Machine Learning for Quantitative Investment

Fintelligence

2022-04-03

Contents

Preface

Welcome to the first edition of Machine Learning for Quantitative Investment (ml4quant)! This is an online book all about machine learning and data science methods in quantitative investment presented by Fintelligence Academy: https://fintelligence-academy.github.io/.

The adoption of Machine Learning and Artificial Intelligence continues to progress from a niche activity to mainstream applications at an ever-accelerating pace in this Big Data era. Here at Fintelligence, we are embracing the ML in financial application.

In this book, we aim to walk through the ….

Having quantitative research in mind, the intended audience is presumed interdisciplinary, fluent in mathematical notation, familiar with basic data science concepts. You may be surprised at the neglect of econometric/finance elements - we are confident to teach you the foundational knowledge of financial market along the reading journey. In addition, this book could be also interesting to those readers who are thinking of joining the quant workforce or data science community.

In another vein, this book also brings the mathematical foundations of basic machine learning concepts to the fore and collects the information in a single place so that a reader could learns both modern quantitative finance and machine learning. We assimilate and benefit core ideas from the two classical machine learning textbooks *An Introduction to Statistical Learning (ISL2)* and *The Elements of Statistical Learning (ESL)*, which are the best to-go reference books to learn machine learning. Here for ml4quant, we strive to strike a balance of difficulty in the middle level between the two books and season the topic with a new flavor of financial data science.

We genuinely hope that both the quantitative researchers and data scientists could enjoy reading this interdisciplinary book and find it helpful.

> "The best time to plant a tree is twenty years ago. The second best time is now." - Chinese Proverb by Poet Wang Bo
>
> " "-

Chapter 1

Introduction

Already, artificial intelligence is all around us, from self-driving cars to virtual assistants and software that translate or invest. Impressive progress has been made in AI in recent years, driven by exponential increases in computing power and by the availability of vast amounts of data, from software used to discover new drugs to recommendation algorithms used to predict our cultural interests.

Machine learning is the latest in a long line of attempts to distill human knowledge and reasoning into a form that is suitable for constructing computer automated systems. The enthusiastic practitioner who is interested to learn more about the magic behind successful machine learning algorithms currently faces a daunting set of pre-requisite knowledge:

- Programming languages and data analysis tools
- Large-scale computation and the associated frameworks
- Mathematics and statistics and how machine learning builds on it

Machine learning builds upon the language of mathematics to express concepts that seem intuitively obvious but that are surprisingly difficult to formalize. Once formalized properly, we can gain insights into the task we want to solve.

An underlying theme is that the acceleration of innovation and the velocity of disruption are hard to comprehend or anticipate and that these drivers constitute a source of constant surprise, even for the best connected and most well informed. Indeed, across all industries, there is clear evidence that the technologies that underpin the Fourth Industrial Revolution are having a major impact on businesses.

collaborative innovation

and relentlessly and continuously innovate.

Chapter 2

Elements of Machine Learning

In a minimalism style, machine learning refers to how computers can "learn" by finding patterns in data and using them to make predictions. Mathematically, given a real-valued output y and predictors vector X containing p variables, we assume a general function f to describe y:

$$y = f(X) + \epsilon \tag{2.1}$$

Where, f is a fixed but unknown function, and ϵ is a zero-mean random error term, which is supposed to be independent of X.

The core of machine learning is a suite of data-driven algorithms for estimating f. ML is based on *statistical learning theory* to design models to understand patterns and employs *optimization algorithms* to train the model to "learn" the pattern using input data.

The foundation of practical machine learning is the data. Data drives everything else. The model can not learn much pattern without enough data and could even have biased behaviour if the data quality is poor. In contrast, with substantial data, the machine learning model could achieve impressive results beyond expectation. A vivid example is to check the google's search engine - it is machine learning algorithm under the hood.

> "All models are wrong, but some are useful." - George E.P. Box

2.1 Linear Regression

Linear regression adopts a linear functionf^{Linear} to equation (2.1) with learnable model parameters $\beta^T = (\beta_0, \beta_1, ..., \beta_p) \in (p, 1)$:

$$y = \beta^T X + \epsilon \tag{2.2}$$

Here we assume the first column of X is all ones as an "intercept feature" and thus β_0 corresponds to the intercept term.

The model given by (2.2) defines the *population regression line*, which is the best linear approximation to the true relationship between X and u. The population regression line is unobserved and we have access to a set of sample observations to compute an sample-based estimate line. Fundamentally, we apply a standard statistical approach of using information from a sample to estimate characteristics of a large population. Here the linear regression focus on the estimate of coefficient $\hat{\beta}$.

The most popular estimation method is *least squares*, also known as Ordinary Least Squares (OLS). OLS regression is an estimated model based on sample data in which we pick the coefficients $\hat{\beta}$ to minimize the residual sum of squares (RSS):

$$RSS(\beta) = \sum_{i=1}^{N} (y_i - f(x_i))^2 = (y - X\beta)^T (y - X\beta) \tag{2.3}$$

To minimize (2.3), we differentiate the term with respect to β and obtain:

$$\frac{\partial RSS}{\partial \beta} = -2X^T(y - X\beta)\frac{\partial RSS}{\partial \beta \partial \beta^T} = 2X^T X$$

Assuming that X has full column rank, hence $X^T X$ is positive definite leading to optimality, we thus could set the first derivative to zero to get the $\hat{\beta}$:

$$X^T(y - X\beta) = 0 \Rightarrow X^T y = X^T X \hat{\beta}$$

We thus solve the unique solution:

$$\hat{\beta} = (X^T X)^{-1} X^T y \tag{2.4}$$

Please note all this derivation makes no assumptions about the validity of model - it simply finds the best linear fit to the data except from ensuring X is of full-rank. The non-full-rank case occurs most often when there is redundancy.

