Machine Learning Asset Pricing Factors in an Emerging Stock Market

Dan Gabriel Anghel
The Bucharest University of Economic Studies

We analyze risk factors in the emerging stock market of Romania using Machine Learning classification models and find novel evidence showing that Liquidity, Conditional Skewness, and Volatility display predictive power over long-term stock returns. In our sample, a portfolio formed using information derived from all three factors earns an average excess return of 12% per year, a result with t≥3 significance. This is adjusted for the risk premiums associated with the Market, Size, and Value factors, which are reconfirmed as being significant in our analysis. Momentum is shown to have no influence, adding to existing evidence pointing toward the same conclusion for markets from the Central and Eastern European region. Besides uncovering potential novel risk factors, our paper shows that Machine Learning models are a useful tool for studying asset pricing in small, emerging stock markets.

Keywords: asset pricing, risk factors, machine learning, emerging stock markets

INTRODUCTION

There are approximately 400 risk factors that have been investigated as possible asset price determinants in international financial markets (Harvey, Liu, and Zhu, 2015). However, only a handful of those have been investigated in small, less developed markets. This paper contributes to the literature by testing some of the previously uninvestigated risk factors in the frontier equity market of Romania, using the rather unconventional approach of Machine Learning (ML). In essence, we use trading histories and past balance sheet information for listed stocks–alongside data on the state of the economy–to fit a ML model that is used to forecast future long-term stock prices and construct portfolios capable of generating significant alphas relative to a benchmark asset pricing model. Comparing the results obtained when incorporating different sets of variables (factors) enables us to conclude about their informational content.

The choice of using ML is motivated by the relatively small size and low liquidity of the market, which severely limits the applicability of standard approaches in the analysis of asset pricing factors. Specifically, hedge portfolios are usually constructed from the intersection of lower-level portfolios that arise from individual factor sorts (see, e.g., Fama and French, 1992, 1993, 2015; Harvey, Liu, and Zhu, 2015). However, this is not possible when the number of factors is large or when the number of stocks in the market is small, as many hedge portfolios end up with very few or even without constituting stocks. Another advantage of ML is that it is well equipped to handle nonlinear relationships in the data, whereas conventional approaches only account for linear relationships. As a disadvantage, ML is not generally seen as a useful tool in the financial economics literature due to its tendency to overfit the data, leading to data snooping bias. However, we take this into account and design a methodology that should appropriately handle data snooping. Specifically, we choose the Random Forest ML technique, which is less prone to data snooping, and combine this with true out-of-sample forecasting when testing results.

This is not the first paper that uses ML methods to investigate asset pricing factors in the stock market. For example, recent papers have analyzed both developed (e.g., French, 2017) and emerging stock markets from Asia (e.g., Cao, Parry, and Leggio, 2011; Jan and Ayub, 2019). Likewise, it is not the first that uses such algorithms to forecast stock market variables in Romania or other emerging markets such as those from Central and Eastern Europe. For example, early attempts, including Anghelache and Trifan (2013), Birău, Ehsanifar, and Mohammadi (2013) or Ruxanda and Badea (2013), focus on predicting stock prices and investing in individual assets, mainly market indexes. However, it is the first paper that combines the two in a way relevant for investigating, "learning" candidate asset pricing factors that have relevant informational content for determining long-term stock price movements. The actual input variables are selected from classes that can be associated with candidate risk factors, such as Market, Growth, Value, Momentum, or Volatility (these have been investigated before). However, we also incorporate candidate factors that have not been considered for Romania and, more generally, for any emerging market in the Eastern European region. Specifically, we use the Conditional Skewness factor of Harvey and Siddique (2000), or the Liquidity factor of Chordia, Subrahmanyam and Anshuman (2001). These are supplemented by other new variables that describe aggregate macroeconomic conditions or the cross-sectional dispersion of stock returns. Finally, we consecutively test all the factors relative to three established, benchmark models, namely the Capital Asset Pricing Model (Sharpe, 1964), the Fama-French 3-factor model (Fama and French, 1993) and the Carhart 4-factor model (Carhart, 1997).

The remainder of the paper is organized as follows. Section 2 discusses related work and the limitations encountered when studying the Romanian emerging market. Section 3 presents the data and testing methodology. Section 4 reports and comments on the test results. Section 5 concludes.

RELATED LITERATURE AND CHALLENGES IN ASSET PRICING

Romania is a medium-sized country situated in Central and Eastern Europe. Following the centralized, communist regime that was established after WWII, the country transitioned to a free market economy after 1989, this culminating with joining the European Union in 2007. The modern history of Romania's stock market starts in the mid-1990s when it was reopened after almost half a century. In the early years, thousands of stocks could have been traded due to the many mandatory listings enforced by the mass privatization program implemented by the government (starting in 1995). However, with deficient financial education of the general population, low levels of market transparency, and a small number of free-floating shares, only a fraction of the companies were traded. After numerous bankruptcies, mergers, and voluntary delistings, only 83 stocks remain listed in the main market segment at the beginning of 2020 [Source: The Bucharest Stock Exchange, https://www.bvb.ro/FinancialInstruments/Markets/Shares. Accessed February 11, 2020]. Romania thus has a relatively small and underdeveloped market, even compared to some of its post-communist neighbors in Central and Eastern Europe (e.g. Poland, the Czech Republic, Hungary). MSCI and FTSE currently rank Romania as a frontier stock market.

Even though the study of asset pricing factors is very important from both theoretical and practical perspectives, Romania's status and associated low financial importance have negatively influenced the amount of research dedicated to the behavior of local stock prices. Among the relatively few existing studies, most focus on testing Fama's (1970) Weak-Form Efficient Market Hypothesis (EMH) using standard statistical tests. Similar to other emerging markets in the region, the results predominantly show that market prices do not follow random walks and are predictable to some extent using past trading information (e.g., Dragotă and Oprea, 2014; Andrieş, Ihnatov and Sprincean, 2017). A few papers test if this predictability is significant from an economic perspective, i.e. if investors can use it to earn cost-adjusted excess returns, the evidence being mixed in this regard (Anghel, 2013, 2015, 2017; Dragotă and Ţilică, 2014).

Only a handful of studies that analyze risk factors on the Romanian market have been completed so far. In the first relevant paper of this kind, Tudor (2009) tests the relationship between above-market returns and beta, size, leverage, book-to-market equity, and earning-price ratios, finding that only book-to-market equity and earning-price ratios are important. In a more comprehensive study, Anghel, Dumitrescu and Tudor (2015) find that the Fama and French (1993) three-factor model significantly outperforms the CAPM (Sharpe, 1964), showing that alongside Book-to-Market, the Size factor is also important. When extending the model with the momentum factor of Jegadeesh and Titman (1993), they find no significant additional predictability, concluding that momentum is not important in Romania, at least when estimated in its standard form. However, although not specifically focusing on asset pricing factors, Anghel (2015) finds that test null rejections of no economic profit tests are correlated with price momentum. Also, Anghel (2017) finds that intraday returns also display momentum at long intervals. Such results hint that a momentum factor may still be responsible for some stock price movements in this market, possibly at different intervals or frequencies compared to what has been tested before.

