Master's Thesis - Quantitative Finance

Predicting option implied volatility features using machine learning models: A comparative study with traditional implied volatility models

Author:
Maud van Lent

Student ID:
483143

Supervisor:
Gustavo Freire

Second Assessor:
Alberto Quaini

June, 2023

Abstract

This paper investigates the predictability of shape features of option implied volatility surfaces (IVS) through a comparative analysis of traditional econometric and machine learning models. The study utilises monthly implied volatility surface data of US equity options (NYSE, AMEX, and NASDAQ), spanning from 1996 to 2021, and includes 94 firm characteristics as predictive features. Three main IVS shape features are explored: level, slope, and curvature. The predictive performance of models is examined using an out-of-sample R^2 measure, and variable importance is analysed to gain insight into the factors driving the predictions of the best-performing models. The results demonstrate that machine learning models, such as Extremely Randomised Trees, Gradient Boosted Regression Trees, and Neural Networks, outperform the traditional Black-Scholes model in predicting IV shape features. Additionally, ensemble techniques exhibit superior performance, providing valuable insights into this relatively unexplored area of IV feature predictability. The research contributes to the field of option implied volatility prediction, offers novel insights into individual equity option IV shape predictability, and advances the understanding of machine learning applications in financial markets.

Contents

List of Tables

List of Figures

1 Introduction

The option markets have experienced significant growth in recent years, with options trading on exchanges worldwide increasing by more than 125% between 2013 and 2020, from $9.42 to $21.22 billion contracts (Bali et al., 2021a). These numbers show an increase in the popularity of option trading by investors. Therefore, forecasting option prices is important to obtain a complete understanding of future trends in the financial market. Often option prices are represented by the implied volatility surface (IVS), which is the implied volatility of an option as a function of the corresponding moneyness and time to maturity (Almeida et al., 2022). The implied volatility (IV) of an option is the level of volatility that is expected by the market based on the option price and is therefore referred to as the market's volatility forecast (Liu et al., 2021). The option implied volatility is computed from the Black-Scholes (BS) model introduced by Black and Scholes, 1973. When all other option parameters in the option pricing formula are known, there is a one-to-one relationship between option prices and the underlying expected asset volatility. This one-to-one mapping is useful as implied volatilities of different options are easier to compare than individual option prices, both in the cross-section and across time. This can be understood by considering that mapping option prices to implied volatilities of options with different characteristics, such as different strike prices, expiry dates, and underlying assets, allows for a fair comparison (Almeida et al., 2022; Audrino and Colangelo, 2010; Zeng and Klabjan, 2019). Therefore, this paper further investigates the predictability of option implied volatility.

Implied volatility holds significant importance in research on derivatives and in the field of risk management (Chen et al., 2023a; Muzzioli, 2010). It not only provides insights into the market price of the underlying asset's risk (Chang et al., 2012), but also comprises the additional compensation investors' require for taking on higher moment risks, such as volatility (Bali et al., 2019; Carr and Wu, 2009; Elyasiani et al., 2020), skewness (Chang et al., 2013; Langlois, 2020; Sasaki, 2016), kurtosis (Diavatopoulos et al., 2012; Dörries, 2021) and variance-of-variance risk premiums (Kaeck, 2018). Apart from being essential in risk management, the most valuable indicators of future market volatility can be derived from option IV. The implied volatility surface's shape provides insight into the risk-neutral distribution of underlying asset returns, enabling a more precise identification of market expectations regarding future price movements (Almeida et al., 2022; Zeng and Klabjan, 2019). Hence, institutional investors rely on implied volatility to determine their option positions (Hull and White, 2017), which is used alongside implied volatility to develop option pricing frameworks (Carr and Wu, 2016). The accurate fitting and prediction of implied volatility is of utmost importance to financial professionals, academics, and traders, emphasising the significance of research in this area.

A vast amount of research has been done on the predictability of IVS (Goncalves and Guidolin, 2006;

Mayhew, 1995). Several features of the IVS as well as stylised facts of financial data have been found to deteriorate the predictions. As discussed by Varian, 2014, Gu et al., 2020 and Bali et al., 2021a, the nonlinearities, complex interactions, and vast set of available explanatory variables in IV data establish it as a suitable choice for the application of machine learning models. Tang et al., 2022 document the transformation in the finance industry caused by the increase in the use of machine learning models. However, this stream of academic literature has addressed little attention to implied volatility forecasting (Christensen et al., 2021).

Therefore, in this research, we want to answer the main research question 'Can machine learning models improve the accuracy of option implied volatility forecasts compared to traditional econometric models?'. To answer this question we perform an extensive comparative analysis on the performance of various traditional (linear) and machine learning methods, in line with Gu et al., 2020, for the prediction of three main implied volatility shape features. To explore which models best capture various aspects of the IVS, we are, to the best of our knowledge, the first to examine the predictive performance of various machine learning models on IV level, slope and curvature. The level shape feature of the IVS denotes the overall magnitude of the volatility. High levels indicate a larger volatility, conversely, lower levels denote relatively stable future option prices. The slope and curvature features contribute to the volatility 'smile'. The slope denotes changes in implied volatility with respect to either varying time to maturity or strike prices, this paper examines the latter. A steep IVS slope indicates a higher expected volatility for out-of-the money options, with strike prices far from the current underlying asset price. A positive curvature of the IVS slope indicates an increase in implied volatility for both deep in-the-money and deep out-of-the-money options relative to at-the-money options.

In contrast to previous studies in the field of implied volatility forecasting, in our paper, we employ a wide range of techniques i.e. a Benchmark model (Black-Scholes), linear models (OLS, OLS-3, OLS-30), regularisation (Elastic net) and tree-based models (Random forest, Extremely randomised forest, Boosted trees). Analysing multiple model configurations allows for fair comparison. Starting with the benchmark models, we incorporate an IV feature prediction model based on the Black and Scholes, 1973 (BS) model assumptions. The BS model or variations of this method are often incorporated in research as a benchmark model for predicting implied volatility (Almeida et al., 2022; Audrino and Colangelo, 2010; Bennell and Sutcliffe, 2004; Ewing, 2010; Freire and Kleen, 2023; Isengildina-Massa et al., 2007; Li, 2005; Poon and Granger, 2003; Zulfiqar and Gulzar, 2021). Apart from IV feature predictions based on a popular traditional model, we incorporate two linear multi-factor models, in line with Gu et al., 2020. For the first model, we perform an ordinary least squares (OLS) optimisation based on all predictive features incorporated in the research. The second model contains a selection of 3 features (market beta, bid-ask spread, book-to-market ratio)

