SECURITIES FUNDAMENTALS AND PRICE OUTCOMES FROM 1990 TO 2010

Elijah Moree, Stone Emory, Zack Marin, and ShaoYang “Derrick” Gu

A Machine Learning project satisfying the requirements of ECON 573.

Chapel Hill

2024

Introduction

Longstanding investment theory has relied primarily on ad-hoc analysis of a company's assets, obligations, and prospects. Figures such as Warren Buffett and Benjamin Graham popularized the philosophy of "Value Investing," seeking to compare all market prices for company stock to the inherent value the analyst ascribes to the stock's particular division of interest in that company. Value investing may be shown at times to be in opposition to quantitative analysis, which does not stress any particular philosophy of the market, but instead seeks data-driven strategies for capitalizing on market anomalies. The object of this paper is to, in some sense, marry the two, and be able to establish and weigh the most pertinent predictors of a company's success using machine learning methods. Specifically, we hypothesize that we can validate fundamental, value-oriented security analysis by analyzing predictors associated with price outcomes.

The interest in this question is twofold. Firstly, the self-interest the reader might have in growing their stock portfolio might lead the reader to understand the predictors of company success. Secondly, we seek to contribute to the proper analysis of company value. Accelerating market efficiency leads to more productive capital allocation, which has obvious positive downstream effects. We help answer the question of the utility of value strategies over extended periods. In either case, whether prevailing wisdom is validated or critiqued, market knowledge increases and our work improves upon the corpus of decision-making in either a positive (the intelligent investor *ought to* do X) or negative (the intelligent investor *ought to think twice about* X) manner.

Our study appears to have marginally confirmed fundamentals-oriented strategies. The data heavily weighs the effect of company size and average monthly trading volume (VolMkt). There were however very large effects to be found in a variety of seemingly value-adjacent factors, such as IdioVolAHT. These results were not particularly compelling however since in what exact direction they informed returns varied widely. We conclude that value investing appears to be supported by our research, considering that the two most consistently relevant predictors (i.e., independent of time period) appear to either control for firms with large amounts of resources (and likely entrenched market positions) and high trading volume (which is a proxy for investor confidence in the liquidity of a company's security).

Literature Review

Multi-factor models for asset pricing have been of academic interest for decades, with Fama and French's 1992 paper (Fama and French, 1992) having defined scholarly research for the past 30 years. Fama and French found that two factors, size and book-to-market value, could accurately predict cross-sectional stock values. The authors sampled data between 1963 and 1990, finding that a reduced set of factors could "absorb the roles of leverage and E/P in average stock returns".

Machine learning analysis of securities prices has built upon the work of Fama and French, analyzing an ever-growing "factor zoo" of predictors. Bagnara has noted that this problem has led to overfitting and multicollinearity, as well as issues with this sort of financial data which leads to the general issue that there are "...weak signal-to-noise ratios, which means that there is a high degree of randomness resulting from forces of competition and profit maximization that wipe out most of predictability" (Bagnara, 2024). A multiplicity of machine learning methods have thus been employed to identify complex relationships that are not observable from simple OLS regression.

There is a further difficulty, however, in finding true predictability. The idiom "hindsight is 20/20" is accurate in this scenario since, as Bagnara notes, an econometrician may find predictive power analyzing the data generally but none out of sample. The efficient market hypothesis would serve to explain this phenomenon, meaning that market forces digest the information and price it into securities at a rapid pace, thus eliminating the effectiveness of any particular investment strategy. This is further supported by the fact that sample testing produces predictability, however, the predictability experiences a "...substantial decay over time". In the same vein, Fama and French improved upon their model, creating a five-factor asset pricing model in 2015 (Fama and French, 2015) to capture a large amount of variance in returns with few predictors. There has been interest in dimension-increasing methods such as Deep Neural Networks (Karolyi, and Stijn Van Nieuwerburgh, 2020), but controlling the "factor zoo" continues as the primary object of research in this area.

Our findings are relevant from both a popular and an institutional level. Retail traders would be well served to understand the general predictors of company success, and institutional investment entities would be able to further parse the implied relative weight of their effect on returns. Value investing posits generally that market forces are near-random in the short term, but that sound investment strategies isolate asymmetric risk and produce returns over longer periods. Thus chunking into 10-year periods offers a reasonable sample size for identifying long-term success. Once we isolate these factors, we will be able to examine them and identify whether they confirm the importance of fundamental analysis, or appear to suggest no correlation between value factors and price outcomes.

The Model

The advantage of our approach is that we make very few initial assumptions about what our data ought to predict. This is a data-driven approach, where the only question is whether the results obtained are consistent with traditional investment theory or not. Factors consistent with value investment theory would however have certain traits. As one analyst put it, "Value investing is more focused on companies that are *well established and are delivering stable revenues and consistent profits*" (Garmhausen, 2023). In other words, value investing wants to identify patterns of dependability on which to base a reasonable evaluation of the company's prospects. In contrast, to falsify these propositions, we may either establish that returns are random concerning "value" factors, or that the most prominent predictors are factors that have very little association with common value investing consensus.

We approach the problem from a slightly different perspective compared to previous econometric machine learning research. Rather than looking to explain variance over all periods through a few isolated predictors, we seek to explain percent returns on price over 3 periods, 1990 to 2000, 2000 to 2010, and the summary of the two periods, 1990 to 2010. We believe that the focus on these three periods allows our methods to produce a set of predictors (and relative weights for them) that are independent of any particular intra-period fluctuation. Whether it be the stock bubble pop of March 2000, or the bull run ending in 2007, we want to understand the predictors that could construct a long-term, sound portfolio.

Data Description

We are using the work of Chen and Zimmerman's Open Source Asset Pricing project.

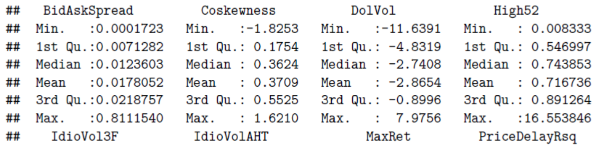
The project was initially an attempt to validate findings of asset pricing literature, finding that 98% of predictor results could be reproduced (Chen and Zimmerman, 2021). Chen and Zimmerman's data collects information drawn from previous meta-studies, providing 319 firm-level characteristics, drawn from CRSP, Compustat, and IBES, among other sources. We utilize the data according to the August 2023 release. Individual securities are identified by PERMNO, which is a permanent identification number utilized by WRDS (Wharton Research Data Services, where Chen and Zimmerman source most of their statistics) to track interest in a company through various legal changes (whether that be a name change, acquisition, etc). The data provided by Chen and Zimmerman extends as far back as the 1930s, however, we use a cross-section of securities and their prices from 1990, 2000, and 2010 (comprising a total of 283,384 observations).