There are also a couple of papers that investigate asset pricing factors at a regional level and also include Romania. More specifically, Zaremba and Konieczka (2015) investigate 11 Central and Eastern European markets and find that value, size, and momentum premiums are significant in a frictionless environment, but that only the value premium is resilient to controlling for illiquidity and transaction costs. Zaremba (2015) then uses a similar sample to further show that stock returns are non-monotonically related to a systematic component of risk and negatively related to an idiosyncratic component of risk, variations being caused by firm size and price momentum. This echoes results obtained in more developed markets (e.g., Ang et al., 2006) and hints that volatility is a relevant asset pricing candidate factor in this region as well. In other related papers, Oprea and Brad (2014) find a positive correlation between changes in consumer confidence and stock market returns in Romania. This result is replicated and extended by Stoykova (2017) and Stoykova, Paskaleva, and Stoykov (2018) on stock markets from the entire South-Eastern European region, hinting that other factors may also exist but have not been discovered so far.

Overall, the literature examining common risk factors in the stock market of Romania and its neighbors mostly shows that the Market, Size and Value factors play an important role in shaping local stock returns, while liquidity and volatility stand out as promising alternatives but have not been investigated in detail. As Harvey, Liu, and Zhu (2016) highlight, the common belief in the international literature is that the number of true factors is rather small, maybe close to five, this being related to a principal component analysis of "statistical" common factors driving time-series variation in equity returns. Because of this, the Fama and French (1993) three-factor model remains the most popular choice among researchers and practitioners, sometimes being augmented by a momentum factor as in Carhart (1997). The Fama and French (2015) five-factor model has also recently gained some visibility. Nevertheless, recent evidence has highlighted the role that other types of factors might have, especially volatility-related ones (Ang et al., 2006). Moreover, Harvey, Liu, and Zhu (2016) review over 300 risk factors have been tested in the international literature, many of them with great success, at least at standard significance levels ($t \geq 2$). Because emerging markets may behave differently from developed ones, some of these risk factors might be important here, while others might not. However, relevant tests have not been performed, this constituting a limitation of the literature. This paper fills this gap by testing additional risk factors in a frontier stock market, namely Romania.

METHODOLOGY

Data

The data sample consists of all stocks traded on the Romanian market, The Bucharest Stock Exchange, in the period March 2005 to January 2020. Daily trading information, consisting of prices, number of trades, volume, and share turnover (volume times average price, expressed in local currency) are collected from the local brokerage firm Tradeville (2005-2015) and Bloomberg (2015-2020). A total of 555,424 unique observations are used. The market portfolio is proxied by the BET index, a value-weighted index of the largest and most liquid companies in the market. Stocks that are traded in less than half of all possible days are filtered out due to insufficient liquidity, leaving a sample of 68 firms. For these, we use market prices to compute log-returns on each trading day, $r_t = \ln(C_t/C_{t-1})$, where C_t

represents the closing price in day *t*. The resulting daily trading data is used to compute variables associated with asset pricing factors at the monthly frequency. These are detailed in the next section. Additional variables are then added each month based on firm-specific balance sheet information and macroeconomic data, which are also collected from Bloomberg. The risk-free interest rate is proxied by the yield on 1-year bonds issued by the Romanian government.

Variables Related to Risk Factors

This paper searches for surplus informational content in some well-chosen variables that represent established and candidate asset pricing factors. The actual variables that we incorporate in the analysis are listed in Table 1. First, we account for the Market factor using the systematic risk coefficient (beta) associated with each stock. The possible influence of time-varying betas is also handled by incorporating distinct estimates for it using several lookback windows. Second, we account for factors related to company fundamentals using balance sheet information. For the Size factor, we use the market capitalization of each firm, while the Value factor is proxied using the Book-to-Market ratio. Third, we account for Momentum using cumulative returns computed using various lookback intervals. Fourth, we account for higher-order moments of stock return distributions and higher-order co-dependence by incorporating volatility and conditional skewness also computed over several distinct lookback windows. Fifth, liquidity is also incorporated using the cumulative turnover over different lookback intervals. Sixth, contemporaneous macroeconomic variables are incorporated to test if they have additional informational content that is not present in the other variables. These later additions can be considered as acting similar to control variables in a standard regression analysis. Finally, the cross-sectional volatility (CSSD) variable is added, which is traditionally used in the investigation of herding behavior (see, e.g. Chang, Cheng and Khorana, 2000).

TABLE 1
LIST OF VARIABLES USED IN THE ANALYSIS

Type	Variable	Description
Market	Beta 1m	Firm-level systematic risk measure, estimated with the regression $R_{i,t} = \alpha_i + \beta_i R_{m,t} + \varepsilon_{m,t}$, using daily observations on stock returns in the previous 1 month.
	Beta 6m	Idem, except that data in the previous 6 months is used.
	Beta 1y	Idem, except that data in the previous 1 year is used.
Size	MkCap	Firm-level market capitalization computed as the current price times the total number of ordinary shares outstanding.
Value	Book-to-Market	Firm-level Book-to-Market ratio computed as total equity divided by market capitalization
Momentum	Mom 1w	Firm-level cumulative return of stock prices in the previous 1 week.
	Mom 2w	Idem, except that data in the previous 2 weeks is used.
	Mom 3w	Idem, except that data in the previous 3 weeks is used.
	Mom 1m	Idem, except that data in the previous 1 month is used.
	Mom 6m	Idem, except that data in the previous 6 months is used.
	Mom 1y	Idem, except that data in the previous 1 year is used.
Volatility	Vol 1w	Firm-level standard deviation of stock returns in the previous 1 week.
	Vol 2w	Idem, except that data in the previous 2 weeks is used.
	Vol 3w	Idem, except that data in the previous 3 weeks is used.
	Vol 1m	Idem, except that data in the previous 1 month is used.
	Vol 6m	Idem, except that data in the previous 6 months is used.
	Vol 1y	Idem, except that data in the previous 1 year is used.
Conditional Skewness	Skew 6m	Firm-level conditional skewness (Harvey and Siddique, 2000) computed over a 6-month lookback interval.
	Skew 1y	Idem, except that data in the previous 1 year is used.
Liquidity	Turn 1w	Firm-level cumulative market turnover in the previous 1 week.
	Turn 2w	Idem, except that data in the previous 2 weeks is used.
	Turn 3w	Idem, except that data in the previous 3 weeks is used.
	Turn 1m	Idem, except that data in the previous 1 month is used.
	Turn 6m	Idem, except that data in the previous 6 months is used.
	Turn 1y	Idem, except that data in the previous 1 year is used.
Macro	EXRATE	Appreciation or depreciation (YoY) of the EUR/RON exchange rate.
	GDP	Real GDP growth rate (YoY).
	INFL	Growth rate of consumer prices in the economy (YoY).
	UNMPL	Unemployment level.
	CAD	Current Account Deficit, expressed as a percent of GDP.
	DEBT	Total country debt, expressed as a percent of GDP.
Other	CSSD	1-month average of the series of Cross Sectional Standard Deviations of returns (Chang, Cheng and Khorana, 2000), computed for all stocks in the filtered sample.

Machine Learning Forecasting Model

We use the Random Forest ML technique (Breiman, 2001) to develop ("train") a model that uses the variables defined in the previous section, called "features", to make predictions about future price movements. A Random Forests is a classification algorithm that uses a "training" sample to construct independent decision trees classifiers, which are then used to make predictions in an independent "test" sample. We use a variation of the algorithm that applies bootstrap aggregating ("bagging") to train decision trees on randomly select observations from the training sample. This procedure helps to improve the stability and accuracy of the model, contributing to reducing its variance and avoiding overfitting. The actual testing procedure goes as follows.