that are found to hold substantial predictive power in previous research on implied volatility (Chen et al., 2023a; Christoffersen et al., 2018; Freire and Kleen, 2023; Geske and Zhou, 2009). In addition to the linear regression models the research incorporates a regularisation technique, the elastic net. This incorporates both an L1 penalty term, which ensures variable selection, and an L2 penalty term, which makes it suitable for working with multicollinearity. To explore the predictive power of machine learning models in the field of IVS forecasting, we make use of various methods that have produced promising results in similar fields of research. The first forest model examined in this paper is the Random Forest model (RF). Due to its robustness to outliers and noise in combination with resistance to overfitting, this model has shown to attain a high predictive accuracy in a wide range of financial applications (Christensen et al., 2021; Gu et al., 2020). A diversification of the RF algorithm is the Extremely Randomised Forest model (Extra Trees). The Extra Trees algorithm shares similarities with the RF algorithm but also offers some potential advantages over it such as a higher diversity of trees, reduced variance and faster training. Furthermore, two extensions on the Gradient Boosting model framework (GBM) are incorporated. In GBM, trees are sequentailly build, therefore, the model is able to capture complex relationships and interactions between features and produced a lower bias compared to random forest models. Firstly, Dropout Additive Regression Trees (Dart) are introduced. Dart enhances the generalisation ability of the ensemble model by incorporating dropout regularisation, which prevents overfitting, therewith, reducing the variance. Secondly, the Extreme Gradient Boosting model (XGBoost) is used because of its improved speed and efficiency. Furthermore, XGBoost incorporates regularisation techniques to prevent overfitting and enhance generalisation. The GBM methods provide favourable results in the field of IVS forecasting (Audrino and Colangelo, 2010; Vrontos et al., 2021). The last model incorporated in this research, is a simple Feedforward Neural Network (NN). The model's architecture is different from the forest models. The interconnected network of nodes allows the model to more effectively capture complex relationships and patterns in the data.

The predictive performance of the models is assessed using an out-of-sample R^2-statistic, in line with Bali et al., 2021a; Chen et al., 2023a; Félix et al., 2020 and Gu et al., 2020. Additionally, the significance of the difference between the predictive performance of various models is examined using a Diebold and Mariano, 2002 test statistic.

Moreover, as a sub-question to what models best predicts the IVS features, we intend to uncover what features drive their predictive performance. Hereto, we investigate the significance of individual features to determine the factors that contribute to the effectiveness of the best-performing models. With this approach, we strive to open the 'black box' and obtain insight into what features drive the IVS movements.

The second part of our research revolves around the research question 'Can an optimal combination of

models outperform individual models in predicting implied volatility features?'. To address this question we employ two ways of combining models (stacking and equally weighted) and suggest three different model ensembles. Ensembles have shown to improve upon the use of individual models both in accuracy and robustness (Clements and Vasnev, 2021; Dietterich, 2000; Genre et al., 2013; Zhang et al., 2021).

The data we use in this research consists of monthly implied volatility surface data of US equity options ranging from January 1996 to December 2021, obtained from *OptionMetrics IvyDB*, in combination with 94 firm characteristics from Gu et al., 2020.

Based on our findings we conclude that machine learning models are able to outperform the traditional Black-Scholes model for IV level and curvature prediction. For IV level predictions all machine learning models significantly outperform a simple BS model, whereas for the curve predictions only the Elastic Net regression produces a significant improvement over the BS predictions. Results from the IV slope prediction also indicate that improvements can be made using machine learning models, however, none of the R^2_{OOS} values are found to be significant. It follows from the comparative analysis that the best-performing models are Extremely randomised trees for the IV level predictions ($R^2_{OOS} = 0.487$), Extremely randomised trees for the IV slope predictions ($R^2_{OOS} = 0.272$), and an Ordinary least squares regression with 40 predictors for the IV curvature predictions ($R^2_{OOS} = 0.262$). The individual features that are found to drive the predictive performance of the best performing machine learning models (Extra Trees, Dart, XGBoost, NN), are *bid-ask spread* for IV level predictions, *idiosyncratic return volatility* for IV slope predictions and *dividend to price ratio* for IV curvature predictions. The analysis also highlights the differences in feature importance among the models. Furthermore, when combining the individual models, the equally weighted ensembles produce satisfactory outcomes, exhibiting superior performance compared to the individual models for IV level ($R^2_{OOS} = 0.491$) and slope ($R^2_{OOS} = 0.274$). For curve predictions the best performing model (OLS-40) is not outperformed by the ensembles.

The contributions of our paper to existing literature are fourfold. Firstly, advances the field of option implied volatility (IV) prediction by examining the effectiveness of various machine learning (ML) models. This improved IV prediction is of relevance for investors in terms of option pricing and predicting future market movements. One of the main objectives is to identify models that perform well in predicting IV features (level, slope and curvature), providing novel insights into the relatively undiscovered area of IV feature predictability of individual equity options. Notably, our findings highlight the potential of extremely randomised forests, gradient boosted regression trees, and neural networks for IV level, slope, and curve prediction, and highlights the variable importance assigned by the models for each of these features, suggesting valuable advancements in the field of IV shape predictions.

Secondly, to the best of our knowledge, this paper is the first to comprehensively explore a range of ensembles for IV predictions, employing a comparative analysis of diverse individual machine learning models. The results indicate the value of including ensembles in the sets of algorithms for novel studies in option IV prediction.

Moreover, this research deviates from the often used S&P 500 index, representing the risk expectations of the aggregate stock market, which is often studied in previous literature. Instead, this paper examines individual equity options, providing insights into the predictability and driving variables of the IV shape of individual equity options.

Lastly, this research contributes to the growing body of literature on machine learning applications for financial problems, addressing the use of machine learning models in the domain of financial markets and providing valuable insights for researchers in this field.

The subsequent parts of this paper are structured as follows. Section 2 describes the position of this paper in the existing field of literature and our potential contributions. This is followed by Section 4, which elaborates on the various linear and machine learning algorithms used for the comparative analysis, accompanied by an explanation of the performance measures. Subsequently, the results of the IV feature predictions based on the models in Section 4, are presented in Section 5. To conclude our findings, a concise conclusion of the results is given in Section 6. This is followed by Section 7, discussing the limitations of our research and future recommendations.

2 Literature

Due to the aforementioned role of implied volatility in option pricing and risk management, a large body of literature has been dedicated to forecasting implied volatility. Our contributions to the literature bridge two streams of research. Firstly, we aim to uncover advancements in implied volatility predictions. Secondly, we contribute to the application of machine learning methods in finance.