The key variable of interest here is price. We assess the price of each security by looking at monthly close prices for each stock, as provided by the WRDS database. In cases where the monthly close price is not available, the average between the bid-ask spread is substituted for the close price. Price is adjusted for stock splits/buybacks via the CFACPR variable provided in the WRDS database. We explain price by looking at particularly important factors over the three years provided; only securities which are maintained across the period are analyzed in respective mappings (ie, only assets present in both 1990 and 2000 would be used to infer conclusions about certain predictors).

Ideally, our data would have values for every month-stock observation, however not all predictors are reported on a month-by-month basis. We employed several different methods to fill in the data. Initially, predictors with more than 15% missing values in their columns were subtracted from the data. This serves two purposes, firstly to minimize predictors that have little data, and secondly to isolate our focus on variables more likely to be commonly held by different securities. Since we only analyze cross-sections of three years, it is important to minimize variables with low reporting as the value of information of a predictor only reported once a year or more will be diminished by the fact that it is exposed to high variance or anomalies.

After isolating this subset we have 18 predictors which fill this minimum level of reporting but may nonetheless have empty values for some months. We solved this problem by filling in empty cells with the last observed value for that security. If a last observed value is not available, we fill in the average of the predictor value more generally (prioritizing their within security average). The average price is calculated for each year, then the percentage change from each period is further computed to create our dependent variables (ie, percentage change between 1990-2000). Considering this is a large dataset, we do not see much difficulty with the interpretability of results as a result of our data cleaning practices, other than perhaps underestimating factor effects by pushing them towards the mean.

For the three different cross-sections, we have the following descriptive statistics:

Summary for period 1990-2000	**Summary for period 2000-2010**
	

```
##   BidAskSpread        Coskewness          DolVol              High52
##  Min.   :0.0001723   Min.   :-1.8253   Min.   :-11.6391   Min.   : 0.008333
##  1st Qu.:0.0071282   1st Qu.: 0.1754   1st Qu.: -4.8319   1st Qu.: 0.546997
##  Median :0.0123603   Median : 0.3624   Median : -2.7408   Median : 0.743853
##  Mean   :0.0178052   Mean   : 0.3709   Mean   : -2.8654   Mean   : 0.716736
##  3rd Qu.:0.0218757   3rd Qu.: 0.5525   3rd Qu.: -0.8996   3rd Qu.: 0.891264
##  Max.   :0.8111540   Max.   : 1.6210   Max.   :  7.9756   Max.   :16.553846
##    IdioVol3F           IdioVolAHT            MaxRet          PriceDelayRsq
##  BidAskSpread        Coskewness         DolVol           High52
## Min.   :0.0003275   Min.   :-1.2115   Min.   :-9.9069   Min.   :0.02083
## 1st Qu.:0.0066828   1st Qu.: 0.4444   1st Qu.:-2.7229   1st Qu.:0.57292
## Median :0.0142155   Median : 0.7971   Median :-0.8413   Median :0.74359
## Mean   :0.0259183   Mean   : 0.7372   Mean   :-0.8287   Mean   :0.71584
## 3rd Qu.:0.0300450   3rd Qu.: 1.0670   3rd Qu.: 1.0104   3rd Qu.:0.87661
## Max.   :1.0213211   Max.   : 2.5138   Max.   :10.1503   Max.   :2.25000
##   IdioVol3F          IdioVolAHT            MaxRet          PriceDelayRsq
## Min.   :-0.54498   Min.   :-0.3912998   Min.   :-2.00000   Min.   :0.0003356
## 1st Qu.:-0.03529   1st Qu.:-0.0372074   1st Qu.:-0.08571   1st Qu.:0.2015135
## Median :-0.02136   Median :-0.0241861   Median :-0.04918   Median :0.4975712
## Mean   :-0.02853   Mean   :-0.0297645   Mean   :-0.07022   Mean   :0.5123223
## 3rd Qu.:-0.01307   3rd Qu.:-0.0155079   3rd Qu.:-0.02817   3rd Qu.:0.8355359
## Max.   : 0.00000   Max.   :-0.0000208   Max.   : 0.10000   Max.   :1.0000000
## PriceDelaySlope     RealizedVol          ReturnSkew        ReturnSkew3F
## Min.   :-821.585   Min.   :-0.63901   Min.   :-4.4772   Min.   :-4.44181
## 1st Qu.:   -0.262   1st Qu.:-0.03989   1st Qu.:-0.5887   1st Qu.:-0.50855
## Median :    0.704   Median :-0.02462   Median :-0.1387   Median :-0.09936
## Mean   :    1.513   Mean   :-0.03209   Mean   :-0.1213   Mean   :-0.10006
## 3rd Qu.:    1.648   3rd Qu.:-0.01529   3rd Qu.: 0.2872   3rd Qu.: 0.29103
## Max.   :5127.974   Max.   : 0.00000   Max.   : 4.4772   Max.   : 4.44181
##     VolMkt              betaVIX              zerotrade       zerotradeAlt1
## Min.   :-5.626369   Min.   :-0.2835276   Min.   : 0.000   Min.   : 0.000
## 1st Qu.:-0.081141   1st Qu.:-0.0049325   1st Qu.: 0.000   1st Qu.: 0.000
## Median :-0.041767   Median :-0.0001711   Median : 0.175   Median : 0.000
## Mean   :-0.071195   Mean   :-0.0003926   Mean   : 2.587   Mean   : 2.710
## 3rd Qu.:-0.021231   3rd Qu.: 0.0041429   3rd Qu.: 3.437   3rd Qu.: 3.652
## Max.   :-0.000047   Max.   : 0.4324130   Max.   :19.529   Max.   :20.087
##   STreversal             Size         percentage_change_1990_2000
## Min.   :-285.7143   Min.   :-18.535   Min.   :  -99.94
## 1st Qu.:  -4.7619   1st Qu.:-12.583   1st Qu.:  -14.19
## Median :   0.6211   Median :-11.130   Median :   61.20
## Mean   :   1.1041   Mean   :-11.250   Mean   :  331.20
## 3rd Qu.:   8.2569   3rd Qu.: -9.749   3rd Qu.:  259.38
## Max.   :  90.6250   Max.   : -3.308   Max.   :59499.30
```

```
## Min.   :-0.65661   Min.   :-0.3909087   Min.   :-2.71429   Min.   :0.0001434
## 1st Qu.:-0.04580   1st Qu.:-0.0530427   1st Qu.:-0.12000   1st Qu.:0.1480921
## Median :-0.02831   Median :-0.0335201   Median :-0.06875   Median :0.3543841
## Mean   :-0.03535   Mean   :-0.0406178   Mean   :-0.09443   Mean   :0.4233121
## 3rd Qu.:-0.01720   3rd Qu.:-0.0223974   3rd Qu.:-0.03896   3rd Qu.:0.6803217
## Max.   : 0.00000   Max.   :-0.0006797   Max.   : 0.13393   Max.   :1.0000000
## PriceDelaySlope      RealizedVol          ReturnSkew        ReturnSkew3F
## Min.   :-7894.131   Min.   :-0.66587   Min.   :-4.3644   Min.   :-4.0185
## 1st Qu.:   -0.211   1st Qu.:-0.05355   1st Qu.:-0.7079   1st Qu.:-0.5935
## Median :    0.634   Median :-0.03276   Median :-0.2446   Median :-0.1697
## Mean   :   -0.729   Mean   :-0.04109   Mean   :-0.2655   Mean   :-0.1882
## 3rd Qu.:    1.404   3rd Qu.:-0.01999   3rd Qu.: 0.1874   3rd Qu.: 0.2260
## Max.   : 1623.414   Max.   : 0.00000   Max.   : 4.3644   Max.   : 3.6722
##     VolMkt               betaVIX              zerotrade        zerotradeAlt1
## Min.   :-18.513107   Min.   :-0.2450775   Min.   : 0.0000   Min.   : 0.0000
## 1st Qu.: -0.167530   1st Qu.:-0.0061017   1st Qu.: 0.0000   1st Qu.: 0.0000
## Median : -0.074313   Median :-0.0002661   Median : 0.0000   Median : 0.0000
## Mean   : -0.193669   Mean   :-0.0010494   Mean   : 0.9208   Mean   : 0.9429
## 3rd Qu.: -0.036150   3rd Qu.: 0.0047991   3rd Qu.: 0.1667   3rd Qu.: 0.0000
## Max.   : -0.000205   Max.   : 0.3564411   Max.   :18.8707   Max.   :20.0870
##   STreversal              Size        percentage_change_2000_2010
## Min.   :-1034.4000   Min.   :-20.18   Min.   : -99.988
## 1st Qu.:   -7.1429   1st Qu.:-13.80   1st Qu.: -48.625
## Median :    0.0000   Median :-12.35   Median :   8.499
## Mean   :   -0.6336   Mean   :-12.48   Mean   :  96.353
## 3rd Qu.:    8.5284   3rd Qu.:-11.00   3rd Qu.: 113.137
## Max.   :   87.1622   Max.   : -4.79   Max.   :7835.852
```

