The Joint Cross Section of Option and Stock Returns Predictability with Big Data and Machine Learning *

Ruslan Goyenko† **Chengyu Zhang**‡

March 22, 2022

Abstract

Which market has leading informational advantage: stocks or options? Using large set of stock and option characteristics, and machine learning, we provide a comprehensive analysis of which characteristics are the first order importance predictors of option and stock returns. First, we find that *option*, rather than stock, characteristics are dominant predictors of *option returns*. Second, *option*, rather than stock, characteristics are also dominant predictors of *stock returns*. Consistent with the argument that an increase in trading activity in derivatives decreases information asymmetry about the underlying, option illiquidity is identified as the most important predictor of both stock and option returns.

Keywords: Machine learning, Option pricing, Stock return predictability
JEL Codes: G10, G12, G13, G14

*We are grateful to Christian Dorion, Benjamin Golez, Bing Han, Dmitriy Muravyev, Neil Pearson, Paul Schultz, Dacheng Xiu, and participants of VDW (Virtual Derivatives Workshop, April, 2021), and SoFiE 2021 UCSD Conference for helpful comments and suggestions.

†Department of Finance, Desautels Faculty of Management, McGill University, 1001 rue Sherbrooke Ouest, Montreal, Quebec, Canada H3A 1G5, e-mail: ruslan.goyenko@mcgill.ca

‡Department of Finance, Desautels Faculty of Management, McGill University, 1001 rue Sherbrooke Ouest, Montreal, Quebec, Canada H3A 1G5, e-mail: chengyu.zhang@mail.mcgill.ca

Electronic copy available at: https://ssrn.com/abstract=3747238

1 Introduction

Which market, stocks or options, has an informational advantage, that is which market leads? This question has been debated in the literature since the inception of options markets.

Theoretical and empirical literature argues that options markets increase price efficiency (Cao (1999), Easley, O'hara, and Srinivas (1998), Chakravarty, Gulen, and Mayhew (2004), Pan and Poteshman (2006)) as options may allow informed agents to obtain leverage more readily. Options trading also decreases information asymmetry overall and contributes to greater price transparency (Cao (1999), Roll, Schwartz, and Subrahmanyam (2009)). To the date, however, there is no clear conclusion in the literature about whether the options market leads the stock market, or whether the option market has any informational advantage or significant complimentarity compared to the stock market.[1]

The literature studying *options* returns predictability is dominated by *stock-based* characteristics as the main predictors, suggesting that the stock market leads.[2] Further, the stock return predictability literature considers stock characteristics alone as the main predictors (Green, Hand, and Zhang (2017)), while there is an extensive co-existing literature identifying informed signals about future stock returns using options data.[3] Yet, option market predictors are not as prioritized.

This is surprising, as according to recent data, the options market is one of the largest markets in the world in terms of trading activity.[4] It has become an investment vehicle for both institutional and retail investors.[5] Given historically high popularity of options among all investors, and availability of bigger and longer cross-sectional and time series data, compared to other studies, we ask the following question. Does an unprecedented rise in options trading allow for better identification of whether options reflect more information about future both option and stock returns than the stock market alone? In other words, with more options data available, and more modern techniques – can we resolve a long lasting debate – which market leads, if any, and why?

[1] Section 2 provides the literature review.

[2] See among others the following predictors of options returns: stock idiosyncratic volatility (Cao and Han (2013)), lottery like stock preferences (Byun and Kim (2016)), stock short-sale constraints (Ramachandran and Tayal (2020)), stock illiquidity (Kanne, Korn, and Uhrig-Homburg (2018)), 10 stock characteristics related to cash flow variance, the cash-to-asset ratio, analyst earning forecast dispersion, 1 - and 5-year changes in shares outstanding, profit margins, profitability, stock price, external financing and Z-score (Cao, Han, Tong, and Zhan (2021)). There are few exceptions ((Goyal and Saretto (2009), Vasquez (2017) and Boyer and Vorkink (2014) who use options characteristics to predict options returns. However, these predictors are highly outnumbered by stock-based predictors.

[3] See among others Pan and Poteshman (2006), An, Ang, Bali, and Cakici (2014)), Cremers and Weinbaum (2010), Xing, Zhang, and Zhao (2010), Johnson and So (2012) and Cremers, Goyenko, Schultz, and Szaura (2019)

[4] As of writing of this paper, in the end of August 2020, single stock options volume made up more than 120% of the volume of shares of underlying stocks in notional value (source, WSJ, September 13 , 2020, "The widely popular trades behind the market's swoon and surge", by G. Zuckerman and G. Banerji). This makes options market the biggest (synthetically) stock market in the world.

[5] The exponentially increasing option volume is largely attributed to the influx of retail investors (Zuckerman and Banerji, WSJ, Sept 13, 2020).

Electronic copy available at: https://ssrn.com/abstract=3747238

To do so, we address two seemingly separate but economically related questions: What are the best predictors of option returns? What are the best predictors of stock returns, given that stock market signals fail to predict stock returns in post-2003 data ((Green, Hand, and Zhang (2017)), after wide adoption of algorithmic trading? If record high trading activity in options allows for better empirical identification of an informational advantage of option markets, then option-based predictors should dominate both stock and option return predictability. Alternatively, if conventional wisdom is true, then stock-based predictors should lead both options and stock returns.[6]

Answering these questions would not be possible without recent advancements of applications of machine learning (ML) in asset pricing pioneered by Gu, Kelly, and Xiu (2020). Leveraging big data, we construct a large set of options "Factor Zoo" predictors, jointly consider it with the stock "Factor Zoo" (Cochrane (2011)), and run a horse-race between options and stock based predictors to identify the primary, first order importance predictors of monthly option and stock returns. ML methods are perfectly suited for processing big data, condensing sparsity among large sets of predictors, reducing data dimensionality, and putting emphasis on non-redundant variable selection.

We follow Gu, Kelly, and Xiu (2020) and consider the following ML algorithms: LASSO of Tibshirani (1996), Elastic Net (EN) of Zou and Hastie (2005), Ridge of Hoerl and Kennard (1970), Random Forest (RF) of Breiman (2001), Feed-forward Neural Network (NN), see Goodfellow, Bengio, and Courville (2016).[7] To the range of these algorithms we also add Sparse Group LASSO (SGL) introduced by Simon, Friedman, Hastie, and Tibshirani (2013). Unlike other methods, SGL can recognize the data structure by combining various covariates into a single group based on their similarity. This is especially suitable for our analysis as it allows to explicitly control for various stock and option characteristic related groups, as well as for the lag structure. Notably, as our objective is not to compare different ML approaches but rather analyze the general results produced by them, we consider and focus our discussions on two ensemble methods: ensemble of penalized linear models, which are easier to train and relate their results to those in the previous literature based on linear techniques, and ensemble of all models, which allows incorporating non-linear predictive relations.

To assess the accuracy of return predictability of various machine learning methods and predictors, we use several criteria: (i) for option returns, we assess consistency of predictability across all optionable stocks and for robustness for the most liquid S&P500 stock options; (ii) we use out-of-sample (OOS) R^2 (Gu, Kelly, and Xiu (2020)) to evaluate the accuracy of the

[6]Consistent with the rest of the literature (Green, Hand, and Zhang (2017), Gu, Kelly, and Xiu (2020)), the informational advantage here includes a combination of information about future cash flows, the extent to which markets efficiently incorporate this information, and exposures to priced risk factors.

[7]We omit Gradient Boosting Regression Tree (GBRT) of Friedman (2001) as we find that its performance is the most sensitive to reductions in sample size, as it often occurs when we analyze the cross-section of S&P500 constituent firms.

Electronic copy available at: https://ssrn.com/abstract=3747238

forecasts; (iii) we use various individual and group variable importance ranks to identify the most important individual features or feature-groups; (iv) we evaluate performance of machine learning portfolios formed on their return forecasts, and the performance of long-short zero investment strategy portfolio and its Sharpe ratio to evaluate economic gains of forecasts.

ML algorithms are known to over-fit in-sample. All our performance statistics are therefore for out-of-sample, OOS. ML methods require relatively long sample periods for training. As our option data are from OptionMetrics, we start the sample with the earliest available observation in 01/1996, with the overall sample from 01/1996 to 12/2019. We use the first 10 years of monthly data as the initial training sample, and the subsequent 2 years as the first validation sample. Our out-of-sample test data are therefore from 03/2008 to 12/2019, which is far past the year 2003 (Green, Hand, and Zhang (2017)).

For options, we focus on predicting the returns of delta-neutral straddles on individual equities, as well as on equities of S&P500 firms as the most actively traded asset classes. At-the-money straddles, which combine similar positions in a put and a call with the same maturity and strike price, are constructed to have their deltas equal to zero. Therefore, these returns are completely invariant to the performance of the underlying stock. The latter is imperative, as if option returns still retain the effect of the underlying stock, the spurious correlation between stock characteristics and expected option returns would be detected via this imperfectly hedged mechanism.[8] Our cross-section of stocks is based on the large cross-section of all optionable stocks, which excludes micro-cap or the smallest stocks that do not have listed options and thus eliminates possibility of capturing anomalies' predictability (Hou, Xue, and Zhang (2020)).

Our main results can be summarized as follows. First, in contrast to conventional views, we find that option predictors dominate stock predictors in forecasting returns of all optionable stocks as well as returns of the most liquid options of S&P500 components. This dominance is consistent across all performance metrics. For example, an ensemble method which incorporates OOS forecasts of all ML algorithms, and uses both options and stock predictors (features) achieves OOS R^2 of 5.51% (3.26%) for all (S&P500) stock options. Very similar OOS R^2s are achieved with only options predictors, 5.48% (3.20%). When we use only stock predictors, OOS R^2s are *statistically* significantly lower, 5.06% (3.03%) . Therefore, options predictors alone provide better accuracy about future option returns.

While this evidence is purely statistical, Sharpe ratios of long-short zero investment strategy formed on these returns forecasts can provide evaluations of economic gains. Each month, t, we form decile portfolios based on the one-month-ahead, $t+1$, out-of-sample forecasts of option returns, as well as high-minus low strategy, with long position in options in the highest and

[8] It is well-known that the delta-hedge is not perfect, and does not eliminate the affect of the underlying stocks completely(See among others Buraschi and Jackwerth (2001), Cetin, Jarrow, Protter, and Warachka (2006), Garleanu, Pedersen, and Poteshman (2008), Christoffersen, Goyenko, Jacobs, and Karoui (2018)). However, our results with delta-hedged returns are qualitatively similar and are available upon request.

Electronic copy available at: https://ssrn.com/abstract=3747238

short position in options in the lowest return forecast decile. The realized gains of this trading strategy in month $t+1$ are aimed to estimate potential profits to an investor who uses machine learning forecasts and monthly holding periods. When we use the same ensemble of all method with both options and stock predictors, and all optionable stocks, the OOS annualized Sharpe ratio of this trading strategy is 3.58. A very similar Sharpe ratio is achieved when we use only options predictors, 3.55, and substantially lower Sharpe ratio is obtained when we use only stock predictors, 2.89.

Moreover, individual features' importance weight analysis ranks the predictive importance of option characteristics substantially higher then stock features. Among option characteristics, option illiquidity (Christoffersen, Goyenko, Jacobs, and Karoui (2018)) is always ranked at the top. Among other option characteristics, implied volatility, and the difference between historical and implied at-the-money volatility (Goyal and Saretto (2009)) are other important top predictors of option returns.

Among important stock characteristics for predicting option returns are momentum and reversals (An, Ang, Bali, and Cakici (2014)). They are followed, in order of relative importance, by stock illiquidity variables: Amihud illiquidity, size, bid-ask spreads, dollar trading volume. Interestingly, stock characteristics related to cash flow variance, the cash-to-asset ratio, profit margins, or overall firms' financial health (Cao, Han, Tong, and Zhan (2021), or lottery like trends (Byun and Kim (2016)) have relatively low importance. Chordia, Huh, and Subrahmanyam (2007) argue that trading activity is driven by the information these variables capture. This explains higher relative importance of illiquidity and trading volumes as proxies for trading activity compared to firms' fundamental signals for predicting options returns.

Overall, our results provide strong evidence that options rather than stock predictors dominate option returns predictability. This raises the question of whether these options predictors also dominate predictability of their underlying stock returns. Given that option prices are forward looking values of the underlying stocks, and option predictors forecast them better than stock predictors, then we expect option predictors to dominate in predicting stock returns as well.

To test this hypothesis, we next run a similar horse-race between option and stock characteristics in predicting one-month ahead excess stock returns. We continue relying on OOS R^2s statistics and the performance of machine learning portfolios formed, as before, based on models forecasts of excess stock returns.

First, we find that OOS R^2s obtained with only options predictors are significantly higher compared to those obtained with only stock predictors, or both options and stock predictors. For example, an ensemble of penalized linear models with only options predictors achieves OOS R^2 of 1.90%. This is statistically significantly higher then OOS R^2s of 1.09% or 1.17% obtained using only stock or both options and stock predictors, respectively.

Second, to evaluate economic gains to investors who use only stock predictors versus only

Electronic copy available at: https://ssrn.com/abstract=3747238

options predictors to forecast stock returns, we evaluate the performance of machine learning portfolios formed on their return forecasts. As an example, the annualized OOS Sharpe ration of long-short zero investment strategy portfolio based on the forecasts using only options variables and ensemble of all models is 1.49. The corresponding Sharpe ratio based on forecasts using only stock predictors is substantially lower, 1.15. Interestingly, Sharpe ratio of the forecasts based on both options and stock predictors is identical to the one which uses only options predictors, 1.49. This clearly indicates informational advantage of options predictors compared to stock predictors, and substantially larger economic gains to investors who rely on information from the option market. This is also consistent with results of Green, Hand, and Zhang (2017) who find that stock characteristic predictability in monthly stock returns deteriorates after 2003.

In terms of feature importance, almost all ML methods unanimously identify option illiquidity as the main predictor of stock returns. Theoretically, illiquidity and trading volume are mutually related (Benston and Hagerman (1974), Stoll (1978a)). It suggests that a constantly increasing popularity of option contracts reflected by consistently raising trading activity and volumes, which are normally captured by market illiquidity, does reflect more information and decreases uncertainty about the underlying (Cao (1999)).

The result that options traders have informational advantage over stock investors goes back to Pan and Poteshman (2006), who show that proprietary option trading volumes significantly predict stock returns after controlling for stock predictors, and this effect persists for several weeks suggesting informed trading on both positive and negative information in the option market first, before the stock market reacts. The authors had perfect identification, as their data on signed opening or closing, long or short option positions were not publicly displayed at the time and hence informed traders did not need to "hide" their trades. Since then, not only similar identification opportunity never appeared, but also their results have been disputed in the literature and the latest testimony is that options market is driven by an overall demand for leverage (Ge, Lin, and Pearson (2016)). With much bigger data, and modern machine learning techniques, we are able to provide a novel supportive evidence for Pan and Poteshman (2006) results.

To understand the economic channel behind the leading role of options markets, we investigate economic determinants of options illiquidity, the leading predictor of both option and stock returns. Traditionally, illiquidity or higher trading costs arise to compensate liquidity providers for adverse selection risk (Kyle (1985), Glosten and Milgrom (1985)), inventory risk (Stoll (1978b), Ho and Stoll (1983)), and the risk of holding inventory while waiting for the offsetting order flows (Grossman and Miller (1988)). Trading activity, or volume plays a critical role as intuitively higher trading volume is associated with lower asymmetric information, lower inventory risk and lower inventory holding time and costs. This results in higher liquidity. In classical literature, liquidity and volume are strongly related (Benston and Hagerman (1974)).

Using similar methodologies as above, we identify the main OOS predictors of options

Electronic copy available at: https://ssrn.com/abstract=3747238

illiquidity for the cross-section of all optionable stocks. Among more than 100 options and stock characteristics, most predictability appears to be driven by at most 10 predictors, as the weights of all other features are close to zero. Besides its own lag, the second most important variable predicting option illiquidity is stock illiquidity, measured by Amihud's illiquidity ratio, and the third is OS (option-to-stock volume ratio). Given the recent dramatic increase in the option volume compared to the stock volume, this result might as well represent investors, and especially retail investors, switching trading activity from stocks to options.[9] The next three important predictors are again related to stock illiquidity and are in order of importance: size, dollar volume, and turnover. Thus option illiquidity seem to capture almost all dimensions of stock illiquidity: price impact captured by illiquidity ratio, trading intensity captured by dollar volume and turnover, and overall visibility captured by size. Among other important predictors are stock *beta* and *beta squared* which can be viewed as proxies for uncertainty (Chordia, Huh, and Subrahmanyam (2007)). The latter is directly related to the theoretical argument of Cao (1999), that an increase in derivatives trading decreases information asymmetry and reduces stocks' price uncertainty. Consistent with this argument, we uncover positive options illiquidity premium in the stock returns: stocks with more illiquid options, which also are the stock-options with lower trading activity and higher price uncertainty, have economically and statistically significantly higher expected returns.

The rest of the paper is organized as follows. In Section 2 we describe our contribution to the related literature. Section 3 reveiws the data and methodologies. Section 4 reports results for predicting option returns. Section 5 reports results for predicting stock returns. In Section 6 we discuss economic channel behind leading role of the option market. Section 7 concludes the paper.

2 Related Literature

The informational role of options markets has been debated since their inception. Given high embedded leverage in options contracts, informed investors should always be attracted to use them (Black (1975)), and hence options markets should lead the stocks. The empirical evidence to support this argument is mixed and rather inconclusive.

For example, to support the argument, using intra-day options trading data, Chakravarty, Gulen, and Mayhew (2004) find that there is a significant fraction of price discovery which takes place in options compared to stocks. In contrast, Muravyev, Pearson, and Broussard (2013) find that at high frequency the options market follows rather than leads the stock market, which rejects the argument.

At the daily frequencies, Hu (2014) shows that the predictive effect of options order flows for

[9]The recent GameStop saga provides perhaps the most extreme example of retail investors in particular preferring options over stocks (source, WSJ, January 27, 2021, "How GameStop's Reddit-and Options-Fueled Stock Rally Happened")

Electronic copy available at: https://ssrn.com/abstract=3747238

stock returns is mostly attributed to options market makers' delta hedging trades, and it disappears after several days, which points to the market structure rather than informational role of the options.

At lower frequencies, weekly or monthly, allowing for longer time for the information to incorporate into asset prices, the evidence is mixed as well. Consistent with the leading informational role of options, Pan and Poteshman (2006) find that signed options volume initiated by end-users predict stock returns several weeks out. Johnson and So (2012) compute ratio of options to stock volume, O/S, and argue that it negatively predicts stock returns due to prevailing trading on the negative information about the underlying stock in the options market. In contrast, Ge, Lin, and Pearson (2016) find that the latter explanation is not entirely accurate as options-to-stock volume ratio similarly predicts both positive and negative stock returns. The authors conclude that the demand for leverage, rather than information, dominates options trading.

More recently, Muravyev, Pearson, and Pollet (2021) further analyze three leading options-based stock returns predictors: the volatility spread (Cremers and Weinbaum (2010)), the volatility skew (Xing, Zhang, and Zhao (2010)), and O/S. The authors find that after removing hard to borrow for short-selling stocks, the arbitrage profits from trading on these signals completely disappear. This suggests that options markets capture market frictions associated with short-selling constraints in the stock market rather than information.

Cremers, Goyenko, Schultz, and Szaura (2019) provide very similar analysis while considering larger set of options based predictors for the stock returns. Using IHS Markit stock borrowing indicative fees to identify hard-to-borrow stocks, the authors too find that all leading option-based predictors of stock returns capture the costs of short-selling, and after accounting for hard-to-borrow stocks, the potential arbitrage profits disappear. In contrast to the previous literature however, they introduce a new option-based measure of informed trading, which captures both, trading on positive and negative information, predicts stock returns several months out, and survives after exclusion of hard to borrow stocks. This supports the argument that informed traders prefer to trade in the options first (Black (1975)). Similar conclusions are also reported by Kacperczyk and Pagnotta (2019) who analyze a hand-collected sample of illegal insiders' trades and demonstrate that substantial fraction of informed trading does take place in options. Here, the identification problem of trading on private information is transparently resolved by using prosecuted insider trades sample.

Overall, to the date, there is no clear, convincing conclusion yet of whether the options market has indeed the leading advantage over the stock market. The results depend a lot on the variables used, and the empirical design which has been relying on the linear relations between predictors and returns in the current options literature. Arguably, the best approach is to use all available information, all possible predictors available in the literature from option and stock markets, and more importantly, to allow for possible non-linearity between predictors and expected returns. This approach should be able to provide a fuller, more comprehensive answer to this long lasting

Electronic copy available at: https://ssrn.com/abstract=3747238

debate, and to gain a better understanding between competing hypotheses. Finally, the dramatic increase in option trading in recent years exceeding in nominal values the stock volume points to more evolved economic story rather than reflection of market frictions associated with overcoming short-selling constraints in the stock market.

How do our findings contribute to the literature?

First, for the first time, to the best of our knowledge, we are able to provide compelling evidence that options markets lead the stock market. We achieve this by running parsimonious horse-race between options' "Factor Zoo" and stocks' "factor Zoo" to identify the most relevant, first order importance predictors for options and stock returns. ML methodologies allow us to identify variables which have the strongest statistical correlation with future returns, and the most influential ones, after controlling for redundancy, are options characteristics.

Second, the literature on stock and options returns predictability largely treats stock characteristics as the main predictors. For example "Factor Zoo" (Cochrane (2011)) refers to stock-based predictors only, even though not all of these predictors survive out-of-sample tests (Harvey, Liu, and Zhu (2016), Green, Hand, and Zhang (2017), Hou, Xue, and Zhang (2020)), and none of them survives in more recent data, post 2003-sample (Green, Hand, and Zhang (2017)). Further, the majority of options returns predictors are also based on stock characteristics.[10] To this literature we add an important insight of much richer informational environment of option-based characteristics, or options "Factor Zoo".

Third, as option literature starts adopting larger sets of stock characteristics to predict option returns (Cao, Han, Tong, and Zhan (2021)), it has never applied "Factor Zoo" (Green, Hand, and Zhang (2017)) approach which has been done in the stock market. Instead of adding one or a few predictors at the time, we run a parsimonious analysis across all possible options and stock predictors to identify which characteristics provide *independent* information. To the best of our knowledge, we are the first to run a cross-market horse race to be able to call the winner. In all our testing scenarios, options characteristics alone provide more accurate predictions than either options and stock characteristics together, or stock characteristics alone.