First, we define a classification variable that the model must fit and then predict. Because we are interested in using the features to make forecasts about future stock returns, we use the 1-month-ahead excess return of each stock compared to the market return. Because we use a classification algorithm, we convert this into a binary variable that takes 1 when the excess return is positive and 0 otherwise. This is convenient as the predictions can be easily used to directly construct hedge portfolios. This resulting target variable is added for each stock and each month alongside the rest of the data sample. Second, we choose an initial subsample of 3 / 8 of the entire sample, which makes May 2011 the initial cutoff date in our case. All observations before this date are added to the training sample, while the observations exactly on this date constitute the first test (prediction) sample. The target variable is eliminated from the prediction samples, to avoid look-ahead bias. Third, we fit a 5000-tree Random Forest model (presented in Appendix A) on the training sample and use it to make a 1-month-ahead, out-of-sample prediction about future returns using the features computed for each stock at the beginning of the current cutoff month. The predictions are then used to form a portfolio at the beginning of the next month, this being held until the next prediction is made. This procedure is repeated for all other months in the sample following May 2011, with the model being re-trained every month.

This estimation procedure can be interpreted as an ML-driven technique for learning an asset pricing model from the raw variable data (French, 2017). Compared to traditional approaches, it has the advantage of not being restricted by linearity assumptions. Also, it enables certain flexibility when defining candidate factors. However, it does present some drawbacks, mainly related to overfitting. Recognizing that data snooping may play a role in the analysis, some methodological steps are adapted to implicitly handle for it. First, the investigated factors (features) are not selected using any kind of optimization algorithm, but are inspired by the results in previous papers and are defined beforehand. Second, all forecasts use only historical information and the returns are then estimated using future data not seen before by the ML algorithm; this amounts to a true out-of-sample test. Third, although other ML algorithms exist and have been used before (the most preeminent example is Artificial Neural Networks), the Random Forest method is specially selected here because its implementation implicitly handles data snooping by fitting the model on a subsample of observations randomly selected using bagging. Not all observations (from the training set) are seen by the optimization algorithm at any one time, leading to better generalization properties.

Although the portfolio resulting from the ML model is estimated and rebalanced at the end of each month, its returns can be computed at a daily frequency. Thus, daily returns on the market portfolio (proxied by the BET index) and hedge portfolios, SMB, HML, and MOM, are also computed contemporaneously, the latter three using the rebalancing procedure of Fama and French (1992, 2015), which was also used by Anghel, Dumitrescu and Tudor (2015). The resulting series enable the evaluation of ML portfolio returns (R_P) relative to some established asset pricing models. Specifically, we use as benchmark models the CAPM (Sharpe, 1964), the Fama-French 3-factor model (Fama and French, 1993), and the Carhart 4-factor model (Carhart, 1997). We use the resulting series aggregated at daily, weekly, and monthly frequencies to estimate each model and compare the results. The actual regressions that we test are defined as follows:

$$R_p - R_f = \alpha + \beta_1(R_m - R_f) + \varepsilon \tag{1}$$

$$R_p - R_f = \alpha + \beta_1(R_m - R_f) + \beta_2 SMB + \beta_3 HML + \varepsilon \tag{2}$$

$$R_p - R_f = \alpha + \beta_1(R_m - R_f) + \beta_2 SMB + \beta_3 HML + \beta_4 MOM + \varepsilon \tag{3}$$

where Eq. (1) represents the CAPM, Eq. (2) represents the Fama-French 3-factor model, and Eq. (3) represents the Carhart 4-factor model. The daily series are also used to estimate models on non-overlapping subsamples of 1 year, this being a useful robustness test that further enables the analysis of possible time-varying nature in results. We repeat the above procedure for various combinations of input features that correspond with the different factors. A comparison of the results enables us to draw inferences about their surplus informational content. Table 2 summarizes the different test runs. The quantities of interest are the estimated alpha (excess returns) and the regression R-squares (Fama and French, 1993).

TABLE 2
TEST RUNS

Run 1	Only variables related to the Market factor (betas) are used.
Run 2	Size and Value variables are added.
Run 3	Momentum variables are added.
Run 4	Volatility, Conditional Skewness, and Liquidity variables are added.
Run 5	Macroeconomic variables are added.
Run 6	Other variables are added.

RESULTS

The test results for all runs and all models are summarized in Table 3. Detailed results are available at request. Several interesting findings are worth noting. First, concerning the benchmark asset pricing models, individual and average regression R-squares show significant improvement from the CAPM to the Fama-French 3-factor model. Also, the loadings associated with the SMB and HML factors are statistically significant in most test runs, independent of data sample or sampling frequency. This implies that the Size and Value factors have significant explanatory power over asset price returns in the emerging Romanian stock market. On the other hand, we find no significant additional explanatory power for the Momentum factor. Specifically, the associated loadings are not statistically significant and R-squares do not improve when adding it, implying that momentum is not an important risk factor in this market. These results strongly resemble those of Anghel, Dumitrescu and Tudor (2015) or Zaremba and Konieczka (2015), which were obtained using more traditional research methods. On the one hand, our results enable us to extend their previous conclusions to an updated sample that spans until the beginning of 2020. On the other hand, this is a testimony that ML, even though unconventional, is a valid method for investigating stock market risk factors.

TABLE 3
TEST RESULTS

Panel A. Alphas

Data Sample (Frequency)	Run 1			Run 2			Run 3		
	CAPM	FF	Carhart	CAPM	FF	Carhart	CAPM	FF	Carhart
2011 (Daily)	-0.0003	-0.0004	0.0002	-0.0003	-0.0004	0.0004	-0.0015	-0.0016	-0.0012
	(0.7387)	(0.7122)	(0.7841)	(0.7623)	(0.7030)	(0.6070)	(0.1059)	(0.0792)	(0.1680)
2012 (Daily)	-0.0013	-0.0012	-0.0010	-0.0006	-0.0006	-0.0005	-0.0008	-0.0006	-0.0003
	(0.0099)	(0.0177)	(0.0247)	(0.2036)	(0.1956)	(0.2983)	(0.2660)	(0.4197)	(0.6151)
2013 (Daily)	-0.0006	-0.0006	-0.0005	-0.0007	-0.0006	-0.0006	0.0000	0.0000	0.0000
	(0.3008)	(0.2829)	(0.2819)	(0.1945)	(0.2133)	(0.2199)	(0.9697)	(0.9403)	(0.9864)
2014 (Daily)	0.0006	0.0004	0.0002	0.0011	0.0010	0.0008	-0.0002	-0.0004	-0.0005
	(0.4732)	(0.6111)	(0.7940)	(0.2710)	(0.2632)	(0.3539)	(0.7272)	(0.4443)	(0.2109)
2015 (Daily)	-0.0004	-0.0001	0.0000	-0.0002	0.0002	0.0003	0.0002	0.0005	0.0005
	(0.5064)	(0.9202)	(0.9469)	(0.7973)	(0.6960)	(0.6534)	(0.7428)	(0.3583)	(0.3475)
2016 (Daily)	0.0009	0.0001	0.0002	0.0011	0.0006	0.0005	0.0010	0.0005	0.0005
	(0.2428)	(0.8967)	(0.8054)	(0.0610)	(0.2737)	(0.3005)	(0.0856)	(0.3213)	(0.3425)
2017 (Daily)	0.0005	0.0004	0.0003	0.0008	0.0007	0.0006	0.0010	0.0009	0.0007
	(0.2648)	(0.3307)	(0.5155)	(0.0597)	(0.0764)	(0.1684)	(0.0751)	(0.1015)	(0.2111)
2018 (Daily)	-0.0011	-0.0010	-0.0010	0.0000	0.0001	0.0002	-0.0010	-0.0009	-0.0009
	(0.0178)	(0.0193)	(0.0238)	(0.9441)	(0.7876)	(0.7064)	(0.0517)	(0.0615)	(0.0735)
2019 (Daily)	0.0006	0.0006	0.0006	0.0006	0.0004	0.0004	0.0002	0.0001	0.0001
	(0.1416)	(0.1759)	(0.1719)	(0.2196)	(0.4130)	(0.3989)	(0.6955)	(0.8865)	(0.8867)
FULL SAMPLE (Daily)	-0.0001	-0.0002	-0.0002	0.0002	0.0001	0.0001	-0.0001	-0.0002	-0.0001
	(0.5022)	(0.3233)	(0.3511)	(0.3799)	(0.5136)	(0.4689)	(0.5960)	(0.4234)	(0.4528)
FULL SAMPLE (Weekly)	-0.0006	-0.0010	-0.0009	0.0009	0.0006	0.0007	-0.0005	-0.0008	-0.0007
	(0.5214)	(0.2635)	(0.3057)	(0.3459)	(0.5033)	(0.4265)	(0.5646)	(0.3317)	(0.3754)
FULL SAMPLE (Monthly)	-0.0030	-0.0049	-0.0045	0.0036	0.0018	0.0024	-0.0025	-0.0039	-0.0034
	(0.4965)	(0.1564)	(0.1809)	(0.4045)	(0.6131)	(0.4780)	(0.5555)	(0.3055)	(0.3584)