Summary for period 1990-2010

```
##  BidAskSpread        Coskewness         DolVol           High52
## Min.   :0.0003275   Min.   :-1.1925   Min.   :-9.9069   Min.   :0.0375
## 1st Qu.:0.0057328   1st Qu.: 0.4584   1st Qu.:-3.1503   1st Qu.:0.6087
## Median :0.0114993   Median : 0.8172   Median :-1.2035   Median :0.7712
## Mean   :0.0221862   Mean   : 0.7513   Mean   :-1.1916   Mean   :0.7394
## 3rd Qu.:0.0250721   3rd Qu.: 1.0864   3rd Qu.: 0.7004   3rd Qu.:0.8919
## Max.   :0.6246707   Max.   : 2.5138   Max.   : 9.9680   Max.   :2.2500
##   IdioVol3F          IdioVolAHT             MaxRet          PriceDelayRsq
## Min.   :-0.43902   Min.   :-2.429e-01   Min.   :-1.50000   Min.   :0.0003356
## 1st Qu.:-0.03163   1st Qu.:-3.384e-02   1st Qu.:-0.07692   1st Qu.:0.1538649
## Median :-0.01904   Median :-2.131e-02   Median :-0.04444   Median :0.4433051
## Mean   :-0.02602   Mean   :-2.717e-02   Mean   :-0.06422   Mean   :0.4820803
## 3rd Qu.:-0.01195   3rd Qu.:-1.421e-02   3rd Qu.:-0.02632   3rd Qu.:0.8107237
## Max.   : 0.00000   Max.   :-2.076e-05   Max.   : 0.01429   Max.   :1.0000000
## PriceDelaySlope     RealizedVol          ReturnSkew        ReturnSkew3F
## Min.   :-340.693   Min.   :-0.47566   Min.   :-4.4772   Min.   :-4.32200
## 1st Qu.:   -0.247   1st Qu.:-0.03604   1st Qu.:-0.5821   1st Qu.:-0.50582
## Median :    0.667   Median :-0.02216   Median :-0.1390   Median :-0.09925
## Mean   :    2.236   Mean   :-0.02944   Mean   :-0.1326   Mean   :-0.10254
```

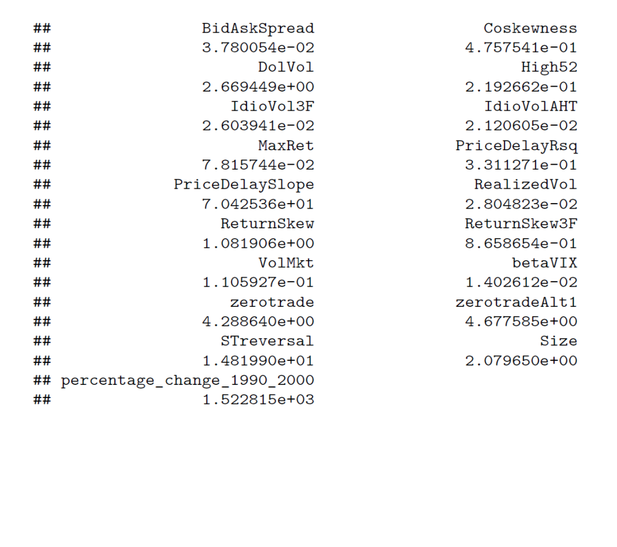
Summary for period 1990-2010 (cont)

```
## 3rd Qu.:    1.543   3rd Qu.:-0.01418   3rd Qu.: 0.2756   3rd Qu.: 0.28832
## Max.   :5127.974   Max.   : 0.00000   Max.   : 4.4772   Max.   : 4.44181
##     VolMkt              betaVIX              zerotrade       zerotradeAlt1
## Min.   :-3.271335   Min.   :-0.2450775   Min.   : 0.000   Min.   : 0.000
## 1st Qu.:-0.075649   1st Qu.:-0.0046242   1st Qu.: 0.000   1st Qu.: 0.000
## Median :-0.039845   Median :-0.0002008   Median : 0.000   Median : 0.000
## Mean   :-0.066634   Mean   :-0.0004988   Mean   : 2.253   Mean   : 2.362
## 3rd Qu.:-0.021045   3rd Qu.: 0.0036913   3rd Qu.: 2.635   3rd Qu.: 2.739
## Max.   :-0.000315   Max.   : 0.2640725   Max.   :19.297   Max.   :20.087
##   STreversal             Size         percentage_change_1990_2010
## Min.   :-277.7778   Min.   :-18.535   Min.   :  -99.90
## 1st Qu.:  -4.7619   1st Qu.:-13.038   1st Qu.:   11.78
## Median :   0.1248   Median :-11.520   Median :  184.06
## Mean   :   0.8951   Mean   :-11.615   Mean   :  608.89
## 3rd Qu.:   7.4286   3rd Qu.:-10.052   3rd Qu.:  574.51
## Max.   :  85.2273   Max.   : -5.472   Max.   :27188.36
```