From empirical research it follows that the assumptions under the seminal Black and Scholes model (Black and Scholes, 1973), of a flat IVS without jumps in the underlying asset returns, contradict the observed IV processes (Ball and Torous, 1985; Bates, 1996; Beckers, 1980). Following these misspecifications, a host of new models have been formulated, relaxing the assumptions imposed by the BS model. In research by Cox and Ross, 1976; Heston, 1993; Hull and White, 1987, the local implied volatility is formulated by a deterministic function. Dumas et al., 1998 smooth out the implied volatility surface. This correction outperforms the local volatility methods. And Carr and Wu, 2016 formulate a specification for the IVS dynamics. On the other hand, Bakshi et al., 1997; Merton, 1973; Pan, 2002 introduce jumps in the underlying stock returns. Despite its limited predictive performance and misspecification, the Black-Scholes model remains widely used as a benchmark model (Almeida et al., 2022; Audrino and Colangelo, 2010; Ewing, 2010; Freire and Kleen, 2023; Isengildina-Massa et al., 2007; Li, 2005; Poon and Granger, 2003; Zulfiqar and Gulzar, 2021).

Besides the BS model formulation and the subsequent relaxations, IVS features and their respective predictability is studied by Goncalves and Guidolin, 2006 and Mayhew, 1995. Goncalves and Guidolin, 2006 find a statistically predictable pattern. However, several features of the IVS have been found to deteriorate the predictions. Financial data presents several well-known stylised facts that have been show to pose serious challenges to standard econometric models, such as strong persistence in autocorrelations, fat tails in return distributions, and nonlinearity. As argued by Corsi, 2009, traditional models such as GARCH and stochastic volatility models are unable to reproduce these features. Another limitation of traditional (linear) models is the use of small datasets. This is because additional covariates would cause conventional models, often relying on linear regressions, to break down when explanatory variables are highly correlated, have a low signal-to-noise ratio, or when the underlying structure is significantly nonlinear (Christensen et al., 2021; Corsi, 2009). As discussed by Varian, 2014, Gu et al., 2020 and Bali et al., 2021a, the nonlinearities, complex interactions, and vast set of available explanatory variables in IV data makes it perfectly suited for the application of machine learning models.

The finance industry is undergoing a significant transformation through the use of machine learning

(Tang et al., 2022). In recent research decision trees, support vector machines, neural nets, deep learning, and many other machine learning models have been employed for various financial applications such as predicting asset returns (Avramov et al., 2023; Bryzgalova et al., 2020; Chen et al., 2023b; Chinco et al., 2019; Gu et al., 2020; Rapach et al., 2013), bond returns (Bali et al., 2020; Bali et al., 2021b; He et al., 2021), portfolio optimisation (Ma et al., 2021; Wang et al., 2020a), credit risk modelling (Galindo and Tamayo, 2000; Wang et al., 2020b). Giglio et al., 2022; Nagel, 2021 and Zaffaroni and Zhou, 2022 present an overview of various studies in the field of finance that employ machine learning techniques.

While there is an extensive body of literature exploring applications of machine learning models in finance, few papers focus on implied volatility forecasting. The literature regarding realised volatility, however, is richer. Luong and Dokuchaev, 2018 and Christensen et al., 2021 make use of random forest models for realised volatility forecasting and Mittnik et al., 2015 find promising results using component-wise gradient boosting. Papers on neural networks (NN) such as feed forward NN (Carr et al., 2019), artificial NN (Donaldson and Kamstra, 1997), heterogeneous autoregressive NN (Fernandes et al., 2014; Hillebrand and Medeiros, 2010) and long short-term memory NN (Bucci, 2020; Kim and Won, 2018; Rahimikia and Poon, 2020), also make up a large part of the research on realised volatility forecasting using machine learning methods. As previously stated, machine learning algorithms are seldom applied for IVS forecasting. We proceed to name a few examples. Malliaris and Salchenberger, 1996 utilise artificial neural networks, considering past volatilities and options market factors, to predict S&P100 implied volatility. Lee et al., 2007 introduce a particle swarm optimisation method, resulting in promising option prices based on the IV estimates. Similarly, Wang et al., 2012 apply backpropagation-trained neural networks to forecast prices under different volatility models. Furthermore, Audrino and Colangelo, 2010 employ a boosting algorithm, based on regression trees, to predict implied volatility surfaces. Recent research by Almeida et al., 2022 demonstrates the superior performance of deep learning methods, highlighting their potential to further improve results when combining parametric and non-parametric approaches.

A larger body of literature studies the closely related option pricing forecasting problem (Ackerer et al., 2020; Amilon, 2003; Bali et al., 2021a; Das and Padhy, 2017; De Spiegeleer et al., 2018; Dugas et al., 2009; Garcia and Gençay, 2000; Hutchinson et al., 1994; Liu et al., 2019; Park et al., 2014). Option prices can be derived from IV, in which IV is the more general measure which indicates market behaviour and is comparable across options (Almeida et al., 2022; Liu et al., 2021). IVS movements are therefore expected to be closely related to the behaviour of option prices and are, in addition, found to depend on the same firm characteristics as the underlying assets (Chen et al., 2023a; Freire and Kleen, 2023). Overall the aforementioned papers assert promising results in the use of machine learning methods for

implied volatility or option price predictions. This further motivates us to examine a diverse selection machine learning applications for predicting IVS.

3 Data

This research makes use of implied volatility surface data of individual US equity options. The data is obtained from *OptionMetrics IvyDB* in Wharton Research Data Services (WRDS), and includes implied volatility surfaces for all US exchange-listed equities and all firms in the NASDAQ. The sample ranges from January 4$^{\text{th}}$ 1996 to December 31$^{\text{st}}$ 2021, in line with Bali et al., 2021a. Therefore, the sample consists of 6541 trading days. The data contains information on the entire U.S. equity option market and includes the interpolated implied volatility (IVS), expiration date, strike price, delta and information on whether the option is a call or a put. We include the IVS of the underlying stock for which the options are traded for more than 10 consecutive years within our time frame to ensure a most liquid crossection of options. To analyse the shape of the IV curve, we examine the 30-day IV curve generated by put options and focus solely on option IVs with a delta of -0.8, -0.5, or -0.2 corresponding to, respectively, in-the-money (ITM), at-the-money (ATM) and out-the-money (OTM) options. Options rarely trade precisely at these deltas on every date. To address this issue, we utilise the interpolated IV curve. This curve estimates the IV for individual equity options of American style using a Cox-Ross-Rubinstein (CRR) binomial tree model.