Data Sample (Frequency)	Run 4			Run 5			Run 6		
	CAPM	FF	Carhart	CAPM	FF	Carhart	CAPM	FF	Carhart
2011 (Daily)	-0.00022	-0.00034	-0.00031	-0.0034	-0.0036	-0.0031	-0.0033	-0.0035	-0.0030
	(0.8548)	(0.7766)	(0.7978)	(0.0442)	(0.0271)	(0.0502)	(0.0547)	(0.0367)	(0.0678)
2012 (Daily)	0.00068	0.00062	0.00061	0.0025	0.0025	0.0026	0.0022	0.0022	0.0023
	(0.2116)	(0.2606)	(0.2723)	(0.0006)	(0.0007)	(0.0003)	(0.0030)	(0.0038)	(0.0023)
2013 (Daily)	0.00105	0.00110	0.00109	0.0012	0.0012	0.0013	0.0008	0.0008	0.0009
	(0.0190)	(0.0137)	(0.0152)	(0.0773)	(0.0609)	(0.0323)	(0.2331)	(0.2065)	(0.1434)
2014 (Daily)	0.00045	0.00042	0.00042	0.0003	0.0001	0.0000	0.0008	0.0006	0.0006
	(0.3243)	(0.3580)	(0.3597)	(0.6826)	(0.8649)	(0.9539)	(0.1195)	(0.2309)	(0.2762)
2015 (Daily)	-0.00042	-0.00039	-0.00042	0.0000	0.0005	0.0005	0.0001	0.0004	0.0005
	(0.6006)	(0.6294)	(0.5938)	(0.9771)	(0.4207)	(0.3567)	(0.8986)	(0.3892)	(0.3319)
2016 (Daily)	0.00114	0.00114	0.00116	0.0012	0.0011	0.0010	0.0014	0.0010	0.0010
	(0.0061)	(0.0060)	(0.0052)	(0.0116)	(0.0150)	(0.0227)	(0.0087)	(0.0377)	(0.0470)
2017 (Daily)	0.00173	0.00177	0.00175	0.0010	0.0009	0.0007	0.0010	0.0009	0.0007
	(0.0006)	(0.0005)	(0.0005)	(0.0224)	(0.0322)	(0.0712)	(0.0281)	(0.0394)	(0.0930)
2018 (Daily)	-0.00013	-0.00014	-0.00015	0.0003	0.0003	0.0003	-0.0004	-0.0003	-0.0002
	(0.8192)	(0.8120)	(0.7992)	(0.5371)	(0.5179)	(0.5580)	(0.5868)	(0.6913)	(0.7400)
2019 (Daily)	0.00063	0.00069	0.00069	0.0000	0.0000	0.0000	0.0005	0.0005	0.0005
	(0.2395)	(0.1985)	(0.1997)	(0.9700)	(0.9596)	(0.9598)	(0.4303)	(0.4532)	(0.4541)
FULL SAMPLE (Daily)	0.00053	0.00052	0.00053	0.0005	0.0004	0.0004	0.0005	0.0005	0.0005
	(0.0082)	(0.0087)	(0.0079)	(0.0380)	(0.0482)	(0.0459)	(0.0401)	(0.0521)	(0.0492)
FULL SAMPLE (Weekly)	0.00253	0.00247	0.00248	0.0023	0.0020	0.0020	0.0023	0.0020	0.0021
	(0.0059)	(0.0074)	(0.0070)	(0.0589)	(0.0827)	(0.0806)	(0.0637)	(0.0936)	(0.0863)
FULL SAMPLE (Monthly)	0.01046	0.00982	0.00984	0.0092	0.0079	0.0085	0.0089	0.0073	0.0077
	(0.0028)	(0.0052)	(0.0053)	(0.1273)	(0.1776)	(0.1396)	(0.1558)	(0.2243)	(0.1991)

Panel B. R-squares

Data Sample (Frequency)	Run 1			Run 2			Run 3		
	CAPM	FF	Carhart	CAPM	FF	Carhart	CAPM	FF	Carhart
2011 (Daily)	0.5623	0.5662	0.5662	0.5718	0.5874	0.5874	0.5493	0.5766	0.5766
2012 (Daily)	0.5016	0.5100	0.5100	0.4559	0.4653	0.4653	0.2380	0.3310	0.3310
2013 (Daily)	0.2804	0.3009	0.3009	0.2800	0.2795	0.2795	0.2557	0.3215	0.3215
2014 (Daily)	0.0477	0.3447	0.3447	0.0655	0.3494	0.3494	0.2754	0.2971	0.2971
2015 (Daily)	0.2109	0.4033	0.4033	0.2166	0.4151	0.4151	0.1900	0.3600	0.3600
2016 (Daily)	0.1125	0.4221	0.4221	0.2689	0.4603	0.4603	0.2563	0.4076	0.4076
2017 (Daily)	0.1798	0.2043	0.2043	0.2118	0.2426	0.2426	0.1978	0.2306	0.2306
2018 (Daily)	0.3497	0.3812	0.3812	0.4970	0.5268	0.5268	0.5389	0.5918	0.5918
2019 (Daily)	0.2435	0.2436	0.2436	0.2339	0.2930	0.2930	0.3040	0.3190	0.3190
FULL SAMPLE (Daily)	0.2529	0.3358	0.3358	0.2912	0.3701	0.3701	0.3241	0.3870	0.3870
FULL SAMPLE (Weekly)	0.2485	0.3955	0.3955	0.3136	0.4154	0.4154	0.3964	0.4895	0.4895
FULL SAMPLE (Monthly)	0.3029	0.5895	0.5895	0.3709	0.5771	0.5771	0.4167	0.5268	0.5268