Forth, there is also a handful of option characteristics which predicts option returns.[11] To the date, there is no comprehensive analysis about whether any of these characteristics survive out-of-sample tests to predict *options* returns. We find that many of them do, with options illiquidity

[10] See for example the pricing of the following stock variables in the cross-section of option returns: idiosyncratic volatility (Cao and Han (2013)), lottery like stock preferences (Byun and Kim (2016)), stock short-sale constraints (Ramachandran and Tayal (2020)), stock illiquidity (Kanne, Korn, and Uhrig-Homburg (2018)), 10 stock characteristics related to cash flow variance, the cash-to-asset ratio, analyst earning forecast dispersion, 1 - and 5-year changes in shares outstanding, profit margins, profitability, stock price, external financing and Z-score (Cao, Han, Tong, and Zhan (2021)).

[11] See for example: Option illiquidity (Christoffersen, Goyenko, Jacobs, and Karoui (2018)), or difference between historical and implied volatilities (Goyal and Saretto (2009)), or level, slope and value factors constructed with the options data (Karakaya (2014)).

Electronic copy available at: https://ssrn.com/abstract=3747238

leading the race.

Fifth, there is also a voluminous literature identifying variables from the option markets which can successfully predict stock returns.[12] These option-based characteristics have never faced an out-of-sample horse-race with plethora of stock characteristics (Green, Hand, and Zhang (2017)) in their relative importance for predicting *stock* returns.

Sixth, we relate to trading and trading costs, market frictions, literature. The common perception is that delta-hedging can elevate all risk concerns of options market makers, and as a result, option prices can change only due to changes in the valuations of underlying stocks. In practice, however, net order flows from end-users, inability to hedge continuously, model and price jump risks of market makers inventories contribute to substantial deviations of option prices from fundamental values (Bollen and Whaley (2004), Garleanu, Pedersen, and Poteshman (2008), Christoffersen, Goyenko, Jacobs, and Karoui (2018)). Therefore, trading, which sparked dramatically in recent years, and associated with it trading costs are on its own, independently from the stock market, can be more important drives of option prices.

Relatedly, stock predictability literature with stock characteristics alone is indecisive in the recent data. Chordia, Roll, and Subrahmanyam (2011) while exploring the sharp uptrend in trading activity in recent years of their sample find that an increase in turnover is strongly associated with an increase in informed trading by institutions. This pattern is also accompanied by decreased cross-sectional return predictability.

Green, Hand, and Zhang (2017) further show that in the sample after-2003, stock-characteristic based predictability disappears. The authors suggest that this shift in the monthly stock return predictability "... presents a meaningful challenge to both past and future research". We take on this challenge by adding options markets characteristics to already large set of 94 stock characteristics of Green, Hand, and Zhang (2017),[13] and our out-of-sample period covers the most recent post-2003 data.

Finally, we relate to illiquidity premium literature. Option illiquidity has been shown to predict options returns (Christoffersen, Goyenko, Jacobs, and Karoui (2018)), while stock illiquidity premium in stock returns greatly diminishes in recent data (Ben-Rephael, Kadan, and Wohl (2015)). Our results on options illiquidity premium in stock returns point to the identification issue. For example, as discussed by Christoffersen, Goyenko, Jacobs, and Karoui (2018), option illiquidity contemporaneously increases with stock illiquidity, probability of informed trading in the underlying stocks (PIN), and exposure to higher volatility shocks or price jumps. Therefore, a more parsimonious measure of illiquidity, e.g. options illiquidity, capturing more than purely transaction costs can have a different result for the significance of illiquidity premium in the stock market.

[12] Options risk neutral measures (Conrad, Dittmar, and Ghysels (2012)), or options-based measures of informed trading, see Cremers, Goyenko, Schultz, and Szaura (2019) for overview and Section 3 below.

[13] This set of characteristics has also been used by Gu, Kelly, and Xiu (2020)

Electronic copy available at: https://ssrn.com/abstract=3747238

3 Data and Methodologies

The machine learning methods are a collection of commonly used regularized linear regressions, non-parametric non-linear models and parametric linear models. More specifically, we implement the following machine learning algorithms: LASSO of Tibshirani (1996), Elastic Net (EN) of Zou and Hastie (2005), Ridge of Hoerl and Kennard (1970), Random Forest (RF) of Breiman (2001), Feed-forward Neural Network (NN), see Goodfellow, Bengio, and Courville (2016) and finally Sparse Group LASSO (SGL) introduced by Simon, Friedman, Hastie, and Tibshirani (2013).

The most common approach in machine learning literature is to "tune" hyperparameters adaptively using the data from the validation sample. Hyperparameters include the penalization parameters in lasso and elastic net, the number of random trees in a forest, and the depth of the trees. Tuning parameters are estimated from the validation sample taking into account estimated model coefficients, where the coefficients are estimated from the training data alone. The third, the testing sub-sample, is used for neither estimation nor tuning, and is truly out of sample evaluation of model's predictive performances.

We analyze the predictive power of machine learning algorithms for both option returns and returns on the underlying stocks. We therefore define a return on an asset in the most general form as:

$$r_{i,t+1} = \mathrm{E}_t\left(r_{i,t+1}\right) + \epsilon_{i,t+1} \tag{3.1}$$

where

$$\mathrm{E}_t\left(r_{i,t+1}\right) = g^{\star}\left(z_{i,t}\right) \tag{3.2}$$

Stocks are indexed as $i = 1, \ldots, N_t$, and months as $t = 1, \ldots, T$. $z_{i,t}$ is P-dimensional vector of predictors which we also discuss in the next section. We follow the approach of Gu, Kelly, and Xiu (2020), where the function $g^{\star}(\cdot)$ maintains the same form over time and across different assets, and leverages information from the entire panel. Appendix B provides a specific description of each of the methods and as well as their implementation details.

Forecast Combination It is common in computer science literature to use ensemble estimations, i.e. re-estimating the same model several times, and then averaging the results. The motivation for using ensemble is that every optimization is initiated with the random seed. As such, there can be small differences between every round of optimization. To diminish the variance and converge to more stable parameters, each model can be re-estimated from 5 to 10 times. Alternatively, as it is done in Gu, Kelly, and Xiu (2020), one can use ensemble of different models to achieve the most precise estimates.

Let $\hat{r}^{(k)}_{i,t+1}$ be asset's i expected return estimated with method k $(k = 1, \ldots, K)$ and K be the number of methods. We consider two types of K combination. Combination 1 $(K = 4)$

Electronic copy available at: https://ssrn.com/abstract=3747238

includes only penalized linear models, LASSO, Ridge, EN, and SGL. Combination 2 includes all 12 methods: PCR, PLS, LASSO, Ridge, EN, SGL, RF, NN1, NN2, NN3, NN4, and NN5. Having two ensemble methods allows comparing forecasting strength of relatively simple linear to more complex non-linear models. The estimates of $\hat{r}_{i,t+1}^{(k)}$ ae equally combined to obtain a new prediction of expected returns:

$$\hat{r}_{i,t+1} = \frac{1}{K}\sum_{k=1}^{K}\hat{r}_{i,t+1}^{(k)} \tag{3.3}$$

The idea behind Equation (3.3) is that combining forecasts of expected returns from different methods can reduce the variance of individual forecasts. The previous literature suggests that the forecast combination method works well for return predictability (Rapach, Strauss, and Zhou (2010)). Combination 2 is of a particular interest since option returns predictability has not been evaluated via non-linear dependencies between predictors and expected returns.

Performance Evaluation Following Gu, Kelly, and Xiu (2020), to assess predictive performance for individual option returns or excess stock returns forecasts, we calculate the out-of-sample R^2 as

$$R^2_{\text{oos}} = 1 - \frac{\sum_{(i,t)\in\mathcal{T}_3}\left(r_{i,t+1} - \widehat{r}_{i,t+1}\right)^2}{\sum_{(i,t)\in\mathcal{T}_3} r^2_{i,t+1}} \tag{3.4}$$

where $\mathcal{T}_3$ refers to the "test" periods. More specifically, we first split our sample (03/1996 to 12/2019) into first 10 years of training sample, 03/1996 to 02/2005, $\mathcal{T}_1$, and two years of validation sample, 03/2006 to 02/2008, $\mathcal{T}_2$, with our first out of sample prediction for 03/2008 to 02/2009, $\mathcal{T}_3$. We than expand the training sample by one year, roll the validation sample by one year, and produce the forecast for the next out of sample year, and so on.

We use Diebold and Mariano (1995) test to make pairwise comparison between the models in terms of out-of-sample predictive accuracy. More specifically, to compare the forecast performance of method (1) versus (2), the modified Diebold-Mariano test statistic is $DM_{12} = \bar{d}_{12}/\widehat{\sigma}_{\bar{d}_{12}}$, where

$$d_{12,t+1} = \frac{1}{n_{3,t+1}}\sum_{i=1}^{n_{3,t+1}}\left(\left(\widehat{e}_{i,t+1}^{(1)}\right)^2 - \left(\widehat{e}_{i,t+1}^{(2)}\right)^2\right) \tag{3.5}$$

$\hat{e}_{i,t+1}^{(1)}$ and $\hat{e}_{i,t+1}^{(2)}$ are the return forecast errors for an individual asset i at time $t+1$ generated by two methods, and $n_{3,t+1}$ is the number of assets in the testing sample (month $t+1$). $\bar{d}_{12}$ and $\widehat{\sigma}_{\bar{d}_{12}}$ are, respectively, the time-series mean and Newey-West standard error of $d_{12,t+1}$ over the testing sample.

Electronic copy available at: https://ssrn.com/abstract=3747238

Data The options data are from OptionMetrics for the period 01/1996 to 12/2019. They include the best closing bid and ask prices, volume, open interest, deltas, and implied volatilities. We use closing bid-ask midpoints as a proxy for option prices. OptionMetrics uses the Cox, Ross, and Rubinstein (1979) binomial tree model to estimate deltas and implied volatilities, therefore allowing for early exercise, and further assuming a constant dividend yield. The monthly stock returns data are from CRSP for the similar period. Options are known for infrequent trading and high illiquidity. It is conventional in options literature to test the robustness of empirical results on more liquid and frequently traded cross-section of options of S&P500 constituents. Therefore, for option returns predictability, we analyze two samples: first, all stocks with tradable options, and second, the most liquid option classes of S&P500 firms. CRSP/Compustat provide identifiers for S&P500 index constituents which we apply to our data on the monthly bases.

Option Returns Similar to Coval and Shumway (2001), we compute monthly returns on delta-neutral straddles, $R^{Straddle}_{t+1,i,j}$, assuming the pay off to a writer of the straddle over one month holding period. Straddle returns are computed for each pair of at-the-money call, i, and put, j, options on the same underlying, with the same strike price and time to expiration. At-the-moneyness is defined as in Bollen and Whaley (2004) for absolute values of deltas between 0.375 and 0.625. The returns on straddles,$R^{Straddle}_{t+1,i,j}$, are computed for positive volume days and non-zero open interest options in the end of month t, i.e. the beginning of interval for which we measure returns. To control for the data quality, we also impose the following filters in the end of month t. Equity option quotes with dollar quoted spreads greater than \$3 are deleted. We also delete illiquid options contracts with the quoted relative bid-ask spreads greater than 70%, and options with mid-point quoted prices below 50 cents in the beginning of intervals we measure returns. We exclude options contracts which violate obvious no-arbitrage bounds: for calls, the price must be less than the current stock price, for puts it must be less than the strike. Finally, we only use contracts with 31 to 180 days to maturity. Straddle returns are computed on a contract level first. We then compute weighted average returns on a class (stock) level across contracts, using open interest in the end of month t as a weight.

Option Characteristics In our analysis we use all option based variables which are identified in the literature to predict the cross-section of stock returns. Cremers, Goyenko, Schultz, and Szaura (2019) argue that most of these variables capture market frictions, like hard to borrow stocks, and do not represent trading opportunities. If investors however overcome short-sale constraints in the stock market by increasing their demand for options, this can affect option prices and expected returns (Ramachandran and Tayal (2020)). Overall we consider nine options characteristics, and fifteen option-implied risk neutral moments, totaling to overall 24 option based variables.

We first describe 9 characteristic variables. Following An, Ang, Bali, and Cakici (2014) we

Electronic copy available at: https://ssrn.com/abstract=3747238

include changes in call and put implied volatilities. Implied volatilities are obtained from the daily implied volatility surface calculated by OptionMetrics. Their empirical analysis uses end-of-month call and put implied volatilities for options with a delta of 0.5 and 30 days to maturity. We remove observations with missing implied volatilities and deltas. Only at-the-money series (abs(delta)=0.5) with 30 days to maturity are retained. We use the last available daily observation of the month for each call/put for each stock to compute the monthly changes. We also estimate the average implied volatility on a stock level by averaging implied volatilities of at-the-money calls and puts of the same underlying.

As in Cremers and Weinbaum (2010) we construct the differences between implied volatilities of calls and implied volatilities of puts. Using all pairs of calls and puts on a stock with the same strike price and expiration date, a daily weighted average difference is calculated across option pairs using the open interest as weights. We than average daily observation for a month for each stock.

Similar to Xing, Zhang, and Zhao (2010) we also construct risk neutral skewness, defined as the difference between the implied volatility of out-of-the-money puts and the implied volatility of at-the-money calls, to predict future stock returns. We estimate skewness as the difference between implied volatilities of puts with a delta of -0.2, and the average implied volatility from put and call contracts with an absolute value of delta of 0.5. Implied volatilities are obtained from the OptionMetrics volatility surface for options with 30 days to maturity. We calculate skewness daily and average daily values to compute a skewness measure for each month.

Muravyev, Pearson, and Pollet (2018) use a non-linear transformation of put-call parity violations to estimate implied stock borrowing fees that short-sellers would expect to pay. Following the procedure outlined in their paper, we use pairs of puts and calls with the same strike price and time to expiration and attribute violations of put-call parity to borrowing fees. Similar to Ofek, Richardson, and Whitelaw (2004), the authors argue that borrowing fees, which capture stock short sale constraint, are a strong predictor of stock return. Relatedly, Ramachandran and Tayal (2020) find that stock short-sale constraints are also predicting option returns.

As in Johnson and So (2012) we also construct O/S ratio, which is the natural logarithm of the ratio of options volume to stock volume. Unlike Johnson and So, who use just short-term options, we use the total option volume across all strikes and maturities. We calculate O/S monthly. We measure stock volume in round lots of 100 to make it comparable to option contracts on 100 shares.

While the above measures have been shown to predict stock returns, the next two measures are being argued to predict option returns. Goyal and Saretto (2009) show that the difference between historical realized volatility and at-the-money implied volatility significantly predicts option returns. Similar to the authors, we compute historical volatility as the standard deviation of daily realized stock returns over the most recent 12 months, and implied volatility as the average of the implied volatilities of calls and puts which are closest to at-the-money, and are one month

Electronic copy available at: https://ssrn.com/abstract=3747238

to maturity. Christoffersen, Goyenko, Jacobs, and Karoui (2018) show that option illiquidity is priced in the cross-section of option returns. The authors use CBOE/LiveVol intra-day trading data to estimate option effective spreads as a measure of illiquidity. LiveVol data start in 2003, while our sample starts in 1996 where only end of day bid and ask quotes are available from OptionMetrics. To keep the length of the sample which is essential for machine learning training, we use percentage end of day bid-ask spreads as a proxy for illiquidity. We first compute it on the last day of each month for each contract with non-zero trading volume on that day, and then we average the estimates across contracts on the stock level.

The final set of measures are option implied risk-neutral moments from Conrad, Dittmar, and Ghysels (2012). The authors demonstrate that these measures robustly predict stock returns by capturing their various risk characteristics. The risk neutral moments are variance, skewness and kurtosis, and similar to the authors we also estimate them for various horizons: 1, 2, 4, 5 and 6 months. Therefore the risk neutral volatility, skewness or kurtosis, are computed for each stocks five times using options with 1, 2, 4, 5 or 6 months to expiration. We follow exactly the procedures outlined in Conrad, Dittmar, and Ghysels (2012).[14]

The summary of variable notations and definitions for option-based variables is provided in Appendix Table A.1.

Stock Characteristics, Macroeconomic Predictors and Data Structure Similar to Gu, Kelly, and Xiu (2020) we rely on a large set of 94 stock-level predictors used by Green, Hand, and Zhang (2017). The abbreviations of these characteristics are provided in the Appendix, Table A.2.[15]

We also use eight macroeconomic predictors following the variable definitions detailed in Welch and Goyal (2008), including dividend-price ratio (dp), earnings-price ratio (ep), book-to-market ratio (bm), net equity expansion (ntis), Treasury-bill rate (tbl), term spread (tms), default spread (dfy), and stock variance (svar).[16].

We also use 68 industry dummies defined by the first two digits of Standard Industrial Classification (SIC) codes. Overall, the total number of characteristics that we use for prediction is 118 (94 stock-based + 24 option based). As in Gu, Kelly, and Xiu (2020), we use these covariates alone as well as their interaction terms. The total number of covariates is thus $118(8+1)+68=1130$. As the machine learning algorithms are quite computationally intensive, we refit models every year, instead of every month. Each time we refit, we increase the training sample by one year. We keep the validation sample of 2 years at all times, and roll it forward by one year to include the most recent 12 months.

[14]We are grateful to Robert Dittmar for providing the code to estimate risk neutral moments.

[15]To construct the variables we use the SAS code available from Jeremiah Green's Web site. As it is a common practice with big data, similar to Green, Hand, and Zhang (2017) and Gu, Kelly, and Xiu (2020) we replace missing characteristics with the cross-sectional median at each month for each stock.

[16]The monthly data for these variables are from Amit Goyal's website

Electronic copy available at: https://ssrn.com/abstract=3747238

It is important to standardize all right hand side variables before inputting them into machine learning algorithms. We perform the analysis first using standardized variable values with mean zero and standard deviation of 1. However, in all tables below we report more conservative results where instead of using the standardized values of the predictors, we rank all characteristics each period and map the ranks into [-1,1] interval.[17]

We also consider forming different predictors in groups by their similarities, and thus imposing the structure on the covariates. Most of machine learning methods do not recognize the group structure, and are not able to accommodate it. The only exception is SGL, where we can explicitly define the groups. We consider seven groups. Group 1 includes recent price trends (stock characteristics), group 2 - liquidity (stock characteristics), group 3 and 4 are risk measures, and valuation ratios and fundamental signals respectively (stock characteristics) (see Table A.2), group 5 is option characteristics (variables 1 to 9, Table A.1) , group 6 is option risk neutral moments (variables 10 to 24, Table A.1), and group 7 is 68 industry dummies.

SGL also allows to control for the lag structure of features. Below we report the results without controlling for lags, as they are quite similar to those with the lag structure from 1 to 3. This suggests that different persistence between options or stock predictors does not affect our out-of-sample inferences.

Option and Stock Returns Summary Statistics. Table 1 provides summary statistics of delta-neutral straddle returns for all optionable stocks, Panel A, and for S&P500 firms, Panel B, as well as returns of all optionable stocks, Panel C. Our sample of all stocks with options listed consists of 4,867 unique firms, out of which for 3,805 firms we are able to estimate delta-neutral straddle returns, Panel A. Similarly, there are 836 unique S&P500 firms in our sample, for 814 of which we are able to compute straddle returns, Panel B. Using the most liquid S&P500 stock-options provides a viable robustness.

In general, options returns are negative on average as option buyers pay variance risk premiums to option sellers. Similar to the previous literature (Cao, Han, Tong, and Zhan (2021)), we thus consider the pay offs to option writers and hence all option returns are being multiplied by -1. The magnitudes of straddle returns are quite similar between All and S&P 500 firms. These statistics are also similar in magnitudes compared to those reported in Goyal and Saretto (2009).

It is worth noting a signal-to-noise ratio (mean-to-standard deviation ratio) of these returns as machine learning algorithms are normally adapted for the data with quite high signal-to-noise ratio. For example, for an image recognition, it is easier to train an algorithm to identify an image of a cat versus a truck, i.e. it is hard to confuse a cat with much bigger and of a different shape truck. Finance data have substantially lower signal to noise ratio, e.g. buy vs sell signal. Gu, Kelly, and Xiu (2020) show that the algorithms we use perform quite successfully in predicting stock returns.

[17]This procedures closely follows Gu, Kelly, and Xiu (2020) and the references therein.

Electronic copy available at: https://ssrn.com/abstract=3747238

The option straddle returns in Table 1 have slightly higher signal-to-noise ratio compared to their stocks. As such, it is expected to be able to predict option returns more accurately which we do in the next section.

4 Predicting Option Returns with Option and Stock Characteristics

Table 2 presents monthly $OOS\ R^2$ for the three sets of predictors of option returns. First we use all option and stock variables, Panel A, second we use only option based variables, Panel B, and finally we use only stock based characteristics, Panel C. Each set of characteristics always includes interaction terms with macroeconomic variables and 68 industry dummies. For example, the number of features in Panel B is smaller, $24(8+1)+68=284$, compared to 1130 in Panel A, or $94(8+1)+68=914$ in Panel C. Each panel presents two sets of results: returns on delta-neutral straddles of all optionable stocks, and, separately, of S&P500 components.

The standard OLS regressions fail to account for the sparsity of big data, tend to overfit in-sample, and produce highly negative $OOS\ R^2$s across all panels. In Panel A, in contrast to OLS, the penalized linear models show huge improvements. For the cross-section of all firms, EN and Lasso provide the highest $OOS\ R^2$ of 5.54% for all, or 3.28% for the cross-section of S&P500 stock-options. The generally lower $OOS\ R^2$s for S&P500 firms suggest that it is more difficult to predict more liquid and more frequently traded option returns.

Consistent with Gu, Kelly, and Xiu (2020), relatively shallow neural nets, NN2, have the next best $OOS\ R^2$ of 5.36%. Neural nets normally preform the best on large data sets. Hence it is not surprising that their performance becomes less competitive compared to penalized linear models on the cross-section of S&P500 stock-options, with NN3 achieving $OOS\ R^2$ of 3.11%

The ensemble methods which are intended to decrease the variances of individual forecasts provide the highest $OOS\ R^2$s. Here, Combinations 1 and 2 provide quite similar estimates of 5.55% (3.32%)and 5.51% (3.26%) respectively for all (S&P500 constituent) stocks.

Panel A, Table 2, presents results obtained using both options and stock predictors. Our main goal here is to determine which set of predictors has the leading role, i.e. independently contributes to higher $OOS\ R^2$. Panels B and C present similar statistics obtained using only options or only stock predictors respectively. It is apparent from comparing all three panels, that the results in Panel B produced by option predictors alone are quite close in magnitudes to those reported in Panel A, while $OOS\ R^2$ statistics in Panel C, provided by stock predictors alone, are by an order of magnitude smaller. For example, for all stock-options, in Panel B, EN's $OOS\ R^2$ is 5.47% (vs. 5.54% in Panel A), while it is only 5.02% in Panel C. The combination methods in Panel B, and especially Combination 2, provide very similar statistics to Panel A as well. For example, for all

Electronic copy available at: https://ssrn.com/abstract=3747238

stock-options, $OOS\ R^2$ of Combination 2 is 5.48% (vs. 5.51% in Panel A), while it is only 5.06% in Panel C. The results are qualitatively similar for the cross-section of S&P500 stock-options.