A natural way to resolve the non-unique representation is to drop redundant columns in X.

Lastly, since $\hat{y} = X\hat{\beta}$, we could form the predictions

$$\hat{y} = X\hat{\beta} = (X(X^TX)^{-1}X^T)y \tag{2.5}$$

We denote $P = X(X^TX)^{-1}X^T$ which is projection matrix. Moreover, we could represent the residual $\hat{\epsilon}$ with P:

$$\hat{\epsilon} = y - X\hat{\beta} = y - (X(X^TX)^{-1}X^T)y = (I_n - P)y \tag{2.6}$$

Note $Q = I_n - P$ is also a projection matrix. There are nice properties of *projection matrices* that we will use in the subsequent sections:

$$Q^T = Q;\ Q^2 = Q$$

2.1.1 Sampling Uncertainty and Statistical Inference

We infer from (2.4) that both $\hat{\beta}$ and $\hat{Y} = X\hat{\beta}$ are linear transformations of y as random variables. If we had collected different sample data of size n, we would to be sure have different estimates, due to *sampling uncertainty.* We want to quantify this uncertainty by using *sampling statistics* and essentially get more control of our estimate coefficient $\hat{\beta}$. Furthermore, the inclusion of uncertainty measure such as standard error enables us to conduct statistical inference.

To pin down the sampling properties of $\hat{\beta}$, we include supplementary assumptions including *uncorrelated residual*, *constant variance of residual* σ^2, and additionally X *is non-random.* Based on the assumption, we infer the variance of y:

$$Var(y) = \sigma^2 I_n \tag{2.7}$$

We could then derive the expectation of residual sum of squares (RSS):

$$RSS = Y^T(1-P)Y \Rightarrow E[Y^T(1-P)Y] = (n-p)\sigma^2$$

Hence we get the unbiased estimate of variance $\hat{\sigma}^2 = \frac{RSS}{n-p}$. We next derive the variance-covariance matrix of the $\hat{\beta}$ using equation (2.7):

$$\begin{aligned} Var(\hat{\beta}) &= Var(X(X^TX)^{-1}X^T) \\ &= X(X^TX)^{-1}X^TVar(y)(X(X^TX)^{-1}X^T)^T \\ &= \sigma^2(X^TX)^{-1} \end{aligned}$$

Therefore, the uncertainty in individual estimate β_j is quantified by its *standard error (SE)*

$$SE(\beta_j) = \sqrt{\frac{\sigma}{(X^TX)^{-1}_{jj}}}$$

2.1.2 Rethinking the Unbiasedness of OLS

We require a property of unbiasedness for estimator $\hat{B}$. Unbiased estimator does not systematically over- or under-estimate the true parameter B.

Consider the decomposition of mean squared error of an estimator $\tilde{\theta}$ in estimating θ:

$$\begin{aligned} MSE(\tilde{\theta}) &= E[\tilde{\theta} - \theta]^2 \\ &= Var(\tilde{\theta}) + [E(\tilde{\theta}) - \theta]^2 \end{aligned}$$

Here the first term $Var(\tilde{\theta})$ is the variance, while the second term $[E(\tilde{\theta}) - \theta]^2$ is the squared bias.

The Gauss-Markov (GM) theorem implies that the least squares estimator has the smallest mean squared error (MSE) of all *linear estimators with no bias* (pay attention to the "no bias" here). However, there may well exist a biased estimator with smaller mean squared error. Such an estimator would *trade a little bias for a larger reduction in variance.* Put bluntly, *biased estimates are commonly used* for better out-of-sample performance (i.e., less MSE in test set). In reality, any method that shrinks or sets to zero some of the least squares coefficients may result in a biased estimate.

2.2 Assumptions of Linear Regression and Violation Implications

The question for the assumption of linear regression is ill-posed and requires more context as we need to specify the desired properties that we want the linear model to hold. Hence, we start with the standard assumptions that guarantee the above GM theorem to hold for OLS regression estimate: 1. There exists an additive linear model for (X, Y) as equation (2.1),

$$Y = X\beta + \epsilon$$

Where, β and X are *non-random* while the randomness stem from ϵ

2. Zero Expected Mean of Residual: $E[\epsilon_i] = 0$
3. Homoscedasticity: Constant Variance of Residual: $Var(\epsilon_i) = \sigma^2 < \infty$
4. Uncorrelated Residual Error: $Cov(\epsilon_i, \epsilon_j) = 0, \ \forall i \neq j$

Furthermore, if the error terms are 5. normally distributed and 6.identically and independently distributed (i.i.d.), we could infer that the OLS estimator becomes the Maximum Likelihood Estimation (MLE).

Problem 1: Non-linearity of the Data

The linear regression model assumes that there is a straight-line relationship between the predictors and the response. If the true relationship is far from linear, then all of the conclusions that we draw from the fit are suspect.

