i}}{n} / \frac{\sum_{i=0}^{n-1} xDw_{t-i}}{n})$
MA_5 (Moving Average)	$(\sum_{i=1}^{5} xC_{t-i+1})/5$
$BIAS_6$	$\frac{C_t - MA_6}{MA_6} x100\%$
PSY_{12}	$PSY_{12} = (A/12)x100\%$
ASY_5	$(\sum_{i=1}^{5} xSY_{t-i+1})/5$
ASY_4	$(\sum_{i=1}^{4} xSY_{t-i+1})/4$
ASY_3	$(\sum_{i=1}^{3} xSY_{t-i+1})/3$
ASY_2	$(\sum_{i=1}^{2} xSY_{t-i+1})/2$
ASY_1	$ASY_1 = SY_{t-1}$

strategy means that we buy the asset when the price trend goes up and sell the asset when the price trend goes down according to the strategy signals.

A trend-following strategy aims to generate profit by finding, joining, and riding the trend. In technical analysis, trends recognize by price actions performing higher highs and higher lows in an uptrend price movement and lower lows and lower highs in a downtrend move. Whereas, in the stable price action, the trend is sideways. However, the profit generation in the market usually needs price movement (excluding option strategies) because the risk-return tradeoff states that the potential return rises with an increase in risk.

Risk refers to an uncertainty of the potential price action of decreasing or increasing. Volatility is a measure of the dispersion of the returns for given security. However, risk or volatility refers to price movement. The higher is the price movement (the longer the trend), the higher is the potential return. In other words, the security prices must move (like in trends) to generate a potential return. Donchian Channel strategy generates trading signals to join the trends upwards or downwards. Some advantages of the Donchian Channel strategy are that it can join and ride along with trends. Moreover, it is usable for any liquid security and any period. A disadvantage of the Donchian Channel strategy is that it may generate losses with false breakouts while searching for a price signal to trigger an entry trade to join a long trend for given security.

The Donchian Channel strategy is a more straightforward implementation of the turtle trading strategy that was the original idea. The turtle trading strategy might be a target for future study.

Donchian channel trend-following strategy modeled after the turtle trading strategy has been utilized on commodities and equity index futures contracts in the literature. Swart (2016)

Donchian's 20 and 40 days breakout rule Kaufman (2019) is following:

Buy when today's high > high of the past 40 days
Sell when today's low < low of the past 40 days
Exit longs when today's low < low of the past 20 days
Exit shorts when todays' high > high of the past 20 days

The Donchian Channel strategy, in this study, is optimized for the SP500 E-mini future. The optimized channel length is 50 days, and the exit length is ten days. Moreover, the optimized channel parameters maximized the cumulative daily returns of the strategy.

The optimized and applied Donchian's breakout rule used in this study is following:

Buy when today's close > close of the past 50 days
Sell when today's close < close of the past 50 days
Exit longs when today's close < mean close of the past ten days
Exit shorts when today's close > mean close of the past ten days.

Donchian Channel Trading Strategy Signals Explained.

A Breakout upper channel (UC) 50 days high, at price 1000 - long position opened.
B Hit 10 days average price, at price 1125 - long position closed, (A-B profit).
C Breakout lower channel (LC) 50 days low, at price 1110 - short position opened.

Figure 17. The Donchian Channel Trading Strategy Signals

D Hit 10 days average price, at price 1120 - short position closed, (C-D loss).

E Breakout upper channel (UC 50 days high, at price 1150 - long position opened.

Table 16. The structure of the Donchan channel trading strategy coding.

The structure of the Donchan channel trading strategy coding		
Nr	Element	Description
1.	Entry signal open long position open short position	Open long if close > the highest close during the 50-days channel, or open short position if close < the lowest close during the 50-days channel.
2.	Exit signal close long position close short position	Close long position when the close < the 10-days mean price, or close short position when the close > the 10-days mean price
3.	No position	After exit and before entry signals.
4.	Resulting in the position side after the entry and before the exit signal.	Position side: long (1), short (-1), and no position (0)
5.	Calculate daily percent returns of the underlying.	One-day percent return e.g. +5% or -5%
6.	Multiply the position side with the daily percent return.	Long-profit (1x+5%), short-profit (-1x-5%), non-profit (0x+/-5%)
7.	Stop-loss is the exit signal	Stop-loss depends on the trading strategy

The coding structure of the Donchian channel trading strategy has an entry signal; open a long position when the close is higher than the highest close during the 50-days channel, open short position when the close is lower than the lowest close during the 50-days channel. Furthermore, an exit signal; close a long position when the close is lower than the 10-days mean price, close short position when the close is higher than the 10-days mean price, and when to stay out of the market; no position between the exit and entry signals.

Table 17. The Trading Strategy Logic for Coding.

The Trading Strategy Logic for Coding						
Nr	Position	Position side	Daily percent return of the underlying	(Position side) x (daily percent return of the underlying)	Daily percent return of the trading strategy	Cumulative return of the trading strategy
			Underlying	Underlying	Strategy	Strategy
1.	Long entry					
2.	Day 1	1	+5%	1x+5%	0,05	0,05
3.	Day 2	1	-2%	1x-2%	-0,02	0,03
4.	Day 3	1	+4%	1x+4%	0,04	0,07
5.	Day 4	1	-1%	1x-1%	-0,01	**0,06**
6.	Long exit					
7.	Short entry					
8.	Day 5	-1	-5%	-1x-5%	0,05	0,05
9.	Day 6	-1	-3%	-1x-3%	0,03	0,08
10.	Day 7	-1	+2%	-1x+2%	-0,02	0,06
11.	Day 8	-1	+3%	-1x+3%	-0,03	**0,03**
12.	Short exit					
13.	No position					
14.	Day 9	0	-5%	0x-5%	0	0
15.	Day 10	0	-3%	0x-3%	0	0
16.	Day 11	0	+2%	0x+2%	0	0
17.	Day 12	0	+3%	0x+3%	0	**0**
The total cumulative return of the trading strategy: long 0,06 + short 0,03 + no pos 0 = **9%**						

The position side means a trading position that can be long, short, or no position. A long trading position occurs when buying the financial instrument expecting increasing in value and selling it later at a higher value - a bullish view. A short trading position opens when selling the asset expecting decreasing in value and covering it later at a lower value - a bearish view. No position (or neutral position) means staying out of the market and taking no trading position.

The programmer specifies the conditions to be evaluated or tested by the program. These conditions lead to a set of states: buy=1, sell=-1, and neutral=0. The position side is between the signals defined, long (1), short (-1), and no position (0). The daily percent return of the underlying S&P500 E-mini futures contract multiplies with the position side, resulting in the trading performance of the strategy. The daily percent returns are cumulatively showing the total percentage performance of the trading strategy at the end of the trading period.

3.5.3 Meta-Labeling

Meta-labeling could belong to the ensemble machine learning method because meta-labeling shares some similarities with stacking while being fundamentally different. Gui (2020) Ensemble method uses multiple single models to construct a hybrid model. A hybrid model generally performs better compared to a single model. Fan et al. (2018) Stacking is an ensemble machine learning algorithm that involves how to best combine the predictions from two or more primary machine learning algorithms. Brownlee (2020b) Whereas meta-labeling utilizes two layers of models with different purposes. (Gui 2020)

de Prado (2020) discusses that the goal of meta-labeling is to train a secondary model on the prediction outcomes of a primary model, where losses are labeled as "0" and gains are labeled as "1". Therefore, the secondary model does not predict the side. Instead, the secondary model predicts whether the primary model succeeds or fails at a particular prediction (a meta-prediction). The probability associated with a "1" prediction can then be used to size the position. (de Prado 2020)

The process of stacking differs from the process of meta-labeling. (Gui 2020)

1.Build the primary base model, get the prediction.
2.Filter the prediction with threshold.
3.Combine the prediction with input as new input.
4.Build a secondary model and train it with the new input.
5.Predict with the secondary model.

Generally, the process of meta-labeling is following. (Gui 2020)

1.Build primary base model, get a prediction.
2.Filter the prediction with threshold.
3.Combine the prediction with x_train as new input.
4.Combine the prediction with y_train as a new label.
5.Build a secondary model and train it with new input and label.
6.Predict with the secondary model.
7.The final prediction results from the intersection of base model prediction and the meta-model prediction.

Gui (2020) argues that meta-labeling differentiates from ensemble method, stacking, by adding the primary model prediction both to the feature set and label, whereas stacking is adding primary model prediction only as a new feature to the secondary model. (Gui 2020)

3.5.4 The Validation of Hybrid Model by Logic Gate

A logic gate model is used for validation of the parts of the hybrid model. A logic gate is a model of computation implementing a Boolean function, a logical operation performed on one or more inputs that produce a single binary output. The binary number system was refined by Gottfried Wilhelm Leibniz (1705), and it combined the principles of arithmetic and logic. Moreover, it was influenced by I Ching's binary system.

Table 18. The Logic Gate Truth Table

The Logic AND Gate Truth Table		
Input	Input	Output
A	B	A AND B
0	0	0
0	1	0
1	0	0
1	1	1

The validation of the hybrid model uses the primary digital logic gate, called the AND gate that implements logical conjunction according to a truth table. The output true (1) results, if and only if, all the inputs of AND gate are valid (1).

3.6 Data Analysis

The input data set is from the original database of the SigTech platform. The data period is over 20 years (5557 closing prices) of 17 rolling front futures. The four financial futures contract categories are the fixed income, FX currency, commodities, and equity index future. The SigTech platform provides the data set ready to use for the trading strategy system without cleaning or modifying. The figure of the target variable shows that the general trend was increasing over time. The volatility figure shows high volatility during the financial crisis and pandemia crisis. To summarize, the data set has an extended time-series period, the financial futures are from different categories, and there is variability in the volatility during the examination period.