Tables A.3, A.4, and A.5 in Appendix provide a statistical, Diebold-Mariano, pairwise comparison of $OOS\ R^2$s in each of Panels A, B and C of Table 2 respectively. Within panel pairwise Diebold-Mariano statistics generally reveal that $OOS\ R^2$s' of ensembles, Combination 1 or 2, methods either insignificantly different from individual methods, or statistically dominate them. In the rest of the analysis we therefore focus on these two ensemble methods' forecasts.

While within panel pairwise statistical comparison of $OOS\ R^2$s is important, comparing statistical differences across Panels is imperative to identify which group of variables is a better predictor in the context of $OOS\ R^2$s performance.

Table 3 reports Diebold-Mariano test statistics for pairwise comparison of Combination 1 and 2 models across Panels A, B, and C of Table 2. We adapt the convention where a positive statistic indicates the column model outperforms the row model, and bold numbers denote significance at the 5% level for each individual test. Panel A reports results for all optionable stocks. *All Variables* identifies Combination 1 and 2 from Panel A, Table 2, and *Option Variables* and *Stock Variables* identify Combination 1 and 2 from Panels B and C, Table 2 respectively. For both, Combination 1 and 2, $OOS\ R^2$s obtained with *All predictors* or only *Option predictors* significantly exceed those obtained with only *Stock predictors* in Panel A. The $OOS\ R^2$s differences between using *All Variables* and only *Option Variables* are statistically insignificant across both methods. Therefore, in terms of out of sample predictive accuracy measured by $OOS\ R^2$ statistics, options predictors alone dominate stock predictors alone. Moreover, using both options and stock predictors does not improve the accuracy of the forecasts compared to using option predictors alone. The results are similar for S&P500 options using Combination 2 ensemble of all ML methods. This provides the first statistical evidence of informational advantage of options predictors over stock predictors for the future options returns. Our results suggest that option characteristics predict option returns better than stock characteristics.

4.1 Which Characteristics Matter?

$OOS\ R^2$ statistics are informative about the precision of the forecasts by different sets of variables, but they are silent about economic sources of the forecasts. Here we aim to identify which option or stock characteristics are the most influential in predicting option returns. To understand relative importance of each covariate, we follow Gu, Kelly, and Xiu (2020) in assessing each variable importance to the predictability. Similar to the authors, to measure the importance of variable j, we compute the reduction in panel predictive R^2 from setting all values of predictor j to zero within each training sample, and then average these into a single importance measure for each predictor. It thus allows identifying covariates that have an important influence on the cross-section of returns while simultaneously controlling for all other predictors.

Electronic copy available at: https://ssrn.com/abstract=3747238

Figure 1 reports the importance of the top 20-characteristics for each method in predicting delta neutral straddle returns of all optionable stocks, where the variable importance within the model is normalized to sum to one across all available option and stock predictors. The characteristics are ordered so that the highest rank is in the top and the lowest is at the bottom. One can easily see that across all models, options illiquidity (*opt_baspread*) (Christoffersen, Goyenko, Jacobs, and Karoui (2018)) is unanimously identified as the top predictor across all ML methods, with highly economically meaningful weights. It follows, in terms of the variable importance, by the difference between historical and implied ATM volatility (hviv) of Goyal and Saretto (2009), and the implied volatility on its own (avg_impl_volatility).

Risk neutral skewness (skewness), (Xing, Zhang, and Zhao (2010)) too is always one of the top predictors of option returns. While options illiquidity and *hviv* have been shown to predict option returns, *skewness* is argued to predict stock returns as a measure of private information. This result shows that expected option returns too capture private information. Besides private information, expected option returns also incorporate risk premiums proxied by the implied volatility (*avg_impl_volatility*).

Among stock based variables which appear in the top 20, the majority represents the recent price trends: industry momentum (indmom), short-term reversal (mom1m), 6-month momentum (mom6m), and 12-month momentum (mom12m). These results are consistent with An, Ang, Bali, and Cakici (2014) who argue that stock price trends, momentum and reversal, predict option returns. While still being important, the individual weights of these predictors, or their relative importance for option return predictability, is substantially smaller compared to the options predictors.

Next in ranking are stock liquidity variables: standard deviation of dollar trading volume (std_dolvol), bid-ask spreads (baspread), dollar volume (dolvol) and Amihud's illiquidity (ill). The economic intuition of these results is that more illiquid underlying stocks contribute to overall options illiquidity and hence expected option returns (Christoffersen, Goyenko, Jacobs, and Karoui (2018)).

Figure A.1 in Appendix reports average importance ranking of all characteristics across all models. The importance of each characteristics is first ranked for each model, and we then sum their ranks. Characteristics are ordered such that the highest total ranks are at the top and the lowest are at the bottom. The color gradient within each each column shows the importance of ranked characteristics from the most important to the least (from the darkest to the lightest respectively). Consistent with Cao and Han (2013), stock idiosyncratic volatility (*idiovol*) is also one of the top 20 predictors on average. Interestingly, $beta$ and $betasq$ which are viewed as proxies for uncertainty in the underlying stocks (Chordia, Huh, and Subrahmanyam (2007)) , are as influential as *idiovol*, as these variables are ranked just below it.

The final set of the top 20 stock variables is from the group of valuation ratios and fundamental

Electronic copy available at: https://ssrn.com/abstract=3747238

signals, and include number of earnings increases (nincr), dividend to price ratio (dy), and industry sales concentration (herf). To this end, Chordia, Huh, and Subrahmanyam (2007) find that uncertainty about firm's fundamental values, proxied by earnings surprises, increases trading activity in the stock market. Our results thus show that this trading activity spills over to the options market as well, and similar to the stock market it also has implications for the cross-section of expected option returns.

The literature also identifies other stock variables aimed at predicting option returns: cash flow variance, cash-to-asset ratio, profit margin, profitability, or probability of bankruptcy measured by Z-score (Cao, Han, Tong, and Zhan (2021)). These variables are quite low in the ranking. For example cash flow variance (sdcf), variables related to profitability such as operating profitability (operprof) and percentage change in gross margin less percentage change in sale (pschgm_pchsale), are ranked at the bottom. The only exception is financial health of a firm, which we measure in our data by Piotroski financial statements score (*ps*). It is ranked just below top 20 predictors. Another example is *max_ret* (Byun and Kim (2016)) which is supposed to be a strong predictor of option returns, but it is ranked at the bottom as well.

To put these results into perspective, the procedure that we use allows identifying the best unique, "irreplicable" predictors, and ranks them higher. The more "replicable" predictors would "share" the weights and more importantly, stronger predictors would be ranked higher. For example, *max_ret* and *idiovol* are highly correlated, with *idiovol* being ranked in the top 20 predictors. This suggests that *idiovol* is a stronger predictor of option returns than *max_ret*. Similar intuition explains low ranks of various firm fundamental and valuation variables used to predict option returns by Cao, Han, Tong, and Zhan (2021). Options trading activity measured by options to stock volume ratio (OS) is shown to be positively associated with firm valuations (Roll, Schwartz, and Subrahmanyam (2009)). OS is highly ranked, 21st, and just falls outside for the top 20 (Figure A.1). Therefore, OS while capturing information about firms' fundamentals is just a stronger predictor then raw firm valuation ratios. Figure 2 and Figure A.2 confirm qualitatively similar results for the option returns of S&P500 constituents. This assures that our findings are nor driven by sub-sample of illiquid, infrequently traded options.

Overall the feature importance results undoubtedly highlight the first order importance of options characteristics in predicting option returns. It holds on the level of all and S&P500 underlying stocks. The importance of stock characteristics in predicting option returns is substantially lower, and depends on the feature group. Stock price trends and illiquidity groups on average substantially dominate the importance of valuation ratios & fundamental signals group.

The results so far are based on statistical evidence. We next explore economic gains which can be associated with these statistical results.

Electronic copy available at: https://ssrn.com/abstract=3747238

4.2 Machine Learning Portfolios

To directly exploit machine learning forecasts, and compare option versus stock characteristics predictive power, we design different sets of machine learning portfolios. At the end of each month we compute the one-month-ahead out of sample option straddle return forecast for each of 14 methods (excluding OLS). We then sort stocks into portfolio deciles based on these return forecasts and construct value-weighted portfolios using all stocks or equally weighted portfolios using S&P500 stocks.[18] We rebalance these portfolios every month and also compute zero net-investment portfolio that buys the highest expected option return (decile 10) and sells the lowest (decile 1). For this high-low strategy we also compute annualized Sharpe ratio.

Table 4 reports the results for all optionable stocks, Panel A, and for the options of S&P500 components, Panel B. Each panel presents three sets of results with : (i) using both options and stock features, (ii) only option features, and (iii) only stock features to predict option returns. As before, we focus on performances of ensemble methods: Combination 1 and 2.

Overall, the average portfolio returns increase monotonically from Low to High. Sharpe ratios of high-minus-low portfolios first show that using ensemble of penalized linear models, Combination 1, provides very close estimates to those obtained with ensemble of all ML methods. Combination 2 however continues to dominate which indicates importance of non-linear relations between predictors and option returns which is accommodated by other methods. For all options, the highest Sharpe ratio is obtained for Combination 2 using all stock and option predictors, 3.58, Panel A. Interestingly, almost the same Sharpe ratio is obtained for Combination 2 using only option predictors, 3.55. Using only stock predictors, the corresponding Sharpe ratio is substantially lower, 2.89. Thus, using option predictors alone, an investors can extract substantially larger economic gains compared to using only stock predictors. This further shows economic gains of informational advantage of options over stock predictors.

The results are qualitatively similar for S&P500 options, Panel B, suggesting against potential illiqidity biases. Overall, we conclude that options characteristics are more informative both statistically and economically for predicting option returns than stock characteristics. This result is new to the literature as it suggests that the options market reflects the information about the underlying first. Moreover, it also demonstrates the importance of non-linear relations between predictors and option returns, which is not entirely unexpected providing generally convex payoffs of option contracts. Finally, this also addresses the long lasting debate about which market leads. Option valuations are forward looking about the values of the underlying. Given that option based predictors reflect this information first and better than stock variables suggests that options can lead overall, both option and stock markets. In the next section we test this hypothesis directly on the cross-section of stock returns.

[18] We use equally-weighted portfolios for S&P 500 firms to avoid over weighting the largest five FAANG firms.

Electronic copy available at: https://ssrn.com/abstract=3747238

5 Predicting Stock Returns with Option and Stock Characteristics

In this section we run another horse race between option and stock predictors in terms of forecasting accuracy of the expected (excess) stock returns. Our prior is that once option predictors dominate option returns predictability as we establish above, it means that the option market reflects forward looking information about the underlying first, and hence these option based predictors should also dominate stock based predictors in predicting stock returns.

We use the same methodology and machine learning methods as in the previous section for option returns. Our results here can be directly compared to those of Gu, Kelly, and Xiu (2020) with a few exemptions as our sample is shorter since we start with the availability of OptionMetrics data, in 1996, and we only use stocks with the options listed.

Table 5 reports $OOS\ R^2$s for the three sets of predictors: (i) option and stock characteristics, *All Variables*; (ii) option characteristics alone, *Option Variables*; and (iii) stock characteristics alone, *Stock variables*.

Across all panels, OLS underperforms with highly negative $OOS\ R^2$s. Even though our cross-section of stocks is smaller, and time series sample is shorter than in Gu, Kelly, and Xiu (2020), the $OOS\ R^2$ results for the monthly stock returns predictability with *Stock variables* are qualitatively similar. The highest $OOS\ R^2$=1.28% is obtained with RF. It is followed by SGL with OOS R^2=1.13%, which suggests that grouping covariates by their similarities improves on simple Lasso or EN with $OOS\ R^2$s of 0.91% and 0.76% respectively. Among neural nets, NN2 and NN3 outperform shallow and deep learners with $OOS\ R^2$s of 0.74% and 0.72% respectively.

The results improve a lot when we use option variables alone, (ii). Here penalized linear, dimension reduction models and RF outperform with substantially higher $OOS\ R^2$s. Ridge regression gives the highest $OOS\ R^2$ of 2%, followed by SGL, 1.95%, and PCR, 1.89%. These magnitudes of R^2s substantially exceed those in scenario (iii). Interestingly, even neural nets which are more difficult to train and require more data, perform better with option predictors alone (OOS R^2 of NN3 is 1.09%) than with stock predictors alone ($OOS\ R^2$ of NN3 is 0.72%). This further advocates informational advantage of options predictors over stock predictors in predicting stock returns.

Finally, using all stock and option predictors, specification (i), deteriorates the performance compared to only using options predictors, specification (ii), but improves performance compared to only using stock predictors, (iii). The highest out of sample R^2 is obtained with SGL, 1.31%, followed by Ridge, 1.28%. It appears that option-based variables alone are better predictors of stock returns than stock characteristics, or both stock and options characteristics.

While these results might be seen at odds with "Factor Zoo" literature in the stock market, which advocates stock based predictors' superiority, they are consistent with the findings reported

Electronic copy available at: https://ssrn.com/abstract=3747238

by Green, Hand, and Zhang (2017). The authors find that after 2003 and onwards, stock return predictability with stock characteristics deteriorates. They suggest that the possible reason could be a decrease in costs of exploiting mispricing, i.e., an increasing automated trading in the stock market leads to fast evaporation of potentially true signals in this period. Our out-of-sample period starts after 2003, in 2008. It is hard to assume that algo-trading can incorporate all option-based variables we use in this paper, or at least not all market participants use information from the options market to identify mispricing in the underlying stocks. This can be a reason that signals from the option market still survive and predict stock returns better compared to stock characteristics. The exploded increase in options trading activity in recent years is another testament of investors preferring make informed trades in options, rather than stocks.

Another interesting observation is that the ensemble of linear penalized methods, Combination 1, outperforms ensemble of all methods, Combination 2, and especially more so while using options predictors only, (ii). Within each specification, Combination 1 and 2 generally either outperform or insignificantly different from other methods (see Diebold-Mariano test statistics in Table A.6 in Appendix), except Ridge. Overall, this suggests that while the effect of stock features on expected stock returns is most likely non-linear, the effect of options features on expected stock returns is most likely linear.

Table 6 reports Diebold-Mariano test statistics for pairwise comparison of Combination 1 and 2 models between three specifications of Table 5. As before, we adapt the convention where a positive statistic indicates the column model outperforms the row model, and bold numbers denote significance at the 5% level for each individual test. *All Variables* identifies Combination 1 and 2 for specifiation (i) which uses both options and stock predictors, and *Option Variables* and *Stock Variables* identify Combination 1 and 2 for specifications (ii) and (iii) respectively. For the best performing Combination 1, *Option Variables* significantly outperform either *All Variables* or *Stock Variables*. Therefore, in terms of forecast accuracy measured by $OOS\ R^2$ statistics, options features contain more information about the underlying stocks compared to stock features alone, or combined with options predictors. This provides further evidence of the leading role of options for predicting stock returns.

5.1 Which Covariates Matter?

When it comes to the most influential stock characteristics in predicting stock returns, the overview of these results is already presented in Gu, Kelly, and Xiu (2020). Here we contribute by showing relative importance of option based versus stock-based predictors.

Figure 3 reports the importance of the top 20-characteristics for each ML method in predicting excess returns of all optionable stocks in our sample using both options and stock predictors. As before, the variable importance within the model is normalized to sum to one across all available predictors. The characteristics are ordered so that the highest rank is in the top and the lowest

Electronic copy available at: https://ssrn.com/abstract=3747238

is at the bottom. The striking difference between results in Gu, Kelly, and Xiu (2020) and ours is that options illiquidity (opt_baspread) is almost unanimously, with the only exception SGL, the most important predictor across all models, and with the highest weight. In Gu, Kelly, and Xiu (2020), price reversals (mom1m), and price trends (mom12m) have the highest weights in predicting stock returns in their sample. Figure A.3 in Appendix present overall ranking of all characteristics across all models. As before, the importance of each characteristics is first ranked for each model, and we then sum their ranks. Characteristics are ordered such that the highest total ranks are at the top and the lowest are at the bottom. The color gradient within each column shows the importance of ranked characteristics from the most important to the least (from the darkest to the lightest respectively). Price trends variables like industry momentum (indmom), 12-(mom12m) or 6-month (mom6m) momentum, or reversals (mom1m) are also in top 20 variables across all methods in our sample. However, we also find that options implied volatility is the forth most important predictor after options illiquidity, industry momentum (indmom) and number of earnings increases (nincr). Moreover, changes in call implied volatility (cvol) and option to stock volume ratio (os) are in the top 20 predictors across all models. Other important stock predictors are related to stock liquidity: size (mve), Amihud illiquidity (ill), bid-ask spreads (baspread), and dollar trading volume (dolvol). Idiosyncratic volatility, which is one of top predictors in Gu, Kelly, and Xiu (2020), after controlling for implied volatility and options illiquidity, comes only in the 32nd place after risk neutral option based 6-month volatility (volq6) and 1-month kurtosis (kurtq1).

One caveat of this exercise is that the feature importance, i.e. to compare the relative importance between various options and stock predictors, can only be done for the specification using *All Variables*. Tables 5 and 6 clearly show that this specification is statistically significantly dominated by using only *Option variables* as predictors. Therefore, even though we confirm the leading roles of options characteristics for the expected stock returns, doing so for the sub-optimal specification does not allow yet establishing their economic dominance. We address this in the next sub-section.

5.2 Machine Learning Portfolios of Stock Returns

The results so far are based on various optimization routines and algorithm structures of machine learning methods we use which do not allow to assess and compare economic magnitudes of predictive power between option and stock covariates.

Here, similar to machine learning portfolios of option returns, we form machine learning portfolios of stock returns. At the end of each month we compute the one-month-ahead out of sample stock return forecast for each of 14 methods (excluding OLS). We then sort stocks into portfolio deciles based on these return forecasts and construct value-weighted portfolios. We rebalance these portfolios every month and also compute zero net-investment portfolio that buys the highest expected stock return (decile 10) and sells the lowest (decile 1). For this high-minus-

Electronic copy available at: https://ssrn.com/abstract=3747238

low strategy we also compute annualized Sharpe ratio. The returns of each portfolio are risk adjusted using Fama-French five factor model augmented with Carhart's momentum factor.

Table 7 present portfolio sorting results when using all stock and option predictors (Panel A), only option predictors (Panel B), and only stock predictors (Panel C). As before, we focus on the results produced by ensemble methods, Combinations 1 and 2.

First, the monotonic increase in portfolio *alphas* form the low to high portfolios is only observed for Combination 2, and only in Panels A and B. Thus using larger ensembles allows obtaining more robust results. As Combination 2 also includes non-linear methods, consistent with Gu, Kelly, and Xiu (2020), we also confirm the importance of considering non-linear predictive relations between option predictors and expected stock returns. Further, neither Combination 1 nor 2 methods can satisfy monotonicity requirement in Panel C, which only uses stock features as predictors.

The advantage of using options, Panel B, or both options and stock, Panel A, predictors is further portrayed via Sharpe ratios. The annualized Sharpe ratio for Combination 2 ensemble is the same, 1.49, in both Panels A and B, while the corresponding Sharpe ratio of using only stock predictors is substantially lower, 1.17, Panel C. This huge differences in Sharpe ratios undoubtedly suggests the informational advantage of options over stock predictors in predicting stock returns.

In the previous section we establish the results that options predictors are more important than stock predictors to predict options returns. In this sections we demonstrate that options predictors are also more important than stock predictors in forecasting stock returns. This allows us to generally conclude that options markets lead the stock market.

6 The Leading Role of Options Market

6.1 Intertemporal Evidence

So far we have provided a substantial evidence that options predictors dominate both option and stock return predictability. Given that option prices are based on forward looking believes about the values of the underlying, and option predictors provide their more accurate statistical and economic forecasts allows us to conclude that option market has informational advantage and thus leads the stock market. While this conclusion supports the very fundamental premise of options (Black (1975)) and a lot of existing literature, it comes at odds with the literature arguing that options either rather follow stocks (Muravyev, Pearson, and Broussard (2013)), or they largely reflect information about short-selling costs in the stock market (Muravyev, Pearson, and Pollet (2021)).

Our analysis separates us from the rest of the literature in several important ways. First, the option trading activity increased dramatically during our sample period compared to the previous studies. This increase in option volumes should be accompanied with richer information

Electronic copy available at: https://ssrn.com/abstract=3747238

environment of the option market and its representative indicators (Cao (1999)). As such, options predictors in our sample can become more informationally portent compared to the previous literature, which in turn mitigates identification issues.

Second, we use a big data approach to compare relative importance of option versus stock predictors. This allows to decrease the risk of having an omitted variable problem as we are capturing all available information from all markets.

Finally, we also allow for the non-linear relations between predictors and expected returns. This non-linear relations has been argued to be important in predicting stock returns (Gu, Kelly, and Xiu (2020)), and more recently mutual fund returns (Kaniel, Lin, Pelger, and Van Nieuwerburgh (2022)).

The intertemporal performance of ML stock portfolios presented in Table 7 provides an ideal ground to demonstrate the validity of the above arguments. First, provided that the option market leads the stock market, and that an increasing options trading enriches option predictors with more information, then the importance of option based predictors should be increasing over time compared to the stock predictors in predicting stock returns. Second, if options predictors mostly identify short-selling restrictions in the underlying which investors overcome by trading in options, then the majority of High-minus-Low ML portfolio profits should be attributed to the short rather than long positions. Third, if non-linearity in predictive relations is what the previous literature has not considered in making conclusions, then we should see that option predictors dominate predictability across all times.

Figure 4 presents cumulative out-of-sample portfolio performances of High-minus-Low strategies, on the left, as well as long (top 10%) and short (bottom 10%) stock ML portfolio performances, on the right, for our out-of-sample, 03/2008 to 12/2019, period, for three sets of predictors: All variables, options and stock predictors, as well as for only option predictors or only stock predictors. The ML portfolios here are based on forecasts of Combination 1 for which the monthly summaries are presented in Table 7, the first row of each panel. The graphs depict the performance of \$100 invested in the beginning of OOS period and its growth through the end of the period.

First, consider High-Low strategy, the left graph. The performance of strategies based on stock predictors alone, or both stock and option predictors, dominate the performance of the strategy that relies only on option predictors approximately till 2012-2013. After that, the performance of the strategy with only stock predictors takes a dive while the one based on only option predictors continues its gradual increase and tops all other graphs. It then continues significantly dominate the only stock predictors strategies through the end of the sample period. Two important observations can be noted. First, consistent with almost exponential growth of options trading activity over past 12 years, option predictors gain more and more information over time related to the future valuations compared to stock predictors alone. Second, the value of this information gradually

Electronic copy available at: https://ssrn.com/abstract=3747238

increases over time as demonstrated by the performance of H-L strategy based on option predictors alone. One can easily see how in the studies which use the sample preceding 2012, and linear methodologies, the mixed results can be obtained about which market is more informative about future valuations of the underlying.