Residual plots are a useful visualization tool for identifying non-linearity. Given a fitted linear regression model, we can plot the residuals, $e_i = y_i - \hat{y}_i$, versus the predictor x_i. In the case of a multivariate regression model, we can plot the residuals versus the fitted values $\hat{y}_i$. Independence assumption implies no discernible pattern. The presence of a pattern may indicate a problem with some aspect of the linear model. For instance, if residuals exhibit a clear U-shape, which provides a strong indication of non-linearity in the data.

Problem 2: Correlation of Error Terms

An important assumption of the linear regression model is that the error terms ϵ_i are uncorrelated. Intuitively, if the errors are uncorrelated, then the fact that $epsilon_i$ is positive provides little or no information about the sign of ϵ_{i+1}. One counterexample is to think about double a set of training observations (X, y), as $X_{Copy} = [X; X], y_{Copy} = [y; y]$.

The standard errors that are computed for the estimated regression coefficients or the fitted values are based on the assumption of uncorrelated error terms. If in fact there is correlation among the error terms, then the estimated standard errors will tend to underestimate the true standard errors. As a result, confidence and prediction intervals will be narrower than expected.

Such residual correlations frequently occur in the context of time series data, which consists of observations for which measurements are obtained at discrete points in time. There is the issue of residual autocorrelation - tracking in the residuals that adjacent residuals may have similar values.

In general, the assumption of uncorrelated errors is extremely important for linear regression as well as for other statistical methods, and good experimental design is crucial in order to mitigate the risk of such correlations.

Problem 3: Outliers

An outlier is a point which is far from the value predicted by the model. Outliers can arise for a variety of reasons, such as incorrect data collection. They are quite common in real dataset.

Residual plots can be used to identify outliers. If we believe that an outlier has occurred due to an error in data collection, then one solution is to simply remove the observation. Shrewd care and revised assumptions should be taken when removing outliers, since an outlier may instead indicate a deficiency with the model, such as a missing predictor. 2008 global financial crisis is a data point that needs justification to be discarded as outlier.

Problem 4: Collinearity

Collinearity refers to the situation in which a group of variables are closely related to one another. The presence of collinearity can pose problems in the regression context, since it can be difficult to separate out the individual effects of collinear variables on the response. Collinearity reduces the accuracy of the estimates of the regression coefficients, it causes the standard error for $\widehat{\beta}_j$ to grow.

Correlation matrix of the predictors is a simple way to detect collinearity. Unfortunately, not all collinearity problems can be detected by inspection of the correlation matrix: it is possible for collinearity to exist between three or more variables even if no pair of variables has a particularly high correlation - this is called *multicollinearity.*

There are in two practical solutions to the problem of collinearity. The first, and more straightforward one, is to *drop* one of the problematic variables from the regression. The second solution is to *aggregate* the collinear variables together into a single composite predictor. For example, apply PCA (Principal Component Analysis) approach to model the highly correlated features group and extract the first PC component as the representative indicator.

2.3 Conclusion

Linear models were largely developed in the precomputer age of statistics, but there are still robust baseline model to apply in today's computer era. They are simple and often provide an adequate and interpretable description of how the inputs affect the output. For prediction purposes they can sometimes outperform fancier nonlinear models, especially in situations with small numbers of training cases, low signal-to-noise ratio or sparse data.

Chapter 3

Modern Factor Investing

We give an overview of modern factor investing in this chapter. It is intended to be an introductory chapter to help readers get familiar with the quantitative finance framework. This chapter is adapted from our Webinar series *Modern Factor Investing: Past, Present, Future.*

Over the last 50 years, academic researchers have made major breakthroughs in advancing classical practice in finance, from fundamental analysis to scientific quantitative research. These include portfolio theory, corporate finance, financial engineering of derivative instruments, and many other applications pertaining to financial markets overall. One of the foundational pillar is the advances in capital market theory in the context of descriptive equilibrium propositions in terms of the risk/return tradeoff and the linear factor model to explain the expected asset price, beginning with William F. Sharpe and the *Capital Asset Pricing Model (CAPM)* (Sharpe, 1964). Many related academic developments provided rich asset pricing insight, including *Arbitrage Pricing Theory (APT)* (Ross, 1976), market efficiency proposition, market anomalies such as the *Fama and French factor model* (Fama and French, 1993), and behavioral finance. These altogether gave birth to a new investment paradigm - *factor investing* and is widely adopted in finance industry - *classical finance textbooks usually end there.*

The saying, *"change is the only constant in life"* is attributed to the Greek philosopher, Heraclitus around 500BC. This is even more prominent in the capital market. The evolution of financial market keeps testing the efficacy of the factor theory and model. Ever since the 2008 global financial crisis, we observe an seemingly irreversible performance decay of the traditional factor model that is solely based on conventional data and linear framework. In academia, there is also a replication crisis (or credibility crisis) raging on as *"most claimed research findings in financial economics are likely false."* (Harvey et al., 2016)

Perhaps the *Fourth Industrial Revolution* - *digitalization* - has already come, and *"AI is the new electricity"*. During this trend, factor models and strategies have not remained static. Factor models and investment strategies evolve gradually over time, reflecting the evolution of capital markets and advances in data science theories, data availability, and investment practice. The current innovation efforts in factor investing focus on exploiting new unstructured data sources - the *alternative data*, applying new data science modeling techniques - *machine learning* - *modern factor investing starts here.*

3.1 Overview of Factor Investing

> One of our central themes is that if assets are priced rationally, variables that are related to average returns, such as size and book-to-market equity, must proxy for sensitivity to common (shared and thus undiversifiable) risk factors in returns — Fama and French (1993)

The question of what drives asset returns has been a staple of neoclassical finance and remains a holly grail. The factor model emerged and became a foundation of asset pricing theory in the 1960s (Lintner, 1969; Mossin, 1966; Sharpe, 1964; and Treynor, 1961). In general, *Factor* can be defined as any underlying characteristic relating a group of assets (e.g., equity, bond) that is linearly significant in explaining their return and risk.