3.7 Conclusions

Two presented processes, the Tail Reaper strategy and the MNIST binary classifier process, were the tools designing the coding system of the SigTech platform. The models were studied, and the data was analyzed. Concluding the methodology, the primary model's technical analysis strategy could be any trading strategy without changing the whole structure, the feature set could be more extensive with a variety of features, including volatility, and different measurement metrics may improve or widen the understanding of the trading strategy performance. For example, the examination of the Sharpe ratio as optimizing metrics in the Donchian Channel strategy instead of maximizing the cumulative daily returns is one way to enhance more understanding of the trading system. Moreover, the secondary model Random Forest can substitute by and test with some other machine learning classification algorithm, like the multinomial Naive Bayes classifier, that performs probability prediction. However, the current data, system, and methodology framework are suitable for the study's research question, scope, and resources.

Table 19. The Trading Strategy System Structure

The Trading Strategy System Structure	
Face	Description
1.	**Features (48):** 17 rolling front futures + 1-day historical returns + 14 technical analysis indicators
2.	**Primary / Base Model (technical analysis model):** INPUT: S&P500 E-mini front future (target feature, 5557 close days, 2000-01-04 to 2021-04-21) MODEL: Donchian Channel (channel 50, exit 10) OUTPUT: daily percent returns (positive and negative)
3.	**Labeling / Binary Classification [0,1]:** Binary classification of Donchian Channel output (1=profit, 0=else) (Metalabels[0,1])
4.	**Secondary Model (machine learning prediction):** INPUT: features + labeled output of the primary model (= target variable: Metalabels[0,1]) MODEL: Random Forest OUTPUT: Prediction (profit=1 or not=0) and probability e.g. (0.35, 0.65) of tomorrow's close price
5.	**Rule** => If both Primary and Secondary model perform true (1): Then positive (do the trade), else negative (no trade)

4 ALGORITMIC DESCRIPTIONS

4.1 Random Forest Classifier

A random forest (RF) classifier or random decision forests is an ensemble classifier that produces multiple decision trees. Belgiu & Drăguţ (2016) The RF is an ensemble learning method for classification, regression, and other tasks. In machine learning, ensemble methods use multiple learning algorithms to obtain better predictive performance than any of those algorithms alone. Kam Ho (1995) developed the first algorithm for random decision forests. A random forest (RF) classifier is an ensemble classifier that produces multiple decision trees using a randomly selected subset of training samples and variables. Belgiu & Drăguţ (2016) A random forest classifier uses averaging to improve the predictive accuracy and control over-fitting. Breiman (2001)

Random forests are ensembles of decision trees. Denisko & Hoffman (2018) Morgan & Sonquist (1963) introduced the decision tree methodology, simplifying the analysis of multiple features during prediction. Our study utilizes the decision tree on time series data as a collection of samples, each described by features (A). The decision tree classifies the samples of each feature through a forking path of decision points (B). Each decision point has a rule determining the valid branch. During the flow of decisions, we stop at each decision point to apply its rule to one of the sample's features until the end of the branch or leaf. The leaf has a class label that assigns the sample to that class (E), concluding the path through the decision tree. (Denisko & Hoffman 2018)

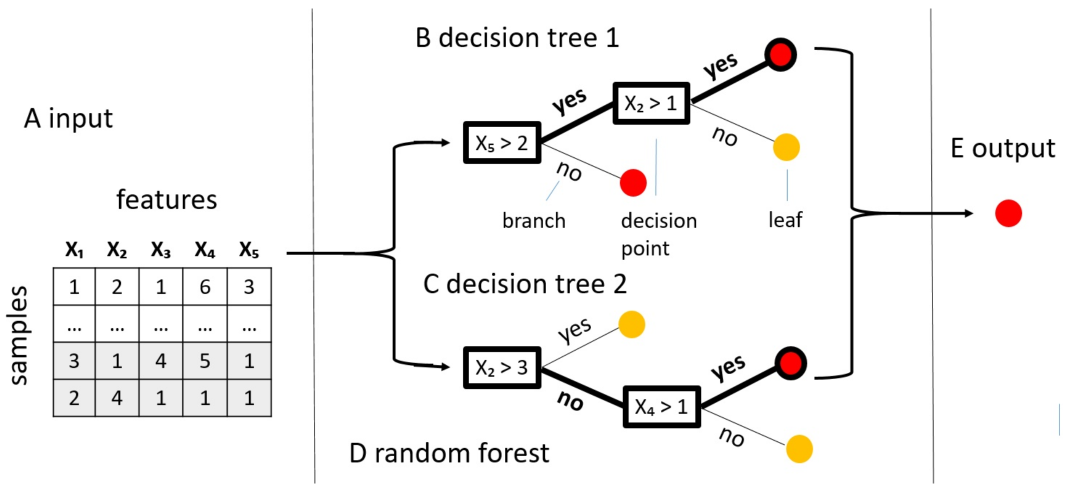

Figure 18. The Random Forest Decision Tree

4.2 Feature Selection

The random forest can perform feature selection, which attempts to improve the transparency and interpretability of machine learning models by ranking the importance of the features used. Man & Chan (2021b) Breiman (2001) explains that first, to measure the feature importance in a data set, fit a random forest to the data. Second, during the fitting, the out-of-bag error (out-of-bag estimate, bagging, measures the prediction error of random forest) for each data point in the data set is recorded and averaged over the forest. Third, to measure the importance of the j-th feature after training, the values of the j-th feature are permuted among the training data, and the out-of-bag error again estimates on this perturbed data set. Fourth, the importance score for the j-th feature calculates by averaging the difference in out-of-bag error both before and after the permutation overall decision trees. Fifth, the score is normalized by the standard deviation of these differences. Finally, the features with large values rank more important than those with small values. (Breiman 2001)

The statistical definition of feature importance measure analyzed by Zhu et al. (2015) A disadvantage of the method in determining feature importance is related to data including categorical features with a different number of levels, random forests is biased in favor of the features with more levels. That problem can be solved by partial permutations (Deng et al. 2011) and growing unbiased trees Strobl et al. (2007). Moreover, if the data includes clusters of correlated features of indistinguishable significance, the smaller clusters are prioritized over, the larger clusters. Toloşi & Lengauer (2011) de Prado (2020) pioneered a cluster-based technique for feature selection called cMDA. The cMDA technique performed superior stability and interpretability results to MDA-selected features when investigated by Man & Chan (2021b)

5 EXPERIMENTS

In the experiments section the study apply the data to the methodology analyzing the trading strategy system performance.

5.1 Primary Model

The primary model input data is S&P500 E-mini front future, including 5557 closing price data from 2000-01-04 to 2021-04-21. The primary model is the technical analysis Donchian Channel trading strategy optimized to the channel of 50 days with the exit of 10 days mean targeting to maximize the cumulative daily returns. The output of the primary model is daily percent returns.

Table 20. The Primary Model - Donchian Channel output, last ten days.

The Primary Model - Donchian Channel output		
Nr	Date	Daily percent return
1.	2021-04-08	0,00467
2.	2021-04-09	0,007461
3.	2021-04-12	0,000188
4.	2021-04-13	0,003036
5.	2021-04-14	-0,003567
6.	2021-04-15	0,010808
7.	2021-04-16	0,003305
8.	2021-04-19	-0,004963
9.	2021-04-20	-0,006977
10.	2021-04-21	0

5.2 Labeling - Binary Classification

The output daily percent returns of the primary model Donchian Channel labels as positive (1) and negative (0). The labeling rule for the daily percent returns is that if the return is higher than zero, it results in a positive label 1, else 0. The labeled output of the primary model is the target variable of the secondary (ML) model prediction. Thus, the secondary (ML) model predicts the output of the primary model.

5.3 Secondary Model

The construction of the secondary model prediction and validation is following. The secondary ML model input data is the labeled daily return output of the primary model. The

Table 21. The Primary Model - Donchian Channel labeled output, last ten days.

The Primary Model - Donchian Channel labeled output			
Nr	Date	Daily percent return	Labeled
1.	2021-04-08	0,00467	1
2.	2021-04-09	0,007461	1
3.	2021-04-12	0,000188	1
4.	2021-04-13	0,003036	1
5.	2021-04-14	-0,003567	0
6.	2021-04-15	0,010808	1
7.	2021-04-16	0,003305	1
8.	2021-04-19	-0,004963	0
9.	2021-04-20	-0,006977	0
10.	2021-04-21	0	0

model is the random forest classifier as predicting machine learning algorithm. The algorithm predicts if the next trade is profit (1) or not (0) and the probability. The probability prediction is sizing the bet of the next trade and showing results by dividing the prediction into a non-profit (0.43) and a profit (0.57). The column on the left side of the probability predicts a non-profit (0.43), and the right side probability refers to a profitable (0.57) trade. If the probability prediction is higher than 0.50, the next trade is profitable. The rule for executing a trade requires both primary and secondary models to agree. The rule decision validates by AND logic gate where the result is true, if and only if both models resulting valid (1). The table of the secondary model prediction results shows five true trades amid the last ten examples.

Table 22. The Secondary Model - Prediction of Profit or Not and Probability, last ten days.

The Secondary Model - Prediction of Profit or Not and Probability						
Nr	Date	Daily percent return	Labeled	Profit 1 or Not 0	Probability	True
			[0,1]	[0,1]	[0,1]	
1.	2021-04-08	0,00467	**1**	**1**	0,46 0,54	true
2.	2021-04-09	0,007461	**1**	**1**	0,48 0,52	true
3.	2021-04-12	0,000188	1	0	0,50 0,50	
4.	2021-04-13	0,003036	**1**	**1**	0,48 0,52	true
5.	2021-04-14	-0,003567	0	1	0,48 0,52	
6.	2021-04-15	0,010808	**1**	**1**	0,44 0,56	true
7.	2021-04-16	0,003305	**1**	**1**	0,47 0,53	true
8.	2021-04-19	-0,004963	0	1	0,46 0,54	
9.	2021-04-20	-0,006977	0	1	0,43 0,57	
10.	2021-04-21	0	0	0	0,61 0,39	

The maximum probability was 0.73 and the minimum probability was 0.27 during the total test period of 917 days.

```
Last 10 prediction probabilities:
  [0 and 1]

[[0.46 0.54]
 [0.48 0.52]
 [0.5  0.5 ]
 [0.48 0.52]
 [0.48 0.52]
 [0.44 0.56]
 [0.47 0.53]
 [0.46 0.54]
 [0.43 0.57]
 [0.61 0.39]]
Max probability:  0.73
Min probability:  0.27
Shape of probability:  (917, 2)
```

Figure 19. The prediction results of the random forest algorithm of the secondary model - probability last ten days, and the maximum and minimum of the total period.