Next, the graph to the right shows that the performance of H-L strategy is mostly driven by the Long rather than short position. As a matter of fact, and investor can gain higher profits by entering only into the Long position. An immediate conclusion is that the option market is not dominated by trading on the negative information. Finally, the performances of Long portfolios based on only options versus only stock predictors are quite close, with the former only marginally outperforming the latter. This again shows how using only linear methods the literature can be inconclusive about which market's variables are the most informative, as the race is quite close.

Figure 5 presents similar cumulative out-of-sample portfolio performances to those of Figure 4 but for Combination 2 forecasts which now also allows for nonlinear relations between predictors and expected returns. Its monthly summaries are presented in Table 7, the second row of each panel. One important observation is that H-L strategy, the graph to the left, using only option predictors consistently dominates all other strategies across all OOS time periods. It suggest that nonlinear relations can be critical for identification purposes, and the option literature should adapt nonlinear methods while studying the informational advantages of the options market. Further, the performance of H-L strategy is entirely driven by Long position as Short position remains largely flat, the graph to the right. Moreover, the performance of Long only portfolio formed based on only option predictors forecasts undisputedly outperforms the one based on only stock predictors forecast. Clearly, option predictors are far superior to stock predictors in predicting the future values of the underlying once non-linearities in predictive relations are allowed. Even with non-linear methods forecast the meaningful outperformance of option predictors alone commences after 2012.

To further validate the argument Figure 6 plots the aggregate ratio of notional value of option trading volume to the dollar stock trading volume. Here, option notional trading volumes are summed up for a day across all traded contracts, and then divided by the sum of overall dollar trading volume of the underlying stocks. We then compute the average of these daily ratios for each month from 01/1996 to 12/2020. First important observation, consistent with the reports in popular press, is that in August 2020 this ratio exceeds 1, implying that trading activity in options for the first time exceeds trading in stocks. It further exceeds 1 in the end of 2020. The other important observation is that this ratio is relatively stable through the end of 2010, and then visibly jumps and starts increasing in 2012. It takes a short dive in 2016, but never to the levels of 1996-2011 period. It then increases dramatically surpassing 1 in the end of the period. The take off of this ratio in post 2012 period also corresponds to the beginning of outperformance of ML portfolios based on option predictors alone forecasts in Figures 4 and 5. This further allows us to conclude that the last

Electronic copy available at: https://ssrn.com/abstract=3747238

decade where options trading activity really sparks provides more informational advantage to the option market, consistent with the theory of Cao (1999).

6.2 Illiquidity and Trading Volume

What kind of information does options trading capture? Figure 3, which reports the most important stock returns predictors for each method consistently identifies option illiquidity (*opt_baspread*) as the most important one. Option illiquidity is also the most important predictor of options returns, Figure 1, with even higher variable importance weights. Given that machine learning methods are able to identify the most non-redundant predictors compared to the pool of all other features, we use this first order importance predictor to analyze the primary economic fundamentals behind the predictability.

Traditionally, illiquidity or higher trading costs arise to compensate liquidity providers for adverse selection risk (Kyle (1985), Glosten and Milgrom (1985)), inventory risk (Stoll (1978b), Ho and Stoll (1983)), and the risk of holding inventory while waiting for the offsetting order flows (Grossman and Miller (1988)). Trading activity, or volume plays a critical role as intuitively higher trading volume is associated with lower asymmetric information, lower inventory risk and lower inventory holding time and costs. This results in higher liquidity. In classical literature, liquidity and volume are strongly related (Benston and Hagerman (1974)).

What kind of information does options illiquidity capture? Christoffersen, Goyenko, Jacobs, and Karoui (2018) document positive *option illiquidity premium* in *option* returns. More illiquid options have higher expected returns due to higher risks and costs of option market makers inventories. The more difficult to delta-hedge, or rebalance delta-hedge positions due to more illiquid underlying stocks with higher volatility, asymmetric information, and probability of price jumps risks, the wider option bid-ask spreads. The authors show that option illiquidity increases with stock illiquidity, probability of informed trading in the underlying stocks (PIN), and interaction of options gamma with the volatility of the underlying stocks. The latter is aimed to capture unhedgable gamma risks associated with more volatile stocks. Goyenko, Ornthanalai, and Tang (2015) further argue that option illiquidity capture a great deal of asymmetric information about the underlying stocks. These studies look at *daily contemporaneous determinants* of options illiquidity.

In contrast, we are interested in identifying important OOS determinants of options illiquidity in our monthly data, and using substantially larger range of both stock and options characteristics not previously considered simultaneously by other studies. Arguably, the main determinants of illiquidity, given its persistence, are also its most important predictors, and out-of-sample approach allows controlling for redundancy. We therefore intend to identify important OOS *monthly predictors* of options illiquidity with big data, i.e. our stock and options 1100+ features, and machine learning.

Electronic copy available at: https://ssrn.com/abstract=3747238

The methodology is similar to previous sections with one exception - our dependent variable is the log of option illiquidity measured by relative bid-ask spreads as before. We use the log transformation to normalize its distribution, since if it is non-transformed, the left-hand side variable would always remain positive. Our main interest here is to identify the main economic drivers of options illiquidity and hence the important information it captures to predict both options and stock returns. Moreover, this will also allow us to understand what information is reflected by an increased trading in options.

In this context, the predictor variable importance weights identification is the most relevant matrix. Figure 7 reports top 20 most important predictors of option illiquidity for each model for the cross-section of all optionable stocks in our sample. As expected, option illiquidity is persistent (Christoffersen, Goyenko, Jacobs, and Karoui (2018)) and its own lag (opt_baspread) is the first important predictor across all methods, except Ridge which places the first weight on beta squared (betasq). Here $betasq$ can be viewed as a proxy for uncertainty (Chordia, Huh, and Subrahmanyam (2007)). Importantly, unlike for option or stock returns, first 10 features capture the most of predictability, as the weights of bottom 10 variables are close to zero. All models are generally in agreement on the top 10 variables, and Figure A.4 in Appendix summarizes the average ranking across all methods and variables. The second most important variable predicting options illiquidity is stock illiquidity (Amihud's ill). This is again not surprising given the discussion above that stock illiquidity increases options market making costs and hence options illiquidity.

The third most important variable is OS (option to stock volume ratio). Given the recent dramatic increase in option volume compared to the stock volume, this result might as well represent the switch of trading activity by investors from stocks to options. An increase in trading of course has an immediate effect on illiquidity, and has applications for the cross-section of returns as well. Empirically, Roll, Schwartz, and Subrahmanyam (2010) find that OS cross-sectionally depends on the costs of trading, the size of the firm, the available degree of leverage in options, institutional holdings, and, to some extent, proxies for the likely availability of private information and the diversity of opinions. These OS determinants should also have pricing implications for stock returns.

Interestingly, the next 3 important predictors are again related to stock illiquidity and are in order of importance: size (mve), dollar volume (dolvol), and turnover (turn). Thus option illiquidity seem to capture almost all dimentionalities of stock illiquidity: price impact captured by ill, trading intensity captured by $dolvol$ and $turn$, and overall visibility captured by size. To this extant, industry adjusted size (mve_ai) is also in the top 10.

The other top 10 predictors are those identified as the main predictors of stock returns in the previous section: number of earnings increases (nincr) and momentum (mom12m). The 10th important predictor is secured debt indicator (securedind) which is a proxy for the credit risk.

Overall, options illiquidity captures a wide range of characteristics about the underlying: stock

Electronic copy available at: https://ssrn.com/abstract=3747238

trading costs and price impacts, uncertainty, information asymmetry, trading intensity, options trading volume relative to stock volume and credit quality of the underlying. Theoretically, a further increase in trading activity should even further reflect this information (Cao (1999)).

Arguably, the variable incorporating information about all these characteristics should be priced, as all ML methods rightly point out. Options illiquidity is highly significantly associated with options trading volume and it has been shown to predict options returns (Christoffersen, Goyenko, Jacobs, and Karoui (2018)). We next explore economic magnitudes of an increased options trading activity captured by options illiquidity for the cross-section of stock returns.

The hypothesis is motivated by the theory (Cao (1999)), where higher trading activity or trading volume in derivatives is associated with better information environment, lower illiquidity, and hence lower expected returns. Alternatively, higher illiquidity, or lower trading activity, should lead to higher expected returns.

Here we adapt a standard portfolio sorting approach as we already establish the robustness of predictability with ML analysis. Table 8 presents monthly decile portfolios' excess stock returns sorted on the previous month values of option illiquidity (op_baspread) for our test sample, 03/2008 to 12/2019. Panel A presents the results for options illiquidity measured as in the main tests using aggregated on the class level relative end-of-day bid-ask spreads in the end of the month, and only using contracts with positive volume. Panel B presents similar results where options illiquidity is measured differently. We still use end of day relative bid-ask spreads but we now aggregate them for a month on a class level, and again using only contracts with positive volume days within a month. As before, all stock portfolios are value-weighted.

The first row in each panel is raw portfolio excess return, followed by risk adjusted returns, Alphas, using Fama-French five factor model augmented with momentum factor, and the last row reports Newey-West t-statistics of portfolio alphas. We also present returns and alphas of zero-investment strategy portfolio of going long the most illiquid decile (10), and shorting the most liquidity decile (1), and its annualized Sharpe ratio.

The first important observation is that the portfolio returns and alphas are increasing almost monotonically with the illiquidity decile in both panels. The alpha of High-minus-Low strategy is 1.63% (t=8.43) per month, and Sharpe Ratio is 2.00 in Panel A, and these numbers are slightly higher in Panel B, with Alpha of 1.78% (t=10.49), and annualized Sharpe ratio of 2.24.

The second important observation is that the results are not driven by the extreme deciles. The individual portfolios' alphas are increasing with portfolio rank and majority of them are highly statistically significant. Since our cross-section of stocks are optionamble stocks which eliminates micro-cap and smaller stocks, we find that the negative alphas of lower decile portfolios are insignificant. Moreover, the positive profits of High-minus-Low position are driven by Long, rather than Short portfolios, which eliminates the argument that the option market is a venue to overcome short-selling frictions in the stock market (D'avolio (2002)).

Electronic copy available at: https://ssrn.com/abstract=3747238

The results point to positive options illiquidity premium in stock returns. This result is new to the literature. Stock illiquidity premium disappears in the stock returns in more recent data once small stocks are excluded (Ben-Rephael, Kadan, and Wohl (2015)).[19] We find that stocks with more illiquid options have higher expected returns. Our analysis suggests that stocks with more illiquid options are also more illiquid themselves, have higher volatility and the risk of price jumps, and also have higher information asymmetry.

We therefore conclude that a dramatic increase in options trading activity leads to an informational advantage of options markets compared to stocks, and option-based predictors, especially those reflecting trading activity, i.e. option illiquidity, dominating both stock and options returns predictability.

7 Conclusion

An explosive increase in options trading activity by institutional and retail investors can lead to options markets having informational advantage over the stock market. If this is the case, then options based predictors, and especially those reflecting trading activity, should dominate both options and stock returns predictability.

We test this hypothesis using large data sets of stock and option predictors, 14 machine learning methods to overcome data sparsity issues and identify non-redundant first order importance predictors.

More specifically, we run a horse race between option and stock based predictors across different dimensions: out-of-sample precision forecast accuracy measured by $OOS\ R^2$; out-of-sample performance of machine learning portfolios formed on their return forecasts, and the performance of long-short zero investment strategy portfolio and its Sharpe ratio to evaluate economic gains of forecasts; and using two different approaches we identify the most influential predictors out of pool of 100+ options and stock features.

First, in contrast to the previous literature, we find that option based predictors dominate stock predictors across all dimensions for predicting options returns. Here, in order of relative importance, option illiquidity, implied volatility and the difference between historical and implied volatility are the top predictors, with options illiquidity dominating overall.

Second, we show that machine learning and large set of option predictors provide statistically and economically more accurate forecasts of stock returns compared to using only stock predictors. Moreover, almost all machine learning methods unanimously identify option illiquidity as the most important predictor of stock returns.

The general conclusion is that options based predictors lead both options and stock return

[19] We also independently performed similar portfolio sorts but conditioning on stock illiquidity. As in Ben-Rephael, Kadan, and Wohl (2015), we find no significant results pointing to the stock market illiquidity premium in our sample.

Electronic copy available at: https://ssrn.com/abstract=3747238

predictability. Among these predictors, options illiquidity is identified as the most important predictor for both stock and options returns.

Arguably, illiquidity and trading activity are highly inter-related. We analyze determinants of options illiquidity and find that it captures the following information about the underlying: stock illiquidity, trading intensity, size, information asymmetry, uncertainty and credit quality. Altogether this demonstrates what kind of information is reflected by an increased options trading. This information should be priced in the stock returns. As a result, we confirm a positive options illiquidity premium in the cross-section of stock returns.

Electronic copy available at: https://ssrn.com/abstract=3747238

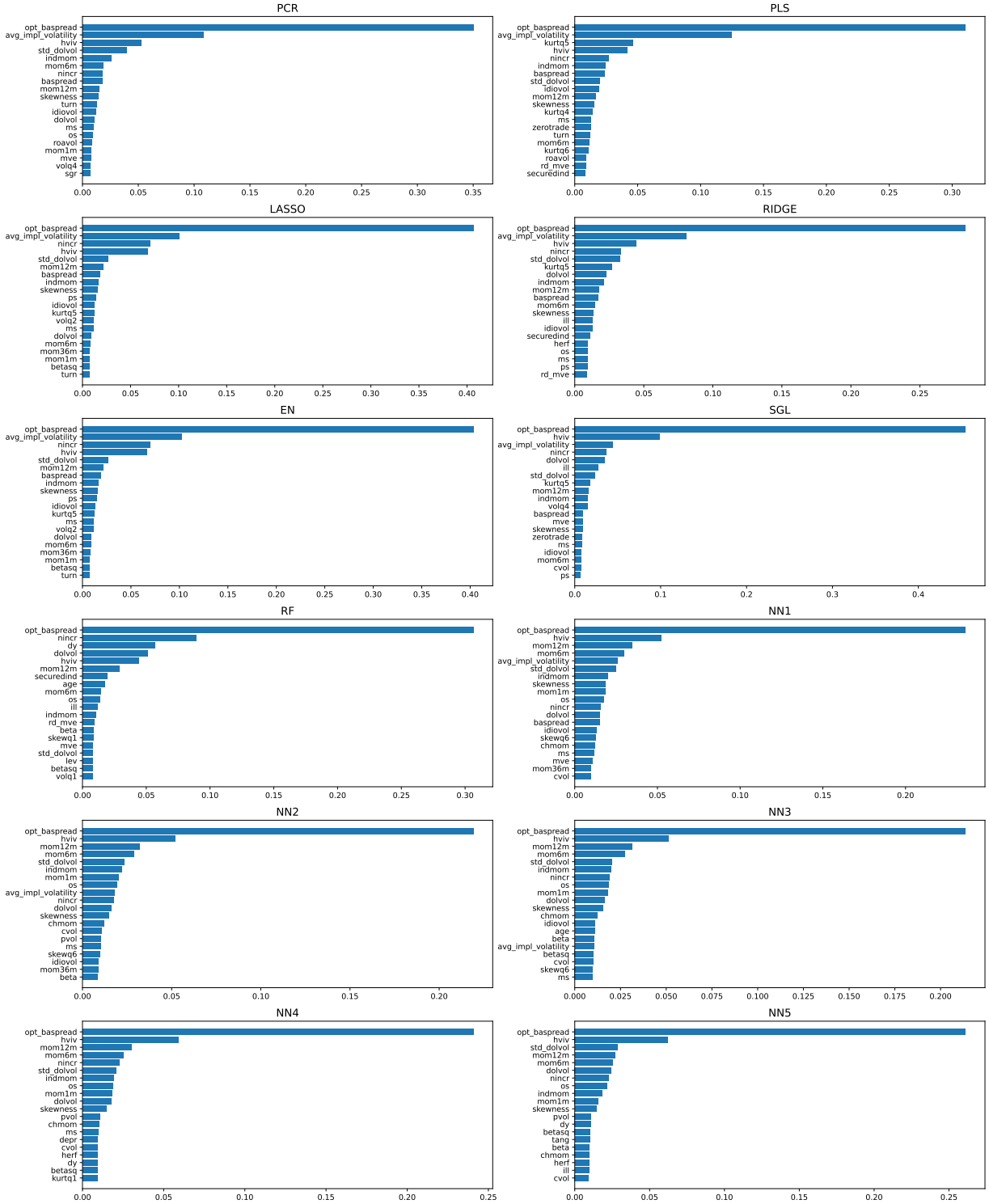

Figure 1: Variable importance by model: straddle returns predictability, all stocks

Notes: The figure demonstrates the variable importance for the top-20 most influential predictors of delta-neutral straddle returns for all optionable stocks. Variable importance in each model is normalized to sum to one.

Electronic copy available at: https://ssrn.com/abstract=3747238

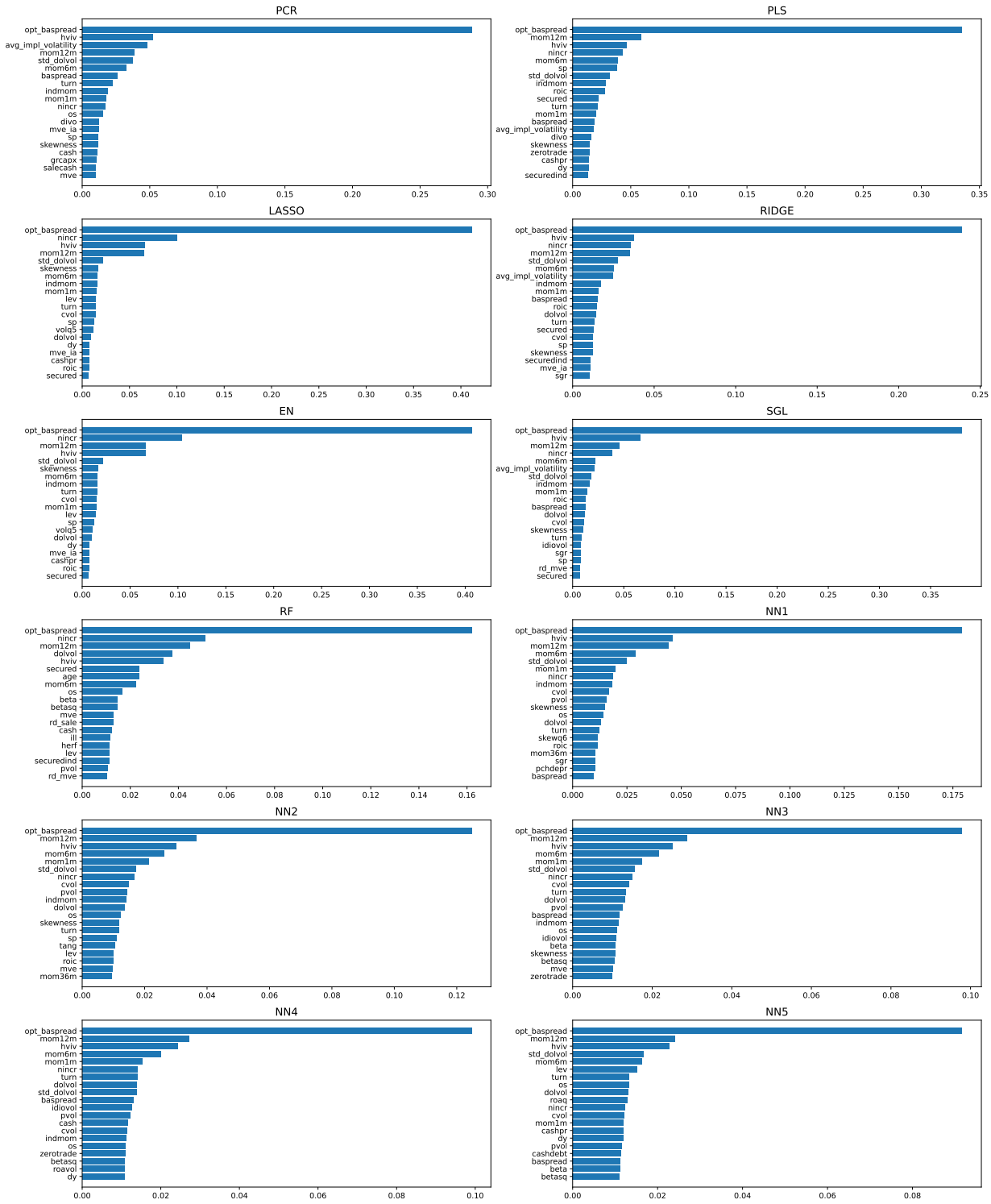

Figure 2: Variable importance by model: straddle returns predictability, S&P 500 stocks

Notes: The figure demonstrates the variable importance for the top-20 most influential predictors of delta-neutral straddle returns for S&P 500 stocks. Variable importance in each model is normalized to sum to one.