Factor investing refer to the general use of factors in the investment process. It is trustworthily a quantitative paradigm in investment where factor models renders return prediction signal, forecast expected asset risk, and drive asset allocation decisions by using the tradable factor portfolios. Practically, factor-driven investment strategies such as smart beta products amount to \$800 billion AUM worldwide (Johnson, 2018) — yes, we are talking about a magnificent investment topic that may affect everyone's saving pension.

As is recognized by Fama and French (1993), firm characteristics (or factors) are associated with the cross-section of expected returns. Assuming individual asset return $r_{i,t}$, factor loadings $z_{i,t}$ and factor return f_t, factor model linearly specifies the relationship:

$$r_{i,t} = z'_{i,t-1} f_t + \alpha_{i,t} + u_{i,t} \tag{3.1}$$

Where, $u_{i,t}$ it the noise term with expected zero return and $\alpha_{i,t}$ is the intercept term of common OLS regression.

Now that we have the mathematical formulation, we delve into the granular categories of the factors:

- **Beta factors** — factors corresponding to f_t are generally referred as "risk premia factors", which have reflected exposure to sources of systematic risk and are supposed to earn a persistent significant risk-compensated premium over long periods (*if all model assumptions work...*).

  In practice, they are called "beta factors" or "style factors" to denote that they are the drivers of systematic return and risk of assets. Furthermore, they are usually associated with the systematic risk that is inherent in market.

  Here the risk premium reward could be thought as selling insurance products to Mr. Market. You undertake the risk to receive a positive reward on average in the long-term, when the risk does happen, Mr Market will come to you for coverage and you suffer a realized loss that should be well anticipated. If you believe the game is fair then you agree with the philosophy of passive investing.

- **Alpha factors** — some other factors fall into the $\alpha_{i,t}$ term as they do not explain the risk well (e.g., *they are not volatile*) and can not be included into the linear terms, yet they do earn persistent return over time with some predictive value to result in $\alpha_{i,t} \neq 0$.

  Alpha factor model are designed to forecast excess return of stocks. If return distribution is characterized by the expected return and the standard deviation, it is often the expected return predicted by alphas that determines whether we buy or sell, overweight or underweight, and the standard deviation driven by betas that determines the size of the portfolio allocations to hedge the undesirable systematic risk - this is the monologue of hedge funds or active investment.

  Efficient market theory forbids the existence of alpha and the academia call them *"anomaly"*, however, hedge fund active managers rely on them and *some HFs seem to live well.* The alpha factors are often called "signal" by practitioners to distinguish from the beta risk premia factors.

  The alpha signals are often proprietary and highly guarded, reflecting creativity as well as superior systems. It is the most important differentiator within the investment firm.

Factors are the bedrock of quantitative investment management and have served the needs of investors for a long time. For equities, in addition to countries and industries, accounting-based fundamental factors documented through empirical research and used extensively in portfolio management include value, size, momentum, volatility, quality, yield, growth, and liquidity. A similar set of common macro drivers has been identified and used across asset classes, including equity, rates, credit, and real assets.

3.2 50 Years of Factor Investing: A Literature Review

Factor research has been prevalent for over 50 years. The oldest and most well-known model of stock returns is the *Capital Asset Pricing Model (CAPM)*, which is in essence a single-factor model with market as the sole factor (see Sharpe (1964)). In other words, the expected return to any stock could be viewed as a function of its beta factor loading to the market . This won William Sharpe a Nobel Prize.

Later, Ross (1976) proposed the *Arbitrage pricing theory (APT)* holds that the expected return of a financial asset can be modeled as a function of various macroeconomic factors or theoretical market indexes, thus introducing the "multi-factor models". Moreover, unlike the CAPM, APT theoretically validated that the number and nature of the APT factors were likely to change over time and vary across markets. Hence, the challenge shifts to build up the empirically sound factor models.

Beyond the market factor, researchers generally look for risk premium factors that are *persistent over time* and have strong explanatory power over a *broad range of stocks*. There are three main categories of factors today: macroeconomic, statistical, and fundamental.

- Macroeconomic factors include measures such as surprises in inflation (Chen et al., 1986)
- Statistical factor models identify factors where the factors are not prespecified in advance (Chamberlain, 1983)
- Fundamental factors are the mostly widely used and credit to Fama and French (1993). They put forward a model explaining US equity market returns with three fundamental actors: the *Market*, the *Size*, and the *Value*. The "Fama-French" model, which today includes Carhart (1997) *Momentum* factor, has become a canon within the finance literature. – Fama also later got a Nobel Prize.

In another vein, Rosenberg (1974) described the importance of the stock fundamental traits in explaining variation of stock returns, leading to the creation of the *multi-factor Barra risk models* which is still the dominant risk model in equity investment.