Moreover, research has shown that the IV is dependent on the same firm characteristics used in predicting financial forecasting problems. Chen et al., 2023a and Freire and Kleen, 2023 find that features used by Green et al., 2017, Gu et al., 2020 and Han et al., 2022 for predicting equity returns, have substantial explanatory power on the IV curve. This implies that the use of these features along with our machine learning techniques would lead to favourable results for IVS forecasting. Therefore, we make use of 94 predictor variables proposed by Gu et al., 2020 for predicting asset returns, which are published on their website[1]. A list of the most important characteristics for this research, and their corresponding descriptions can be found in Table 3, in Appendix A. For extensive details on all characteristics, we refer to Gu et al., 2020. The dataset contains monthly data and exhibits missing values, as indicated in Figure 11 in Appendix B, which displays the percentage of missing values for each month. Similar to Freire and Kleen, 2023 and Gu et al., 2020, we handle missing values in the characteristics by taking the cross-sectional median of observed predictor variables for other stocks within each month. By only making use of the characteristic values given in the same month, we avoid look-ahead bias. We transition from daily option data to monthly data

[1]Retrieved from https://dachxiu.chicagobooth.edu on March 4, 2023

by retaining the end-of-month dates present in the characteristics dataset, thus incorporating option data from the last day of each month. By means of a linking table provided by *WRDS*, which contains the stock PERManent Numbers (*permno*) and corresponding SECurity IDentifier codes (*secid*), we are able to merge the characteristics and option datasets. The implied volatility surface is composed of interpolated IVS data which does not contain missing values. However, through the process of linking options to their corresponding firm characteristics, missing values emerge. This is due to the fact that the characteristics dataset provided by Gu et al., 2020 does not contain information for all stocks present in the *WRDS* dataset. Table 13 in Appendix B shows the percentage of missing values for each month after merging the options and stock characeristics. The gaps that emerge from linking the characteristic's *permno* to the option's *secid*, are also filled using a crossectional median approach. By adopting this approach, we do not remove options from the dataset and our models can work with a more comprehensive set of information as more data is preserved.

Furthermore, since machine learning models are found to perform better in standardised data we perform a data transformation. In line with Freire and Kleen, 2023 and Gu et al., 2020 we standardise the characteristics on a monthly basis by mapping the cross-sectional ranks into the $[-1,1]$ interval.

3.1 IV feature computation

To comprehensively evaluate the predictive performance of the models included in this paper, we consider three fundamental shape characteristics: level, slope, and curvature. It is essential to obtain predictions for the shape elements of the IVS in order to perform an in-depth examination of the predictability of these characteristics and to examine whether the models are able to accurately capture the distinctive IVS features.

We compute the shape characteristics based on 30-day IV curves, in line with Chen et al., 2023a. The level characteristic is based on the IV corresponding to an ATM option delta of -0.5 ($IV^l = IV^{ATM}$). The slope of the IV curve is estimated by analysing the IVs of a put spread option strategy involving the simultaneous buying and selling of put options at different delta levels ($IV^s = IV^{OTM} - IV^{ATM}$). Lastly, the curvature of the IV curve is approximated by analysing a butterfly spread options strategy that involves buying out-the-money (IV^{OTM}) and in-the-money (IV^{ITM}) options while selling at-the-money (IV^{ATM}) options: $IV^c = \frac{IV^{OTM}+IV^{ITM}}{2} - IV^{ATM}$. For every stock we compute the option implied volatility level (IV^l) slope (IV^s) and curvature (IV^c).

The resulting dataset contains 680,867 observations in total, for every option IVS shape feature. Since every IVS shape feature is found for the options corresponding to a specific stock, the number of observations can be seen as the number of stocks, for which the options are traded, over different time periods. On average, there are 26,187 observations per year, with a maximum of 34,182 and a minimum of

Table 1: Summary Statistics of IVS features

Level	Slope	Curvature
0.4241 (0.192, 0.728)	0.0727 (0.000, 0.178)	0.0512 (-0.006, 0.150)

This table presents the average values of implied volatility features level, slope, and curvature with the 10% and 90% quantiles.

11,858 observations per year. Each month contains an average of 2,182 different stocks, and the maximum and minimum number of stocks per month is respectively 2,866 and 895. Figure 14 in Appendix B displays a histogram of the number of stocks, corresponding to the number of option IVS, for every month. Table 1 reports summary statistics of our implied volatility data for slope level and curvature. The average values found for the IV slope and curvature are positive, indicating, as anticipated, that the option IV curve smiles (Chen et al., 2023a). Complete distributions of the IVS features can be found in Figure 15 in Appendix B.

3.2 Sample Splitting & Tuning

In line with Gu et al., 2020 we use an expanding rolling window which is commonly used for tuning machine learning models for prediction problems. It involves, recursively, dividing the data into a training ($\mathcal{T}_1$), validation ($\mathcal{T}_2$), and a test ($\mathcal{T}_3$) set, with $\mathcal{T}_1$ gradually expanding at each iteration, while $\mathcal{T}_2$ and $\mathcal{T}_3$ roll over with the size of the out-of-sample $\mathcal{T}_3$. To apply the expanding rolling window method, we start by selecting a fixed window size for the initial training set $\mathcal{T}_1$. We then fit a (machine learning) model to this training set and evaluate its performance on a validation set $\mathcal{T}_2$ that immediately follows the training set. The model configuration that is found to perform optimally, when tested on $\mathcal{T}_2$, is consequently used for the out-of-sample IVS predictions using the predictor variables in $\mathcal{T}_3$. Section D.2 in the Appendix, further elaborates on the procedure of obtaining the optimal model configuration, with an optimal set of hyperparameters. Next, we expand $\mathcal{T}_1$ and repeat the process of model fitting and evaluation on a new set $\mathcal{T}_2$ that follows the expanded training set. This process is repeated until the end of the dataset is reached. We take the initial train ($\mathcal{T}_1$), validate ($\mathcal{T}_2$) and out-of-sample samples ($\mathcal{T}_3$) to be 11, 5 and 1 years respectively. This is based on ratios similar to Gu et al., 2020. The total out-of-sample period is 10 years. For the first out of sample period 2012, the testing sample $\mathcal{T}_1$ ranges from 1996 to 2006, and the validation sample $\mathcal{T}_2$ ranges from 2007 to 2011. The models are retrained after one year. For the next out-of-sample period (2012), $\mathcal{T}_1$ is increased by one year while the size of $\mathcal{T}_2$ remains the same. A schematic overview of this procedure can be found in Figure 16 in Appendix D.