The test accuracy of the random forest algorithm results in 0.555, and the train set shows 1.0 accuracy. Furthermore, the last prediction of the period date 2021-04-21, next trade profit resulting a zero (0) meaning a loss with 0.61 probability meaning ignore trading.

```
RF train accuracy: 1.000
RF test accuracy: 0.555
Prediction, the next trade is Profit (1) or Not (0):  [0.]
Probability of next trade [0 , 1]:  [[0.61 0.39]]
```

Figure 20. The prediction results of the random forest algorithm in the secondary model - accuracy, profit or not, and probability.

The random forest classifier results a feature importance table where the features are sorted by their score. The feature score informs how much information a feature contributes to the machine learning model. The score measures the more important and less important features and suits for further features selection decisions.

The results of the feature importance table show that the RSI technical indicator is the most important feature in this run. The RSI is a momentum indicator that provides bullish

Table 23. The Secondary Model Features Sorted By Their Scores.

The secondary model features sorted by their scores			
Nr	Score Feature	Score Feature	Score Feature
1.	0,0271 RSI	0,0223 NV_1D	0,0196 NX
2.	0,0269 LC_1D	0,0222 NQ_1D	0,0192 Disp5
3.	0,0249 CT_1D	0,0222 Disp10	0,0184 BP
4.	0,0246 PSY12	0,0222 BP_1D	0,0184 BO
5.	0,0244 ROC	0,0217 LC	0,0181 QS
6.	0,0243 QS_1D	0,0217 ASY3	0,0181 NQ
7.	0,0240 Momentum	0,0212 US_1D	0,0178 EC
8.	0,0240 HO_1D	0,0211 ASY1	0,0178 AD
9.	0,0235 CD_1D	0,0207 HO	0,0177 CD
10.	0,0235 ASY5	0,0206 ES_1D	0,0174 NV
11.	0,0234 NX_1D	0,0205 AD_1D	0,0168 MA
12.	0,0232 BO_1D	0,0203 ASY4	0,0164 OSCP
13.	0,0231 JY_1D	0,0202 BIAS6	0,0162 Close
14.	0,0228 ASY2	0,0201 CT	0,0160 US
15.	0,0226 EC_1D	0,0200 JY	0,0156 TY
16.	0,0223 TU_1D	0,0199 TY_1D	0,0150 TU

and bearish price momentum signals. Thus, it makes sense that it contributes information in predicting the Donchian Channel trend following strategy. The lowest feature score has the US two-year note future (TU).

The cumulative return figure is comparing the trading strategy performances during the test period of 917 days. The underlying instrument S&P500 E-mini future performs 0.61 times cumulative return with the buy and hold trading strategy. The primary model shows 0.38 times cumulative return performance with the Donchian Channel trading strategy. The secondary model predicting profit or not results in 2.80 times cumulative return with the random forest machine learning supported trading strategy. Moreover, adding the probability prediction of sizing the bet in the secondary model shows 3.15 times the cumulative returns.

To summarize, the result of the underlying buy and hold strategy beat the Donchian Channel trend-following strategy. The secondary machine learning model shows significant performance compared to the underlying strategy and the primary model. In this study, the value-added of the machine learning model is about five times the primary model (3.15 vs. 0.61). The secondary model profit or not prediction is a hybrid of the primary and

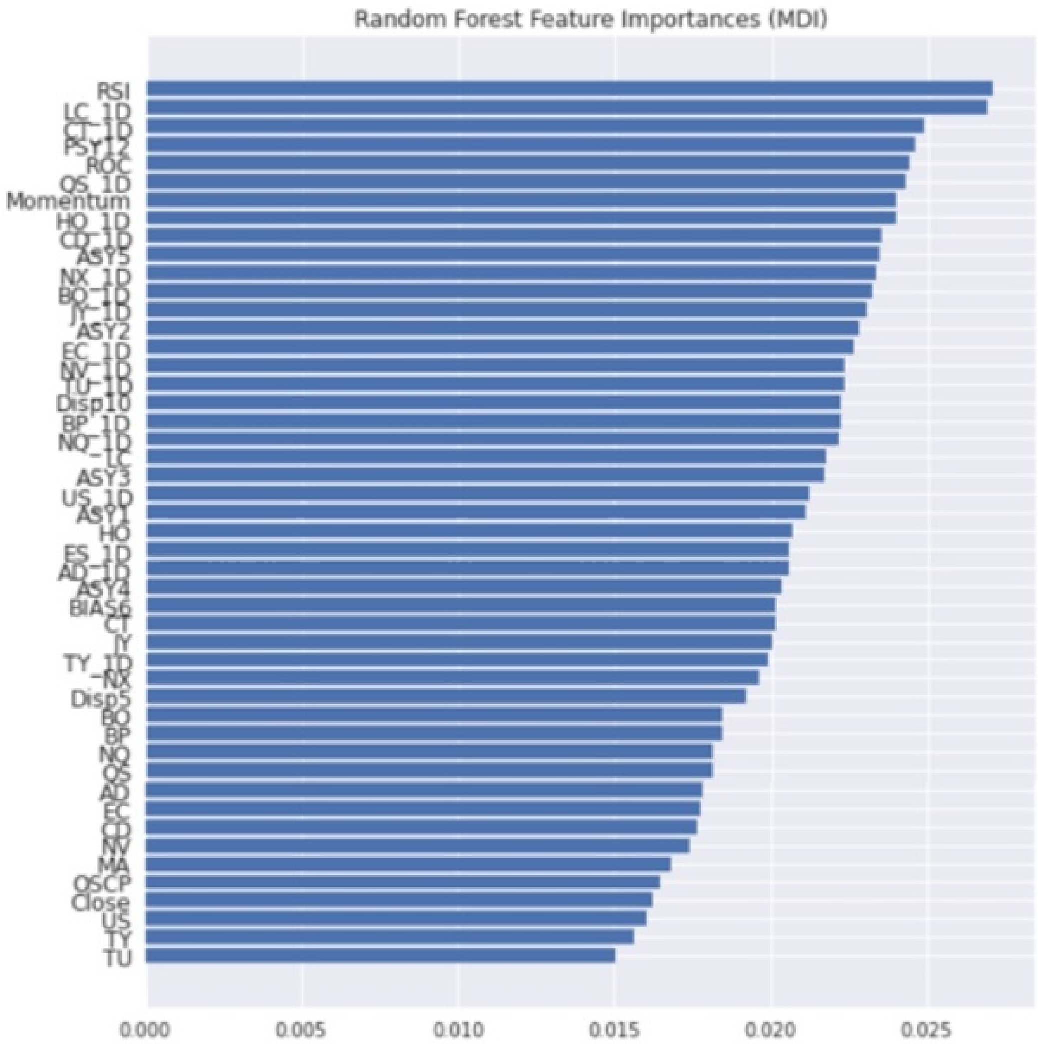

Figure 21. The secondary model random forest feature importance table.

secondary model because both must agree to valid a trade. A critical finding from the result details is that the machine learning model prediction ignores all the non-profit trades. To conclude, it is precious to know when staying out of the market and not trading.

The receiver operating characteristic curve presents the area under the curve (AUC) of 0.56 in value. The result implicates that there is a slight value in the secondary model probability prediction compared to the 0.50 value of no skill.

The confusion matrix shows the values of the secondary model, profit or not, prediction; the true negative (tn) 133, the false positive (fp) 273, the false-negative (fn) 135, and the true positive (tp) 376 of the 917 test days totally. The precision is 0.57 in value explaining how many selected items are relevant. The recall is 0.73 in value implicating how many

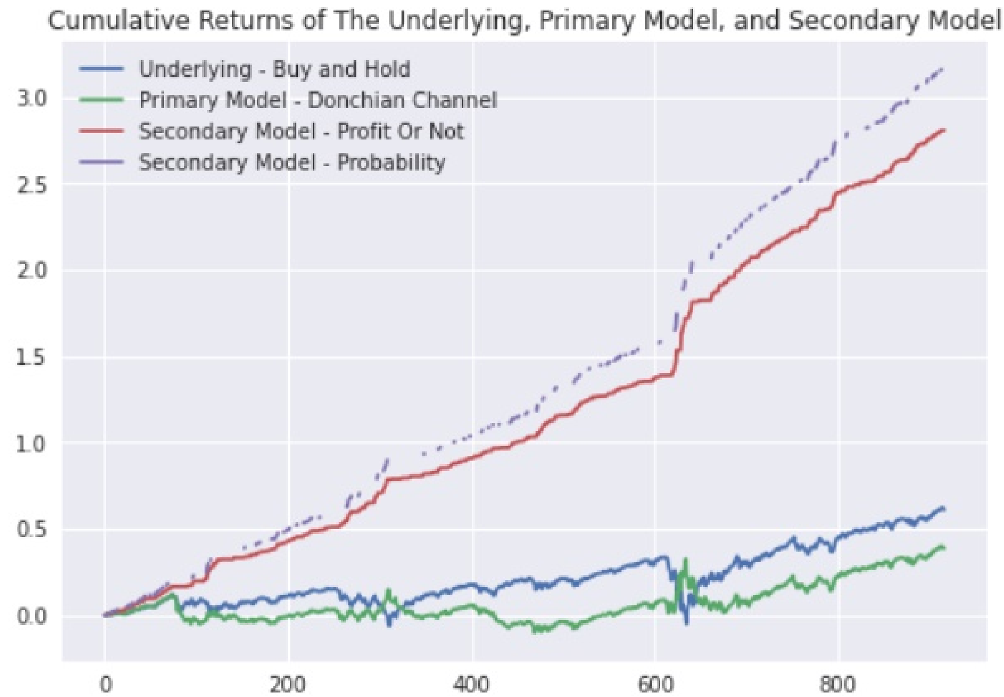

Figure 22. The cumulative return of the underlying, the primary model and the secondary model during test period of 917 days.