3.3 Limitedness of Classical Factor Model and the Factor Zoo

The enduring success of fundamental factors stems from the fact that they are guided by academic theory, and, importantly, are supported by empirical evidence and reflect investment practice, i.e., they make money. However, the story of factors turn into a horror fiction ever since the 2008 Global Financial Crisis. It is not only that some well-acclaimed factors that outperform in the past start to underperform (Fama-French's Value factor is lacklustre, see Figure 3.1), but also there are some lurking black swan events that causing extreme outliers where the model could fail drastically. In statistics, this is usually associated with the *heavy-tailed* distribution and we could remove or winsorize those data points as outliers when fitting a model. Nevertheless, in reality, these outliers may mean you are doomed all at once.

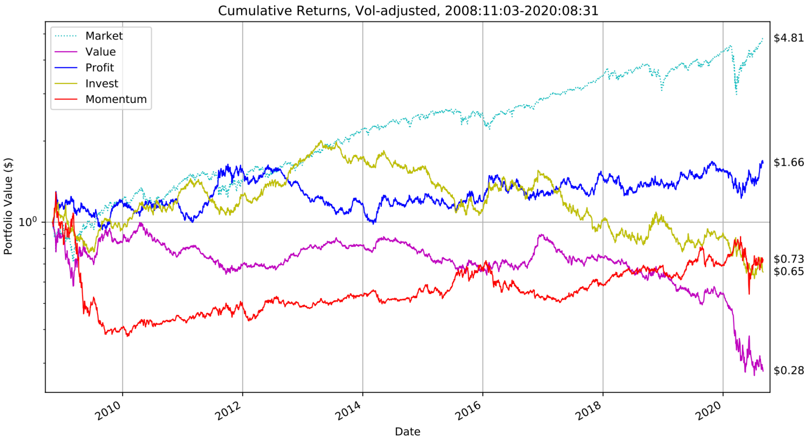


Figure 3.1: Recent performance decay of Fama-French factor portfolios in US market

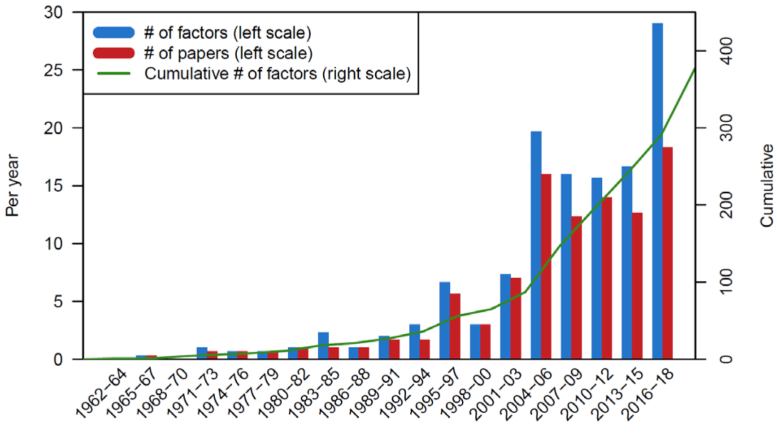
To make a concrete example, lets remember David Viniar's, once the CFO of Goldman Sachs, famous *25 sigma* when Goldman's flagship GEO hedge fund had lost 27% in 2007: *"We were seeing things that were 25-standard deviation moves, several days in a row."*. One commentator wryly noted:

> That Viniar. What a comic. According to Goldman's mathematical models, August, Year of Our Lord 2007, was a very special month. Things were happening that were only supposed to happen once in every 100,000 years. Either that ... or Goldman's models were wrong (Bonner, 2007b).

A direct implication is that the normality assumption of linear models does not hold in financial market, thereby significantly weakening the *statistical inference* power in quantitative fiance. The p-value and relevant test statistics should thus be interpreted with extreme caution or neglected altogether. We empirically recommend the latter.

More importantly, the underlying linearity assumption itself is an unrealistic assumption and we have witness a great amount of evidence on *non-linearity* breaking this fundamental assumption. One of the straightforward counterexample is the interaction effect as one factor's contribution to the asset's expected return may rely on other factors. Furthermore, there could be more complex pattern that certain factor exerts on the asset return and risk. You should forget about the theoretical unbiasedness credit to linear model when entering the quantitative finance.

Another interesting trend is the "factor zoo" (Harvey and Liu, 2019), which refers to a metaphor for the growing number of investment factors proposed by both academics and practitioners. Not considering the less famous journals or the mysterious industry, there are already hundreds of factors claimed to be effective and published in top academic journals.



* Journals published through December 2018. Data collection in January 2019.

Figure 3.2: Out of control factor production (Harvey, 2019)

It further points out to the multiple testing problem: *with so many factors tried, some will appear "significant" purely by chance*, and further indicates the peril of data mining, which may or may not be intentional. From a data science perspective, this also leads to the *high-dimension* nature of the modern factor investing: we now have a "zoo" of available factor candidates and the traditional OLS linear regression is incompetent.

This is not to overlook the merits of linear model such as its transparency and interpretability. Given careful treatment and enhacncement such as regularization, linear model remains a competiteive baseline, especially in set-up with insufficient data input. However, its performance upper bound should not be inflated in the current data science world, with so many alternative and more expressive machine learning algorithms available: linear model is at most a parsimonious underfitted compromise (*think about those hundreds of factors!*) in face of the high noise-to-signal ratio financial data.