3.3 Data Screening

Research shows that the predictive performance of machine learning methods can be enhanced by removing a selection of variables from the complete set of characteristics using a feature selection technique in order to reduce noise, therewith reducing chances of overfitting and improving accuracy (Brownlee, 2020; Chandrashekar and Sahin, 2014; Guyon and Elisseeff, 2003). By reducing the number of features the covariates holding the most predictive power remain, while the influence of irrelevant features is minimised. Chen et al., 2023a use lasso regularisation in their research to identify the characteristics that hold the most predictive power. In line with their research, we also opt for a regularisation technique. Zou and Hastie, 2005, however, find elastic net to outperform regular lasso in feature selection for regression problems, additionally Vrontos et al., 2021 make use of Elastic Net for feature selection in IV sign prediction. We implement this method, since it is known to be able to handle multicollinearity, and obtain a selection of features that are found to have the most predictive power based on the elastic net estimation. When incorporating the elastic net regression, variables are filtered out by putting their corresponding coefficients to zero when minimising the loss function. The elastic net loss functions, containing both an L1 and L2 penalty term, is formulated as follows:

$$\mathcal{L}(\boldsymbol{\theta};\alpha;\rho)=\frac{1}{N}\sum_{i=1}^{N}\left(IV_{i,t}^{m}-\hat{IV}_{i,t}^{m}(\mathbf{z}_{i,t};\boldsymbol{\theta})\right)^2+\alpha\rho\sum_{j=1}^{P_f}|\theta_j|+\frac{1}{2}\alpha(1-\rho)\sum_{j=1}^{P_f}\theta_j^2,\quad m=l,s,c, \tag{1}$$

here $i=1,...,N$ denotes the corresponding stock at time $t=1,...,T$, for shape feature m, representing the IV level (l), slope (s) and curvature (c). The vector of characteristics is denoted by $z_{i,t}$ for stock i at time t. $\boldsymbol{\theta}$ is the resulting set of coefficients of which some are reduced to zero, removing the influence of the corresponding predictors $j=1,...,P_f$ from the regression. The elastic regression incorporates regression coefficients $\rho\in[0,1]$ and $\alpha>0$.

In the variable selection process, we first examine the number of variables that can be filtered out for each feature before a significant decline in predictive performance occurs. To filter out features from the complete set of variables, we estimate the elastic net regression, as described in equation 1, using the initial training dataset $\mathcal{T}_1$ (1996 - 2006). By adjusting parameters α and ρ, different numbers of predictors are filtered out. The performance of each selection of variables is then evaluated by predicting the IVS features of the initial validation set $\mathcal{T}_2$ (2007 - 2011) using the elastic net regression. We examine the number of selected features for which the predictive performance does not exhibit a significant decline. For this process, the features are selected based on the first set of training and validation years from the first prediction loop to avoid look-ahead bias.

Once a general number of variables to select is determined for each IVS feature, the elastic net regression is re-estimated using a combination of the initial training and validation sets ($\mathcal{T}_1$ and $\mathcal{T}_2$). Parameters α and ρ are adjusted such that the optimal selection of variables is identified for each IVS feature, based on the previously determined number of predictors.

4 Methodology

In this paper several models are computed to obtain estimates of IVS features. In the models we use the following notations: The models $g(.)$ depend on characteristics $z_{i,t}$, for stock $i=1,...,N$ in month $t=1,...,T$, and parameter vector $\boldsymbol{\theta}$, and predict implied volatility features $IV_{i,t}^m$ for $m=l,s,c$ denoting respectively the level, slope and curvature of the IVS. A full list of (model) abbreviations that are used throughout this paper can be found in Appendix C.

4.1 Linear benchmark models

4.1.1 Black-Scholes

The simplest form of IVS prediction model is the Black Scholes (BS) model introced by Black and Scholes, 1973. To date, this model is widely used in research on implied volatility (Bates, 1996; Bennell and Sutcliffe, 2004; Ewing, 2010; Poon and Granger, 2003; Zulfiqar and Gulzar, 2021). Under the BS model, the implied volatility is constant. In line with this assumption, we introduce our benchmark model which assumes a constant level factor and zero slope and curvature. The corresponding model for the level factor is, therefore, simply denoted as the average of former implied volatilities:

$$IV_{\mathcal{T}+k}^l = \frac{1}{\mathcal{T}} \sum_{t=1}^{\mathcal{T}} \left(IV_t^l \right), \qquad k=1,...,\mathcal{T}_{oos}, \tag{2}$$

here $\mathcal{T}$ is the size of the dataset of all previous periods, and $\mathcal{T}_{oos}$ is the size of the dataset used for the out-of-sample predictions. In line with the assumption of a flat IV surface, IV^s and IV^c are set to 0.

4.1.2 Ordinary Least Squares

A second benchmark model, introduced to predict the IVS shape based on the proposed characteristics, is the Ordinary Least Squares (OLS) model. In line with Gu et al., 2020 we formulate two OLS models. For the first model, all 94 characteristics are included and for the latter, a selection of features is employed to reduce noise in the predictions. By adding features, Gu et al., 2020 find that the regression efficiency progres-

sively reduces. Therefore, we expect an improvement of the OLS-4 model over the model including all predictions. However, overall we do not expect these methods to produce satisfactory results due to their inability to capture complex nonlinear interactions among predictors. The four factors that are chosen for the OLS-4 model are *market beta* (f_1), *bid-ask spread* (f_2), *book-to-market ratio* (f_3) and *leverage* (f_4). Research by Chen et al., 2023a, Dennis and Mayhew, 2002 and Christoffersen et al., 2018 show that the *beta* holds significant explanatory power in predicting respectively IVS shape, option volatility skew and option prices. In this research, we employ the industry-adjusted market beta. The liquidity measures (*bid-ask spread*) are found to be one of the most important features by Freire and Kleen, 2023, for explaining differences in IVS. Chen et al., 2023a, amongst others, show that *book-to-market ratio* can influence the IV. Lastly, Geske and Zhou, 2009 find that firm leverage has significant statistical and economic effects on option returns.

The OLS-all and OLS-4 model are respectively denoted by:

$$g_{OLS-all}(z_{i,t};\boldsymbol{\theta}) = z'_{i,t}\boldsymbol{\theta}, \qquad g_{OLS-4}(f_1,f_2,f_3,f_4;\boldsymbol{\theta}) = f^1_{i,t}\cdot\theta^1 + f^2_{i,t}\cdot\theta^2 + f^3_{i,t}\cdot\theta^3 + f^4_{i,t}\cdot\theta^4, \tag{3}$$

in which $z_{i,t}$ is a vector including all of the characteristics incorporated in this research and $\boldsymbol{\theta}$ is a vector of the corresponding coefficients.