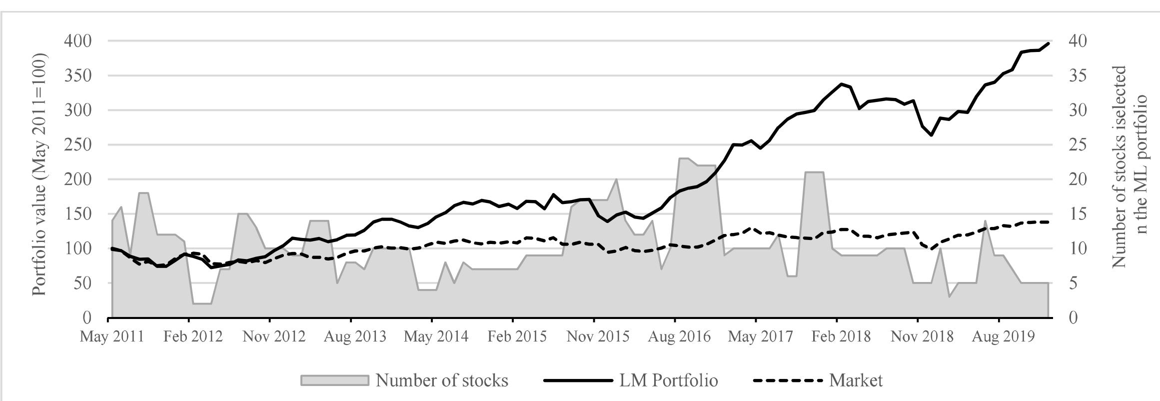
Data Sample (Frequency)	Run 4			Run 5			Run 6		
	CAPM	FF	Carhart	CAPM	FF	Carhart	CAPM	FF	Carhart
2011 (Daily)	0.5354	0.5458	0.5458	0.5354	0.5838	0.5838	0.2828	0.3492	0.3492
2012 (Daily)	0.5258	0.5231	0.5231	0.5258	0.5474	0.5474	0.2680	0.2797	0.2797
2013 (Daily)	0.3407	0.3471	0.3471	0.3407	0.3362	0.3362	0.2258	0.2327	0.2327
2014 (Daily)	0.4206	0.4194	0.4194	0.4206	0.4162	0.4162	0.3297	0.3352	0.3352
2015 (Daily)	0.1750	0.1690	0.1690	0.1750	0.3828	0.3828	0.2725	0.4874	0.4874
2016 (Daily)	0.5054	0.5049	0.5049	0.5054	0.5078	0.5078	0.4249	0.4214	0.4214
2017 (Daily)	0.1633	0.1584	0.1584	0.1633	0.1717	0.1717	0.2420	0.2790	0.2790
2018 (Daily)	0.3702	0.3660	0.3660	0.3702	0.4745	0.4745	0.6638	0.6618	0.6618
2019 (Daily)	0.3274	0.3273	0.3273	0.3274	0.3293	0.3293	0.4357	0.4381	0.4381
FULL SAMPLE (Daily)	0.3796	0.3791	0.3791	0.3796	0.4005	0.4005	0.3337	0.3501	0.3501
FULL SAMPLE (Weekly)	0.4390	0.4385	0.4385	0.4390	0.4706	0.4706	0.3208	0.3647	0.3647
FULL SAMPLE (Monthly)	0.5967	0.6045	0.6045	0.5967	0.6273	0.6273	0.3491	0.3803	0.3803

Note. This table reports the alpha (α) and Adjusted R-squared from regressing ML portfolio returns (R_p) on three benchmark models namely,
CAPM: $R_p - R_f = \alpha + \beta_1(R_m - R_f) + \varepsilon$;
Fama and French 3-factor model: $R_p - R_f = \alpha + \beta_1(R_m - R_f) + \beta_2 SMB + \beta_3 HML + \varepsilon$;
Carhart 4-factor model: $R_p - R_f = \alpha + \beta_1(R_m - R_f) + \beta_2 SMB + \beta_3 HML + \beta_4 MOM + \varepsilon$.
P-values are reported in round parenthesis.

Second, except for a few sporadic results, the first three runs that incorporate Market, Size, Value, and Momentum-related information (features) in the ML forecasting model do not generate portfolios capable of earning positive and statistically significant alphas over any benchmark asset pricing model. This is rather expected, as the benchmarks account for these factors. Interestingly, adding all 6 momentum-related variables does not lead to positive alphas even compared to the CAPM and FF3 models. Given that these variables are more diverse compared to the classical definition of momentum, the results enable us to rule out Momentum as a risk factor for asset prices in Romania. Given the enduring significance and popularity of this factor in international markets, this shows that emerging stock markets, such as the ones in Central and Eastern Europe, do behave differently in some aspects compared to their more developed counterparts.

Third, we find that the results significantly change starting in Run 4 (detailed results are presented in Table B1 in Appendix B). Specifically, adding the Volatility, Conditional Skewness and Liquidity-related features increases the performance of the ML portfolio and generates positive and statistically significant alphas. Figure 1 shows the out-of-sample performance of this portfolio compared to the market. The annualized risk premium (alpha) that is earned by it amounts to 12.55%-13.25% when estimated versus the CAPM and 11.5%-12.5% when estimated versus the Fama-French 3-factor model (using the Carhart 4-factor model generates similar results), depending on the data frequency used. In the context of the Fama-French 3-factor model, this result is significant at the 98.89% level when daily data is used, which increases to 99.55% when monthly data is used. This is well above the $t \geq 3$ limit discussed by Harvey, Liu and Zhu (2016) for new discoveries in asset pricing and leads to the overall conclusion that Volatility, Conditional Skewness and Turnover constitute significant risk factors in the emerging stock market of Romania. Interestingly, adding one factor at a time or two factors at a time (results are reported in Table B2, Appendix B) does not lead to statistically significant results. Among the bunch, considering Turnover alone leads to borderline significant results (p-values range between 0.11 and 0.18); adding Conditional Skewness increases the statistical significance of the alphas (significance is attained at the 5% level); while adding Volatility increases it even more (significance is attained at the 1% level). Thus, the Turnover factor is the dominant one among the bunch, but it seems that it only outperforms the benchmarks when accompanied by the other factors, which contribute by providing complementary information to the ML model.

FIGURE 1
ML PORTFOLIO (RUN 4) PERFORMANCE VS. MARKET (BET INDEX)



Forth, we find evidence of the ML algorithm learning the new asset pricing factors as more observations are added to the training sample. This arises from looking at the tests performed in Run 4 on non-overlapping yearly subsamples. Specifically, the alphas in the early years are insignificant and mostly negative, but turn out positive and statistically significant in the later years. This reinforces the validity of the newly discovered risk factors but additionally shows that they have been present in the data waiting to be discovered. In this context, ML is shown to be a useful tool for this purpose.

Finally, we find that the overall results from Run 4 do not improve in the following Runs when adding Macroeconomic and the CSSD factors. We do find some years for which these latter Runs lead to significant positive results but those are accompanied by other years that post significant negative results. Also, the overall stability of predictions significantly decreases in Runs 5 and 6, as inferred from looking at the R-squares. These results rather imply that macroeconomic and CSSD information are already incorporated in the other variables and that adding the variables introduces excessive noise that disrupts the ML algorithm, leading to suboptimal results.

CONCLUSIONS

This paper analyzes asset pricing factors in the emerging stock market of Romania using the unconventional approach of Machine Learning algorithms. Based on input variables linked to a combination of previously investigated and new risk factors, ML is used to forecast future price movements and construct portfolios capable of generating significant excess returns relative to popular asset pricing models. Besides accounting for nonlinear relationships, this approach is especially useful in an emerging market that has very few listed stocks. Romania is one such example.