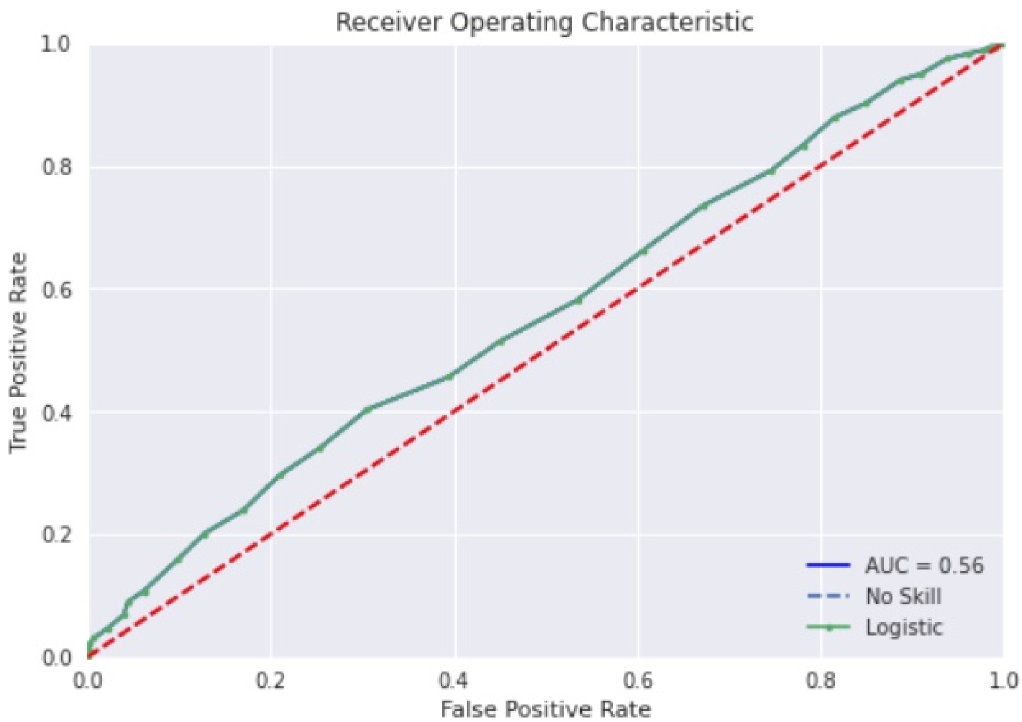

Figure 23. The Receiver Operating Characteristic (ROC) curve of the probability prediction.

relevant items are selected.

```
Confusion matrix:
 [[133 273]
 [135 376]]
```

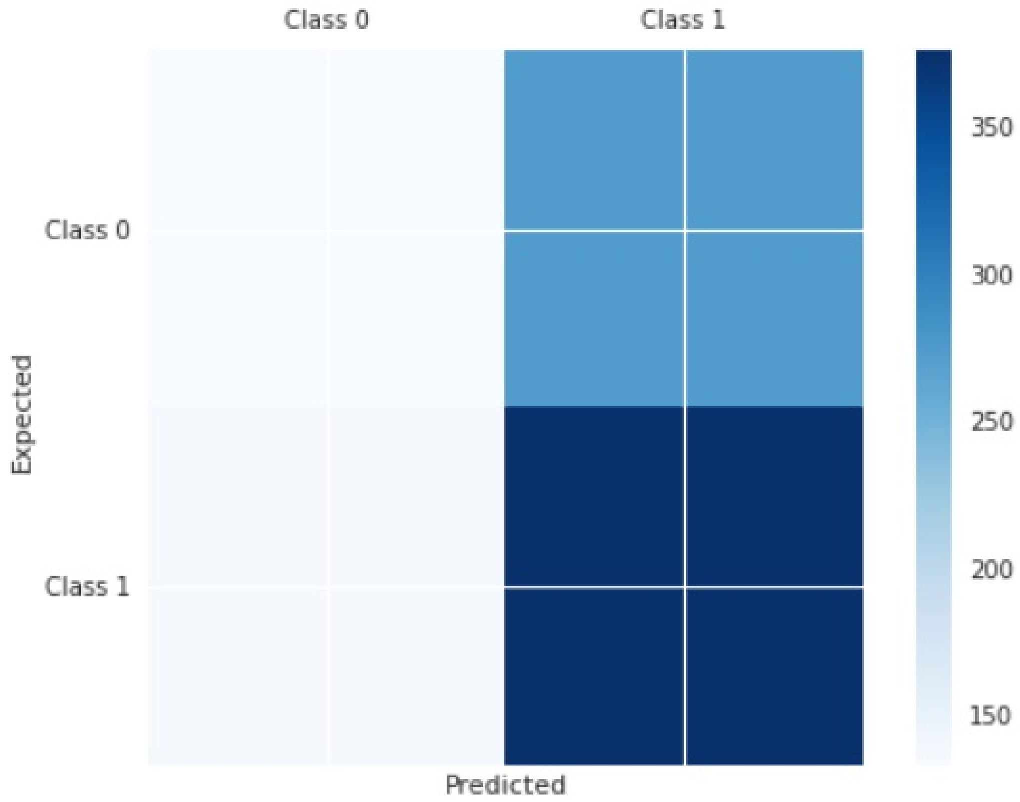

Figure 24. The confusion matrix of the secondary model profit or not prediction.

```
tn, fp, fn, tp
133 273 135 376
Precision = how many selected items are relevant? (positive predictive value (PPV), tp / (tp+fp)
0.5793528505392912
# Recall = how many relevant items are selected? (the true positive rate or sensitivity), tp / (tp+fn)
0.735812133072407
```

Figure 25. The confusion matrix of the profit or not prediction - extracting true positives.

The F1 score of the secondary model prediction, profit or not, is resulting in 0.648 estimations. The F1 score performs the harmonic mean of the precision and recall.

Table 24. The f1 Score of Secondary Model Prediction - Profit or Not.

The F1 Score of Secondary Model Prediction - Profit or Not		
Nr	F1 Score Metrics	Value
1.	F1 score macro:	0.5214
2.	F1 score micro:	0.5550
3.	F1 score weighted:	0.5359
4.	F1 score average=None:	[0.3946 0.6482]
5.	F1 score zero division=1:	0.6482
6.	F1 score:	0.6482

6 RESULTS

The research hypothesis is that machine learning can improve any trading strategy. The investigated trading strategy system shows significantly improving trading performance when testing the hybrid model, including machine learning in the study.

The research questions include finding a machine learning-supported trading strategy that will solve binary classification problems to predict if the next trade is profit or not and the probability of that outcome. The predicted probability utilizes sizing the bet of the next trade.

The research questions are basing the construction of the trading strategy system—the system covering a simple trend-following trading strategy as the primary model. Then, the binary classification problem solves by labeling the primary model's output for targeting the secondary model. Finally, the random forest algorithm predicts the profit or not solution, with the probability in the secondary model. The prediction results validated with the logic gate before the cumulative return analysis show that the system implementation performs successfully with the elements and construction in the study.

The trading strategy system design follows the Tail Reaper and MNIST processes introducing in the literature section. A critical element is the labeling to be able to combine the primary and the secondary model. An essential point is choosing a machine learning algorithm that predicts both the binary classification and the probability; otherwise, the system lacks the tool for sizing the bet. Finally, the logic gate validates the hybrid results of both primary and secondary models; both models must agree. It is the combination of both models of the hybrid trading strategy system that enables great performance.

A character of the random forest algorithm is that the result varies with every run. For example, the accuracy and the number of true positives are changing run by run inside the same day. The resulting level remains relatively stable, but a slight variation occurs.

The study's novelty is the SigTech platform for trading strategy system development, utilizing meta-labels to combine models, and the idea of the hybrid model using machine learning and finally validating with the logic gate. Furthermore, the input data of the study is basing totally on the SigTech data source. Therefore, it was saving resources from cleaning and modifying the data.

Comparing the trading strategy system random forest (RF) results with the previous related research is shown in table 21. Man & Chan (2021b) study the cluster-based MDA (cMDA) method introduced by de Prado (2020) for two financial datasets (S&P500 and Tail Reaper trading strategy) improving predictive performance.

Table 25. Prediction performance comparison to related study.

Prediction Performance Comparison To Related Study				
Method	F1	AUC	Accuracy	Research
cMDA	0.576	0.779	0.583	S&P500 Dataset
MDA	0.508	0.716	0.517	S&P500 Dataset
cMDA	0.658	0.672	0.614	Tail Reaper Strategy
MDA	0.602	0.537	0.529	Tail Reaper Strategy
RF	**0.648**	**0.560**	**0.555**	**This Study**

6.1 Descriptive Statistics

The statistics show that the buy and hold trading strategy's cumulative return (0.61), the CAGR (8.56), and the Sharpe ratio (0.79) of the underlying instrument is higher than the primary model (Donchian channel) cumulative return (0.38), the CAGR (6.52) and the Sharpe ratio (0.50). The cumulative return, CAGR, and the Sharpe ratio estimate from the daily percent returns. In addition, the general equity trend has been increasing during the test period of 917 days. Therefore, the secondary (ML) model statistics are significantly higher than the underlying or the primary model. Moreover, the probability prediction shows a higher cumulative return and Sharpe ratio but not a higher CAGR than the profit or not prediction of the secondary model.

6.2 Inferential Statistics

The validation of the results performing the historical trades, according to the trading signals, with the initial capital of 100 000 monetary units shows 61 070, 38 700, and 280 600 profit on the underlying, primary and secondary model trading strategy, respectively. In addition, the probability prediction for sizing the bet presents an increasing profit of a total of 315 800 in monetary units. The Kelly criterion formula calculates, for example, with a 0.54 probability prediction two times minus one (0.54x2-1=.08), resulting 8% of

Table 26. The cumulative return, CAGR, and the Sharpe ratio.