4.2 Machine Learning models

This section introduces the machine learning models used for predicting implied volatility shape features. Since in recent research Bali et al., 2021a find forest models to outperform neural network (NN) models in predicting option returns, and we expect similar behaviour of implied volatility predictions, we, therefore, solely focus on various forest models in the comparative analysis. We do, however, incorporate an NN model in our ensemble of machine learning techniques. Due to the difference in the configuration of the trees and NN's, the latter is expected to capture different characteristics of the IVS and therefore provides a valuable addition to the predictions.

This section introduces the machine learning models used for predicting implied volatility shape features. To ensure comprehensive coverage, a diverse set of models is examined, based on their past performances in IV prediction or related financial prediction problems. The categories of models investigated include regularisation techniques, forest models and a neural networks. We expect all of these categories to capture distinct characteristics of the IVS data. Therefore, to enhance the predictive capabilities further, we combine the strengths of these different models by creating ensembles of machine learning techniques. This is expected to capture various distinct characteristics of the IVS and therefore provides a valuable

addition to the predictions.

4.2.1 Elastic Net

Regularisation techniques have yielded promising outcomes in forecasting of realised volatility, whether employing ridge regression (Carr et al., 2019) or lasso regression (Audrino and Knaus, 2016; Audrino et al., 2020; Caporin and Poli, 2017). Through the integration of these methods, optimal qualities of both methods are combined. Due to the inclusion of two penalty terms in the loss function, the model is able to reduce the noisy covariates through variable selection demonstrates the ability to handle multicollinearity in predictive terms.

The estimation of the model is done by performing a simple OLS regression on the features that are selected by the feature screening. In contrast to regular OLS, the loss function now contains two penalty terms which result in the formulation given by equation 1.

4.2.2 Random forest

The second machine learning method examined in this paper is the random forest (RF), introduced by Breiman, 2001. This regression tree ensemble is found to perform well for various financial applications. RF performs optimal in predicting asset prices, surpassing models such as support vector regressions, artificial neural networks (Patel et al., 2015), AdaBoosting, K-Nearest neighbors (Ballings et al., 2015), long short-term neural networks, convolutional neural networks (Ma et al., 2021), regularisation and gradient boosting (Gu et al., 2020). More closely related to this paper's problem, Luong and Dokuchaev, 2018 successfully utilise RF for predicting realised volatility. Similarly, Christensen et al., 2021, found RF to be preferred for realised volatility prediction among a wide range of algorithms.

RF is a machine learning algorithm that uses an ensemble of decision trees to compute predictions. Each decision tree (b) is constructed independently. At every node a random subset of features (P_f) is drawn and an optimal feature ($z \in P_f$) with corresponding threshold value α is estimated, based on which the data is split. The optimal split ($s^*(z,\alpha)$) is obtained by minimising the loss function:

$$\mathcal{L}(C(s),C_{left}(s),C_{right}(s)) = \frac{1}{|C(s)|} \sum_{z_{i,t} \in C_{left}(s)} (IV^m_{i,t+1} - \theta_{t,left})^2 + \frac{1}{|C(s)|} \sum_{z_{i,t} \in C_{right}(s)} (IV^m_{i,t+1} - \theta_{t,right})^2, \quad (4)$$

In which θ_t is the average of the observations in the corresponding data sample C. A visualisation of branching out of the regression tree is given in Figure 18 in Appendix E. New observations are assigned to different terminal nodes k and predictions are formed using the corresponding coefficients θ_k. The

final prediction is obtained by aggregating the predictions of all the B trees:

$$g_{RF}(z_{i,t};\boldsymbol{\theta}) = \frac{1}{B}\sum_{b=1}^{B}\sum_{k=1}^{K}\theta_{k,b}I_{z_{i,t}\in C_{k,b}}, \tag{5}$$

here K is the number of terminal nodes in the individual trees. This approach helps to reduce overfitting and improve the accuracy of the model.

4.2.3 Extremely randomised trees

The Extremely Randomised Trees (Extra Trees), is a machine learning algorithm that is similar to Random Forest but uses a different method for constructing decision trees. The financial literature offers limited examples of applications for this model. Nevertheless, previous academic research has shown promising results in utilising the Extra Trees algorithm. Notably, Polamuri et al., 2019 and Sadorsky, 2022 provide evidence supporting its effectiveness in stock price prediction. Furthermore, Ghosh and Sanyal, 2021 published a paper focusing on predicting market volatility in India. Their findings reveal that, even though, extreme gradient boosting emerges as the superior model, Extra Trees surpasses deep and long short-term neural networks in terms of predictive performance.

The Extra Trees algorithm was first introduced by Geurts et al., 2006 and is a variation on the previously mentioned RF model. Apart from randomly selecting a subset of features at every node, the Extra Trees model also chooses a random selection of split points ($s(z,\alpha)$). For each selected feature, a random threshold (α) is drawn from a uniform distribution within the range of the feature values. Similar to the RF algorithm, the optimal split is chosen based on minimal loss following equation (4). Consequently, the final predictions are computed by averaging the predictions of individual trees, as is done in equation (5), emulating the RF algorithm. The additional randomisation allows the Extra Trees algorithm to generate a larger number of diverse trees, which can help to reduce the variance of the model and improve its predictive performance.

4.2.4 Gradient Boosting

An alternative approach to improve upon forest ensembles is through boosting, one of the most powerful learning ideas (Krauss et al., 2017). Gradient Boosting regression model (GBM) is a machine learning algorithm based on the concept of combining weak learners to form a robust model. Unlike Random Forest, which takes the average of all weak learners, GBM progressively builds the model by minimising a differentiable loss function. The algorithm was first introduced by Friedman, 2001 and works by sequentially adding weak learners to the model, each of which corrects the errors made by the previous

ones. In each iteration, a new weak learner (g_b^m) is trained on features $z_{i,t}$, and the inverse of the squared loss between the prior predictions ($\hat{IV^m}_b$) and the actual values (IV^m) denoted by $\varepsilon_{i,t+1}$, moving the model opposite the direction of the loss:

$$\varepsilon_{i,t+1} = -\frac{\partial \mathcal{L}(IV^m_{t+1}, \hat{IV^m}_b)}{\partial \hat{IV^m}_b}\bigg|_{\hat{IV^m_b}=g^m_{b-1}(z_{i,t})} \qquad m=l,s,c, \tag{6}$$

in which $\hat{IV}^m_b(z_{i,t})$ is the prediction for $z_{i,t}$ which follows from the model constructed using weak learners from the previous $b-1$ iterations $g^m_{b-1}(z_{i,t})$. The final model is formed by combining the weak learners using a weighted sum. GBM is known for its high accuracy and ability to handle complex interactions between features. Qin et al., 2013 and Mittnik et al., 2015 employ gradient boosting for, respectively, stock returns and market volatility. In the field of IVS forecasting, Audrino and Colangelo, 2010 finds the boosting procedure applied to regression trees to improve the performance over the use of individual models. Furthermore, GBM outperforms RF in predicting IV directions (Vrontos et al., 2021). Although this is a classification problem, it still endorses the ability of GBM to capture the IVS patterns. We implement two variations on the standard GBM, which have shown to perform well in financial predictive problems; the Gradient boosted regression tree with dropout (Dart) and Extreme Gradient Boosting (XGBoost).