Our results support previous conclusions derived from the related literature analyzing Romania and, more generally, markets in the Central and Eastern European region. Specifically, we find that the Market, Size, and Value factors are significant determinants of long-term asset prices, while Momentum is not. Our main contribution to the literature is showing that the Turnover, Conditional Skewness, and Volatility factors are also significant, generating average excess returns of roughly 12% per year. This is a $t \geq 3$ results, which is significant at 99%, well above the threshold discussed by Harvey, Liu and Zhu (2016) for new discoveries. Interestingly, the three factors independently contribute to portfolio overperformance but seem to work best together.

Overall, our results have important implications for investors and, more generally, stakeholders in emerging stock markets. On the one hand, they show that significant long-term risk premiums can be earned when selecting stocks based on liquidity, conditional skewness, and volatility. On the other hand, they show that ML can be a relevant tool for making investment decisions, and, more generally, it can help researchers study important research questions, such as uncovering asset pricing factors in difficult setups.

ACKNOWLEDGEMENTS

This paper was co-financed from the Human Capital Operational Program 2014-2020, project number POCU / 380/6/13/125245 no. 36482 / 23.05.2019 “Excellence in interdisciplinary PhD and post-PhD research, career alternatives through entrepreneurial initiative (EXCIA)”, coordinator The Bucharest University of Economic Studies.

REFERENCES

Andrieş, A.M., Ihnatov, I., & Sprincean, N. (2017). Do seasonal anomalies still exist in central and eastern European countries? A conditional variance approach. *Romanian Journal of Economic Forecasting*, *20*(4), 60-83.

Ang, A., Hodrick, R.J., Xing, Y., & Zhang, X. (2006). The cross-section of volatility and expected returns. *The Journal of Finance*, *61*(1), 259-299.

Anghel, D.G. (2013). How reliable is the moving average crossover rule for an investor on the Romanian stock market? *The Review of Finance and Banking*, *5*(2), 89-115.

Anghel, D.G. (2015). Market efficiency and technical analysis in Romania. *International Journal of Financial Research*, *6*(2), 164-177.

Anghel, D.G. (2017). Intraday Market Efficiency for a Typical Central and Eastern European Stock Market: The Case of Romania. *Romanian Journal of Economic Forecasting*, *20*(3), 88-109.
Anghel, A., Dumitrescu, D., & Tudor, C. (2015). Modeling portfolio returns on Bucharest Stock exchange using the Fama-French multifactor model. *Romanian Journal of Economic Forecasting*, *17*(1), 22-46.
Anghelache, G.V., & Trifan, A.L. (2013). Forecasting the investors behavior on the capital market in Romania: Trading strategies based on technical analysis versus Artificial Intelligence techniques. *International Journal of Business and Social Research*, *3*(2), 114-121.
Birău, R., Ehsanifar, M., &Mohammadi, H. (2013). Forecasting the Bucharest Stock Exchange BET-C Index based on Artificial Neural Network and Multiple Linear Regressions. In *1st WSEAS International Conference on Mathematics, Statistics &Computer Engineering*. Dubrovnik, Croatia.
Breiman, L. (2001). Random forests. *Machine Learning*, *45*(1), 5-32.
Cao, Q., Parry, M.E., & Leggio, K.B. (2011). The three-factor model and artificial neural networks: predicting stock price movement in China. *Annals of Operations Research*, *185*(1), 25-44.
Carhart, M.M. (1997). On persistence in mutual fund performance. *The Journal of Finance*, *52*(1), 57-82.
Chang, E.C., Cheng, J.W., & Khorana, A. (2000). An examination of herd behavior in equity markets: An international perspective. *Journal of Banking & Finance*, *24*(10), 1651-1679.
Chordia, T., Subrahmanyam, A., & Anshuman, V.R. (2001). Trading activity and expected stock returns. *Journal of Financial Economics*, *59*(1), 3-32.
Dragota, V., & Oprea, D.S. (2014). Informational efficiency tests on the Romanian stock market: a review of the literature. *The Review of Finance and Banking*, *6*(1), 15-28.
Dragotă, V., & Ţilică, E.V. (2014). Market efficiency of the Post Communist East European stock markets. *Central European Journal of Operations Research*, *22*(2), 307-337.
Fama, E.F. (1970). Efficient capital markets: A review of theory and empirical work. *The Journal of Finance*, *25*(2), 383-417.
Fama, E.F., & French, K.R. (1992). The cross-section of expected stock returns. *The Journal of Finance*, *47*(2), 427-465.
Fama, E.F., & French, K.R. (1993). Common risk factors in the returns on stocks and bonds. *Journal of Financial Economics*, *33*, 3-56.
Fama, E.F., & French, K.R. (2015). A five-factor asset pricing model. *Journal of Financial Economics*, *116*(1), 1-22.
French, J. (2017). The time traveller's CAPM. *Investment Analysts Journal*, *46*(2), 81-96.
Géron, A. (2019). *Hands-On Machine Learning with Scikit-Learn, Keras, and TensorFlow: Concepts, Tools, and Techniques to Build Intelligent Systems*. O'Reilly Media.
Jagannathan, R., & Wang, Z. (1996). The conditional CAPM and the cross-section of expected returns. *The Journal of Finance*, *51*(1), 3-53.
Harvey, C.R., Liu, Y., & Zhu, H. (2016). … and the cross-section of expected returns. *The Review of Financial Studies*, *29*(1), 5-68.
Harvey, C.R., & Siddique, A. (2000). Conditional skewness in asset pricing tests. *The Journal of Finance*, *55*(3), 1263-1295.
Jegadeesh, N., & Titman, S. (1993). Returns to buying winners and selling losers: Implications for stock market efficiency. *The Journal of Finance*, *48*(1), 65-91.
Jan, M.N., & Ayub, U. (2019). Do the FAMA and FRENCH Five-Factor model forecast well using ANN? *Journal of Business Economics and Management*, *20*(1), 168-191.
Oprea, D.S., & Brad, L. (2014). Investor sentiment and stock returns: evidence from Romania. *International Journal of Academic Research in Accounting, Finance and Management Sciences*, *4*(2), 19-25.
Ruxanda, G., & Badea, L.M. (2014). Configuring artificial neural networks for stock market predictions. *Technological and Economic Development of Economy*, *20*(1), 116-132.

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

Stoitsova-Stoykova, A. (2017). Relationship Between Public Expectations and Financial Market Dynamics in South-East Europe Capital Markets. *Economic Alternatives*, (2), 237-250.

Stoykova, A.Q., Paskaleva, M.G., & Stoykov, D.Z. (2018). Impact of sentiment indicators on the capital market dynamics and default probability. *International Journal of Contemporary Economics and Administrative Sciences*, *8*(2), 129-153.

Tudor, C. (2009). Price Ratios and the Cross-section of Common Stock Returns on Bucharest Stock Exchange: Empirical Evidence. *Romanian Journal of Economic Forecasting*, *10*(2), 132-46.

Zaremba, A. (2015). Low risk anomaly in the CEE stock markets. *Romanian Journal of Economic Forecasting*, *18*(3), 81-102.

Zaremba, A., & Konieczka, P. (2015). Are value, size and momentum premiums in CEE emerging markets only illusionary? *Finance a úvěr-Czech Journal of Economics and Finance*, *65*(1), 84-104.