The cumulative return, CAGR, and the Sharpe ratio (917 days)				
Nr	Trading Strategy	Cumulative Return	CAGR	Sharpe
1.	Underlying (Buy and hold)	0.6107	8.5654	0.7933
2.	Primary Model (Donchian)	0.3870	6.5200	0.5021
3.	Secondary Model (Profit or not)	2.8060	12.2226	6.9602
4.	Secondary Model (Probability)	3.1580	11.6903	13.2348

increase in position size. The machine learning of the secondary model shows significant value-added compared to the underlying and the primary model generating about five (517%) to eight (816%) times higher profits during the 917 days test period.

Table 27. Validation of Results

Validation of Results (917 days)					
Trading Strategy	Profit	Value to Initial Capital	Value to Primary Model	Value to Underlying	Value of Probability Prediction
1.Initial Capital (100 000)	-	-	-	-	-
2.Underlying (Buy and hold)	61 070	61,07 %	-	-	-
3.Primary Model (Donchian channel)	38 700	38,70 %	-	-	-
4.Secondary Model (Profit or Not)	280 600	280,60 %	$\frac{280600}{38700} = 725{,}06\%$	$\frac{280600}{61070} = 459{,}47\%$	-
5.Secondary Model (probability)	315 800	315,80 %	$\frac{315800}{38700} = 816{,}02\%$	$\frac{315800}{61070} = 517{,}11\%$	$\frac{315800}{280600} = 112{,}54\%$
6.Optimal Bet Size (Kelly criterion)	315 800	315,80 %	$\frac{315800}{38700} = 816{,}02\%$	$\frac{315800}{61070} = 517{,}11\%$	$\frac{315800}{280600} = 112{,}54\%$

To summarize, the research results are in line with the hypothesis, the research question, and the related research. The hypothesis is valid in this study, meaning that machine learning can improve any trading strategy. The research question contributed to solving the binary classification problems of predicting profit or not and the probability of the next trade adding value to sizing the bet. The previous related research shows similar results referring to financial data set, trading strategy, and feature importance predictions. The related machine learning prediction research compared with the metrics of the F1 score, the AUC, and the accuracy. Finally, the machine learning-based hybrid trading strategy system shows significant value-added compared to the underlying and the primary technical analysis-based trading strategy model.

7 CONCLUSIONS

The study aims to find a machine learning-supported trading system—the trading system coded with the Python programming language in the SigTech platform utilizing their data. The introduction section presents the research hypothesis that machine learning can improve any trading strategy, the research question of the binary classification problem predicting profitable trades with probability, with sizing the bet.

The related section introduces the financial theory of information efficiency and efficient market hypothesis. First, the literature review explains the technical analysis, trend following, trading strategies, and the fixed fraction asset allocation, including bet sizing by Kelly criterion. Then a discussion of the futures contracts briefly. Next, the literature review body describes predicting the conditional probability of the next trade's profit. Also, another ML-supported process shows including binary prediction utilizing meta-labeling for combining predictions as a hybrid of a primary and secondary model, and finally, the logic gate for validation.

The methodology section explains the setting of the trading system hierarchy, including theory, system, strategy, process, and signals. After presenting the SigTech platform, the study explains the descriptive statistics of the data and the features construction. Next, the Tail Reaper and MNIST models contribute to designing the final version of the coded trading system. The trading system structure consists of the primary model (Donchian channel), the labeling (binary classification), the secondary ML model prediction of profitable next trades with probabilities sizing the bet. The logic gate rule validates to make the trades according to the designed trading system when both parts of the hybrid model perform validly (1) results. The metric CAGR makes it possible to compare the performance of trading strategy returns increasing or decreasing over time, and the Sharpe ratio measures the return of the investment with risk adjusting. Finally, the procedure section describes the technical analysis indicators used as features, the Donchian channel trading strategy, meta-labeling, and the logic gate method for validation of the hybrid model.

The algorithmic description presents the random forest classifier that produces multipole decision trees leading finally to simple output of binary classification prediction with the probability prediction. The feature selection is a way of improving the transparency and the interpretability of the random forest ML algorithm. The feature importance table shows the ranking of the features and enables feature analysis and selection to improve.

The experiments section analyses the results of the trading system. The primary model outputs are daily percent returns that labels by binary classification to one if positive, else zero. The labeled output of the primary model is the target label in the secondary model prediction. The trades is done if the hybrid model both parts primary and secondary confirm it by the output of one. The feature importance table shows the highest ranking for the RSI momentum indicator and the lowest for the US 2-year note (TU). The maximum probability prediction is 0.73 and the minimum is 0.27 for the whole test period of 917 days.

In the results section, the secondary model's random forest shows the F1 score of 0.648, the AUC of 0.560, and the test accuracy of 0.555 compared with the related research of cMDA testing 0.576 (F1), 0.779 (AUC), and the accuracy of 0.583. The cumulative return of the trading strategies resulting the underlying of 0.61, the primary model 0.38, the secondary model (profit or not) of 2.80, and the secondary model probability of 3.15. The validation of the results performing the historical trades, according to the trading signals, with the initial capital of 100 000 monetary units shows 61 070, 38 700, and 280600 profit on the underlying, primary and secondary model trading strategy, respectively. The machine learning of the secondary model shows significant value-added compared to the underlying and the primary model generating about five(517%) to eight (816%) times higher profits during the 917 days test period.

7.1 Discussion

This thesis collects elements of how to utilize the power of machine learning in improving a trading strategy performance by testing a hybrid model. The study explains the components of meta-labeling gluing the models, the hybrid model validated by the logic gate, and the SigTech platform for programming and providing reliable and high-quality data. The beginning of the journey felt complex, improving over time, and finally aiming towards a simple and straightforward trading strategy system resulting in ideas for future study.

Research question 1. Exploring how to find and construct a machine learning supported trading strategy system predicting the next trading opportunity?

The fundamental elements for a trading strategy construction are an entry signal (open long/short position), an exit signal (close long/short position), and when to stay out of the market (no position). The position side is between the signals defined for example, long (1), short (-1), and no position (0). The daily percent returns of the underlying financial instrument multiply with the position side, resulting in the trading performance of the strategy.

Table 28. The construction of a trading strategy.

The construction of a trading strategy		
Nr	Element	Description
1. 2. 3.	When to get an entry signal? When to get an exit signal? When to be without a position?	Open long or short position Close long or short position No position
4.	Resulting in the position side between the signals.	Position side: long (1), short (-1), and no position(0)
5.	Calculate daily percent returns of the underlying.	One-day percent return +5% or -5%
6.	Multiply the position side with the daily percent returns.	Long-profit (1x+5%), short-profit (-1x-5%), non-profit (0x+/-5%)
7.	Should a stop-loss level be defined?	Depends on the trading strategy

The position side means a trading position that can be long, short, or no position. A long trading position occurs when buying security. A short trading position opens when selling the asset. No position or neutral position means staying out of the market and taking no trading position. The programmer specifies the conditions to be evaluated or tested by the program. These conditions lead to a set of states: buy=1, sell=-1, and neutral=0.

For example, a long (1) position side profiting when daily percent return is positive (1 x +5%) and a short position side profiting when daily percent return is negative (-1 x -5%). No position shows neutral profit (0 x +/-5%) for the trading strategy performance.

Table 29. The Trading Strategy Logic for Coding.

The Trading Strategy Logic for Coding						
Nr	Position	Position side	Daily percent return of the underlying	(Position side) x (daily percent return of the underlying)	Daily percent return of the trading strategy	Cumulative return of the trading strategy
			Underlying	Underlying	Strategy	Strategy
1.	Long entry					
2.	Day 1	1	+5%	1x+5%	0,05	0,05
3.	Day 2	1	-2%	1x-2%	-0,02	0,03
4.	Day 3	1	+4%	1x+4%	0,04	0,07
5.	Day 4	1	-1%	1x-1%	-0,01	**0,06**
6.	Long exit					
7.	Short entry					
8.	Day 5	-1	-5%	-1x-5%	0,05	0,05
9.	Day 6	-1	-3%	-1x-3%	0,03	0,08
10.	Day 7	-1	+2%	-1x+2%	-0,02	0,06
11.	Day 8	-1	+3%	-1x+3%	-0,03	**0,03**
12.	Short exit					
13.	No position					
14.	Day 9	0	-5%	0x-5%	0	0
15.	Day 10	0	-3%	0x-3%	0	0
16.	Day 11	0	+2%	0x+2%	0	0
17.	Day 12	0	+3%	0x+3%	0	**0**
The total cumulative return of the trading strategy: long 0,06 + short 0,03 + no pos 0 = **9%**						

Therefore, the trading strategy construction allows replacing the primary model trading strategy with any trading strategy in the studied trading system.

The construction of the secondary model prediction and validation is following. The secondary ML model input data is the labeled daily return output of the primary model. The model is the random forest classifier as predicting machine learning algorithm. The algorithm predicts if the next trade is profit (1) or not (0) and the probability. The probability prediction is sizing the bet of the next trade and showing results by dividing the prediction into a non-profit (0.43) and a profit (0.57). The column on the left side of the probability predicts a non-profit (0.43), and the right side probability refers to a profitable (0.57) trade. If the probability prediction is higher than 0.50, the next trade is profitable. The rule for executing a trade requires both primary and secondary models to agree. The rule decision validates by AND logic gate where the result is true, if and only if both models

Table 30. The Secondary Model - Prediction of Profit or Not and Probability, last ten days.