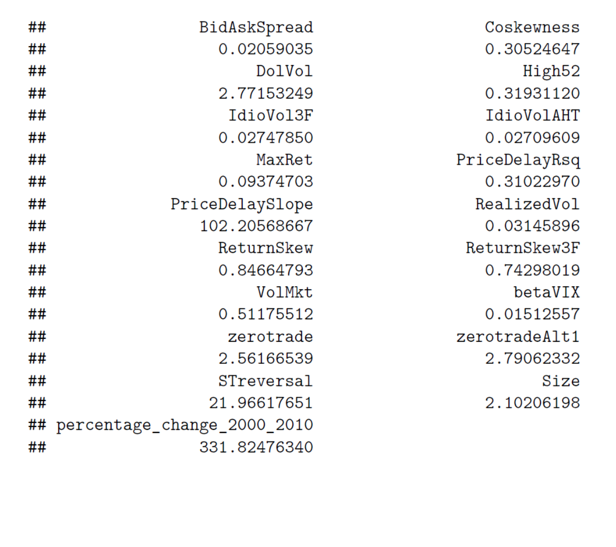
We can also see standard deviations for each dataset's variables:

Standard deviations for period 1990-2000

Standard deviations for period 1990-2000

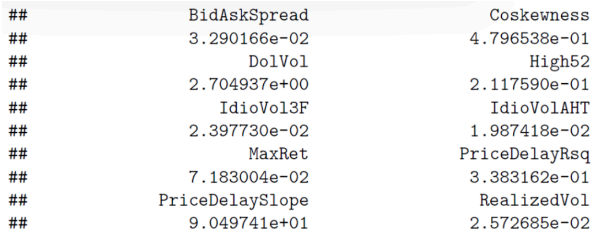


```
##                BidAskSpread           Coskewness
##                3.780054e-02         4.757541e-01
##                      DolVol               High52
##                2.669449e+00         2.192662e-01
##                  IdioVol3F           IdioVolAHT
##                2.603941e-02         2.120605e-02
##                      MaxRet        PriceDelayRsq
##                7.815744e-02         3.311271e-01
##             PriceDelaySlope          RealizedVol
##                7.042536e+01         2.804823e-02
##                  ReturnSkew          ReturnSkew3F
##                1.081906e+00         8.658654e-01
##                      VolMkt              betaVIX
##                1.105927e-01         1.402612e-02
##                   zerotrade        zerotradeAlt1
##                4.288640e+00         4.677585e+00
##                  STreversal                 Size
##                1.481990e+01         2.079650e+00
## percentage_change_1990_2000
##                1.522815e+03
```



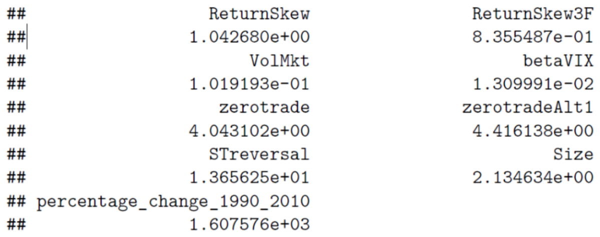
```
##                BidAskSpread           Coskewness
##                  0.02059035           0.30524647
##                      DolVol               High52
##                  2.77153249           0.31931120
##                   IdioVol3F           IdioVolAHT
##                  0.02747850           0.02709609
##                      MaxRet        PriceDelayRsq
##                  0.09374703           0.31022970
##             PriceDelaySlope          RealizedVol
##                102.20568667           0.03145896
##                  ReturnSkew         ReturnSkew3F
##                  0.84664793           0.74298019
##                      VolMkt              betaVIX
##                  0.51175512           0.01512557
##                   zerotrade        zerotradeAlt1
##                  2.56166539           2.79062332
##                  STreversal                 Size
##                 21.96617651           2.10206198
## percentage_change_2000_2010
##                331.82476340
```

Standard deviations for period 1990-2000



```
##            BidAskSpread           Coskewness
##            3.290166e-02         4.796538e-01
##                  DolVol               High52
##            2.704937e+00         2.117590e-01
##               IdioVol3F           IdioVolAHT
##            2.397730e-02         1.987418e-02
##                  MaxRet        PriceDelayRsq
##            7.183004e-02         3.383162e-01
##         PriceDelaySlope          RealizedVol
##            9.049741e+01         2.572685e-02
```

Standard deviations for period 1990-2000 (cont)



```
##                  ReturnSkew         ReturnSkew3F
##                1.042680e+00         8.355487e-01
##                      VolMkt              betaVIX
##                1.019193e-01         1.309991e-02
##                   zerotrade        zerotradeAlt1
##                4.043102e+00         4.416138e+00
##                  STreversal                 Size
##                1.365625e+01         2.134634e+00
## percentage_change_1990_2010
##                1.607576e+03
```

Predictor names are provided by Chen and Zimmerman and can be translated using Signal documentation available on their website, but generally follow a pattern of interpretability (ie, RealizedVol is realized total volatility).

Econometric Model

As discussed, price is the dependent variable. The efficient market hypothesis suggests that once information becomes widely available, the market quickly responds and prices it into the exchange rate of the security. Thus, price serves in some sense as a proxy variable for current sentiment concerning a company's "value" as determined through quasi-democratic market forces. Price is however dependent on a particular company's total value, and the number of free-floating shares available. Thus we chose to contextualize returns as the change in average price between the respective years. It should be noted that each of the machine learning methods that are used in this paper is intended to inform the reader about the potency of particular predictors, which serves to answer the more esoteric question of how well the data conforms to value investing philosophy. We have chosen not to stagger any of the methods (i.e., such as shrinking the variables via Lasso and then running Regression Trees) since this would more serve to produce an investment strategy rather than inform the question at hand.

We use a variety of different regression prediction methods, both supervised and unsupervised. Linear regression will be utilized to establish particularly relevant variables. LASSO is further utilized to shrink predictors that are inconsequential toward our return price. Regression trees and random forests are then used to find particular cross-sections that produce high/low/medium returns. These unsupervised methods are especially useful in creating "rules" that might inform an investor's portfolio decisions. Structured Vector Machines are applied to similarly cluster data into identifiable, predictable sets. Finally, we utilize k Nearest Neighbors to cluster data in a supervised manner and locate prominent variables.

Results

A generalized Linear Regression model was run on all three periods with varying results. For the period 1990-2000, Significant predictors at the 0.1% level include IdioVolAHT (-), PriceDelayRsq (-), DolVol (-), and BidAskSpread (-), indicating negative correlations with the dependent variable. Additional significant predictors at the 1% or 5% levels are RealizedVol (-), Size (+), STreversal (-), IdioVol3F (+), VolMkt (+), Coskewness (-), zerotradeAlt1 (+), and High52 (+). Despite the presence of several high-significance predictors, the model's adjusted R-squared is relatively low at 0.81%, showing it explains a minor fraction of the variation in returns.