3.4 Modern Factor Investing in the Wave of Big Data and Machine Learning

What is Big Data? A wide array of information sensing devices as well as the cloud service technique has led to an exponential growth in available big datasets. Starting from handheld mobiles, cameras, microphones, Internets, digitalised map, Internet of Things (IoT), to LiDAR (Light Detection and Ranging) sensors and satellite imaging. The amount of data that the world collects has experienced explosive growth. The rapid development of informative technologies has paved the way for a transition to digital world.

The rapid advancement in the Internet and social media sectors has far reaching effects not only on our day to day lives; but also, for finance researchers and investors. Quantitative finance is no longer a subclass of finance but arguably the whole finance is evolving into a data-driven scientific subject. There comes a fashionable term in the finance domain - *alternative data*.

Alternative data is data not typically used by researchers and model builders and can be sourced from a kaleidoscope of data sources ranging from text, audio, image, graph, etc. These data sets typically have minimal aggregation and processing making them more *difficult* to access and use. In the last several years, we have seen an explosion in the availability of new alternative data sources.

- Some of the alternative data sets are market data that were previously difficult to access (e.g., analyst forecasts, insider transactions, options trading information, credit default swaps, and hedge fund positions). These data sets are available in numerical format and can be sourced from commercial vendors.

- Other new data sets come in new atypical formats (e.g., text, audio, and video), and they can only be extracted from internet sites and social media. Investors may gain additional insights by processing and understanding the information contained in unstructured alternative data sets such as news reports, product reviews, employee reviews, job postings, regulatory filings, call transcripts, satellite images, etc.

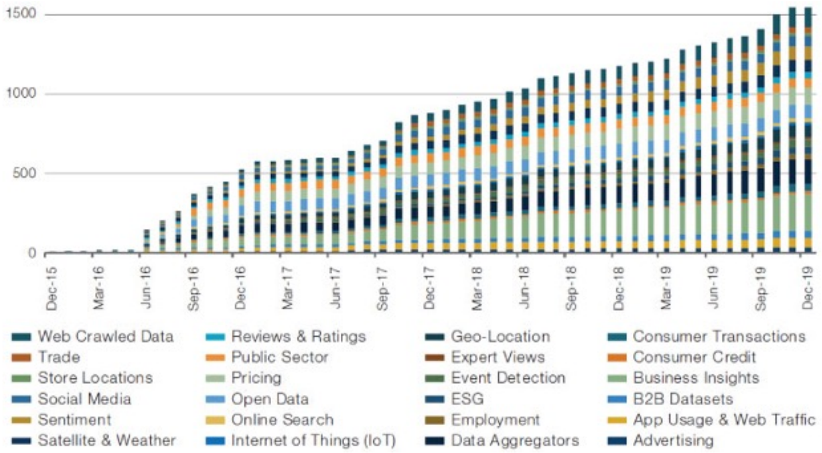


Figure 3.3: Increasing number of alternative data vendors (Man Institute, 2019)

The new formats require new modeling techniques and powerful computers to extract information. Innovations in information technology such as cloud computing and bespoke microprocessors provide the computational firepower necessary to apply *machine learning* and other computationally intensive data science techniques on the new unstructured alternative data sets.

Natural language processing (NLP) is one of the most exciting frontiers to bring alternative value to the factor investing framework. In essence, we could now leverage deep-learning based NLP model to extract value from text data in the financial domain. Financial sector accumulates large amount of financial communication text. It is only the recent machine learning breakthrough of general *Natural Language Understanding (NLU)* language model such as BERT (Devlin et al., 2018) that makes the unstructured text data become a goldmine to generate value for quantitative finance. Due to its unique data source, NLP-based factors tend to have a merit of *being orthogonal to existing fundamental factors* with less multicollinearity issue. It is thus pivotal to note that NLP-based factors could actually complement the existing factors, instead of the disrupting them overall. We will give a detailed introduction of Financial NLP in the latter sections, which well deserve a chapter in its own right.

In another vein, much of the efforts have been focused on employing modern high-dimensional machine learning techniques to better model the asset returns now given the affluence of data source. Applying machine learning routines to financial data has been implicitly motivated by the AFA (American Finance Association) presidential address of Cochrane (2011), who suggests that in the presence of a vast collection of noisy and highly correlated return predictors, there is a need for other methods beyond cross-sectional linear regressions. Indeed, *machine learning* offers a natural and theoretical-sound way to accommodate a high-dimensional predictor set and flexible functional forms, and it

Table 3.1: Factor machine learning papers, up till December 2021

ML.Models	Linearity	Factor.Model	Predictors	Paper	Name
Linear Shrinkage	Linear	Pricing Kernel	Firm	NKS, 2020	Shrinking the cross-section
IPCA	Linear	Beta Pricing	Firm	KPS, 2019	Characteristics are covariances: A
Auto Encoder	Nonlinear	Beta Pricing	Firm	GKX, 2019	Autoencoder asset pricing models
Reg Tree	Nonlinear	Reduced Form	Firm + Macro	GKX, 2020	Empirical asset pricing via machi
GNN	Nonlinear	Pricing Kernel	Firm + Macro	CPZ, 2019	Deep learning in asset pricing
Transformer	Nonlinear	Residual Term	Firm	JMG, 2021	Deep learning statistical arbitrage

employs “regularization” methods to select models, mitigate overfitting biases, and uncover complex patterns and hidden relationships.

There has been an emerging body of pioneer academic work that reports phenomenal investment profitability based on signals generated by machine learning methods, let alone the long data science enthusiasm from investment industry. A funny observed phenomenon to echo this trend is that nowadays Machine Learning PhDs are more likely to land a researcher job in quantitative Hedge Funds than those pure Finance PhDs.