The Secondary Model - Prediction of Profit or Not and Probability						
Nr	Date	Daily percent return	Labeled	Profit 1 or Not 0	Probability	True
			[0,1]	[0,1]	[0,1]	
1.	2021-04-08	0,00467	**1**	**1**	0,46 0,54	true
2.	2021-04-09	0,007461	**1**	**1**	0,48 0,52	true
3.	2021-04-12	0,000188	1	0	0,50 0,50	
4.	2021-04-13	0,003036	**1**	**1**	0,48 0,52	true
5.	2021-04-14	-0,003567	0	1	0,48 0,52	
6.	2021-04-15	0,010808	**1**	**1**	0,44 0,56	true
7.	2021-04-16	0,003305	**1**	**1**	0,47 0,53	true
8.	2021-04-19	-0,004963	0	1	0,46 0,54	
9.	2021-04-20	-0,006977	0	1	0,43 0,57	
10.	2021-04-21	0	0	0	0,61 0,39	

resulting valid (1). The table of the secondary model prediction results shows five true trades amid the last ten examples.

The learnings of the programming:

1. In large, complicated programming projects, avoid overwriting global variables. It may have disastrous consequences.

2. The time-series data is susceptible, and the slightest mistake can completely invalidate a model's result.

3. It is infinitely easier to review and verify specific phases of the process in large programming projects if the process divides into its self-contained modules.

Research question 2. How can metalabeling enable and improve the integration of machine learning in a trading strategy?

The methodology is basing on the process that is utilizing meta-labeling. The meta-labeling method trains a secondary model on the prediction outcomes of a primary model, where losses label as zeros and gains label as one. The meta-labeling method is possible for any trading strategy that outputs daily percent returns. The most exciting finding of the process is that the labeling of the primary model output allows the adding of the machine learning model as the secondary model on the top. Thus, a machine learning algorithm can improve any trading strategy by labeling. Therefore, the secondary model predicts the primary model's output and not the original input data. Hence, any trading strategy or technical analysis model can be the primary model, label the output and improve it with a machine learning algorithm. Thus, it gives the process added value in the form of scalability.

Research question 3. What is a hybrid model that combines the elements of the trading strategy system?

The trading strategy system designed for coding at the SigTech platform based on the Tail Reaper and the MNIST examples. The hybrid model combines the primary model and the secondary model. The hybrid model elements are glued by the meta-labeling method and validated by the logic gate method.

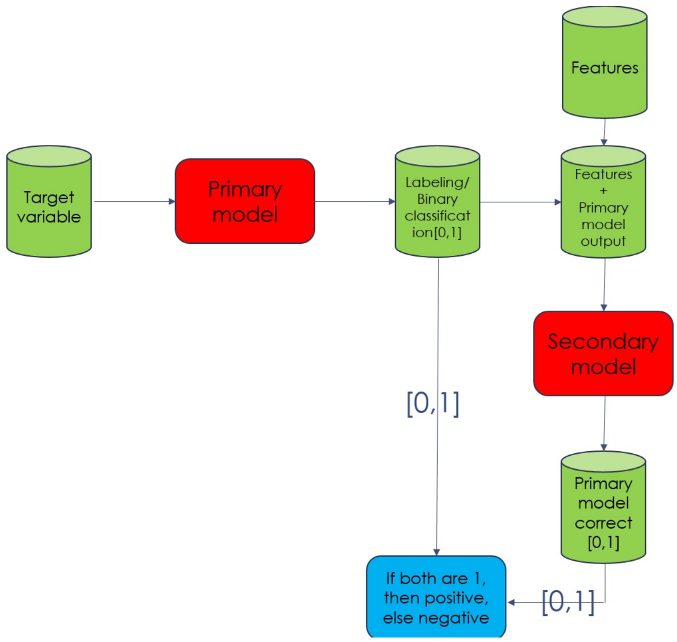

Figure 26. The Trading Strategy System.

The design of the trading strategy system using the hybrid model is following. The primary model is inputting a target variable and the output labels by binary classification of profitable (1) and non-profitable (0). The combination of the features and labeled output of the primary model are the inputs to the secondary model. The secondary machine learning model predicts the labeled output of the primary model if the output is correct (1) or not (0). The AND logic gate method validates the output of the primary model and the prediction of the secondary model. If and only if both models are resulting in a valid

(1), the trade is executed, else not.

Table 31. The Trading Strategy System Structure

The Trading Strategy System Structure	
Face	Description
1.	**Features (48):** 17 rolling front futures + 1-day historical returns + 14 technical analysis indicators
2.	**Primary / Base Model (technical analysis model):** INPUT: S&P500 E-mini front future (target feature, 5557 close days, 2000-01-04 to 2021-04-21) MODEL: Donchian Channel (channel 50, exit 10) OUTPUT: daily percent returns (positive and negative)
3.	**Labeling / Binary Classification [0,1]:** Binary classification of Donchian Channel output (1=profit, 0=else) (Metalabels[0,1])
4.	**Secondary Model (machine learning prediction):** INPUT: features + labeled output of the primary model (= target variable: Metalabels[0,1]) MODEL: Random Forest OUTPUT: Prediction (profit=1 or not=0) and probability e.g. (0.35, 0.65) of tomorrow's close price
5.	**Rule** => If both Primary and Secondary model perform true (1): Then positive (do the trade), else negative (no trade)

Research question 4. What are the empirical results from the implemented trading strategy system?

The empirical results of the implemented trading strategy system show in the cumulative return figure comparing the trading strategy performances during the test period of 917 days. The underlying instrument S&P500 E-mini future performs 0.61 times cumulative return with the buy and hold trading strategy. The primary model shows 0.38 times cumulative return performance with the Donchian Channel trading strategy. The secondary model predicting profit or not results in 2.80 times cumulative return with the random forest machine learning supported trading strategy. Moreover, adding the probability prediction of sizing the bet in the secondary model shows 3.15 times the cumulative returns.

The empirical results of the trading strategy system are in line with the related studies. Comparing the random forest (RF) results with the previous related research is shown

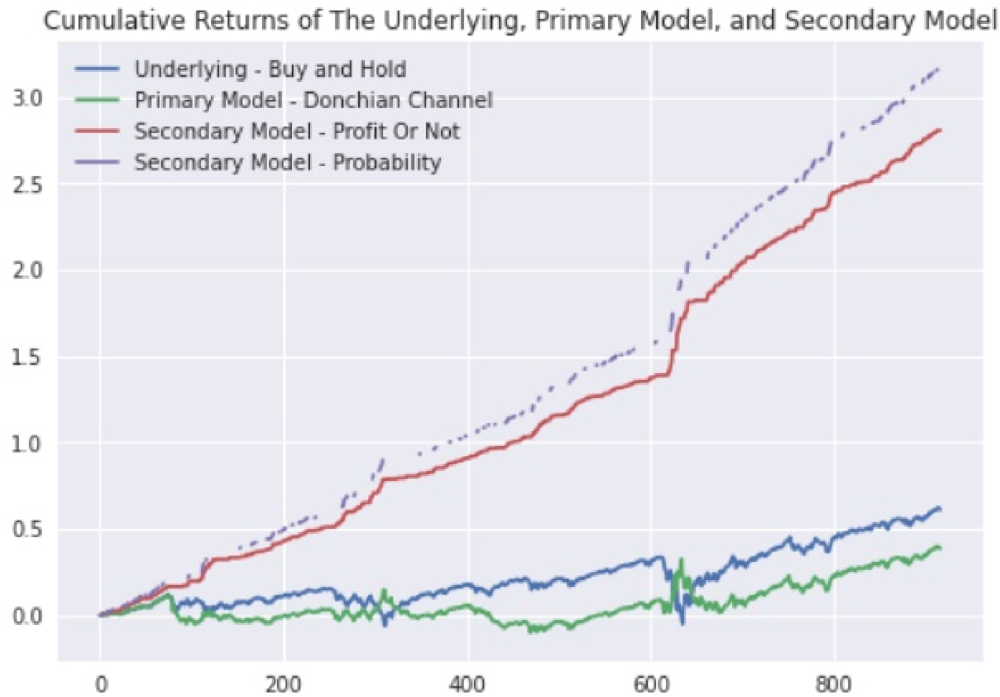

Figure 27. The cumulative return of the underlying, the primary model and the secondary model during test period of 917 days.

in the comparison table. Man & Chan (2021b) study the cluster-based MDA (cMDA) method introduced by de Prado (2020) for two financial datasets (S&P500 and Tail Reaper trading strategy) improving predictive performance.

Table 32. Prediction performance comparison to related study.

Prediction Performance Comparison To Related Study				
Method	F1	AUC	Accuracy	Research
cMDA	0.576	0.779	0.583	S&P500 Dataset
MDA	0.508	0.716	0.517	S&P500 Dataset
cMDA	0.658	0.672	0.614	Tail Reaper Strategy
MDA	0.602	0.537	0.529	Tail Reaper Strategy
RF	**0.648**	**0.560**	**0.555**	**This Study**

The validation of the empirical results performing the historical trades, according to the trading signals, with the initial capital of 100 000 monetary units shows 61 070, 38 700, and 280 600 profit on the underlying, primary and secondary model trading strategy, respectively. In addition, the probability prediction for sizing the bet presents an increasing profit of a total of 315 800 in monetary units. The Kelly criterion formula calculates, for example, with a 0.54 probability prediction two times minus one (0.54x2-1=.08), resulting 8% of increase in position size. The machine learning of the secondary model shows significant value-added compared to the underlying and the primary model generating about five (517%) to eight (816%) times higher profits during the 917 days test period.