Electronic copy available at: https://ssrn.com/abstract=3747238

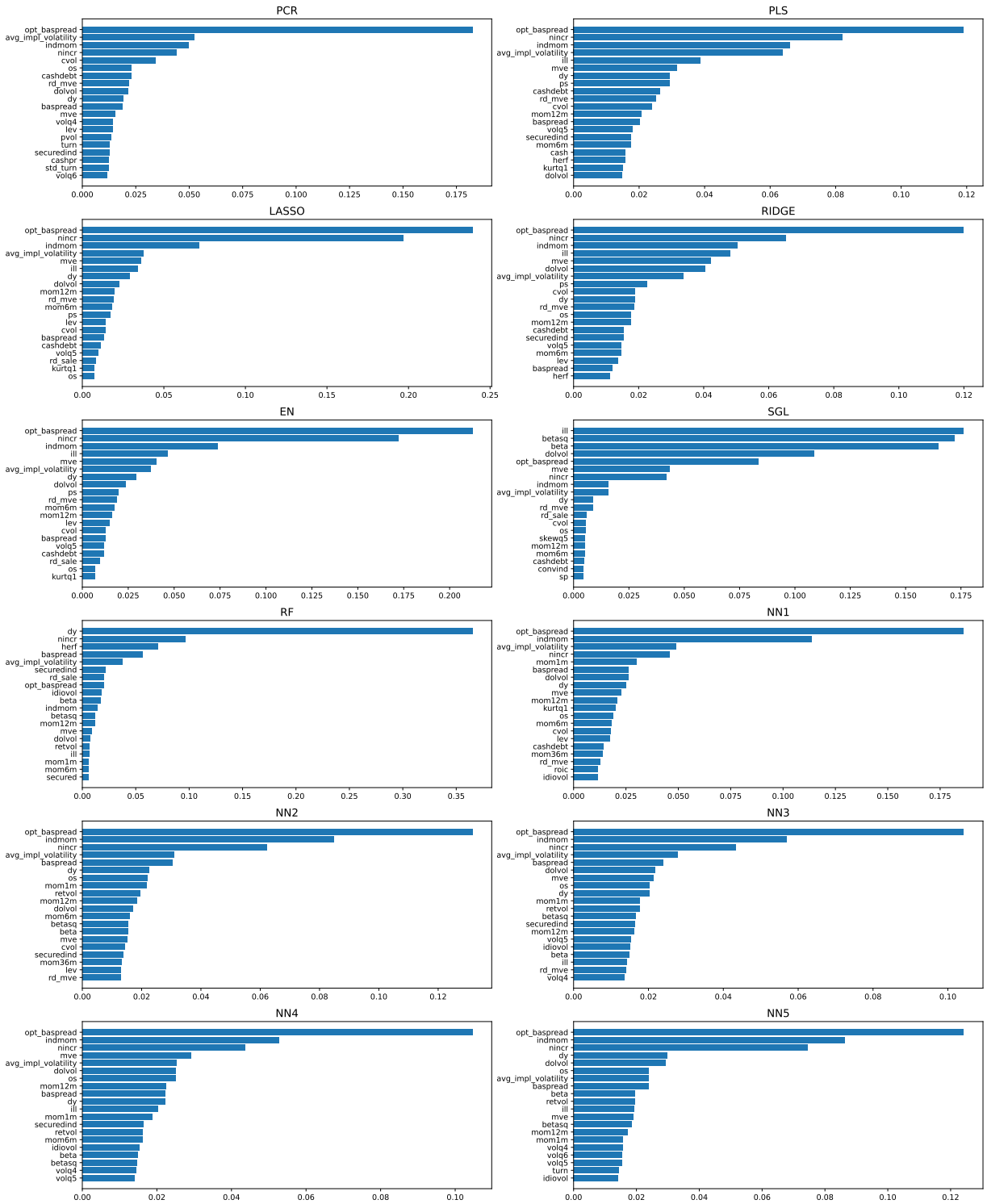

Figure 3: Variable importance by model: stock excess return predictability, all stocks

Notes: The figure demonstrates the variable importance for the top-20 most influential predictors of excess stock returns. Variable importance in each model is normalized to sum to one.

Electronic copy available at: https://ssrn.com/abstract=3747238

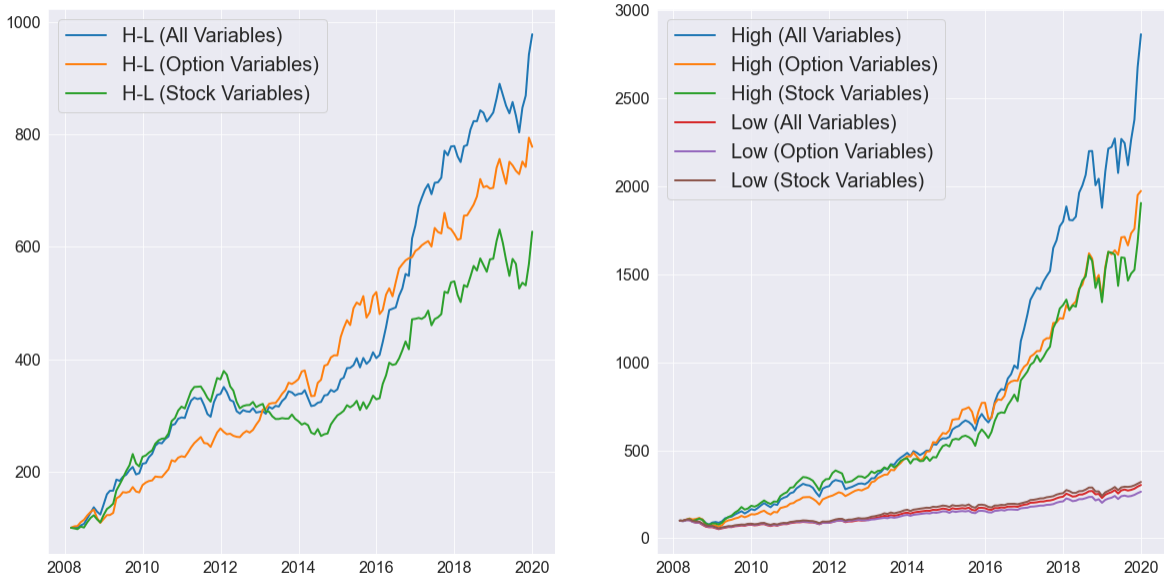

Figure 4: Intertemporal Performance of stock ML portfolios: Combination 1

Notes: The figure demonstrates the cumulative returns of High minus Low, left, and top versus bottom, right, stock ML portfolios constructed based on the forecasts of Combination 1 method for the three sets of predictors: all variables, which includes both option and stock predictors, only option predictors, and only stock predictors.

Electronic copy available at: https://ssrn.com/abstract=3747238

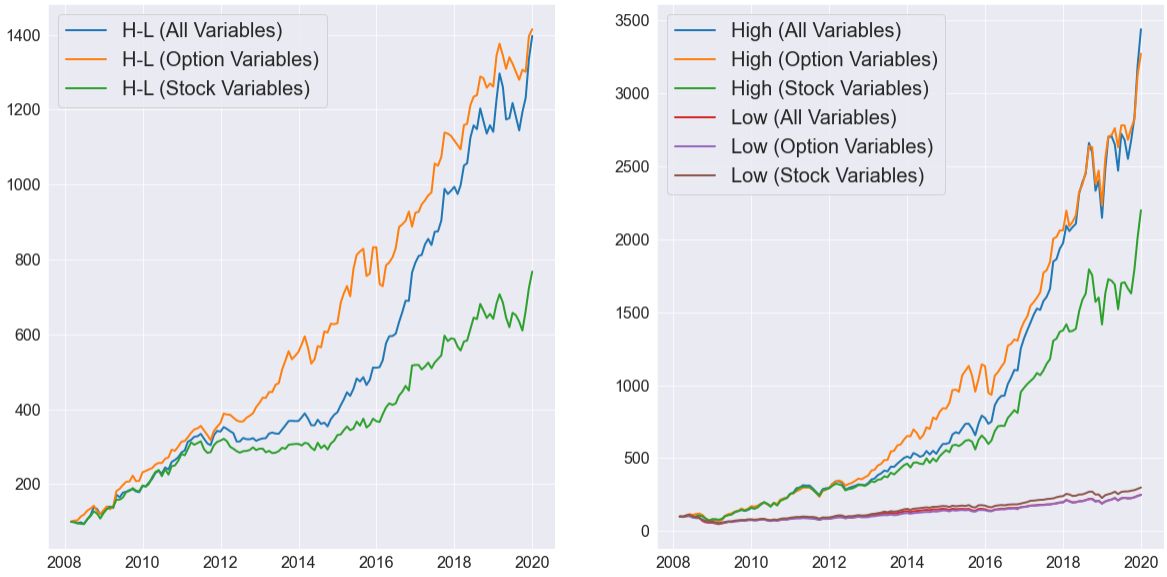

Figure 5: Intertemporal Performance of stock ML portfolios: Combination 2

Notes: The figure demonstrates the cumulative returns of High minus Low, left, and top versus bottom, right, stock ML portfolios constructed based on the forecasts of Combination 1 method for the three sets of predictors: all variables, which includes both option and stock predictors, only option predictors, and only stock predictors.

Electronic copy available at: https://ssrn.com/abstract=3747238

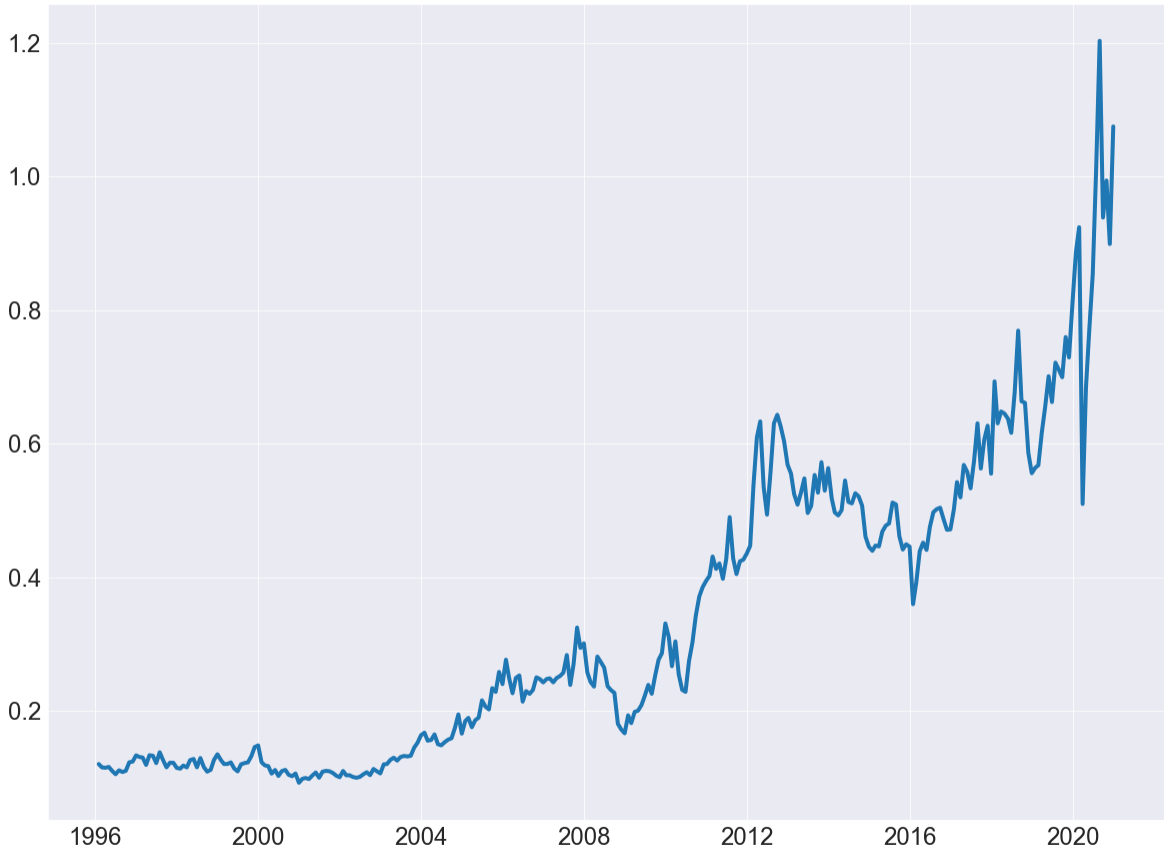

Figure 6: Aggregate Notional Option to Dollar Stock Volume Ratio

Notes: The figure plots the aggregate notional option to dollar stock volume ratio for 01/1996 to 12/2020 time period.

Electronic copy available at: https://ssrn.com/abstract=3747238

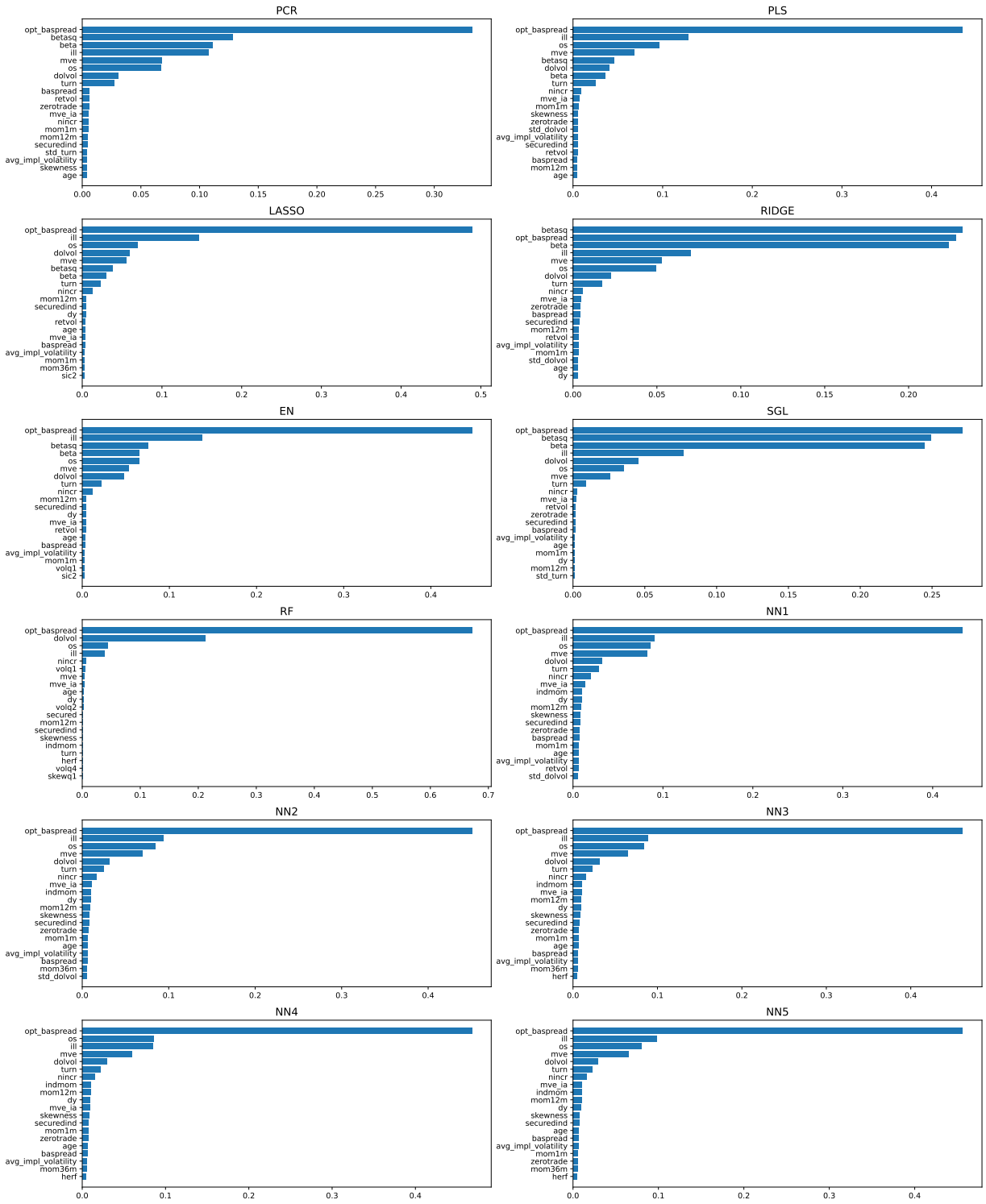

Figure 7: Variable importance by model: option illiquidity predictability, all stocks

Notes: The figure demonstrates the variable importance for the top-20 most influential predictors of option illiquidity for all optionable stocks in our sample.

Electronic copy available at: https://ssrn.com/abstract=3747238

Panel A. Option Returns, All Firms								
# of Obs	# of Unique Firms	Mean	Std	P10	P25	P50	P75	P90
121049	3805	6.06%	28.20%	-27.56%	-5.02%	11.00%	23.25%	34.84%
Panel B. Option Returns, S&P 500 Firms								
# of Obs	# of Unique Firms	Mean	Std	P10	P25	P50	P75	P90
66291	814	5.06%	28.25%	-29.37%	-6.10%	10.68%	22.55%	33.35%
Panel C. Stock Returns								
# of Obs	# of Unique Firms	Mean	Std	P10	P25	P50	P75	P90
246031	4867	1.55%	13.88%	-12.63%	-5.20%	1.07%	7.41%	15.45%

Table 1: Summary Statistics

Notes: The table reports descriptive summary statistics of monthly options delta-neutral straddle returns as well as returns of all optionable underlying stocks. Panels A and B present pooled summary returns to a straddle writing strategy for all and S&P500 stocks respectively. Panel C describes the returns of the underlying stocks. Each panel also includes total number of observations in each sample and the number of unique firms. The sample period is from 02/1996 to 12/2019.

Electronic copy available at: https://ssrn.com/abstract=3747238

Panel A. All Variables															
	OLS	PCR	PLS	Lasso	Ridge	Elastic Net	SGL	RF	NN1	NN2	NN3	NN4	NN5	Combination1	Combination2
All Straddle Return	-39.42	5.21	5.30	5.54	5.34	5.54	5.51	5.22	4.99	5.36	5.27	4.98	4.95	5.55	5.51
S&P 500 Straddle Return	-45.43	3.31	3.18	3.28	3.33	3.28	3.08	2.46	2.71	2.85	3.11	2.91	2.13	3.32	3.26
Panel B. Option Variables															
	OLS	PCR	PLS	Lasso	Ridge	Elastic Net	SGL	RF	NN1	NN2	NN3	NN4	NN5	Combination1	Combination2
All Straddle Return	-3.93	5.25	5.24	5.47	5.33	5.47	5.37	5.45	5.07	5.27	5.21	5.34	5.14	5.45	5.48
S&P 500 Straddle Return	-7.94	3.15	3.20	3.17	3.23	3.17	3.19	3.04	2.93	3.00	2.80	2.78	2.49	3.22	3.20
Panel C. Stock Variables															
	OLS	PCR	PLS	Lasso	Ridge	Elastic Net	SGL	RF	NN1	NN2	NN3	NN4	NN5	Combination1	Combination2
All Straddle Return	-34.80	5.01	5.02	5.03	5.01	5.02	4.81	4.91	4.69	4.90	4.78	4.84	4.66	5.04	5.06
S&P 500 Straddle Return	-36.05	3.05	2.92	3.00	3.03	3.00	3.00	2.81	2.38	2.86	2.75	3.07	1.70	3.05	3.03

Table 2: Monthly out-of-sample option returns prediction performance (percentage $OOS\ R^2$)

Notes: The table reports monthly $OOS\ R^2$ for option return predictability for the entire panel of stocks first using both option and stock predictors, Panel A, second, only option predictors, Panel B, and, third, only stocks predictors, Panel C. The machine learning methods are described in Section 2. The results are reported for all optionable stocks, as well as for stocks of S&P500 index components.

Electronic copy available at: https://ssrn.com/abstract=3747238

Panel A. All Straddle Returns					
	Combination1			Combination2	
	Option Variables	Stock Variables		Option Variables	Stock Variables
All Variables	-1.11	**-5.53**	All Variables	-0.47	**-6.66**
Option Variables		**-3.18**	Option Variables		**-5.50**
Panel B. S&P 500 Straddle Returns					
	Combination1			Combination2	
	Option Variables	Stock Variables		Option Variables	Stock Variables
All Variables	-0.77	**-3.29**	All Variables	-0.70	**-3.09**
Option Variables		-1.29	Option Variables		**-1.84**

Table 3: Diebold-Mariano test across variable sets

Notes: The table reports Diebold-Mariano test statistics for pairwise comparison of Combination 1 and 2 models across Panels A, B, and C of Table 2. Positive statistic indicates the column model outperforms the row model, and bold numbers denote significance at the 5% level for each individual test. Each Panel uses Combination 1 and 2 forecasts based on *All Variables (both options and stock predictors)*, only *Option Variables*, and only *Stock Variables* of Table 2.

Electronic copy available at: https://ssrn.com/abstract=3747238

Table 4: Performance of the machine learning portfolios: Option returns, Option and Stock predictors

Notes: The table reports the monthly performance of prediction-sorted portfolios for 03/2008 to 12/2019 out-of-sample testing period. All stocks are sorted into deciles based on their predicted option returns for the next month using stock and option predictors, only options predictors, or only stock predictors. SR is annualized Sharpe ratio. The reported returns are in percentages. All stock portfolios are value-weighted, and S&P500 stock portfolios are equally-weighted.

Panel A. All Straddle Returns, Option and Stock Predictors

	Low	2	3	4	5	6	7	8	9	High	H-L	H-L t-stats	H-L SR
Combination1	1.24	3.31	4.37	4.39	5.49	6.89	7.51	7.90	9.17	10.93	9.69	12.33	3.33
Combination2	1.55	2.83	4.06	4.76	5.49	6.55	7.43	8.43	8.78	11.36	9.81	12.59	3.58

All Straddle Returns, Only Option Predictors

	Low	2	3	4	5	6	7	8	9	High	H-L	H-L t-stats	H-L SR
Combination1	2.28	2.56	3.38	4.79	5.54	7.49	7.40	8.08	8.83	10.87	8.59	12.27	3.28
Combination2	2.62	2.34	3.34	4.71	5.60	6.68	7.92	7.76	8.98	11.29	8.67	12.99	3.55

All Straddle Returns, Only Stock Predictors

	Low	2	3	4	5	6	7	8	9	High	H-L	H-L t-stats	H-L SR
Combination1	1.95	4.10	4.97	5.48	5.93	6.97	7.24	7.45	8.39	8.75	6.80	9.17	2.58
Combination2	1.71	3.92	4.74	5.65	5.83	6.83	6.88	7.65	8.44	9.57	7.86	11.25	2.89

Panel B. S&P 500 Straddle Returns, Option and Stock Predictors

	Low	2	3	4	5	6	7	8	9	High	H-L	H-L t-stats	H-L SR
Combination1	1.10	2.94	3.04	3.52	3.63	4.21	5.88	6.97	7.46	9.50	8.39	10.07	2.69
Combination2	1.20	2.47	2.62	3.73	3.53	4.65	5.61	6.66	8.24	9.54	8.34	12.00	2.95

S&P 500 Straddle Returns, Only Option Predictors

	Low	2	3	4	5	6	7	8	9	High	H-L	H-L t-stats	H-L SR
Combination1	2.26	2.63	2.64	3.21	3.65	4.73	5.64	7.20	7.37	8.91	6.65	9.16	2.36
Combination2	2.27	2.82	2.29	3.05	3.98	4.50	5.96	6.84	7.45	9.10	6.83	9.60	2.55

S&P 500 Straddle Returns, Only Stock Predictors

	Low	2	3	4	5	6	7	8	9	High	H-L	H-L t-stats	H-L SR
Combination1	1.25	3.40	4.03	3.72	4.73	5.59	5.59	6.47	6.42	7.03	5.77	7.83	2.12
Combination2	1.41	3.05	3.58	3.80	4.69	5.27	5.02	6.72	7.19	7.47	6.06	8.49	2.28

Electronic copy available at: https://ssrn.com/abstract=3747238

All Stocks, Excess Returns, *OOS* R^2															
	OLS	PCR	PLS	Lasso	Ridge	Elastic Net	SGL	RF	NN1	NN2	NN3	NN4	NN5	Combination1	Combination2
All Variables	-69.68	-0.23	1.06	0.90	1.28	0.76	1.31	1.20	-0.87	0.08	0.75	0.58	0.63	1.17	1.12
Option Variables	-14.13	1.89	1.77	1.73	2.00	1.74	1.95	1.76	-0.89	1.00	1.09	0.41	1.51	1.90	1.63
Stock Variables	-63.79	-0.29	0.65	0.91	1.24	0.76	1.13	1.28	-2.52	0.74	0.72	0.41	0.48	1.09	0.92

Table 5: Monthly out-of-sample stock-level stock returns prediction performance (percentage $OOS\ R^2$)

Notes: The table reports monthly $OOS\ R^2$ for excess stock return predictability for the entire panel of all optionable stocks first using both option and stock predictors, *All variables*, second, only option predictors, *Option Variables*, and, third, only stocks predictors, *Stock Variables*. The machine learning methods are described in Section 2.