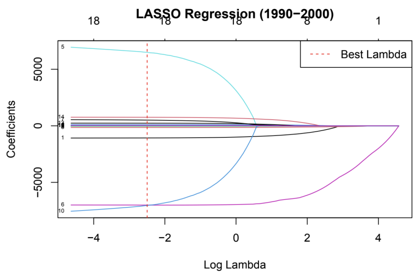
The negative coefficients for volatility metrics IdioVolAHT(Idiosyncratic Volatility) and RealizedVol (Realized Volatility) are notably intriguing during the dotcom bubble era, a time marked by intense growth and speculative investment in tech stocks, which were often highly volatile. These negative coefficients indicate that, even amidst widespread market enthusiasm, stocks exhibiting high idiosyncratic and realized volatility ought to be less favored by investors. In other words, the negative coefficient on IdioVolAHT and RealizedVol indicates that stocks with higher realized volatility tend to have lower returns.

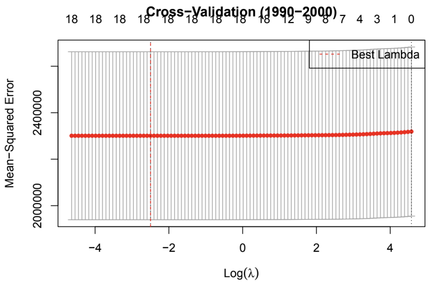
The negative coefficients on liquidity measures (PriceDelayRsq, DolVol, BidAskSpread) suggest that, between 1990 and 2000, stocks that rapidly incorporated market-wide information (indicated by lower PriceDelayRsq), exhibited high trading volumes (higher DolVol), and had smaller bid-ask spreads (lower trading costs) tended to perform better than those with slower price adjustments, lower trading activity, and higher trading costs. In other words, the more liquid stocks outperformed less liquid ones. This is surprising since theoretical models typically argue that investors should require a higher return for holding less liquid assets. Nevertheless, during the dotcom boom, it might have been the case that investors were primarily seeking quick gains, gravitating towards stocks with higher liquidity and driving up their prices. Additionally, the observed negative liquidity premium could be explained by a flight-to-liquidity effect, where investors favor more liquid stocks in periods of market turmoil.

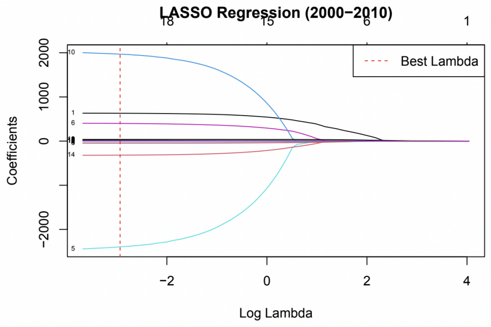
Other noteworthy predictors also played significant roles in stock performance during the dot-com era. The positive coefficient with Size reflects the superior performance of large-cap stocks, with major tech companies like Microsoft and Intel leading the surge and significantly appreciating value. In times of uncertainty, large-cap stocks are typically viewed as more secure investments because of their well-established operations and financial robustness. Conversely, the negative coefficient on STreversal (short-term reversal) indicates that stocks that recently performed well continued to do so. At the same time, those that underperformed remained weak, highlighting the momentum effect that was particularly strong during this period as investors chased trending stocks. The positive coefficients on IdioVol3F (idiosyncratic volatility adjusted for the Fama-French 3-factor model) and VolMkt (market volatility) suggest that stocks with higher idiosyncratic volatility and greater market sensitivity outperformed, possibly reflecting a risk premium for holding undiversifiable risks or the speculative nature of the market, with investors seeking high returns from volatile stocks. The negative coefficient on coskewness indicates that stocks that tend to perform poorly when the market declines also underperform. This may be because investors were particularly cautious about downside risk, favoring stocks that provided some protection against market downturns.

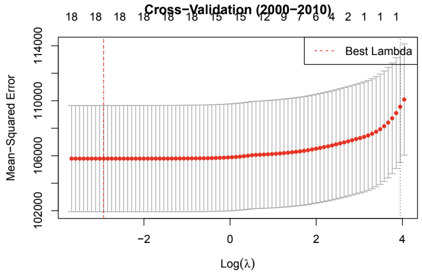
For the period 2000-2010 the behavior of some predictors was the opposite of what we saw from 1990 to 2000. The volatility indicators (IdioVol3F, IdioVolAHT, RealizedVol) now show mixed results, suggesting a more complex relationship. Similarly, liquidity measures (PriceDelayRsq, BidAskSpread) still affect returns, but their signs reversed. The mixed signals from volatility indicators (IdioVol3F, IdioVolAHT, RealizedVol) suggest a complex relationship between volatility and stock returns during this period, contrasting with the simpler relationship observed from 1990 to 2000. IdioVol3F (Idiosyncratic Volatility - Fama-French 3-Factor) tracks the volatility in a stock's returns that isn't explained by the market, size, and value factors. IdioVolAHT (Idiosyncratic Volatility - Ali, Hwang, and Trombley), which uses a different method and timeframe, shows a positive coefficient, suggesting that stocks with higher volatility by this measure performed better. RealizedVol (Realized Volatility) measures a stock's total volatility from *actual* returns relative to market indices, with a positive coefficient that suggests that more volatile stocks generally offer better returns.

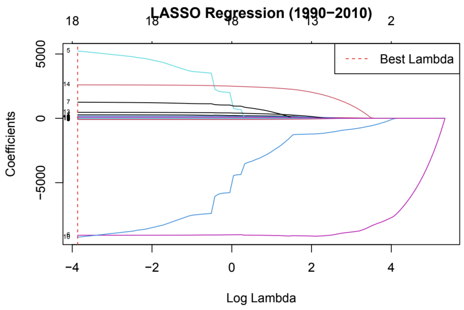
The increased adjusted R-squared of 4.00% for this period indicates that these predictors accounted for a greater share of the variability in stock returns compared to the 1990-2000 period. The relationship between coskewness and stock returns has changed over time, with a negative coefficient in the 1990-2000 period and a positive one in the 2000-2010 and full 1990-2010 periods. This shows how investor preferences regarding the higher moments of return distributions can evolve due to changing market conditions. Nonetheless, it also shows that a significant amount of the variation in returns remains unexplained by these factors. There are a few reasons for our low R-squared. Firstly, the model assumes a linear relationship between predictors and returns, which might not hold if these relationships are non-linear. This mismatch would lead to underestimating the relationships, thus contributing to a low R-squared. Secondly, Stock returns are subject to numerous influencing factors, making them inherently volatile and unpredictable. The low R-squared may reflect this high degree of randomness, hindering the identification of consistent predictor-return relationships. Furthermore, this model probably needs to include other significant variables affecting stock returns. This could encompass macroeconomic factors such as GDP growth or inflation rates, company-specific attributes like earnings quality or management skills, or wider market forces including investor sentiment or geopolitical events. Lastly, the influence of predictors on returns may vary across different periods. If the significance or effect of these predictors changes over time, a single linear model might not capture such dynamics effectively, contributing to the low R-squared. Thus we chose to further parse the data with various other methods to identify patterns unnoticed by linear regression.