Table 33. Validation of Results

Validation of Results (917 days)					
Trading Strategy	Profit	Value to Initial Capital	Value to Primary Model	Value to Underlying	Value of Probability Prediction
1.Initial Capital (100 000)	-	-	-	-	-
2.Underlying (Buy and hold)	61 070	61,07 %	-	-	-
3.Primary Model (Donchian channel)	38 700	38,70 %	-	-	-
4.Secondary Model (Profit or Not)	280 600	280,60 %	$\frac{280600}{38700} = 725,06\%$	$\frac{280600}{61070} = 459,47\%$	-
5.Secondary Model (probability)	315 800	315,80 %	$\frac{315800}{38700} = 816,02\%$	$\frac{315800}{61070} = 517,11\%$	$\frac{315800}{280600} = 112,54\%$
6.Optimal Bet Size (Kelly criterion)	315 800	315,80 %	$\frac{315800}{38700} = 816,02\%$	$\frac{315800}{61070} = 517,11\%$	$\frac{315800}{280600} = 112,54\%$

To summarize, the empirical research results of the implemented trading strategy system are in line with the hypothesis, the research questions, and the related research. The hypothesis is valid in this study, meaning that machine learning can improve any trading strategy. The research questions contributed to solving the binary classification problems of predicting profit or not and the probability of the next trade adding value to sizing the bet. The previous related research shows similar results referring to financial data set, trading strategy, and feature importance predictions. The related machine learning prediction research compared with the metrics of the F1 score, the AUC, and the accuracy. Finally, the machine learning-based hybrid trading strategy system shows significant value-added compared to the underlying and the primary technical analysis-based trading strategy model.

Covel (2009) summarized the sum of the parts as if building a system that gives an entry and exit, tells *how much* to bet along the way, and adjusts to your current capital and current market volatility at all times, and no more analysis is needed.Covel (2009) However, there is an additional advantage of the machine learning prediction power in today's world.

7.2 Summary of Limitations

The limitation of the financial product is to rolling front futures contract because of liquidity and the minimal transaction costs. The time limitation is 20-years historical daily closing data. The primary model limits to a trend-following trading strategy. The secondary model limits to a random forest algorithm. The machine learning algorithm must predict probability. The study limits the feature selection examination while knowing the importance of the subject. The study ignores the transaction cost (because they are low in futures contracts), the taxes, and the risk-free interest rate assuming that the investor gains no yield leaving the capital on the savings account.

7.3 Recommendations For Future Research

Meta-labeling connecting the primary and the secondary elements of the hybrid model trading system and the significant performance of the random forest machine learning model in the secondary model prediction identified the study channel the areas for future research.

i) The empirical examination of the trading system.

The trading system implementation results encourage empirical investigation to find the practical opportunities and obstacles. The SigTech platform has build-in features to connect the trading system with the real time market for further research.

ii) Changing scalable parameters.

The trading system hybrid model further explores the time scale, financial instruments, and the various trading strategies replacing the daily closing prices, S&P500 E-mini futures contract, and the Donchian channel primary model.

iii) Studying feature selection.

Feature selection method cluster-based MDA (cMDA) improves predictive performance, feature stability, and model interpretability. de Prado (2020) The study of Man & Chan (2021b) shows that the stability and the interpretability of the cMDA selected features are superior to MDA selected features. Therefore, feature selection research may further develop and enhance the hybrid trading strategy system performance.

7.4 Concluding Discussion

The results of the research show that the hypothesis is valid, meaning that a machine learning algorithm can improve a trading strategy. In this study, the machine learning model predicts a binary classification (profit or not) and a probability of next trade. Integrating a machine learning model to a trading strategy by a similar system structure as in this study, machine-learning adds value, meaning improving a trading strategy.

To summarize, based on the implemented trading strategy system, results show that by exploring the ways of constructing a hybrid model it is possible to achieve an explainable solution with possibilities for scalability, repeatability, and interpretability. The results show that the solution achieves a state of the art accuracy that enables interpretation and constructs a repeatable process. The key finding is that the process is scalable in the following ways. First, we can choose any trading strategy or technical analysis model as the primary model, label the output accordingly and then further improve with a machine learning algorithm. Second, we can choose any liquid financial instrument as input data instead of the futures contracts. Third, the data period is scalable from daily to monthly and yearly. The period may even be hours or minutes. All these parameters can be changed and tested without impacting the implemented design of the trading strategy system based on meta-labeling and hybrid modeling using the SigTech Platform. The most critical finding is that labeling is the essential tool that allows the machine learning algorithm to improve any trading strategy. A valuable finding referring to the machine learning algorithms is that the secondary model, the machine learning model, also predicts when not to open a trading position.

"Knowing when not to bet is as important as knowing what bets are probably worth making." - Ray Dalio

REFERENCES

Alliance, Population Health. 2021, *Meaning Hybrid Model.* Available: https://dictionary.university/hybrid%20model.

Anderson, John A & Faff*, Robert W. 2004, Maximizing futures returns using fixed fraction asset allocation, *Applied Financial Economics*, vol. 14, no. 15, pp. 1067–1073.

Anson, Mark JP; Fabozzi, Frank J & Jones, Frank J. 2010, *The handbook of traditional and alternative investment vehicles: investment characteristics and strategies*, vol. 194, John Wiley & Sons.

on Artificial Intelligence, (AI HLEG) High-Level Expert Group. 2020, "Ethics guidelines for trustworthy AI," European Commission, Text, Apr. 2019. Accessed: Oct. 26, 2020. Available: https://digital-strategy.ec.europa.eu/en/library/ethics-guidelines-trustworthy-ai.

Bailey, David H; Borwein, Jonathan; Lopez de Prado, Marcos & Zhu, Qiji Jim. 2014, Pseudo-mathematics and financial charlatanism: The effects of backtest overfitting on out-of-sample performance, *Notices of the American Mathematical Society*, vol. 61, no. 5, pp. 458–471.

Belgiu, Mariana & Drăguţ, Lucian. 2016, Random forest in remote sensing: A review of applications and future directions, *ISPRS journal of photogrammetry and remote sensing*, vol. 114, , pp. 24–31.

Belgiu, Mariana & Drăguţ, Lucian. 2016, Random forest in remote sensing: A review of applications and future directions, *ISPRS Journal of Photogrammetry and Remote Sensing*, vol. 114, , pp. 24–31. Available: https://www.sciencedirect.com/science/article/pii/S0924271616000265.

Benhamou, Eric. 2016, Trend Without Hiccups-A Kalman Filter Approach, *Available at SSRN 2747102*.

Breiman, Leo. 2001, Random forests, *Machine learning*, vol. 45, no. 1, pp. 5–32.

Brown, Tom B; Mann, Benjamin; Ryder, Nick; Subbiah, Melanie; Kaplan, Jared; Dhariwal, Prafulla; Neelakantan, Arvind; Shyam, Pranav; Sastry, Girish; Askell, Amanda et al.. 2020, Language models are few-shot learners, *arXiv preprint arXiv:2005.14165*.

Brownlee, Jason. 2020a, *Data Preparation for Machine Learning: Data Cleaning, Feature Selection, and Data Transforms in Python*, Machine Learning Mastery.

Brownlee, Jason. 2020b, Stacking Ensemble Machine Learning With Python. Available: https://machinelearningmastery.com/stacking-ensemble-machine-learning-with-python/.

Brundage, Miles; Avin, Shahar; Wang, Jasmine; Belfield, Haydn; Krueger, Gretchen; Hadfield, Gillian; Khlaaf, Heidy; Yang, Jingying; Toner, Helen; Fong, Ruth et al.. 2020, Toward trustworthy AI development: mechanisms for supporting verifiable claims, *arXiv preprint arXiv:2004.07213*.

Chan, Ernest P. 2019, US nonfarm employment prediction using RIWI Corp. alternative data. Available: http://epchan.blogspot.com/2019/12/us-nonfarm-employment-prediction-using.html.

Chan, Ernest P. 2020a, What is the probability of profit of your next trade? (Introducing PredictNow.Ai). Available: http://epchan.blogspot.com/2020/07/what-is-probability-of-profit-of-your.html.

Chan, Ernest P. 2020b, Why does our Tail Reaper program work in times of market turmoil? Available: http://epchan.blogspot.com/2020/03/why-does-our-tail-reaper-program-work.html.

Chan, Ernie. 2013, *Algorithmic trading: winning strategies and their rationale*, vol. 625, John Wiley & Sons.

Chen, James. 2021a, *What are Donchian Channels?* Available: https://www.investopedia.com/terms/d/donchianchannels.asp.

Chen, James. 2021b, *What Is a Trading Strategy?* Available: https://www.investopedia.com/terms/t/trading-strategy.asp.

Chen, James. 2021c, *What Is an Investment Strategy?* Available: https://www.investopedia.com/terms/i/investmentstrategy.asp.

Covel, Michael W. 2009, *The Complete TurtleTrader*, HarperBusiness.

Cowles 3rd, Alfred. 1933, Can stock market forecasters forecast?, *Econometrica: Journal of the Econometric Society*, pp. 309–324.

Dahlquist, Julie R. 2005, Trading currencies using moving averages–trend following versus contrarian strategies, *Managerial Finance*.

De Prado, Marcos Lopez. 2018, *Advances in financial machine learning*, John Wiley & Sons.

Deng, Houtao; Runger, George & Tuv, Eugene. 2011, Bias of importance measures for multi-valued attributes and solutions, In: *International conference on artificial neural networks*, Springer, pp. 293–300.

Denisko, Danielle & Hoffman, Michael M. 2018, Classification and interaction in random forests, *Proceedings of the National Academy of Sciences*, vol. 115, no. 8, pp. 1690–1692.

Dictionary, Oxford English. 2021, *Meaning of System in English*. Available: https://www.lexico.com/definition/system.

Douglas, Mark J. 2000, *Trading in the zone: master the market with confidence, discipline and a winning attitude*, Penguin.

Faith, Curtis M. 2019, *Way of the Turtle*, MGH, New York.

Fama, Eugene F. 2021, *Efficient Capital Markets A Review of Theory and Empirical Work*, University of Chicago Press, pp. 76–121. Available: https://doi.org/10.7208/9780226426983-007.

Fama, Eugene F & Blume, Marshall E. 1966, Filter rules and stock-market trading, *The Journal of Business*, vol. 39, no. 1, pp. 226–241.

Fan, Xing; Lung, Chung-Horng; Ajila, Samuel A et al.. 2018, Using Hybrid and Diversity-Based Adaptive Ensemble Method for Binary Classification, *International Journal of Intelligence Science*, vol. 8, no. 03, p. 43.