Electronic copy available at: https://ssrn.com/abstract=3747238

	Combination1			Combination2	
	Option Variables	Stock Variables		Option Variables	Stock Variables
All Variables	**1.71**	-1.19	All Variables	1.46	-1.58
Option Variables		**-2.02**	Option Variables		**-2.33**

Table 6: Diebold-Mariano test across variable sets: Stock Returns Predictability

Notes: The table reports Diebold-Mariano test statistics for pairwise comparison of Combination 1 and 2 models of Table 5. Positive statistic indicates the column model outperforms the row model, and bold numbers denote significance at the 5% level for each individual test. Combination 1 and 2 forecasts are based on using both stock and options features, denoted *All Variables*, only options features, denoted *Option Variables*, and only stock features, denoted *Stock Variables* in Table 5.

Electronic copy available at: https://ssrn.com/abstract=3747238

Table 7: Performance of the machine learning portfolios: Stock returns, Option and Stock predictors

Notes: The table reports the monthly performance of prediction-sorted portfolios for 03/2008 to 12/2019 out-of-sample testing period. All stocks are sorted into deciles based on their predicted excess stock returns for the next month using both option and stock predictors. SR is annualized Sharpe ratio. The reported returns are in percentages. All stock portfolios are value-weighted. Each Panel reports raw portfolio excess returns, Alphas computed using Fama-French five factor model augmented with Carhart's momentum factor, and Newey-West t-statistics adjust for 3 auto-correlation lags.

Panel A. All Stocks, value-weighted machine learning portfolio returns, Option and Stock predictors

	Low	2	3	4	5	6	7	8	9	High	H-L	H-L SR
Combination1	0.88	0.83	0.90	1.02	1.03	1.39	1.37	1.60	1.80	2.56	1.68	1.59
Combination2	0.77	0.94	0.93	0.96	1.09	1.41	1.44	1.42	2.22	2.74	1.97	1.49

Portfolios Alphas, Fama-French five factors plus Momentum risk adjustment

	Low	2	3	4	5	6	7	8	9	High	H-L
Combination1	0.02	-0.07	-0.01	0.16	0.11	0.54	0.43	0.69	0.83	1.63	1.60
Combination2	-0.15	0.05	0.04	0.07	0.19	0.46	0.48	0.35	1.20	1.72	1.88

Portfolio Alphas' t-stats

	Low	2	3	4	5	6	7	8	9	High	H-L
Combination1	0.25	-0.85	-0.13	2.36	0.98	4.64	2.93	3.62	3.79	6.26	5.69
Combination2	-0.79	0.54	0.40	0.78	1.99	4.72	2.95	1.93	4.97	7.77	5.75

Panel B. All Stocks, value-weighted machine learning portfolio returns, Only Option predictors

	Low	2	3	4	5	6	7	8	9	High	H-L	H-L SR
Combination1	0.78	0.94	0.81	1.05	1.11	1.23	1.33	1.54	1.90	2.30	1.52	1.42
Combination2	0.74	0.88	0.94	0.99	1.05	1.34	1.30	1.48	1.87	2.72	1.98	1.49

Portfolios Alphas, Fama-French five factors plus Momentum risk adjustment

	Low	2	3	4	5	6	7	8	9	High	H-L
Combination1	-0.06	0.01	-0.08	0.20	0.20	0.35	0.48	0.66	1.00	1.36	1.42
Combination2	-0.13	0.00	0.01	0.10	0.18	0.46	0.31	0.59	0.88	1.72	1.86

Portfolio Alphas' t-stats

	Low	2	3	4	5	6	7	8	9	High	H-L
Combination1	-0.84	0.07	-0.81	2.18	2.56	2.49	3.59	5.00	7.52	8.53	7.02
Combination2	-1.50	0.03	0.08	0.79	1.40	4.02	1.82	3.77	5.20	9.88	8.32

Panel C. All Stocks, value-weighted machine learning portfolio returns, Only Stock predictors

	Low	2	3	4	5	6	7	8	9	High	H-L	H-L SR
Combination1	0.92	0.93	1.08	1.03	1.25	1.37	1.31	1.56	2.19	2.30	1.39	1.15
Combination2	0.87	0.94	1.02	0.89	0.96	1.15	1.34	1.53	1.81	2.42	1.55	1.17

Portfolios Alphas, Fama-French five factors plus Momentum risk adjustment

	Low	2	3	4	5	6	7	8	9	High	H-L
Combination1	0.09	0.05	0.16	0.08	0.24	0.38	0.34	0.49	1.19	1.29	1.21
Combination2	0.01	0.03	0.12	0.04	-0.02	0.20	0.26	0.48	0.84	1.33	1.32

Portfolio Alphas' t-stats

	Low	2	3	4	5	6	7	8	9	High	H-L
Combination1	1.14	0.73	1.58	0.70	1.93	3.12	2.39	2.85	5.79	4.85	4.18
Combination2	0.11	0.37	1.18	0.32	-0.21	1.55	1.56	3.73	4.56	4.89	4.04

Electronic copy available at: https://ssrn.com/abstract=3747238

Panel A. Options Illiquidity Measured in the end of a month												
	Low	2	3	4	5	6	7	8	9	High	H-L	H-L SR
Portfolio Return	0.80	0.91	0.86	1.11	0.95	1.25	1.43	1.54	1.84	2.35	1.55	2.00
Alpha	-0.06	0.00	-0.01	0.19	0.14	0.40	0.56	0.68	0.99	1.56	1.63	
Alpha t-stats	-0.94	0.04	-0.14	2.37	1.81	3.66	5.44	5.84	7.69	10.10	8.43	
Panel B. Average Options Illiquidity for a month												
	Low	2	3	4	5	6	7	8	9	High	H-L	H-L SR
Portfolio Return	0.84	0.84	0.90	1.19	1.00	1.19	1.23	1.57	1.79	2.50	1.66	2.24
Alpha	-0.04	-0.04	-0.01	0.35	0.14	0.29	0.33	0.77	1.08	1.74	1.78	
Alpha t-stats	-0.78	-0.45	-0.11	4.01	1.52	2.55	3.45	6.72	9.39	11.7	10.49	

Table 8: Option Illiquidity Premium in the Stock Market

Notes: The table presents monthly stock portfolio sorts results based on lagged values of option illiquidity for the out out-of-sample test period, 03/2008 to 12/2019. Each panel presents stock portfolio excess returns, Alpha computed using Fama-French five factor model augmented with Carhart's momentum factor, and Newey-West t-statistics adjust for 3 auto-correlation lags. SR is annualized Sharpe ratio. The reported returns are in percentages. All stock portfolios, Panels A and B, are value-weighted. Panel A uses end-of-month relative options bid-ask spreads as a measure of options illiquidity. Panel B uses the monthly average of end-of-day options relative bid-ask spreads as a measure of options illiquidity.

Electronic copy available at: https://ssrn.com/abstract=3747238

References

AN, B.-J., A. ANG, T. G. BALI, AND N. CAKICI (2014): "The joint cross section of stocks and options," *The Journal of Finance*, 69(5), 2279–2337.

BEN-REPHAEL, A., O. KADAN, AND A. WOHL (2015): "The diminishing liquidity premium," *Journal of Financial and Quantitative Analysis*, pp. 197–229.

BENSTON, G. J., AND R. L. HAGERMAN (1974): "Determinants of bid-asked spreads in the over-the-counter market," *Journal of Financial Economics*, 1(4), 353–364.

BLACK, F. (1975): "Fact and fantasy in the use of options," *Financial Analysts Journal*, 31(4), 36–41.

BOLLEN, N. P., AND R. E. WHALEY (2004): "Does net buying pressure affect the shape of implied volatility functions?," *The Journal of Finance*, 59(2), 711–753.

BOYER, B. H., AND K. VORKINK (2014): "Stock options as lotteries," *The Journal of Finance*, 69(4), 1485–1527.

BREIMAN, L. (2001): "Random forests," *Machine learning*, 45(1), 5–32.

BURASCHI, A., AND J. JACKWERTH (2001): "The price of a smile: Hedging and spanning in option markets," *The Review of Financial Studies*, 14(2), 495–527.

BYUN, S.-J., AND D.-H. KIM (2016): "Gambling preference and individual equity option returns," *Journal of Financial Economics*, 122(1), 155–174.

CAO, H. H. (1999): "The effect of derivative assets on information acquisition and price behavior in a rational expectations equilibrium," *The Review of Financial Studies*, 12(1), 131–163.

CAO, J., AND B. HAN (2013): "Cross section of option returns and idiosyncratic stock volatility," *Journal of Financial Economics*, 108(1), 231–249.

CAO, J., B. HAN, Q. TONG, AND X. ZHAN (2021): "Option return predictability," in *Review of Financial Studies, forthcoming*.

CETIN, U., R. JARROW, P. PROTTER, AND M. WARACHKA (2006): "Pricing options in an extended Black Scholes economy with illiquidity: Theory and empirical evidence," *The Review of Financial Studies*, 19(2), 493–529.

CHAKRAVARTY, S., H. GULEN, AND S. MAYHEW (2004): "Informed trading in stock and option markets," *The Journal of Finance*, 59(3), 1235–1257.

Electronic copy available at: https://ssrn.com/abstract=3747238

CHORDIA, T., S.-W. HUH, AND A. SUBRAHMANYAM (2007): "The cross-section of expected trading activity," *The Review of Financial Studies*, 20(3), 709–740.

CHORDIA, T., R. ROLL, AND A. SUBRAHMANYAM (2011): "Recent trends in trading activity and market quality," *Journal of Financial Economics*, 101(2), 243–263.

CHRISTOFFERSEN, P., R. GOYENKO, K. JACOBS, AND M. KAROUI (2018): "Illiquidity premia in the equity options market," *The Review of Financial Studies*, 31(3), 811–851.

COCHRANE, J. H. (2011): "Presidential address: Discount rates," *The Journal of finance*, 66(4), 1047–1108.

CONRAD, J., R. DITTMAR, AND E. GHYSELS (2012): "Ex Ante Skewness and Expected Stock Returns, forthcoming," *Journal of Finance*.

COVAL, J. D., AND T. SHUMWAY (2001): "Expected option returns," *The journal of Finance*, 56(3), 983–1009.

COX, J. C., S. A. ROSS, AND M. RUBINSTEIN (1979): "Option pricing: A simplified approach," *Journal of financial Economics*, 7(3), 229–263.

CREMERS, M., R. GOYENKO, P. SCHULTZ, AND S. SZAURA (2019): "Do Option-Based Measures of Stock Mispricing Find Investment Opportunities or Market Frictions?," *Ruslan and Schultz, Paul and Szaura, Stephen, Do Option-Based Measures of Stock Mispricing Find Investment Opportunities or Market Frictions*.

CREMERS, M., AND D. WEINBAUM (2010): "Deviations from put-call parity and stock return predictability," *Journal of Financial and Quantitative Analysis*, pp. 335–367.

D'AVOLIO, G. (2002): "The market for borrowing stock," *Journal of financial economics*, 66(2-3), 271–306.

DIEBOLD, F. M., AND R. MARIANO (1995): "R.(1995). Comparing predictive accuracy," *Journal of Business & economic statistics*, 20(1).

EASLEY, D., M. O'HARA, AND P. S. SRINIVAS (1998): "Option volume and stock prices: Evidence on where informed traders trade," *The Journal of Finance*, 53(2), 431–465.

FRIEDMAN, J. H. (2001): "Greedy function approximation: a gradient boosting machine," *Annals of Statistics*, pp. 1189–1232.

GARLEANU, N., L. H. PEDERSEN, AND A. M. POTESHMAN (2008): "Demand-based option pricing," *The Review of Financial Studies*, 22(10), 4259–4299.

Electronic copy available at: https://ssrn.com/abstract=3747238

GE, L., T.-C. LIN, AND N. D. PEARSON (2016): "Why does the option to stock volume ratio predict stock returns?," *Journal of Financial Economics*, 120(3), 601–622.

GLOSTEN, L. R., AND P. R. MILGROM (1985): "Bid, ask and transaction prices in a specialist market with heterogeneously informed traders," *Journal of financial economics*, 14(1), 71–100.

GOODFELLOW, I., Y. BENGIO, AND A. COURVILLE (2016): *Deep learning*. MIT press.

GOYAL, A., AND A. SARETTO (2009): "Cross-section of option returns and volatility," *Journal of Financial Economics*, 94(2), 310–326.

GOYENKO, R., C. ORNTHANALAI, AND S. TANG (2015): "Options illiquidity: Determinants and implications for stock returns," *Rotman School of Management Working Paper*, (2492506).

GREEN, J., J. R. HAND, AND X. F. ZHANG (2017): "The characteristics that provide independent information about average us monthly stock returns," *The Review of Financial Studies*, 30(12), 4389–4436.

GROSSMAN, S. J., AND M. H. MILLER (1988): "Liquidity and market structure," *the Journal of Finance*, 43(3), 617–633.

GU, S., B. KELLY, AND D. XIU (2020): "Empirical asset pricing via machine learning," *The Review of Financial Studies*, 33(5), 2223–2273.

HARVEY, C. R., Y. LIU, AND H. ZHU (2016): "... and the cross-section of expected returns," *The Review of Financial Studies*, 29(1), 5–68.

HO, T. S., AND H. R. STOLL (1983): "The dynamics of dealer markets under competition," *The Journal of finance*, 38(4), 1053–1074.

HOERL, A. E., AND R. W. KENNARD (1970): "Ridge regression: Biased estimation for nonorthogonal problems," *Technometrics*, 12(1), 55–67.

HOU, K., C. XUE, AND L. ZHANG (2020): "Replicating anomalies," *The Review of Financial Studies*, 33(5), 2019–2133.

HU, J. (2014): "Does option trading convey stock price information?," *Journal of Financial Economics*, 111(3), 625–645.

JOHNSON, T. L., AND E. C. SO (2012): "The option to stock volume ratio and future returns," *Journal of Financial Economics*, 106(2), 262–286.

KACPERCZYK, M., AND E. S. PAGNOTTA (2019): "Chasing private information," *The Review of Financial Studies*, 32(12), 4997–5047.

Electronic copy available at: https://ssrn.com/abstract=3747238

KANIEL, R., Z. LIN, M. PELGER, AND S. VAN NIEUWERBURGH (2022): "Machine-learning the skill of mutual fund managers," Discussion paper, National Bureau of Economic Research.

KANNE, S., O. KORN, AND M. UHRIG-HOMBURG (2018): "Stock illiquidity and option returns," Discussion paper, Working paper „Karlsruhe Institute of Technology (KIT) and University of ....

KARAKAYA, M. M. (2014): *Characteristics and expected returns in individual equity options*. Citeseer.

KYLE, A. S. (1985): "Continuous auctions and insider trading," *Econometrica: Journal of the Econometric Society*, pp. 1315–1335.

MURAVYEV, D., N. D. PEARSON, AND J. P. BROUSSARD (2013): "Is there price discovery in equity options?," *Journal of Financial Economics*, 107(2), 259–283.

MURAVYEV, D., N. D. PEARSON, AND J. POLLET (2021): "Why Does Options Market Information Predict Stock Returns?," *Working paper*.

MURAVYEV, D., N. D. PEARSON, AND J. M. POLLET (2018): "Understanding returns to short selling using option-implied stock borrowing fees," *Available at SSRN 2851560*.

OFEK, E., M. RICHARDSON, AND R. F. WHITELAW (2004): "Limited arbitrage and short sales restrictions: Evidence from the options markets," *Journal of Financial Economics*, 74(2), 305–342.

PAN, J., AND A. M. POTESHMAN (2006): "The information in option volume for future stock prices," *The Review of Financial Studies*, 19(3), 871–908.

RAMACHANDRAN, L. S., AND J. TAYAL (2020): "Mispricing, short-sale constraints, and the cross-section of option returns," *Journal of Financial Economics (JFE), Forthcoming*.

RAPACH, D. E., J. K. STRAUSS, AND G. ZHOU (2010): "Out-of-sample equity premium prediction: Combination forecasts and links to the real economy," *The Review of Financial Studies*, 23(2), 821–862.

ROLL, R., E. SCHWARTZ, AND A. SUBRAHMANYAM (2009): "Options trading activity and firm valuation," *Journal of Financial Economics*, 94(3), 345–360.

——— (2010): "O/S: The relative trading activity in options and stock," *Journal of Financial Economics*, 96(1), 1–17.

SIMON, N., J. FRIEDMAN, T. HASTIE, AND R. TIBSHIRANI (2013): "A sparse-group LASSO," *Journal of Computational and Graphical Statistics*, 22(2), 231–245.

Electronic copy available at: https://ssrn.com/abstract=3747238

STOLL, H. R. (1978a): "The pricing of security dealer services: An empirical study of NASDAQ stocks," *The journal of finance*, 33(4), 1153–1172.

——— (1978b): "The supply of dealer services in securities markets," *The Journal of Finance*, 33(4), 1133–1151.

TIBSHIRANI, R. (1996): "Regression shrinkage and selection via the lasso," *Journal of the Royal Statistical Society: Series B (Methodological)*, 58(1), 267–288.

VASQUEZ, A. (2017): "Equity volatility term structures and the cross section of option returns," *Journal of Financial and Quantitative Analysis*, 52(6), 2727–2754.

XING, Y., X. ZHANG, AND R. ZHAO (2010): "What does the individual option volatility smirk tell us about future equity returns?," *Journal of Financial and Quantitative Analysis*, pp. 641–662.

YUAN, M., AND Y. LIN (2006): "Model selection and estimation in regression with grouped variables," *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 68(1), 49–67.

ZOU, H., AND T. HASTIE (2005): "Regularization and variable selection via the elastic net," *Journal of the Royal Statistical Society: Series B (statistical methodology)*, 67(2), 301–320.

Electronic copy available at: https://ssrn.com/abstract=3747238

APPENDIX A

No.	Acronym	Description
		Option Characteristics (Group 5)
1	avg_impl_volatility	Average ATM implied volatility
2	cvol	Change in ATM Call implied volatility (An, Ang, Bali, and Cakici (2014))
3	cw	CW (Cremers and Weinbaum (2012))
4	hviv	Historical volatility - Implied ATM volatility (Goyal and Saretto (2009))
5	med_hq	Option implied fee (Muravyev, Pearson, and Pollet (2018))
6	opt_baspread	Option average bid-ask spread
7	os	OS ratio (Johnson and So (2012))
8	pvol	Change in ATM Put implied volatility (An, Ang, Bali, and Cakici (2014))
9	skewness	Volatility surface skewness (Xing, Zhang, and Zhao (2010))
		Option-implied Risk-Neutral Moments (Group 6)
10	volq1	Risk-neutral volatility (1 month maturity)
11	volq2	Risk-neutral volatility (2 month maturity)
12	volq4	Risk-neutral volatility (4 month maturity)
13	volq5	Risk-neutral volatility (5 month maturity)
14	volq6	Risk-neutral volatility (6 month maturity)
15	skewq1	Risk-neutral skewness (1 month maturity)
16	skewq2	Risk-neutral skewness (2 month maturity)
17	skewq4	Risk-neutral skewness (4 month maturity)
18	skewq5	Risk-neutral skewness (5 month maturity)
19	skewq6	Risk-neutral skewness (6 month maturity)
20	kurtq1	Risk-neutral kurtosis (1 month maturity)
21	kurtq2	Risk-neutral kurtosis (2 month maturity)
22	kurtq4	Risk-neutral kurtosis (4 month maturity)
23	kurtq5	Risk-neutral kurtosis (5 month maturity)
24	kurtq6	Risk-neutral kurtosis (6 month maturity)

Table A.1: Option based characteristics: Variable Definitions

Notes: The table presents definitions of option based variables used as an input into machine learning algorithms. Groups numbering is referred to related variables' groups used in SGL (sparse group LASSO).

Electronic copy available at: https://ssrn.com/abstract=3747238

No.	Acronym	Description	Group	No.	Acronym	Description	Group
1	absacc	Absolute Accruals	4	48	mom36m	36-month momentum	1
2	acc	Working Capital accruals	4	49	mom6m	6-month momentum	1
3	aeavol	Abnormal earnings announcement volume	4	50	ms	Mohanram financial statement score	4
4	age	# years since first Compustat coverage	4	51	mve	Size	2
5	agr	Asset growth	4	52	mve_ia	Industry-adjusted size	2
6	baspread	Bid-ask spread	2	53	nincr	Number of earnings increases	4
7	beta	Beta	3	54	operprof	Operating profitability	4
8	betasq	Beta squared	3	55	orgcap	Organizational capital	4
9	bm	Book-to-market	4	56	pchcapx_ia	Industry adjusted percentage change in capital expenditures	4
10	bm_ia	Industry-adjusted book to market	4	57	pchcurrat	Percentage change in current ratio	4
11	cash	Cash holdings	4	58	pchdepr	Percentage change in depreciation	4
12	cashdebt	Cash flow to debt	4	59	pchgm_pchsale	Percentage change in gross margin less percentage change in sales	4
13	cashpr	Cash productivity	4	60	pchquick	Percentage change in quick ratio	4
14	cfp	Cash flow to price ratio	4	61	pchsale_pchinvt	Percentage change in sales less percentage change in inventory	4
15	cfp_ia	Industry-adjusted cash flow to price ratio	4	62	pchsale_pchrect	Percentage change in sales less percentage change in A/R	4
16	chatoia	Industry-adjusted change in asset turnover	4	63	pchsale_pchxsga	Percentage change in sales less percentage chagne in SG&A	4
17	chcsho	Change in shares outstanding	4	64	pchsaleinv	Percentage change in sales-to-inventory	4
18	chempia	Industry-adjusted change in employees	4	65	pctacc	Percent accruals	4
19	chinv	Change in inventory	4	66	pricedelay	Price delay	4
20	chmom	Change in 6-month momentum	1	67	ps	Piotroski financial statements score	4
21	chpmia	Industry-adjusted change in profit margin	4	68	quick	Quick ratio	4
22	chtx	Change in tax expense	4	69	rd	R&D increase	4
23	cinvest	Corporate investment	4	70	rd_mve	R&D to market capitalization	4
24	convind	Convertible debt indicator	4	71	rd_sale	R&D to sales	4
25	currat	Current ratio	4	72	realestate	Real estate holdings	4
26	depr	Deprecation/PP&E	4	73	retvol	Return volatility	3
27	divi	Dividend initiation	4	74	roaq	Return on assets	4
28	divo	Dividend omission	4	75	roavol	Earnings volatility	4
29	dolvol	Dollar trading volume	2	76	roeq	Return on equity	4
30	dy	Dividend to price	4	77	roic	Return on invested capital	4
31	ear	Earnings announcement return	4	78	rsup	Revenue surprise	4
32	egr	Growth in common shareholder equity	4	79	salecash	Sales to cash	4
33	ep	Earnings to price	4	80	saleinv	Sales to inventory	4
34	gma	Gross profitability	4	81	salerec	Sales to receivables	4
35	grcapx	Growth in capital expenditures	4	82	secured	Secured debt	4
36	grltnoa	Growth in long-term net operating assets	4	83	securedind	Secured debt indicator	4
37	herf	Industry sales concentration	4	84	sgr	Sales growth	4
38	hire	Employee growth rate	4	85	sin	Sin stocks	4
39	idiovol	Idiosyncratic return volatility	3	86	sp	Sales to price	4
40	ill	Amihud Illiquidity	2	87	std_dolvol	Volatility of liquidity (dollar trading volume)	2
41	indmom	Industry momentum	1	88	std_turn	Volatility of liquidity (share turnover)	2
42	invest	Capital expenditures and inventory	4	89	stdacc	Accrual volatility	4
43	lev	Leverage	4	90	stdcf	Cash flow volatility	4
44	lgr	Growth in long-term debt	4	91	tang	Debt capacity/firm tangibility	4
45	maxret	Maximum daily return	1	92	tb	Tax income to book income	4
46	mom12m	12-month momentum	1	93	turn	Share turnover	2
47	mom1m	1-month momentum	1	94	zerotrade	Zero trading days	2

Table A.2: Stock based characteristics: Variable Definitions

Notes: The table presents definitions of Green, Hand, and Zhang (2017) 94 stock based variables used as an input into machine learnin algorithms. Groups numbering is referred to related variables' groups used in SGL (sparse group LASSO). The detailed description of e variable is outlined in Green, Hand, and Zhang (2017).