We list some most influential machine learning in factor investing papers in the table 3.1 as a roadmap of the recent modern factor investing. The predictors column refer to the input features the research use to predict the stock returns, where *Firm* stands for typical firm-level factors and *Macro* represents the macro indicator. In the forthcoming sections we will deep dive into some most representative ones (e.g., IPCA by Kelly et al. (2019)) alongside introducing the corresponding machine learning algorithm.

3.5 Chapter Conclusion

Factor investing is a *paradigm in quantitative finance* of more than 50 years history by linearly describing the asset return and risk using the effective and intuitive asset characteristics - *factors*. It enjoys academic acclaim and is fundamentally adopted by industry.

The recent performance decay of conventional factors points to the limitedness of the classical linear factor model. Moreover, the continued growing number of investment factors proposed by both academics and practitioners leads to a high-dimensional “factor zoo” in current factor world and the traditional linear regression thus become incompetent.

Current innovation efforts in factor investing focus on exploiting new unstructured data sources, applying new data science modeling techniques such as machine learning to harness the high-dimensional factors and identify the evolving complex market payoff pattern.

> “Change is the only constant in financial market.” - Fintelligence

Chapter 4

Tree-based Models

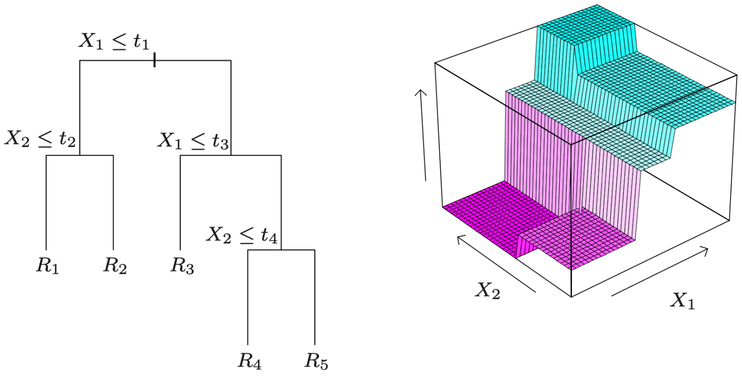
We give an overview of tree-based machine learning model in this chapter, which is currently the *GOAT* prediction model for tabular dataset. Neural networks dominate image recognition and language processing tasks, while tree-based methods are often the method of choice for tabular data such as Kaggle competition. This chapter is cuurently working in progress.

4.1 Introduction to Machine Learning Tree

Tree-based methods partition the feature space into a set of rectangles, and then fit a simple model (like a constant) in each one. They are conceptually simple yet powerful.

Mathematically, given two predictors variables X_1,X_2, a tree model $\hat{f}_{tree}$ is formulated as a sequence of K recursive binary partitions based on the feature X to fit the target variable Y. The corresponding regression model predicts Y with a constant c_m in rectangle region R_m as:

$$\hat{f}_{tree}(x) = \sum_{m=1}^{K} c_m I\{(x_1, x_2) \in R_m\}$$



We now turn to the question of how to *grow* a regression tree. The algorithm needs to automatically figure out the splitting variables, split points, as well as what topology (shape) the tree should have. Given a partition into K regions $R_1, R_2, ..., R_K$, we could start from use a constant c_m in each region to represent the shape:

$$f_{tree}(X) = \sum_{m=1}^{K} c_m I\{x \in R_m\}$$

Adopting the sum of squares $\sum(y_i - f_{tree}(x_i))^2$ as objective to minimize, the best c_m will be the average of y_i within the region R_m:

$$\hat{c}_m = avg(y_i | x_i \in R_m)$$

Finding the best binary partition in terms of minimum sum of squares is generally computationally infeasible. Hence we proceed with a greedy algorithm. Starting with all of the data, consider a splitting variable j and split point s, and define the pair of half-planes $R_L(j,s) = X|X_j \le s,\ R_R(j,s) = X|X_j > s$. We seek the splitting variable j and split point s that solves

$$\min_{j,s}[\sum_{x_i \in R_L(j,s)} (y_i - \hat{c}_L) + \sum_{x_i \in R_R(j,s)} (y_i - \hat{c}_R)]$$

Having found the best split, we partition the data into the two resulting regions and repeat the splitting process on each of the two regions. Then this process is repeated on all of the resulting regions until a stopping criterion is triggered.

How large should we grow the tree? A very large tree might overfit the data, while a small tree might not capture the important structure. Tree size is a tuning parameter governing the model's complexity, and the optimal tree size should be adaptively chosen from the data. The preferred strategy is to grow a large tree T, stopping the splitting process only when some minimum node size (e.g. 6 data point nodes) is reached. This large tree is next reversely (bottom-up) *pruned* by using *cost-complexity pruning*.

Tree pruning refers to collapse any number of its internal (non-terminal) nodes from tree T_0 into a single subtree $T \subset T_0$. Pruning a tree thus removes the split and helps to prevent overfitting the training data. We denote the number of terminal nodes in T as $|T|$, next we derive the cost complexity criterion with hyperparameter α to determine the T:

$$C_\alpha(T) \sum_{m=1}^{|T|} \sum_{x_i \in R_m} (y_i - \hat{c}_m)^2 + \alpha|T|$$

The idea is to find a tree T_α that could minimize $C_\alpha(T)$. Here the tuning α governs the tradeoff between tree size and its goodness of fit to the data. Small values of α result in larger trees T_α. With $\alpha = 0$, the solution is the full tree T_0.