APPENDIX A

Implementation of Machine Learning Forecasting Model

The Machine Learning (ML) classification model for forecasting long-term stock returns is implemented in *Phyton* using the following function:

TABLE A1
PYTHON FUNCTION USED TO IMPLEMENTS RANDOM FOREST MODEL

Name	Python Package	Implementation
Random Forest	sklearn	RandomForestClassifier(bootstrap=True, ccp_alpha=0.0, class_weight=None, criterion='gini', max_depth=None, max_features='auto', max_leaf_nodes=None, max_samples=None, min_impurity_decrease=0.0, min_impurity_split=None, min_samples_leaf=1, min_samples_split=2, min_weight_fraction_leaf=0.0, n_estimators=5000, n_jobs=None, oob_score=False, random_state=7, verbose=0, warm_start=False)

Géron (2019) provides details regarding the role of the different hyperparameters and other properties of ML forecasting algorithms. Note that bagging is implicitly implemented in the *RandomForestClassifier* class.

APPENDIX B

Additional Results

TABLE B1
DETAILED RESULTS OBTAINED IN RUN 4

Data Sample (Frequency)	Obs.	CAPM			Fama-French 3 factor model					Carhart 4-factor model					
		α	β_1	Adj. R^2	α	β_1	β_2	β_3	Adj. R^2	α	β_1	β_2	β_3	β_4	Adj. R^2
2011 (Daily)	149	-0.0002	0.9794	0.5354	-0.0003	1.0949	0.2061	0.0118	0.5458	-0.0003	1.0780	0.1918	0.0067	-0.0367	0.5458
		(0.8548)	(0.0000)		(0.7766)	(0.0000)	(0.0312)	(0.8581)		(0.7978)	(0.0000)	(0.0463)	(0.9188)	(0.2578)	
2012 (Daily)	250	0.0007	0.8664	0.5258	0.0006	0.8755	0.0282	0.0020	0.5231	0.0006	0.8832	0.0259	0.0058	0.0147	0.5231
		(0.2116)	(0.0000)		(0.2606)	(0.0000)	(0.4553)	(0.9474)		(0.2723)	(0.0000)	(0.4938)	(0.8457)	(0.3662)	
2013 (Daily)	251	0.0011	0.6517	0.3407	0.0011	0.6552	-0.0387	0.0248	0.3471	0.0011	0.6502	-0.0366	0.0223	-0.0072	0.3471
		(0.0190)	(0.0000)		(0.0137)	(0.0000)	(0.1671)	(0.2731)		(0.0152)	(0.0000)	(0.1950)	(0.3300)	(0.5220)	
2014 (Daily)	250	0.0005	0.7871	0.4206	0.0004	0.7998	0.0051	0.0265	0.4194	0.0004	0.7999	0.0052	0.0266	0.0004	0.4194
		(0.3243)	(0.0000)		(0.3580)	(0.0000)	(0.8421)	(0.2403)		(0.3597)	(0.0000)	(0.8415)	(0.2606)	(0.9808)	
2015 (Daily)	251	-0.0004	0.7169	0.1750	-0.0004	0.7303	0.0294	0.0149	0.1690	-0.0004	0.6967	-0.0188	-0.1633	-0.1521	0.1690
		(0.6006)	(0.0000)		(0.6294)	(0.0000)	(0.6733)	(0.8023)		(0.5938)	(0.0000)	(0.7812)	(0.0182)	(0.0000)	
2016 (Daily)	254	0.0011	0.7399	0.5054	0.0011	0.6999	-0.0080	-0.0598	0.5049	0.0012	0.6954	-0.0257	-0.0701	-0.0331	0.5049
		(0.0061)	(0.0000)		(0.0060)	(0.0000)	(0.8698)	(0.1965)		(0.0052)	(0.0000)	(0.6065)	(0.1335)	(0.1365)	
2017 (Daily)	224	0.0017	0.4526	0.1633	0.0018	0.4448	0.0051	-0.0128	0.1584	0.0018	0.4500	0.0135	-0.0104	-0.0349	0.1584
		(0.0006)	(0.0000)		(0.0005)	(0.0000)	(0.9246)	(0.7275)		(0.0005)	(0.0000)	(0.8025)	(0.7775)	(0.3424)	
2018 (Daily)	249	-0.0001	0.5953	0.3702	-0.0001	0.6025	0.0241	0.0109	0.3660	-0.0002	0.5967	0.0183	0.0088	0.0595	0.3660
		(0.8192)	(0.0000)		(0.8120)	(0.0000)	(0.5626)	(0.7310)		(0.7992)	(0.0000)	(0.6616)	(0.7813)	(0.1661)	
2019 (Daily)	249	0.0006	0.6613	0.3274	0.0007	0.6149	-0.0384	-0.0581	0.3273	0.0007	0.6144	-0.0385	-0.0576	0.0025	0.3273
		(0.2395)	(0.0000)		(0.1985)	(0.0000)	(0.4638)	(0.1610)		(0.1997)	(0.0000)	(0.4643)	(0.1717)	(0.9539)	
FULL SAMPLE (Daily)	2152	0.0005	0.7579	0.3796	0.0005	0.7632	0.0090	0.0069	0.3791	0.0005	0.7578	0.0080	0.0001	-0.0234	0.3791
		(0.0082)	(0.0000)		(0.0087)	(0.0000)	(0.5405)	(0.5560)		(0.0079)	(0.0000)	(0.5893)	(0.9911)	(0.0013)	
FULL SAMPLE (Weekly)	449	0.0025	0.7551	0.4390	0.0025	0.7730	0.0372	0.0269	0.4385	0.0025	0.7685	0.0364	0.0210	-0.0176	0.4385
		(0.0059)	(0.0000)		(0.0074)	(0.0000)	(0.2631)	(0.2400)		(0.0070)	(0.0000)	(0.2731)	(0.3705)	(0.2772)	
FULL SAMPLE (Monthly)	105	0.0105	0.8674	0.5967	0.0098	0.9341	0.0967	0.0979	0.6045	0.0098	0.9312	0.0965	0.0947	-0.0051	0.6045
		(0.0028)	(0.0000)		(0.0052)	(0.0000)	(0.1778)	(0.0467)		(0.0053)	(0.0000)	(0.1804)	(0.0777)	(0.8777)	

Note. This table presents the estimated coefficients and the results of the associated statistical significance tests in Run 4 (see Table 2), when the volatility, liquidity and conditional skewness factors are added alongside the market, size, value, and momentum factors, in order to train the ML model and generate portfolio returns. P-values are reported in round parenthesis.