Electronic copy available at: https://ssrn.com/abstract=3747238

Panel A. All Straddle Returns

	PCR	PLS	Lasso	Ridge	Elastic Net	SGL	RF	NN1	NN2	NN3	NN4	NN5	Combination1	Combination2
OLS	**2.30**	**2.31**	**2.31**	**2.31**	**2.31**	**2.31**	**2.30**	**2.31**	**2.31**	**2.30**	**2.30**	**2.28**	**2.31**	**2.31**
PCR		0.49	**2.67**	0.77	**2.66**	**2.37**	-0.28	-0.32	0.77	0.46	-0.58	-0.89	**2.79**	**1.97**
PLS			1.40	0.70	1.39	1.17	-0.51	-0.69	0.60	0.00	-1.17	-1.33	**1.74**	**1.66**
Lasso				-1.15	-1.25	-0.38	-1.59	-1.05	-0.60	-1.54	**-1.72**	**-2.52**	0.43	-0.07
Ridge					1.14	0.92	-0.65	-0.89	0.42	-0.39	**-1.73**	**-1.85**	1.53	**1.71**
Elastic Net						-0.34	-1.58	-1.05	-0.59	-1.53	**-1.72**	**-2.51**	0.49	-0.06
SGL							-1.42	-0.96	-0.47	-1.35	**-1.61**	**-2.37**	0.68	0.11
RF								-0.16	0.72	0.49	-0.28	-0.47	1.57	1.34
NN1									1.24	0.65	-0.12	-0.21	1.16	1.32
NN2										-0.94	**-2.87**	**-3.29**	0.77	1.12
NN3											-1.58	**-2.97**	**1.97**	**3.82**
NN4												-0.28	**1.99**	**2.73**
NN5													**2.90**	**4.77**
Combination1														-0.28

Panel B. S&P 500 Straddle Returns

	PCR	PLS	Lasso	Ridge	Elastic Net	SGL	RF	NN1	NN2	NN3	NN4	NN5	Combination1	Combination2
OLS	**2.67**	**2.67**	**2.66**	**2.66**	**2.66**	**2.67**	**2.65**	**2.67**	**2.65**	**2.64**	**2.65**	**2.62**	**2.66**	**2.66**
PCR		-0.57	-0.09	0.29	-0.10	-1.55	**-1.77**	-0.98	-1.31	-0.56	-1.47	**-2.65**	0.23	-0.15
PLS			0.48	0.84	0.48	-0.55	-1.24	-1.04	-1.19	-0.07	-1.09	**-2.45**	0.71	0.50
Lasso				0.53	-0.01	-1.20	**-1.67**	-0.93	-1.34	-0.72	-1.54	**-2.69**	0.95	-0.11
Ridge					-0.53	-1.39	**-1.72**	-1.11	**-1.77**	-0.99	**-2.31**	**-3.07**	-0.24	-0.72
Elastic Net						-1.20	**-1.67**	-0.93	-1.34	-0.71	-1.54	**-2.69**	0.96	-0.11
SGL							-1.34	-0.65	-0.68	0.26	-0.61	**-2.22**	1.56	0.97
RF								0.38	0.73	1.32	0.89	-0.49	**1.75**	1.58
NN1									0.37	0.66	0.39	-1.40	1.03	1.10
NN2										0.91	0.27	**-3.18**	1.56	**2.17**
NN3											-1.05	**-2.56**	0.91	0.67
NN4												**-2.79**	**1.87**	**2.98**
NN5													**2.87**	**3.54**
Combination1														-0.43

Table A.3: Comparison of monthly out-of-sample option returns prediction using Diebold-Mariano tests: Option and Stock Predictors

Notes: The table reports pairwise Diebold-Mariano test statistics comparing the out-of-sample stock-level prediction performance among sixteen models, and using both stock and option predictors. Positive numbers indicate the column model outperforms the row model. Bold font indicates the difference is significant at 5% level or better for individual tests. Panels A presents the results for all optionable stocks, and Panel B presents the corresponding statistics for S&P500 stocks.

Electronic copy available at: https://ssrn.com/abstract=3747238

Panel A. All Straddle Returns

	PCR	PLS	Lasso	Ridge	Elastic Net	SGL	RF	NN1	NN2	NN3	NN4	NN5	Combination1	Combination2
OLS	**2.31**	**2.35**	**2.34**	**2.33**	**2.34**	**2.32**	**2.32**	**2.32**	**2.32**	**2.31**	**2.29**	**2.26**	**2.34**	**2.34**
PCR		-0.20	**2.75**	1.43	**2.70**	**1.76**	1.12	-0.19	0.32	0.14	0.71	-0.36	**2.76**	**1.67**
PLS			1.55	1.05	1.52	1.04	0.94	-0.16	0.40	0.22	0.67	-0.26	1.51	1.44
Lasso				**-1.84**	**-1.78**	**-2.66**	-0.23	-0.66	-0.42	-0.75	-0.46	-1.32	-1.32	0.25
Ridge					**1.79**	0.67	0.66	-0.40	0.06	-0.21	0.35	-0.80	**1.89**	1.44
Elastic Net						**-2.57**	-0.21	-0.65	-0.41	-0.73	-0.44	-1.31	-1.14	0.28
SGL							0.42	-0.47	-0.11	-0.39	0.09	-0.95	**2.80**	1.05
RF								-0.61	-0.32	-0.68	-0.30	-1.54	0.08	0.49
NN1									0.88	0.48	0.62	-0.01	0.62	0.92
NN2										-0.82	0.26	-1.15	0.35	0.84
NN3											0.79	-0.97	0.68	1.56
NN4												**-1.91**	0.33	1.48
NN5													1.27	**2.54**
Combination1														0.47

Panel B. S&P 500 Straddle Returns

	PCR	PLS	Lasso	Ridge	Elastic Net	SGL	RF	NN1	NN2	NN3	NN4	NN5	Combination1	Combination2
OLS	**2.56**	**2.57**	**2.56**	**2.56**	**2.56**	**2.58**	**2.55**	**2.54**	**2.54**	**2.56**	**2.51**	**2.44**	**2.57**	**2.57**
PCR		1.26	0.27	1.43	0.29	0.44	-0.42	-0.46	-0.43	-1.51	-1.34	-1.54	1.12	0.47
PLS			-0.41	0.47	-0.40	-0.18	-0.60	-0.61	-0.64	**-1.70**	-1.56	**-1.67**	0.25	0.08
Lasso				1.52	1.08	0.24	-0.55	-0.54	-0.54	-1.58	-1.59	**-1.69**	**2.60**	0.43
Ridge					-1.51	-0.50	-0.78	-0.70	-0.76	**-1.75**	**-1.73**	**-1.79**	-0.33	-0.15
Elastic Net						0.22	-0.55	-0.54	-0.55	-1.58	**-1.60**	**-1.70**	**2.63**	0.42
SGL							-0.59	-0.64	-0.66	**-1.88**	**-1.67**	**-1.81**	0.53	0.24
RF								-0.15	-0.04	-0.92	-0.85	-1.38	0.76	0.82
NN1									0.37	-0.63	-0.63	**-2.95**	0.70	0.91
NN2										-1.07	-1.09	**-2.92**	0.76	1.17
NN3											0.09	-0.88	**1.82**	**2.09**
NN4												-1.03	**1.81**	**2.52**
NN5													**1.83**	**2.42**
Combination1														-0.06

Table A.4: Comparison of monthly out-of-sample option returns prediction using Diebold-Mariano tests: Option Predictors only

Notes: The table reports pairwise Diebold-Mariano test statistics comparing the out-of-sample stock-level prediction performance among sixteen models, and using only option predictors. Positive numbers indicate the column model outperforms the row model. Bold font indicates the difference is significant at 5% level or better for individual tests. Panels A presents the results for all optionable stocks, and Panels B present the corresponding statistics for S&P500 stocks.

Electronic copy available at: https://ssrn.com/abstract=3747238

Panel A. All Straddle Returns

	PCR	PLS	Lasso	Ridge	Elastic Net	SGL	RF	NN1	NN2	NN3	NN4	NN5	Combination1	Combination2
OLS	**2.28**	**2.29**	**2.28**	**2.28**	**2.28**	**2.29**	**2.27**	**2.29**	**2.28**	**2.27**	**2.27**	**2.26**	**2.28**	**2.28**
PCR		0.13	0.25	0.19	0.25	-1.32	-0.85	-0.70	-0.25	-1.03	-1.02	-1.30	0.38	0.53
PLS			0.02	0.05	0.02	-1.39	-0.66	-1.00	-0.48	-1.05	-0.96	**-1.76**	0.10	0.40
Lasso				0.01	-0.13	-1.21	-1.24	-0.80	-0.42	-1.44	-1.47	**-1.71**	0.21	0.53
Ridge					-0.02	-1.25	-0.94	-0.91	-0.59	-1.52	-1.46	**-2.47**	0.13	0.77
Elastic Net						-1.21	-1.24	-0.79	-0.42	-1.45	-1.49	**-1.71**	0.23	0.55
SGL							0.48	-0.19	0.59	0.09	0.26	-0.31	1.52	1.43
RF								-0.40	0.23	-0.48	-0.34	-0.91	1.31	1.47
NN1									0.92	0.26	0.33	-0.06	0.89	1.12
NN2										-0.90	-0.53	**-1.68**	0.53	1.08
NN3											0.41	-0.77	1.57	**2.79**
NN4												-1.00	1.63	**2.76**
NN5													**1.99**	**3.32**
Combination1														0.55

Panel B. S&P 500 Straddle Returns

	PCR	PLS	Lasso	Ridge	Elastic Net	SGL	RF	NN1	NN2	NN3	NN4	NN5	Combination1	Combination2
OLS	**2.58**	**2.60**	**2.58**	**2.58**	**2.58**	**2.57**	**2.57**	**2.61**	**2.58**	**2.56**	**2.57**	**2.54**	**2.58**	**2.58**
PCR		-0.37	-0.42	-0.15	-0.42	-0.41	**-1.83**	-0.94	-1.09	-1.72	0.11	**-2.55**	0.04	-0.09
PLS			0.23	0.33	0.23	0.24	-0.40	-1.34	-0.34	-0.53	0.41	**-2.96**	0.38	0.44
Lasso				0.34	-0.09	0.08	-1.35	-0.89	-0.84	-1.39	0.38	**-2.48**	1.08	0.26
Ridge					-0.34	-0.54	-1.49	-0.92	-0.93	**-2.01**	0.22	**-2.66**	0.36	0.02
Elastic Net						0.08	-1.35	-0.89	-0.83	-1.40	0.38	**-2.47**	1.09	0.26
SGL							-1.39	-0.86	-0.80	-1.64	0.40	**-2.47**	1.03	0.21
RF								-0.56	0.25	-0.19	1.55	**-2.03**	**1.69**	1.46
NN1									0.75	0.55	0.98	-1.47	0.95	1.09
NN2										-0.47	1.15	**-2.33**	1.05	1.27
NN3											**2.08**	**-2.36**	**1.91**	**2.88**
NN4												**-2.40**	-0.10	-0.21
NN5													**2.60**	**3.06**
Combination1														-0.15

Table A.5: Comparison of monthly out-of-sample option returns prediction using Diebold-Mariano tests: Stock Predictors only

Notes: The table reports pairwise Diebold-Mariano test statistics comparing the out-of-sample stock-level prediction performance among sixteen models, and using only stock predictors. Positive numbers indicate the column model outperforms the row model. Bold font indicates the difference is significant at 5% level or better for individual tests. Panels A presents the results for all optionable stocks, and Panels C presents the corresponding statistics for S&P500 stocks.

Electronic copy available at: https://ssrn.com/abstract=3747238

Panel A. All Variables

	PCR	PLS	Lasso	Ridge	Elastic Net	SGL	RF	NN1	NN2	NN3	NN4	NN5	Combination1	Combination2
OLS	**2.03**	**2.00**	**2.01**	**2.01**	**2.00**	**2.01**	**2.01**	**2.04**	**2.01**	**2.01**	**1.99**	**1.98**	**2.01**	**2.01**
PCR		1.04	1.03	1.20	0.87	1.22	1.09	-0.75	0.44	0.90	0.58	0.54	1.15	1.24
PLS			-0.61	1.13	-1.08	1.16	0.32	-1.08	-1.24	-0.68	-0.60	-0.41	0.51	0.21
Lasso				**1.83**	**-1.84**	**1.93**	0.89	-1.04	-1.30	-0.41	-0.37	-0.22	**2.01**	1.04
Ridge					**-2.21**	0.39	-0.50	-1.16	-1.57	-1.46	-1.17	-0.82	-1.37	-0.67
Elastic Net						**2.25**	1.19	-0.94	-1.00	0.04	-0.13	-0.04	**2.50**	1.37
SGL							-0.63	-1.18	-1.62	-1.57	-1.23	-0.86	-1.60	-0.79
RF								-1.10	-1.48	-1.36	-1.10	-0.70	-0.01	-0.19
NN1									0.77	0.99	0.75	0.69	1.13	1.20
NN2										1.40	0.67	0.56	1.51	**1.93**
NN3											-0.23	-0.07	1.21	**2.41**
NN4												0.16	0.90	0.93
NN5													0.61	0.56
Combination1														-0.21

Panel B. Option Variables

	PCR	PLS	Lasso	Ridge	Elastic Net	SGL	RF	NN1	NN2	NN3	NN4	NN5	Combination1	Combination2
OLS	**1.99**	**2.00**	**2.01**	**2.01**	**2.01**	**2.00**	**1.98**	**2.11**	**2.01**	**2.00**	**2.02**	**1.95**	**2.01**	**2.02**
PCR		-0.72	-0.83	**1.66**	-0.80	1.09	-1.26	-1.45	-1.45	-1.53	-1.57	-1.45	0.11	-0.92
PLS			-0.30	**1.78**	-0.26	1.24	-0.29	-1.49	-1.60	**-1.67**	-1.63	-0.84	1.13	-0.83
Lasso				**1.77**	**1.90**	1.35	-0.06	-1.50	-1.63	**-1.76**	**-1.72**	-0.66	**1.70**	-0.81
Ridge					**-1.75**	-1.62	**-2.38**	-1.53	**-1.74**	**-1.86**	**-1.75**	**-1.95**	**-1.83**	-1.48
Elastic Net						1.33	-0.09	-1.50	-1.63	**-1.77**	**-1.73**	-0.68	**1.67**	-0.85
SGL							**-1.99**	-1.49	-1.62	**-1.72**	**-1.67**	**-1.78**	-0.77	-1.23
RF								-1.40	-1.31	-1.35	-1.49	-0.74	1.63	-0.42
NN1									1.41	1.38	1.23	1.26	1.51	1.53
NN2										0.65	-1.40	0.88	**1.69**	**1.91**
NN3											-1.48	0.85	**1.81**	**2.21**
NN4												1.19	**1.73**	**1.82**
NN5													1.49	0.25
Combination1														-1.31

Panel C. Stock Variables

	PCR	PLS	Lasso	Ridge	Elastic Net	SGL	RF	NN1	NN2	NN3	NN4	NN5	Combination1	Combination2
OLS	**2.01**	**1.97**	**1.98**	**1.98**	**1.98**	**1.98**	**1.98**	**1.98**	**1.96**	**1.97**	**1.95**	**1.95**	**1.98**	**1.98**
PCR		0.79	1.03	1.18	0.89	1.10	1.12	**-2.64**	0.76	0.84	0.49	0.51	1.12	1.06
PLS			0.64	1.63	0.17	1.54	1.25	**-1.67**	0.18	0.13	-0.25	-0.17	1.27	0.63
Lasso				**1.88**	**-1.91**	1.24	1.13	**-1.87**	-0.26	-0.41	-0.62	-0.52	**1.90**	0.14
Ridge					**-2.30**	-1.22	0.03	**-1.91**	-1.21	-1.40	-1.24	-1.22	**-1.71**	-1.36
Elastic Net						**1.80**	1.43	**-1.77**	0.07	0.01	-0.38	-0.28	**2.54**	0.70
SGL							0.43	**-1.85**	-0.91	-1.06	-1.04	-1.02	-0.51	-0.85
RF								**-1.88**	-1.41	-1.59	-1.35	-1.45	-0.62	-1.17
NN1									**1.66**	**1.80**	1.47	1.43	**1.89**	**1.93**
NN2										-0.18	-0.60	-0.63	0.74	0.47
NN3											-0.50	-0.42	0.96	0.94
NN4												0.22	0.95	0.82
NN5													0.88	0.70
Combination1														-0.76

Table A.6: Comparison of monthly out-of-sample stock returns prediction using Diebold-Mariano tests: All Stocks

Notes: The table reports pairwise Diebold-Mariano test statistics comparing the out-of-sample stock-level prediction performance among sixteen models for all optionable stocks in our sample. Positive numbers indicate the column model outperforms the row model. Bold font indicates the difference is significant at 5% level or better for individual tests. Panels A, B and C refer to the forecast based on all stock and options predictors, *All Variables*, only option-based predictors, *Option Variables*, and only stocks predictors, *Stock Variables*, respectively in Table 5.

Electronic copy available at: https://ssrn.com/abstract=3747238

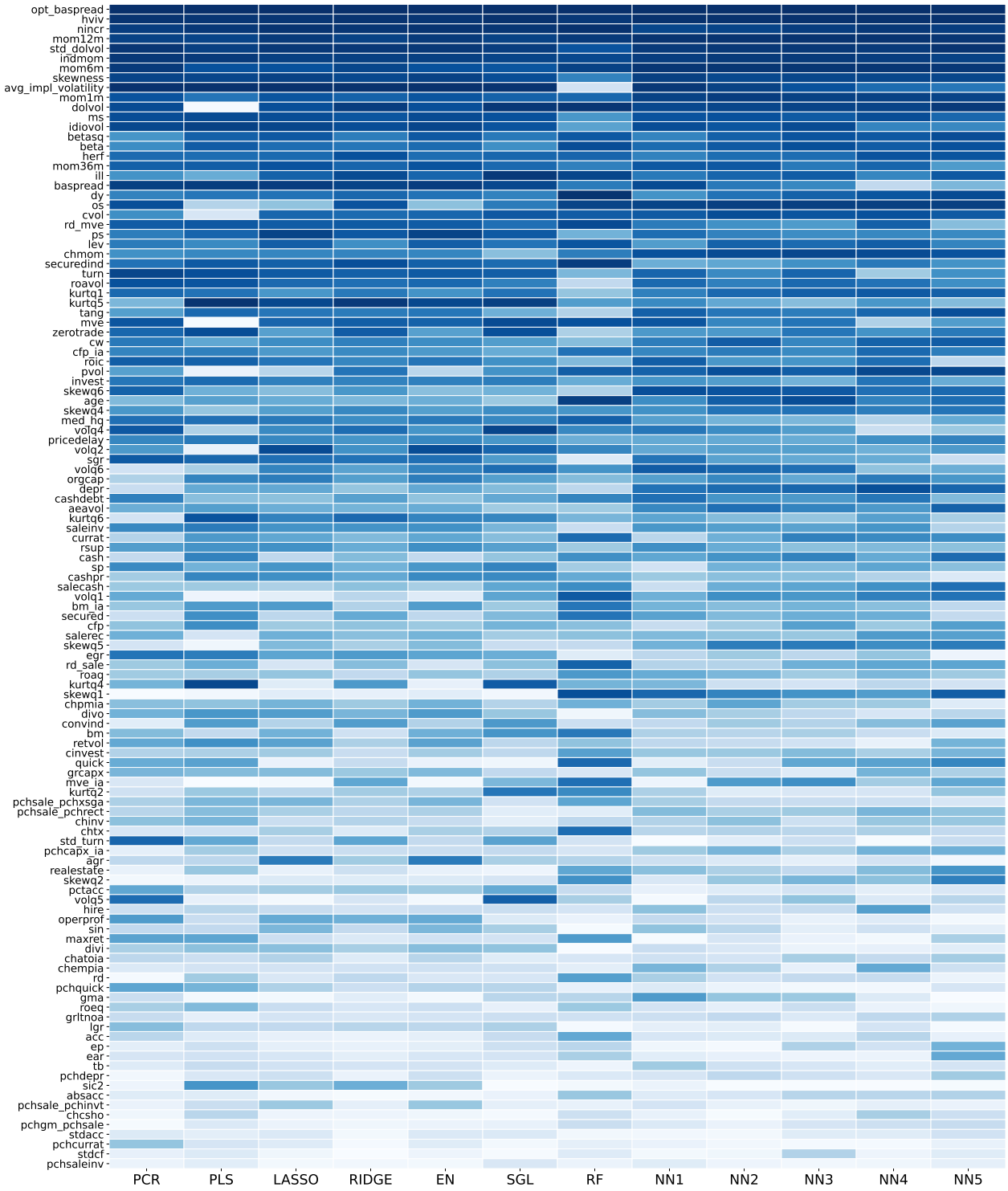

Figure A.1: Characteristic importance: delta-neutral straddle returns, all stocks

Notes: Rankings of ninety-four stock-level characteristics, the industry dummy (sic2), and 24 option characteristics in terms of overall model contribution for predicting delta-neutral straddle returns of all optionable stocks. Characteristics are ordered based on the sum of their ranks over all models, with the most influential characteristics on the top and the least influential on the bottom. Columns correspond to the individual models, and the color gradients within each column indicate the most influential (dark blue) to the least influential (light/white) variables.