First, we opt for the Dart algorithm since Bali et al., 2019 demonstrate its superior performance compared to regular gradient boosting, in predicting option returns. The Dart model, adopted from Vinayak and Gilad-Bachrach, 2015, deviates from the general GBM in two ways. Firstly, the gradient of the loss is computed based on a subset of the thus far constructed weak learners within the ensemble as opposed to the complete set of weak learners ($\sigma_b(z_{i,t}) = \sum_{l \in L} IV^m_l(z_{i,t})$, $L \subset \{1,...,1-b\}$). Secondly, normalisation is performed so that the new tree ($IV^m_b(z_{i,t})$) has the same order of magnitude as the dropped trees.

An efficient variation on GBM is the XGBoost algorithm. In literature, this model is employed in a variety of financial applications, from stock returns to credit risk prediction (Basak et al., 2019; Li et al., 2020; Liu et al., 2022; Ye and Schuller, 2021). The XGBoost algorithm differs from the GBM both in model formulation and optimisation. Contrary to the GBM, the choice for specific splits is not based on the mean squared error between the predicted and actual value (as is done in equation (4), but is determined by the similarity score and gain, which contain a regularisation parameter to prevent overfitting. In terms of optimisation, several techniques are implemented to increase the efficiency of the model. Due to this enhanced computational power, it possesses greater potential for successfully handling

our large dataset. One such technique is histrogram-based-approximation, which entails that instead of considering all possible split points, the model groups the feature values into discrete bins and only considers split points at the boundaries between the bins. A parallel learning technique allows XGBoost to split the data into smaller datasets and run processes in parallel, reducing training time and allowing for larger datasets to be processed. Furthermore, sparsity-aware split finding allows for missing values in the dataset, and cache-aware access stores gradients and is used to compute similarity scores faster.

4.2.5 Neural Networks

Besides forest models, another branch of machine learning models covers Neural Networks (NN). The NN models have a configuration which is very different from the forest models. Due to the presence of different activation functions and interconnected layers, it is possible to capture even more complex linear relations in the data. However, the intricate architecture of neural networks makes computation more challenging and renders the model less interpretable compared to forest models. Gu et al., 2020 find a simple feed forward neural network (FFN) to outperform all other models in predicting asset returns. While these results appear promising, Bali et al., 2019 do not find NNs to outperform forest models in their research on option pricing. A number of papers have succesfully used NNs for predicting implied volatility (features) and option pricing purposes (Ackerer et al., 2020; Almeida et al., 2022; Amilon, 2003; Dugas et al., 2009; Garcia and Gençay, 2000; Itkin, 2015; Yang et al., 2017; Zheng, 2017).

In this paper we incorporate a 'shallow' FFN. The network architecture consists of an input layer, which contains the predictor variables, a hidden layer in which the predictors are transformed, and an output layer that computes a forecast based on the output of the last hidden layer. Activation functions within a layer are responsible for modifying the input, either linearly or non-linearly, to create the output that is then transmitted to the subsequent hidden layer. The predictor variables $z_{i,t}$ form the input of the model. Before entering a new layer $(l+1)$ the input is amplified by a factor $\boldsymbol{\theta}_l$, which contains an intercept and a weight for each feature in $z_{i,t}$. This results in the signal $\theta_l^0 + \sum_{j=1}^{N} z_{i,t}^j \theta_l^j$. The parameter vector θ_l is estimated by minimising a penalised loss function. Specifically, we add a L2 penalty term to the loss function based on biases and to the activation of neurons. The signal is transformed within the hidden layer using an activation function, $f_{l+1}(\theta_l^0 + \sum_{j=1}^{N} z_{i,t}^j \theta_l^j)$. Lastly the signal is multiplied by a last parameter vector $\boldsymbol{\theta}_L$ and is collected into a forecast using a linear transformation. A visualisation of the NN model formulation is given in Figure 19 in Appendix E. Our algorithm is optimised using the well-known Adam optimiser (Kingma and Ba, 2014).

4.3 Ensembles

Machine learning ensembles have become a staple in modern machine learning because of their improved accuracy and robustness (Bianchi et al., 2021; Chowdhury et al., 2020; Ganaie et al., 2022; Hung and Chen, 2009; Rapach et al., 2010; Sylvester Walusala et al., 2017). Instead of relying on a single model, ensembles combine the outputs of multiple models to make a final prediction. This approach has been found to produce more accurate predictions compared to predictions produced by individual models. Each model is designed to capture a different aspect of the data and make predictions based on that aspect. By combining the outputs of multiple models, the ensemble can leverage the strengths of each individual model and mitigate the weaknesses. The improvement in accuracy resulting from ensembles can be attributed to several factors. Dietterich, 2000 identifies three reasons why ensembles are effective in machine learning. Firstly, they provide a statistical advantage by reducing the risk of selecting the wrong classifier by averaging multiple hypotheses. Secondly, ensembles offer a computational advantage by spreading out the chances of being stuck in local optima due to different optimisation algorithms. Thirdly, they provide a representational advantage by increasing the solution space of representative functions through the combination of hypotheses, potentially including the true unknown function.

In this paper, we analyse different strategies used to combine machine learning models. Following Bali et al., 2021a; Clements and Vasnev, 2021; Félix et al., 2020 and Krauss et al., 2017, we use an equally weighted ensemble model. Clements and Vasnev, 2021 find dramatic improvements in the accuracy of realised volatility predictions when employing this simple average forecast. Bali et al., 2019 demonstrate its superior performance in option pricing compared to regularisation methods, boosting algorithms, random forest and feed forward neural networks. According to Genre et al., 2013, who study various forecast combinations schemes, the simple equally weighted average forecast is rarely outperformed. Predictions of the equally weighted ensemble model are formed by taking the average of the predictions of the models within the ensemble.