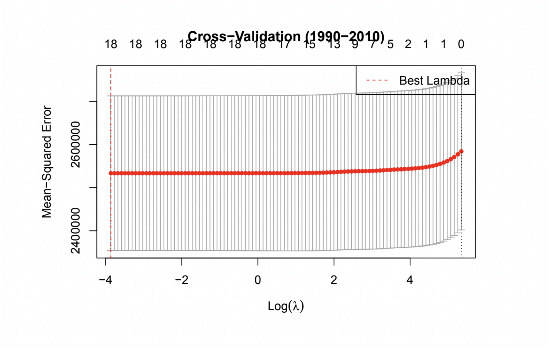












The changing coefficients on IdioVolAHT, DolVol, and betaVIX across the periods are particularly intriguing, as they suggest that the relationships between these factors and stock returns are unstable over time.

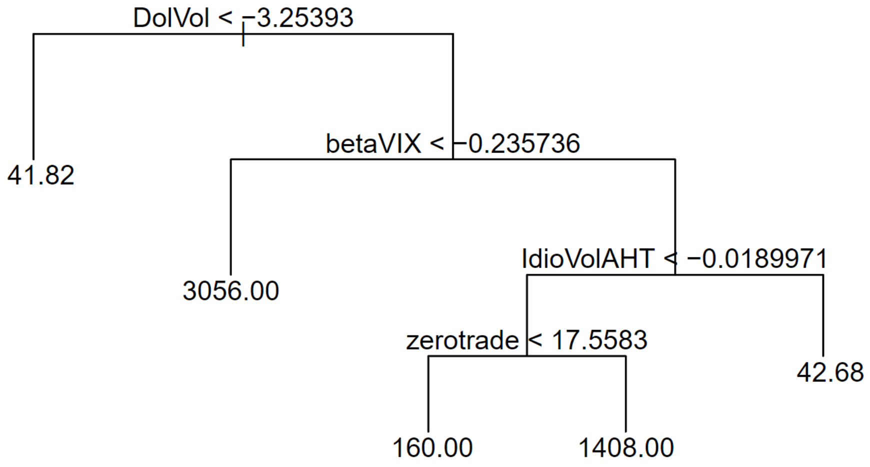
We found that the negative coefficient of IdioVolAHT between 1990-2000 and 1990-2010 supports the idea that investors expect to be compensated for taking on the risk related to holding stocks with high idiosyncratic risk. Nonetheless, the positive coefficient for 2000-2010 shows that stocks with greater idiosyncratic volatility did better. This phenomenon may have occurred because investors preferred stocks with lower risk levels during market instability in the early and late 2000s. Furthermore, DolVol's negative coefficient for 1990-2000 and 1990-2010 indicates that, in line with a liquidity premium, equities with lower trading volume did better than those with larger volumes. Liquidity premium is intended to incentivize investments in assets that are difficult or time-consuming to turn into cash at fair market value. For example, because of its relative illiquidity, a long-term bond will have a higher interest rate than a short-term bond. However, between 2000 and 2010, the positive coefficient points to a preference for more liquid stocks, maybe due to investor demand for flexibility during bear markets. According to the 1990–2010 betaVIX positive coefficient, stocks that were more susceptible to market volatility and thus more sensitive to fluctuations in VIX had higher returns, which is consistent with the concept of premium for volatility risk. On the other hand, the negative coefficient for the years 2000–2010 suggests that stocks with lower VIX sensitivity performed better during this period, perhaps due to investors trying to avoid the volatile market after the experience of the dot-com bubble.

Furthermore, the models' lack of predictors also sheds light on how significant they were during each time period. BidAskSpread is a frequently used proxy for liquidity, which calculates the difference between the highest price a buyer is prepared to pay and the lowest price a seller is willing to accept for a stock. Since BidAskSpread was not included in the 2000-2010 model, it is possible that liquidity had less of an impact on stock returns during this decade. Improvements in market efficiency brought about by developments in trading technology may be what caused the decreased impact of liquidity on stock prices. Additionally, there were notable developments in the 2000s that might have changed investors' preferences and perceptions of risk. Investor attention may have shifted away from liquidity during times of market stress and toward other variables like robust fundamentals or market mood. Finally, the declining significance of liquidity during the 2000s may have something to do with how some industries performed. For instance, if the technology industry, which is generally linked to greater liquidity, underperformed during this time, liquidity may not have been as important in explaining results overall.

We ran three separate regression trees for the sets ranging from 1990-2000, 2000-2010, and then from 1990-2010, respectively. The method used to create our trees was the basic tree function in R from the trees library. Along with this, we used a 90% training set to be able to perform cross-validation on our outputs. Once we specified our dependant variable as “Percent_change” we included all our covariates into the model for potential pruning and decision-making.

Our first regression tree from 1990 to 2000 was not able to be plotted due to there only being one terminal node. This result indicated there is only the intercept in our regression function, and that all outputs are predicted to be that one value. Our terminal nodes of these three regression trees are the raw percentage change in whole numbers. The single node in the tree from 1990 to 2000 has a value of 333 which essentially means there is a 333% increase in return on securities over 10 years, which would correspond to an annualized return of 12.7%. This particular model predicts that the annual increase for every firm would be a 12.7% increase in security prices. This single node tree is potentially the result of the tech bubble popping in early 2000, which may have eliminated a lot of predictability on the ends of the spectrum of returns.

The second model from 2000 to 2010 is more involved than the first tree, having five terminal nodes. There are four splits in this tree involving the variables in the order of DolVol, betaVIX, IdioVoIAHT, and zerotrade. These variables mean, two months of lagged trading volumes, the coefficient of daily change of VIX, the standard deviation of residuals from CAPM regression, and in each month the count of the number of days with no trading, respectively. The tree insinuates that the four most important variables are the ones we have listed here. When performing the cross-validation to the 2000 to 2010 data we discovered the MSE is lowest at 5 terminal nodes. The range of annualized increases goes from 3.5% (41.82 percent increase from start to end year) to 64% annualized increases (total of 1408% increase over period) depending on where the covariates lie for a particular observation from 2000-2010, as can be seen in the below graph.