Electronic copy available at: https://ssrn.com/abstract=3747238

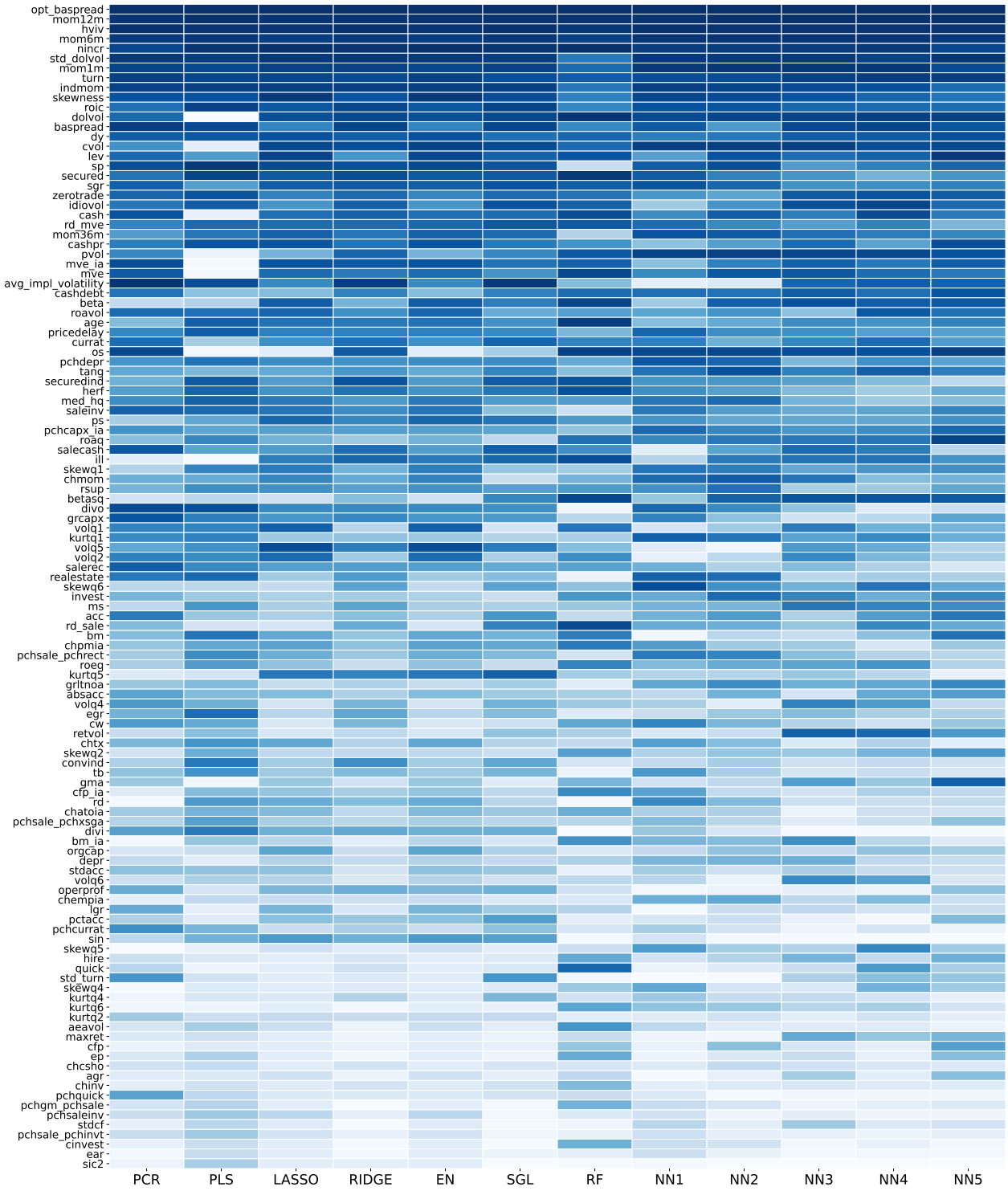

Figure A.2: Characteristic importance: delta-neutral straddle returns, S&P500 stocks

Notes: Rankings of ninety-four stock-level characteristics, the industry dummy (sic2), and 24 option characteristics in terms of overall model contribution for predicting delta-neutral straddle returns of all optionable stocks. Characteristics are ordered based on the sum of their ranks over all models, with the most influential characteristics on the top and the least influential on the bottom. Columns correspond to the individual models, and the color gradients within each column indicate the most influential (dark blue) to the least influential (light/white) variables.

Electronic copy available at: https://ssrn.com/abstract=3747238

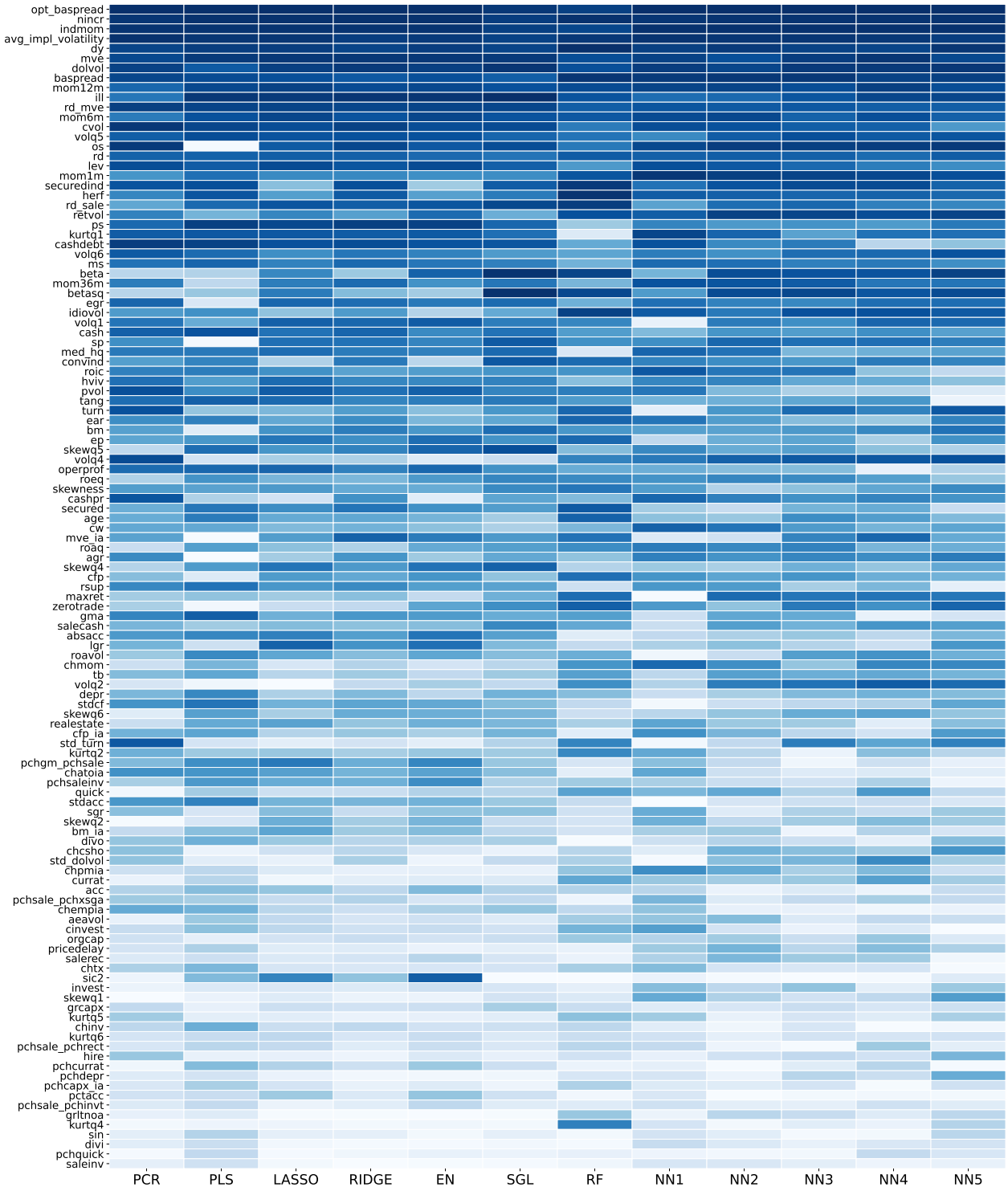

Figure A.3: Characteristic importance: stock excess returns, all stocks

Notes: Rankings of ninety-four stock-level characteristics, the industry dummy (sic2), and 24 option characteristics in terms of overall model contribution for predicting excess returns of all optionable stocks. Characteristics are ordered based on the sum of their ranks over all models, with the most influential characteristics on the top and the least influential on the bottom. Columns correspond to the individual models, and the color gradients within each column indicate the most influential (dark blue) to the least influential (light/white) variables.

Electronic copy available at: https://ssrn.com/abstract=3747238

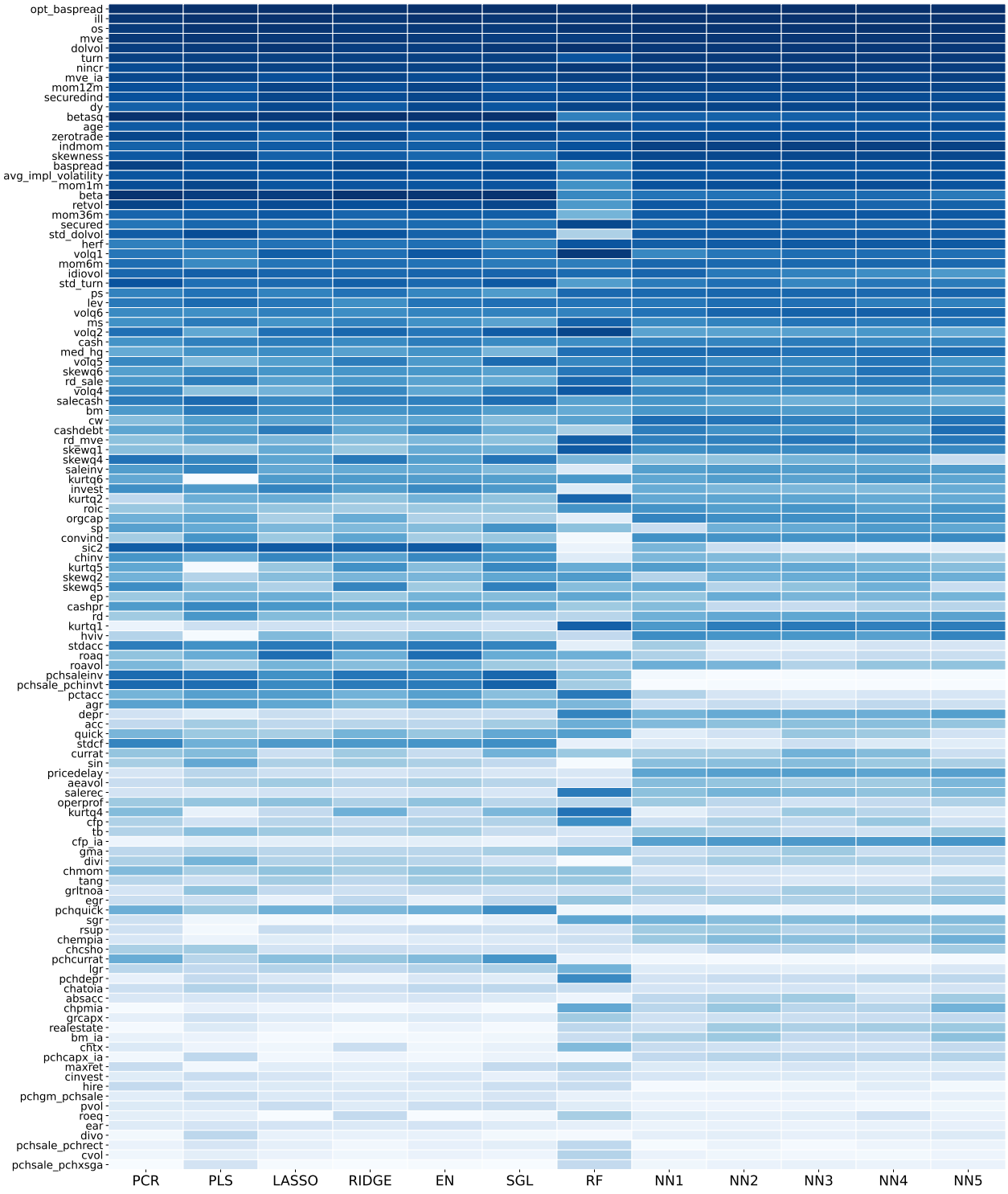

Figure A.4: Characteristic importance: log option illiquidity, all stocks

Notes: Rankings of ninety-four stock-level characteristics, the industry dummy (sic2), and 24 option characteristics in terms of overall model contribution for predicting option illiquidity for all optionable stocks in our sample.

Electronic copy available at: https://ssrn.com/abstract=3747238

APPENDIX B: Machine Learning Methods

Penalized Linear Models The simple linear model imposes that conditional expectations of $g^{\star}(\cdot)$ can be approximated by a linear function of raw predictors and the parameter vector θ,

$$g\left(z_{i,t};\theta\right)=z_{i,t}'\theta \tag{B.0.1}$$

The base line linear model uses a standard least squares objective function:

$$\mathcal{L}(\theta)=\frac{1}{NT}\sum_{i=1}^{N}\sum_{t=1}^{T}\left(r_{i,t+1}-g\left(z_{i,t};\theta\right)\right)^{2} \tag{B.0.2}$$

Penalized methods generally differ by a penalty appended to the original loss function such as:

$$\mathcal{L}(\theta;\cdot)=\mathcal{L}(\theta)+\phi(\theta;|\cdot) \tag{B.0.3}$$

A prominent example is the Elastic Net penalty which is of the form:

$$\phi(\theta;\lambda,\rho)=\lambda(1-\rho)\sum_{j=1}^{P}|\theta_j|+\frac{1}{2}\lambda\rho\sum_{j=1}^{P}\theta_j^2\equiv\lambda(1-\rho)\|\theta\|_1+\frac{1}{2}\lambda\rho\|\theta\|_2$$

Elastic Net convexly combines l_1 regularization and l_2 regularization of the regression coefficients. The ratio between l_1 regularization and l_2 regularization is controlled by ρ. When $\rho=0$, there is only l_2 penalty and EN is reduced to LASSO, and when $\rho=1$, there is only l_1 penalty and EN is reduced to Ridge.[20]

The LASSO estimator offers an appealing convex relaxation of a difficult non-convex best subset selection problem. By construction, it produces sparse parsimonious models zeroing-out a large number of the estimated parameters. The model selection is not free and comes at a price that can be high in the low signal-to-noise environment with heavy-tailed dependent data. However, the LASSO does not recognize that covariates at different lags are temporally related. It also does not allow to assemble all lag dependent variables into a group. Other group structures could be considered, for instance combining various covariates into a single group, as we do below in the empirical section. The sparse-group LASSO (SGL) (Simon, Friedman, Hastie, and Tibshirani (2013)) , allows us to incorporate such structure into the estimation procedure. In contrast to the group LASSO, see Yuan and Lin (2006), the SGL promotes sparsity between

[20] To minimize the potential over-fitting, we set $\rho=0.5$ and optimize λ from $\{10^{-3},10^{-2.9},10^{-2.8},\ldots,10^{3.8},10^{3.9},10^{4}\}$. In LASSO, the l_1 regularization strength is selected from $\{10^{-4},10^{-3.9},10^{-3.8},\ldots,10^{3.8},10^{3.9},10^{4}\}$; in Ridge, the l_2 regularization strength is selected from $\{10^{-1},10^{-0.9},10^{-0.8},\ldots,10^{7.8},10^{7.9},10^{8}\}$.

Electronic copy available at: https://ssrn.com/abstract=3747238

and within groups, and allows us to capture the predictive information from each group, such as approximating functions form the specific covariates from each group. With K_g groups and the parameters pertaining a specific member k as θ^k, the penalty function is:

$$\phi(\theta;\alpha) = \alpha\|\theta\|_1 + (1-\alpha)\sum_{k=0}^{K_g}\|\theta^k\|_2 \tag{B.0.4}$$

Setting $\alpha = 1$, we obtain the LASSO estimator while setting $\alpha = 0$ leads to the group LASSO estimator of Yuan and Lin (2006). Note also that with a single group (K_g = 1), the penalty resembles the Elastic Net penalty. Therefore, with a single group, the SGL achieves similar to the Elastic Net regularization goals. In practice, groups are defined by a particular problem, while α can be fixed or selected in a data-driven way.[21]

Dimension reduction: PCR and PLS Based on Equations 2.1-2.3, the excess return can be written as:

$$r_{i,t+1} = z'_{i,t}\theta + \varepsilon_{i,t+1} \tag{B.0.5}$$

Using matrix notation it can be reorganized as

$$R = Z\theta + E \tag{B.0.6}$$

where R is the $NT \times 1$ vector of $r_{i,t+1}$, Z is the $NT \times P$ matrix of stacked predictors $z_{i,t}$, and E is a $NT \times 1$ vector of residuals $\epsilon_{i,t+1}$. There are two popular dimension reduction techniques: principal components regression (PCR) and partial least squares (PLS). They both aim to reduce the set of predictors from dimension P to a much smaller number K of linear combinations of predictors. The general form of the forecasting model is

$$R = (Z\Omega_K)\,\theta_K + \tilde{E} \tag{B.0.7}$$

Ω_K is $P \times K$ matrix with columns $w_1, w_2, \ldots, w_K$. PCR chooses the linear combination of weights w_j recursively by solving:

$$w_j = \arg\max_w \mathrm{Var}(Zw), \quad \text{s.t. } w'w = 1, \quad \mathrm{Cov}\,(Zw, Zw_l) = 0, \quad l = 1, 2, \ldots, j-1 \tag{B.0.8}$$

Thus PCR chooses the K linear combinations of Z that approximate the best the full predictor set. In contrast, PLS seeks for the best K linear combinations of Z that have the highest correlation

[21] We first fit the models with $\alpha \in \{0.05, 0.15, 0.25, \ldots, 0.85, 0.95\}$, taking a variety of convex combination of LASSO and sparse group LASSO into consideration. We also experiment with different lag structures and find that our conclusions are not altered with inclusions of lags.

Electronic copy available at: https://ssrn.com/abstract=3747238

with the forecast target by solving:

$$w_j = \arg\max_w \text{Cov}^2(R, Zw), \quad \text{s.t.} \quad w'w = 1, \quad \text{Cov}\left(Zw, Zw_l\right) = 0, \quad l = 1, 2, \ldots, j-1 \tag{B.0.9}$$

Random Forests Unlike linear models, trees are fully non-parametric. They are designed to find groups of observations that behave similarly to each other. Once a group is determined, a tree starts "growing" in a sequence of "branches". Each new branch, or partition, sorts the data leftover from the preceding step into bins based on one of the predictor variables. The specific predictor variable upon which a branch is based, and its value where the branch is split, are chosen via optimization to minimize forecast error. Each tree can be classified as having K_n terminal nodes (called "leaves") with a depth of K_d. The prediction of a given tree then can be stated as within-leave averages of predictors. A set of branches for a given partition can be represented as a product of indicators for sequential branches. The loss associated with the forecast error for a given branch is called "impurity". It describes how similarly observations behave on each specific side of the split. We use the most popular ℓ_2 impurity for each branch of the tree. We employ the standard Breiman (2001) algorithm to estimate random forest models which consist of 300 trees. The depth of each tree, and the random subset of predictors, or features, that are considered at each potential split within a tree are the key tuning parameters. The number of features is allowed to vary from $2, 4, 8, 16, 32 \ldots$, to the max number of predictors, and the depth is between 1 and 6. The random forest prediction is then the bootstrapped average at any prediction point across trees.

Neural Networks Our final non-linear method is traditional feed-forward Neural Network (NN). These consist of an "input layer" of raw predictors, one or more "hidden layers" that interact and nonlinearly transform the predictors, and an "output layer" that aggregates hidden layers into an ultimate outcome prediction. The number of units in the input layer is equal to the dimension of the predictors, and each of the predictor signals is amplified or attenuated according to the parameter vector, θ. For example, each neuron n in the first hidden layer transforms inputs into an output as, $x_n^{(1)} = f(\theta_{n,0}^{(0)} + \sum_{j=1}^{P} z_j \theta_{n,j}^{(0)})$. The final prediction $g(z;\theta)$ is typically a linear combination of the last hidden layer output. We omit all the details here, see for example Goodfellow, Bengio, and Courville (2016) for a comprehensive textbook treatment. To structure a neural network one has to makes choices about the number of hidden layers, or the number of neurons in each layer. Similar to Gu, Kelly, and Xiu (2020) we consider architectures with up to five hidden layers. We consider first the shallowest neural network, NN1, with a single hidden layer of 32 neurons. Further, NN2 with 2 hidden layers and 32 and 16 neurons respectively; NN3 with three hidden layers, and 32, 16 and 8 neurons respectively; NN4 with four hidden layers, and 32, 16, 8 and 4 neurons respectively; and finally NN5 with five hidden layers, and 32, 16, 8, 4 and 2 neurons respectively.

Electronic copy available at: https://ssrn.com/abstract=3747238

For each layer, the kernel is initialized with standard random uniform distribution and the bias is initialized with the value 0. We apply batch normalization on the hidden layer to even the variability and magnitude of each unit. The neurons of the hidden layer then go through an activation function. We use the same activation function at all nodes, and choose a popular functional form in recent literature known as the rectified linear unit (ReLU). The neural network is optimized by the "Adam" algorithm with learning rate shrinking from 0.01 to 0.0001, batch size of 10,000, a total of 100 epochs, and ensemble of 10. Following Gu, Kelly, and Xiu (2020), we also adopt the "early stopping" algorithm that ends the optimization process when validation sample errors show no improvements for 5 consecutive epochs.[22] It is used alongside ℓ_1 penalization on the hidden layer to shrink the kernel coefficients, and λ of 0.0001.

[22] We are grateful to Dacheng Xiu for his guidance in tuning the NN algorithm

Electronic copy available at: https://ssrn.com/abstract=3747238