Fernando, Jason. 2021, *What Is Compound Annual Growth Rate – CAGR?* Available: https://www.investopedia.com/terms/c/cagr.asp.

Freedman, Lawrence. 2015, *Strategy: A history*, Oxford University Press.

Ganti, Akhilesh. 2021, *What Is Asset Management?* Available: https://www.investopedia.com/terms/a/assetmanagement.asp.

Gehm, F. 1983, *Commodity Money Market Management*.

Gui, Ke. 2020, Meta-labeling and Stacking, How to boost your machine learning score. Available: https://towardsdatascience.com/meta-labeling-and-stacking-f17a7f9804ec.

Gut, Allan. 2013, *Probability: a graduate course*, vol. 75, Springer Science & Business Media.

Haigh, Michael S & List, John A. 2005, Do professional traders exhibit myopic loss aversion? An experimental analysis, *The Journal of Finance*, vol. 60, no. 1, pp. 523–534.

Harvey, Campbell R; Liu, Yan & Zhu, Heqing. 2016, … and the cross-section of expected returns, *The Review of Financial Studies*, vol. 29, no. 1, pp. 5–68.

Hu Baogang, Beijing, a professor in Chinese Academy of Sciences. 2017, Is there any hybrid model in machine learning? Available: https://www.quora.com/Is-there-any-hybrid-model-in-machine-learning.

Jegadeesh, Narasimhan & Titman, Sheridan. 1993, Returns to buying winners and selling losers: Implications for stock market efficiency, *The Journal of finance*, vol. 48, no. 1, pp. 65–91.

Kam Ho, Tin. 1995, Random decision forests, In: *Proceedings of 3rd international conference on document analysis and recognition*, pp. 278–282. Available: https://web.archive.org/web/20160417030218/http://ect.bell-labs.com/who/tkh/publications/papers/odt.p

Kaufman, P.J. 2019, *Trading Systems and Methods*, Wiley Trading, Wiley. Available: https://books.google.fi/books?id=6YuzDwAAQBAJ.

Kelly, J. L. 1956, A new interpretation of information rate, *The Bell System Technical Journal*, vol. 35, no. 4, pp. 917–926. Available: https://ieeexplore.ieee.org/document/6771227.

Kirkpatrick Charles, D & Dahlquist Julie, R. 2007, *Technical Analysis*.

Languages, Oxford & Google. 2021, *Definition of Process in English*. Available: https://languages.oup.com/google-dictionary-en/.

Lill, Dominic. 2021, *What is financial trading? And should you do it?* Available: https://www.talk-business.co.uk/2017/05/04/financial-trading/.

Lundberg, Scott & Lee, Su-In. 2017a, A unified approach to interpreting model predictions, *arXiv preprint arXiv:1705.07874*.

Lundberg, Scott M & Lee, Su-In. 2017b, A Unified Approach to Interpreting Model Predictions, In: I. Guyon; U. V. Luxburg; S. Bengio; H. Wallach; R. Fergus; S. Vishwanathan & R. Garnett, eds., *Advances in Neural Information Processing Systems 30*, Curran Associates, Inc., pp. 4765–4774. Available: http://papers.nips.cc/paper/7062-a-unified-approach-to-interpreting-model-predictions.pdf.

Luo, Yin. 2014, Seven Sins of Quantitative Investing. Available: https://newyork.qwafafew.org/wp-content/uploads/sites/4/2015/10/Luo_20150128.pdf.

Man, Xin & Chan, Ernest. 2020, The best way to select features?, *arXiv preprint arXiv:2005.12483*.

Man, Xin & Chan, Ernest P. 2021a, The Best Way to Select Features? Comparing MDA, LIME, and SHAP, *The Journal of Financial Data Science*, vol. 3, no. 1, pp. 127–139. Available: https://jfds.pm-research.com/content/3/1/127.

Man, Xin & Chan, Ernest P. 2021b, Cluster-based Feature Selection. Available: http://epchan.blogspot.com/2021/01/the-amazing-efficacy-of-cluster-based.html.

McComas, William F. 2003, A textbook case of the nature of science: Laws and theories in the science of biology, *International Journal of Science and Mathematics Education*, vol. 1, no. 2, pp. 141–155.

Miner, Robert C. 2008, *High probability trading strategies: Entry to exit tactics for the forex, futures, and stock markets*, vol. 328, John Wiley & Sons.

Mitchell, Tom M. 1997, Does machine learning really work?, *AI magazine*, vol. 18, no. 3, pp. 11–11.

Molnar, Christoph. 2019, *Interpretable Machine Learning*, https://christophm.github.io/interpretable-ml-book/.

Moreira, Gonçalo. 2021, “System Modeling. Available: https://learningcenter.fxstreet.com/education/learning-center/unit-3/chapter-1/definition-of-a-trading-sys

Morgan, James N & Sonquist, John A. 1963, Problems in the analysis of survey data, and a proposal, *Journal of the American statistical association*, vol. 58, no. 302, pp. 415–434.

News, Market Business. 2021, *Active portfolio strategy – definition and meaning*. Available: https://marketbusinessnews.com/financial-glossary/active-portfolio-strategy/.

de Prado, Marcos Lopez. 2020, *Machine Learning for Asset Managers*, Cambridge University Press.

Pring, Martin J. 2002, *Study guide for technical analysis explained*, vol. 5, McGraw-Hill New York.

Psichogios, Dimitris C & Ungar, Lyle H. 1992, A hybrid neural network-first principles approach to process modeling, *AIChE Journal*, vol. 38, no. 10, pp. 1499–1511.

Qiu, Mingyue & Song, Yu. 2016, Predicting the direction of stock market index movement using an optimized artificial neural network model, *PloS one*, vol. 11, no. 5, p. e0155133.

Sharpe, William F. 1966, Mutual Fund Performance, *The Journal of Business*, vol. 39, no. 1, pp. 119–138. Available: http://www.jstor.org/stable/2351741.

Sharpe, William F. 1994, The sharpe ratio, *Journal of portfolio management*, vol. 21, no. 1, pp. 49–58.

Steve Ward, Simon Cotterill. 2021, *Traders Mind Journal*, https://www.tradersmindjournal.com/.

Strobl, Carolin; Boulesteix, Anne-Laure & Augustin, Thomas. 2007, Unbiased split selection for classification trees based on the Gini index, *Computational Statistics & Data Analysis*, vol. 52, no. 1, pp. 483–501.

Swart, Justin-Niall. 2016, *Testing a price breakout strategy using Donchian Channels*, Master's thesis, University of Cape Town.

Thames, Hudson &. 2020, Meta Labeling (A Toy Example). Available: https://hudsonthames.org/meta-labeling-a-toy-example/.

Thompson, Michael L & Kramer, Mark A. 1994, Modeling chemical processes using prior knowledge and neural networks, *AIChE Journal*, vol. 40, no. 8, pp. 1328–1340.

Toloşi, Laura & Lengauer, Thomas. 2011, Classification with correlated features: unreliability of feature ranking and solutions, *Bioinformatics*, vol. 27, no. 14, pp. 1986–1994.

Vince, Ralph. 1990, *Portfolio management formulas: mathematical trading methods for the futures, options, and stock markets*, vol. 1, John Wiley & Sons.

Ward, Steve. 2010, *High Performance Trading: 35 Practical Strategies and Techniques to Enhance Your Trading Psychology and Performance*, Harriman House Limited.

Wasserstein, Ronald L & Lazar, Nicole A. 2016, *The ASA statement on p-values: context, process, and purpose*.

Wikipedia contributors. 2021a, *Classification — Wikipedia, The Free Encyclopedia*. [Online; accessed 16-May-2021]. Available: https://en.wikipedia.org/wiki/Statistical_classification.

Wikipedia contributors. 2021b, *Nonfarm payroll employment — Wikipedia, The Free Encyclopedia*. [Online; accessed 15-May-2021]. Available: https://en.wikipedia.org/wiki/Nonfarm_payrolls.

Wikipedia contributors. 2021c, *Portfolio — Wikipedia, The Free Encyclopedia*. [Online; accessed 15-May-2021]. Available: https://en.wikipedia.org/wiki/Portfolio.

Wikipedia contributors. 2021d, *Precision Recall Table — Wikipedia, The Free Encyclopedia*. [Online; accessed 11-May-2021]. Available: https://en.wikipedia.org/wiki/Precision_and_recall.

Wikipedia contributors. 2021e, *Return — Wikipedia, The Free Encyclopedia*. [Online; accessed 15-May-2021]. Available: https://en.wikipedia.org/wiki/Rate_of_return.

Wikipedia contributors. 2021f, *Strategy — Wikipedia, The Free Encyclopedia*. [Online; accessed 11-May-2021]. Available: https://en.wikipedia.org/wiki/Strategy.

Wikipedia contributors. 2021g, *Trend Following — Wikipedia, The Free Encyclopedia*. [Online; accessed 11-May-2021]. Available: https://en.wikipedia.org/wiki/Trend_following.

Wilcox, Cole & Crittenden, Eric. 2005, Does Trend Following Work on Stocks?, *The Technical Analyst*, vol. 14, , pp. 1–19.

Wu, Mu-En & Chung, Wei-Ho. 2018, A novel approach of option portfolio construction using the Kelly criterion, *IEEE Access*, vol. 6, , pp. 53044–53052.

Zhu, Ruoqing; Zeng, Donglin & Kosorok, Michael R. 2015, Reinforcement learning trees, *Journal of the American Statistical Association*, vol. 110, no. 512, pp. 1770–1784.

Zhu, Yingzi & Zhou, Guofu. 2007, Technical analysis and theory of finance, In: *EFA 2007 Ljubljana Meetings Paper*.

Zicari, Roberto V; Brodersen, John; Brusseau, James; Düdder, Boris; Eichhorn, Timo; Ivanov, Todor; Kararigas, Georgios; Kringen, Pedro; McCullough, Melissa; Möslein, Florian et al.. 2021, Z-Inspection®: A Process to Assess Trustworthy AI, *IEEE Transactions on Technology and Society*.