TABLE B2
RESULTS OBTAINED WHEN COMBINING VOLATILITY, LIQUIDITY, AND CONDITIONAL SKEWNESS IN VARIOUS WAYS

Data Sample (Frequency)	Run3 + VOL			Run3 + COSKEW			Run3 + TURN			Run3 + VOL + COSKEW			Run3 + VOL + TURN			Run3 + COSKEW + TURN		
	CAPM	FF	Carhart	CAPM	FF	Carhart	CAPM	FF	Carhart	CAPM	FF	Carhart	CAPM	FF	Carhart	CAPM	FF	Carhart
2011 (Daily)	-0.00164	-0.00171	-0.00114	-0.00151	-0.00158	-0.00118	-0.00128	-0.00134	-0.00072	-0.00075	-0.00083	-0.00029	-0.00084	-0.00093	-0.00061	-0.00095	-0.00101	-0.00060
	(0.1150)	(0.0972)	(0.2319)	(0.0920)	(0.0708)	(0.1556)	(0.1599)	(0.1339)	(0.3590)	(0.4613)	(0.4055)	(0.7509)	(0.4909)	(0.4330)	(0.6060)	(0.2734)	(0.2270)	(0.4485)
2012 (Daily)	-0.00059	-0.00030	-0.00004	-0.00052	-0.00038	-0.00010	0.00081	0.00101	0.00107	-0.00045	-0.00015	0.00010	0.00028	0.00037	0.00048	0.00077	0.00084	0.00083
	(0.3775)	(0.6366)	(0.9396)	(0.4691)	(0.5822)	(0.8690)	(0.3228)	(0.2088)	(0.1833)	(0.5279)	(0.8209)	(0.8713)	(0.6435)	(0.5401)	(0.4161)	(0.1417)	(0.1016)	(0.1090)
2013 (Daily)	0.00038	0.00038	0.00041	-0.00002	0.00000	0.00003	0.00068	0.00062	0.00065	0.00069	0.00070	0.00074	0.00054	0.00057	0.00060	0.00066	0.00061	0.00064
	(0.4724)	(0.4708)	(0.4064)	(0.9686)	(0.9938)	(0.9338)	(0.1629)	(0.1980)	(0.1618)	(0.2285)	(0.2085)	(0.1646)	(0.3345)	(0.3041)	(0.2664)	(0.1462)	(0.1770)	(0.1326)
2014 (Daily)	-0.00017	-0.00049	-0.00070	0.00095	0.00073	0.00052	-0.00023	-0.00030	-0.00040	-0.00005	-0.00043	-0.00062	0.00023	0.00003	-0.00004	0.00018	0.00012	0.00000
	(0.7587)	(0.3726)	(0.1352)	(0.3989)	(0.4043)	(0.5317)	(0.6870)	(0.5962)	(0.4656)	(0.9275)	(0.3995)	(0.1641)	(0.6062)	(0.9413)	(0.9187)	(0.7592)	(0.8362)	(0.9941)
2015 (Daily)	-0.00039	0.00002	0.00003	0.00000	0.00033	0.00031	-0.00024	0.00015	0.00016	0.00005	0.00045	0.00045	-0.00001	0.00046	0.00051	-0.00003	0.00042	0.00045
	(0.5542)	(0.9716)	(0.9599)	(0.9995)	(0.5527)	(0.5710)	(0.7340)	(0.8053)	(0.7825)	(0.9311)	(0.3931)	(0.3940)	(0.9919)	(0.5373)	(0.4963)	(0.9665)	(0.5095)	(0.4827)
2016 (Daily)	0.00105	0.00048	0.00044	0.00112	0.00061	0.00056	0.00077	0.00078	0.00072	0.00126	0.00075	0.00070	0.00101	0.00102	0.00093	0.00099	0.00101	0.00094
	(0.0911)	(0.3735)	(0.4087)	(0.0718)	(0.2727)	(0.3135)	(0.0546)	(0.0520)	(0.0714)	(0.0368)	(0.1569)	(0.1839)	(0.0164)	(0.0143)	(0.0214)	(0.0160)	(0.0147)	(0.0211)
2017 (Daily)	0.00068	0.00063	0.00042	0.00078	0.00068	0.00036	0.00099	0.00085	0.00063	0.00071	0.00062	0.00035	0.00118	0.00109	0.00096	0.00127	0.00116	0.00093
	(0.2044)	(0.2436)	(0.4310)	(0.1756)	(0.2286)	(0.5087)	(0.0929)	(0.1343)	(0.2591)	(0.2406)	(0.2929)	(0.5529)	(0.0385)	(0.0523)	(0.0885)	(0.0176)	(0.0256)	(0.0709)
2018 (Daily)	-0.00039	-0.00030	-0.00025	-0.00083	-0.00065	-0.00061	-0.00021	-0.00014	-0.00008	-0.00023	-0.00012	-0.00008	-0.00040	-0.00029	-0.00022	-0.00024	-0.00015	-0.00009
	(0.4470)	(0.5243)	(0.5967)	(0.1663)	(0.2375)	(0.2665)	(0.7658)	(0.8205)	(0.9030)	(0.6580)	(0.7983)	(0.8596)	(0.5760)	(0.6664)	(0.7431)	(0.6823)	(0.7842)	(0.8652)
2019 (Daily)	0.00039	0.00031	0.00031	0.00014	0.00003	0.00003	0.00080	0.00071	0.00071	0.00037	0.00033	0.00033	0.00083	0.00070	0.00070	0.00081	0.00076	0.00076
	(0.4039)	(0.5147)	(0.5127)	(0.8108)	(0.9577)	(0.9576)	(0.1517)	(0.2012)	(0.1964)	(0.5272)	(0.5758)	(0.5764)	(0.1343)	(0.2054)	(0.1871)	(0.1773)	(0.2056)	(0.2043)
FULL SAMPLE (Daily)	-0.00005	-0.00009	-0.00008	0.00006	-0.00001	0.00000	0.00032	0.00029	0.00029	0.00019	0.00015	0.00016	0.00036	0.00033	0.00033	0.00044	0.00040	0.00041
	(0.8200)	(0.6378)	(0.6807)	(0.8058)	(0.9445)	(0.9952)	(0.1280)	(0.1647)	(0.1556)	(0.3521)	(0.4557)	(0.4146)	(0.0906)	(0.1180)	(0.1141)	(0.0248)	(0.0333)	(0.0319)
FULL SAMPLE (Weekly)	-0.00020	-0.00050	-0.00041	0.00027	-0.00011	-0.00001	0.00150	0.00130	0.00134	0.00095	0.00067	0.00077	0.00170	0.00144	0.00147	0.00205	0.00187	0.00189
	(0.8382)	(0.5747)	(0.6405)	(0.8089)	(0.9113)	(0.9915)	(0.1103)	(0.1518)	(0.1394)	(0.3250)	(0.4601)	(0.3802)	(0.0788)	(0.1156)	(0.1080)	(0.0249)	(0.0359)	(0.0336)
FULL SAMPLE (Monthly)	-0.00129	-0.00267	-0.00210	0.00069	-0.00130	-0.00080	0.00600	0.00525	0.00541	0.00356	0.00214	0.00269	0.00674	0.00555	0.00553	0.00850	0.00774	0.00787
	(0.7564)	(0.4610)	(0.5405)	(0.8840)	(0.7453)	(0.8362)	(0.1280)	(0.1747)	(0.1634)	(0.3801)	(0.5474)	(0.4236)	(0.0776)	(0.1162)	(0.1194)	(0.0327)	(0.0481)	(0.0454)

Note. This table presents the estimated alphas (α) and Adjusted R-squares when the volatility, liquidity and conditional skewness factors are added in different combinations alongside the market, size, value, and momentum factors, in order to train the ML model and generate portfolio returns (see Table 2). The ML portfolio returns (R_p) are actually regressed using three benchmark models namely,

CAPM: $R_p - R_f = \alpha + \beta_1(R_m - R_f) + \varepsilon$;

Fama and French 3-factor model: $R_p - R_f = \alpha + \beta_1(R_m - R_f) + \beta_2 SMB + \beta_3 HML + \varepsilon$;

Carhart 4-factor model: $R_p - R_f = \alpha + \beta_1(R_m - R_f) + \beta_2 SMB + \beta_3 HML + \beta_4 MOM + \varepsilon$.

P-values are reported in round parenthesis.