4.2 Boosting and Additive Trees

Boosting is a predictive model paradigm developed for additive expansions by building models $f_{(b)}(x)$ sequentially. This involves starting with a null model and then repeatedly fitting a small model to the residuals of the current model and updating the current model with a down-weighted or ‘shrunken’ version of that weak models.

$$\hat{f}^{boost}(x) = \sum_{b=1}^{B} \beta^b f_{(b)}(x; \theta_b) \tag{4.1}$$

4.3 Gradient Boosting Tree

The most popular variations of boosting are gradient boosting tree (GBT) and its modern variations (e.g. LightGBM, XGBoost, CatBoost). Decision trees $f_{(b)}^{tree}(x; \theta_b)$ are put as standard building block in the gradient boosting tree. GBT grows the tree sequentially by iteratively fitting the $f_{(b)}^{tree}(x)$ through a gradient descent optimization manner with a shrinkage regularization parameter learning rate $\lambda \in (0, 1)$ to avoid overfitting.

$$\hat{f}^{GBT}(x) = \sum_{b=1}^{B} \lambda^b f_{(b)}(x; \theta_b) \tag{4.2}$$

4.3.1 Algorithm

We show the current standard algorithm to train a gradient boosting tree.

Algorithm: Gradient Boosting Tree

Input: sample predictors X, target response y of total data sample size N

Parameters: B: the number of boosting stages to perform, λ: learning rate shrinks the contribution of each tree, η: the fraction of subsamples to be used for fitting the individual base learner tree, $d \in (1, 2, ...)$: the maximum depth of the individual regression estimators

1. Set $r_0 = y$ as initial point and fit a initial tree model $f_{(0)}^{GBT}(x; d)$ on sample training data (X, r_0)
2. For $b = 1, 2, ..., B$, repeat:
   (a) [*functional gradient descent*] For each data point $j = 1, 2, ..., N$, compute the gradient descent $r_b = (r_{b,1}, r_{b,2}, ...)^T$ at stage b based on the pre-specified loss function $\mathcal{L}$:
   $$r_{b,j} = -\left[\frac{\partial \mathcal{L}(y_j, f(x_j))}{\partial f(x_j)}\right]_{f=f_{(b-1)}^{GBT}}$$
   (b) [*stochastic subsampling*] sample a fraction η of the training observations to growth the next trees using the subsample $\tilde{X}, \tilde{r}_b$
   (c) Fit a base regression tree $f_{(b)}$ with d splits to the subsample training data $(\tilde{X},\ \tilde{r}_b)$
   (d) [*shrunk weighting*] Update $f_{(b)}^{GBT}(x) = f_{(b-1)}^{GBT}(x) + \lambda f_{(b)}(x)$

End the training

Output: the boosted model $\hat{f}^{GBT}(x) = f_{(B)}^{GBT}(x) = \sum_{b=0}^{B} \lambda^b f_{(b)}(x; d)$

4.4 Chapter Conclusion

While the forecasting performance of individual trees is notoriously poor, ensembles of trees built by boosting are currently considered the most successful approach to classification and regression for tabular data.

> “Finance is not a plug-and-play subject as it relates to machine learning.” - Marcos Lopez de Prado (2018), Advances in Financial Machine Learning

Bibliography

Carhart, M. M. (1997). On persistence in mutual fund performance. *The Journal of finance*, 52(1):57–82.

Chamberlain, G. (1983). Arbitrage, factor structure and mean-variance analysis in large asset markets. *Econometrica*, 51:1305–1324.

Chen, N.-F., Roll, R., and Ross, S. A. (1986). Economic forces and the stock market. *Journal of business*, pages 383–403.

Devlin, J., Chang, M.-W., Lee, K., and Toutanova, K. (2018). Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

Fama, E. F. and French, K. R. (1993). Common risk factors in the returns on stocks and bonds. *Journal of Financial Economics*, 33:3–56.

Harvey, C. R. and Liu, Y. (2019). A census of the factor zoo. *Available at SSRN 3341728*.

Harvey, C. R., Liu, Y., and Zhu, H. (2016). … and the cross-section of expected returns. *The Review of Financial Studies*, 29(1):5–68.

Johnson, B. (2018). Strategic beta etf market still expanding. *Morningstar Research*.

Kelly, B. T., Pruitt, S., and Su, Y. (2019). Characteristics are covariances: A unified model of risk and return. *Journal of Financial Economics*, 134(3):501–524.

Lintner, J. (1969). The valuation of risk assets and the selection of risky investments in stock portfolios and capital budgets: A reply. *The review of economics and statistics*, pages 222–224.

Mossin, J. (1966). Equilibrium in a capital asset market. *Econometrica: Journal of the econometric society*, pages 768–783.

Rosenberg, B. (1974). Extra-market components of covariance in security returns. *Journal of Financial and Quantitative Analysis*, 9(2):263–274.

Ross, S. A. (1976). The arbitrage theory of capital asset pricing. *Journal of Economic Theory*, 13(3):341–60.

Sharpe, W. F. (1964). Capital asset prices: A theory of market equilibrium under conditions of risk. *The journal of finance*, 19(3):425–442.

Treynor, J. L. (1961). *Toward a theory of market value of risky assets*.