The third model ran on the model from 1990 to 2010, there was only one split based on the value of IdioVoIAHT, or rather the standard deviation of residuals from CAPM regression. The terminal nodes are 804% (23% annual increase in securities) for values of CAPM less than -0.02076, and 400 (14% annual increase in securities) for values of CAPM greater than -0.02076. The cross-validation for this final tree has a minimized MSE for the one terminal node.

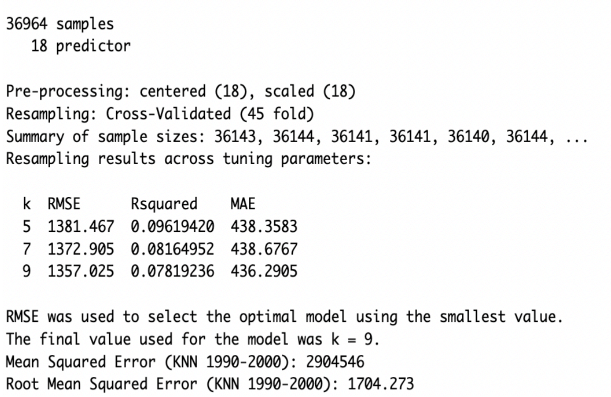
The regression tree function runs tests that ensure that overfitting does not occur. This is why our first regression tree had only one terminal node and the third tree only had one split with two terminal nodes. There are other pre-pruning methods that we could have implemented to add more nodes. One such method is the rpart function in Rstudio which allows you to alter the critical point (cp) to change the sensitivity of the model to create a model with more terminal nodes by having fewer observations in each node. The reason we did not end up using this function was that the cross-validation methods were more intensive and often skewed from what we produced with the tree function. Overall, the trees provided us the opportunity to look further at which values offered the best decision boundaries for regression output. Variables such as IdioVoIAHT, DolVol, and betaVIX appear to be most relevant for our analysis.

The process of random forests is to bring together a conglomeration of different regression trees for a particular dataset in hopes of creating better prediction accuracy for our prospective model. Each dataset was partitioned with 15% of the data for training due to the size of the data. In hopes of reducing the statistical variance from our previous regression tree models, we aggregated multiple samples of regression trees using an argument called "mtry" in the randomforest function which allows us to consider only a certain number of predictors in creating the trees for our forest. When we pass the argument "mtry=p", meaning passing the number of predictors, we can employ bagging with our random forest analysis. In our code we had a p of 18, so we passed the argument "mtry=6" to only account for one-third of our predictors to properly perform random forests. The four variables that lower the MSE percentage the most, or increase the MSE the least would be betaVIX, STreversal Coskewness, and High52.

In plotting out the expected prices of securities with the actual test prices, we are looking for our values to be plotted as close to a line with a slope of 1 and an intercept of 0 to show that the two values are the same. We found in our output that the MSE of our models from 1990-2000 and 1990-2010 are quite large with values of 1.3 million and 1.9 million respectively. We believe that this is correlated to the poor performance of the regression trees for these two datasets. However, the performance of the random trees model for the dataset from 2000-2010 was somewhat better, with an MSE of 88,700. All three of these models have large MSE, and possible methods such as pre-pruning as mentioned before, and tuning of parameters such as lowering mtry could have positive impacts on the MSE.

With our datasets, we attempted to use Support Vector Machines (SVM) for regression. We applied it to our datasets, hoping to be able to predict the percentage change variable in each of the three datasets. Each dataset was partitioned with 15% of the data for training due to the size of the data. The results we obtained using SVM were inconclusive. For each dataset, the accuracy was 0 and the MSE for each was unusually large. This is a potential area for further study and modification of the raw data since there could be a need for better feature engineering and more parameter tuning in our code. However, due to the exorbitant amount of time it takes to run SVM on datasets such as these specific three, which are already shortened, our preliminary analysis has determined SVM is likely not the best model choice for this data.

After an unsuccessful time with SVM, we applied boosting to each of our three datasets. Our goal was to reduce bias and variance in building a series of models to improve prediction accuracy. The boosting method is proficient at reducing errors due to the nature of the method. Boosting attempts to do this by combining multiple weak models into a singular strong predictor. Across our three different datasets, we implemented boosting to try and reduce bias and variance. By using a generalized boosted model (GBM) on our data, the boosting process is also more robust to overfitting. The rel.inf column stands for relative influence, which is a measure of how much each predictor contributes to the model. The value is a proportion of a feature's predictive power. The features that are the main drivers in the model for percentage change from 1990 to 2000 are PriceDelayRsq, PriceDelaySlope, and VolMkt. These same three have the highest relative inference for the other two datasets as well, suggesting that they have the greatest influence on the percentage change of stock in all three time periods.

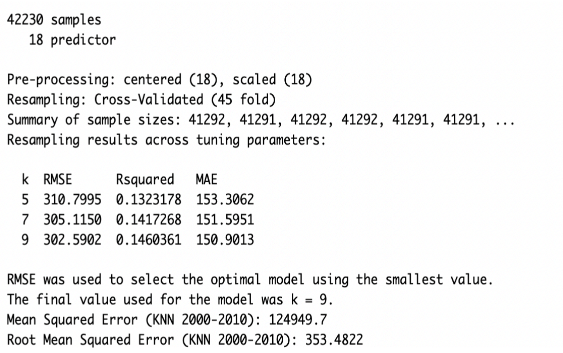


```
36964 samples
   18 predictor

Pre-processing: centered (18), scaled (18)
Resampling: Cross-Validated (45 fold)
Summary of sample sizes: 36143, 36144, 36141, 36141, 36140, 36144, ...
Resampling results across tuning parameters:

  k  RMSE      Rsquared    MAE
  5  1381.467  0.09619420  438.3583
  7  1372.905  0.08164952  438.6767
  9  1357.025  0.07819236  436.2905

RMSE was used to select the optimal model using the smallest value.
The final value used for the model was k = 9.
Mean Squared Error (KNN 1990-2000): 2904546
Root Mean Squared Error (KNN 1990-2000): 1704.273
```

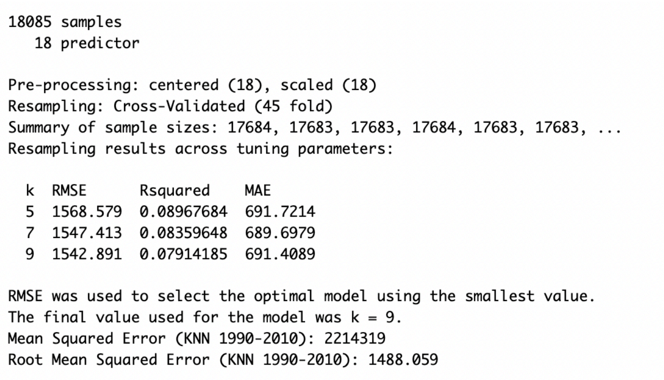


```
42230 samples
   18 predictor

Pre-processing: centered (18), scaled (18)
Resampling: Cross-Validated (45 fold)
Summary of sample sizes: 41292, 41291, 41292, 41292, 41291, 41291, ...
Resampling results across tuning parameters:

  k  RMSE      Rsquared   MAE
  5  310.7995  0.1323178  153.3062
  7  305.1150  0.1417268  151.5951
  9  302.5902  0.1460361  150.9013

RMSE was used to select the optimal model using the smallest value.
The final value used for the model was k = 9.
Mean Squared Error (KNN 2000-2010): 124949.7
Root Mean Squared Error (KNN 2000-2010): 353.4822
```



```
18085 samples
   18 predictor

Pre-processing: centered (18), scaled (18)
Resampling: Cross-Validated (45 fold)
Summary of sample sizes: 17684, 17683, 17683, 17684, 17683, 17683, ...
Resampling results across tuning parameters:

  k  RMSE      Rsquared    MAE
  5  1568.579  0.08967684  691.7214
  7  1547.413  0.08359648  689.6979
  9  1542.891  0.07914185  691.4089

RMSE was used to select the optimal model using the smallest value.
The final value used for the model was k = 9.
Mean Squared Error (KNN 1990-2010): 2214319
Root Mean Squared Error (KNN 1990-2010): 1488.059
```

The R-squared value for the 1990-2010 model (0.07914185) is similar to that for the 1990-2000 model (0.07819236) but lower than that for the 2000-2010 model (0.1460361). This suggests that the predictors' explanatory power varies across the subperiods. Meanwhile, it shows the 2000-2010 period having a relatively better performance. While the R-squared value has improved for the 2000-2010 model, it still indicates that a significant portion of the variance in stock returns remains unexplained by the model. The 2000-2010 period may have been influenced by specific market conditions, economic events, or regulatory changes that the predictors explicitly capture. These factors could have a significant impact on stock returns during this period. Furthermore, there could be other reasons for the low explanatory power of the predictions, for example, non-linear relationships, irrelevant predictors, and time-varying relationships.

Additionally, the RMSE and MSE values for the 1990-2010 model (RMSE: 1542.891, MSE: 2,214,319) are lower than those for the 1990-2000 model (RMSE: 1704.273, MSE: 2,904,546) but higher than those for the 2000-2010 model (RMSE: 302.5902, MSE: 124,949.7). This indicates that the model's predictive accuracy differs across these periods, with the 2000-2010 period exhibiting better performance than the other periods. The high RMSE and MSE values indicate that the KNN model predictions have a significant average deviation from the actual stock returns. These metrics measure the average magnitude of the prediction errors. In other words, the high RMSE or MSE value suggests that the model's predictions are inaccurate and have a large margin of error.

The high prediction errors could be due to several reasons. For example, the 18 predictors used in the model may not be sufficient to predict stock returns accurately. On top of this, the presence of outliers or extreme values in the stock returns data can affect the model's performance. KNN is sensitive to outliers because it considers the nearest neighbors for making predictions. Outliers can distort the distance calculations and lead to higher prediction errors. Lastly, the KNN has its own limitations, it depends on the assumption that similar instances based on the predictors will have similar target values. If this assumption does not hold true for the data, the model's predictions may be inaccurate, resulting in high RMSE and MSE values.

Although the RMSE and MSE values throughout the 2000-2010 period have improved in terms of prediction accuracy, the values are still relatively high, representing that there is still room for improvement in the model's predictive power. It is worth noting that the optimal k value of 9 remains constant throughout these periods, showing the 9 nearest neighbors provide the best balance between bias and variance for the KNN model in both periods. This consistency in the optimal k value could indicate that the underlying structure or patterns in the data are similar across the two periods.

Conclusion

Our data-driven methodology found that the predictor most independent of any particular market period or fluctuation appeared to be size, which in a variety of the methods retained a positive coefficient with relation to returns. This is one of the main findings of the original Fama and French 1992 paper, which studied the period immediately before ours (1963-1990) We theorize that there is a significant Pareto principle, in other words, the large-cap stocks (of which there are statistically fewer) will produce the majority of predictable returns. This also seems to conform strongly with the general tendency for companies who achieve market dominance to retain it for lengthy periods, creating, as Warren Buffet would say, a competitive “moat”.

Implied volatility predictors (ie, risk sentiment variables) were very often very significant between periods but would change signs depending on the period or vary widely in strength, An example of this would be IdioVol3F and IdioVolAHT, which is significant in all periods, but switches from positive in 1990-2000 to negative in 2000-2010 when using Lasso. Volatility measures that corresponded to actual securities, however (such as VolMkt of RealizedVol), could semi-consistently serve as a factor in predicting stock returns. This variation emphasizes how changes in economic and market conditions can alter the impact of market volatility on asset pricing. Following major market declines, there could be a higher preference for stability, causing less volatile stocks to become more appealing. Moreover, the increasing influence of institutional investors, who frequently prioritize long-term risk prevention, may have also contributed to this pattern.

We may consider a few drawbacks to the data, which provide opportunities for further research. Firstly, our model only included companies that were present in both periods, since mapping the percentage change between stocks that drop to zero before the 10-year mark would significantly skew results towards losses. A portfolio approach would solve this issue, however, the scope of this paper did not allow for assembling baskets of securities before analysis. Secondly, our dataset is heavily skewed towards US securities, which may not accurately reflect market conditions globally, especially during events that have localized weight (tech boom of the late 90s, 2008 crash).

Finer tuning parameters could also be used to marginally increase the accuracy of many of the methods. However, regression trees experienced particular difficulty. We found that our initial prepruning methods resulted in a single node tree for the period 1990-2000. This may be because of the distortion of the 2000 tech bubble pop, meaning the mapping of certain securities onto their equivalent ten years later may have been obfuscated by a bearish market correction. Attempts to soften the pruning methods resulted in large decreases in the interpretability of our results in other periods, thus for the scope of this paper we chose to accept the less-than-optimal results and focus on identifying patterns holistically.

Works Cited

Bagnara, M., “Asset Pricing and Machine Learning: A critical review”, *Journal of Economic Surveys*, 2024.

Chen, Andrew Y., and Tom Zimmermann. "Open Source Cross-Sectional Asset Pricing." *Critical Finance Review*, vol. 27, no. 2, 2022, pp. 207-264.

Karolyi, G. Andrew, and Stijn Van Nieuwerburgh,"New Methods for the Cross-Section of Returns", *The Review of Financial Studies*, Volume 33, Issue 5, May 2020,

Fama, Eugene F., and Kenneth R. French. "A five-factor asset pricing model", *Journal of Financial Economics*, Volume 116, Issue 1, 2015.

Fama, Eugene F., and Kenneth R. French. "The Cross-Section of Expected Stock Returns", *The Journal of Finance*, 1992.

Garmhausen, Steve, "What Is Value Investing?", *Wall Street Journal*, 2023.