Furthermore, we employ a stacking method. Both Zhang et al., 2021 and Pasupulety et al., 2019, who, respectively, examine option implied volatility and stock price prediction, find the stacking ensemble to produce desirable accuracies. Stacking, introduced by Witten and Frank, 2002, is a method in which base learners within the ensemble are fit to data and a new model is trained based on the predictions made by the previous models. This meta-learner, consequently, produces the ensemble predictions.

We propose three different combinations of models for both types of ensembles. The first ensemble is created using three different categories of models (regularisation, regression tree, neural network), with the idea of including the specific strengths of each model. The second ensemble consists only of the three best-performing models that are ranked highest based on their individual predictions with the hope of further

improving upon the already satisfying results. To include information that is possibly missed by the optimal model configuration from the comparative analysis, a simple neural network is added to the third ensemble.

4.4 Evaluation methods

The model performance is assessed by means of the out-of-sample R^2 statistic for individual IVS predictions, introduced by Campbell and Thompson, 2008 Different from Gu et al., 2020 and Bali et al., 2021a, we chose not to use a naive forecast of zero in the denominator when evaluating the IV level predictions, but instead we make use of historical average IV as a naive forecast. This formulation is more in line with the Black-Scholes assumptions for a constant IVS level, different from zero. Félix et al., 2020 also make use of this method in evaluating IV forecasts. The R^2_{oos} formulation is given by:

$$R^2_{OOS} = 1 - \frac{\sum_{(i,t)\in\mathcal{T}_3}(IV^m_{i,t+1} - \hat{IV}^m_{i,t+1})^2}{\sum_{(i,t)\in\mathcal{T}_3}(IV^m_{i,t+1} - \bar{IV}^m_{i,t+1})^2}, \quad m = l,s,c. \tag{7}$$

Here, $\mathcal{T}_3$ refers to the testing sample, which is the out-of-sample data used in calculating the R^2 statistic. Furthermore, $\hat{IV}^m_{i,t+1}$ are the IV feature predictions and $IV^m_{i,t+1}$ the actual IV values, for every combination of stock i at time t in $\mathcal{T}_3$, for IV features level (l), slope (s) and curvature (c). $\bar{IV}^m_{i,t+1}$ is the historical average of the IV observations in the training sample ($\mathcal{T}_1 + \mathcal{T}_2$) used to obtain predictions $\hat{IV}^m_{i,t+1}$. Based on the formulation of the BS model, $\hat{IV}^m_{i,t+1}$ and $\bar{IV}^m_{i,t+1}$ are the same, resulting in an R^2_{oos} of zero. We employ the data in the training sample for $\bar{IV}^m_{i,t+1}$ to avoid look-ahead bias in the historical average used for the test statistic. For slope and curvature prediction evaluation we do make use of the adjusted R^2 and use a zero forecast for the benchmark model found in the denominator, changing $\bar{IV}^m_{i,t+1}$ to zero. This is in line with the Black-scholes model assumption of a flat IVS.

We employ the Diebold and Mariano (1995) test statistic, in line with Gu et al., 2020, to conduct pairwise comparisons of the forecast accuracy among different models:

$$DM^{(1,2)} = \bar{d}^{(1,2)} / \hat{\sigma}_{\bar{d}^{(1,2)}}, \qquad d^{(1,2)}_{t+1} = \frac{1}{\mathcal{N}_{\mathcal{T}_3,t+1}} \sum_{i=1}^{\mathcal{N}_{\mathcal{T}_3,t+1}} ((\hat{e}^{(1)}_{i,t+1})^2 - (\hat{e}^{(2)}_{i,t+1})^2), \tag{8}$$

here $\hat{e}^{(1)}_{i,t+1}$ and $\hat{e}^{(2)}_{i,t+1}$ show the prediction errors of respectively model (1) and model (2). Every value of $d^{(1,2)}_{t+1}$ is obtained using the prediction errors in the IV features for every observation i in the out-of-sample year $t+1$. $\mathcal{N}_{\mathcal{T}_3,t+1}$ is the number of out-of-sample observations for every year $t+1$, in test sample $\mathcal{T}_3$. $\bar{d}^{(1,2)}$ and $\hat{\sigma}_{\bar{d}^{(1,2)}}$ denote the time-series average and Newey-West standard error of the differences $d^{(1,2)}_t$ over the test sample.

5 Results

The Results are split into three parts. First, in section 5.1, feature selection is performed, filtering out a selection of variables from the initial set. Second, we aim to find an answer to the first research question by analysing and comparing the predictive performance of the aforementioned models in predicting IVS features in section 5.2.2. To uncover the variables driving the predictions we examine the feature importance of the best performing models in section 5.2.3. In light of our second research question, section 5.3 examines in detail the possible improvements in IVS feature predictions by combining models in ensembles and analysing their predictive performance for different time frames.

5.1 Feature Selection

Through the process discussed in section 3.3, we chose the number of features to select using the elastic net screening. Since we find that a smaller number of variables is required to both work better with machine learning models as well as for computational purposes, we reduced the features to a number that would decrease the variance and size of the dataset without diminishing predictive performance. As mentioned in section 3.3, the reduction to a specific number of features is performed using the initial training set ($\mathcal{T}_1$), while the consequences for the predictive power of the resulting elastic net regression are examined for predictions of the initial validation set ($\mathcal{T}_2$). It was found that for the IVS level 30 features sufficed, while more features (40) were required for optimal prediction of IVS slope and curvature. It was also found that smaller values for ρ are required for the IVS slope and curve screening since these were shown to be more sensitive to changes in the regularisation parameters. The final selection of features is obtained from an elastic net regression using a combination of the training and validation set. The selection of features are presented in Figures 1, 2 and 3 for respectively level, slope and curve predictions. The selection of features is employed in the following sections to train the machine learning models and examined in a simple OLS regression.

The resulting set of features for IVS level, slope, and curvature show similarities. For example, beta squared (*betasq*) is present in all three figures. This is to be expected since it has shown to hold significant predictive power for the prediction of the IVS shape (Chen et al., 2023a) and option returns (Christoffersen et al., 2018).

Other features that appear in all three selections are return volatility (*retvol*), industry-adjusted size *mve_ia*, share turnover *turn*, dollar trading volume *dolvol*, idiosyncratic return volatility *idiovol* and secured debt indicator *securedind*. These findings are substantiated by research from Freire and Kleen, 2023, who also find *retvol*, *mve_ia* and *idiovol* to be amongst the selected variables for predicting option prices.

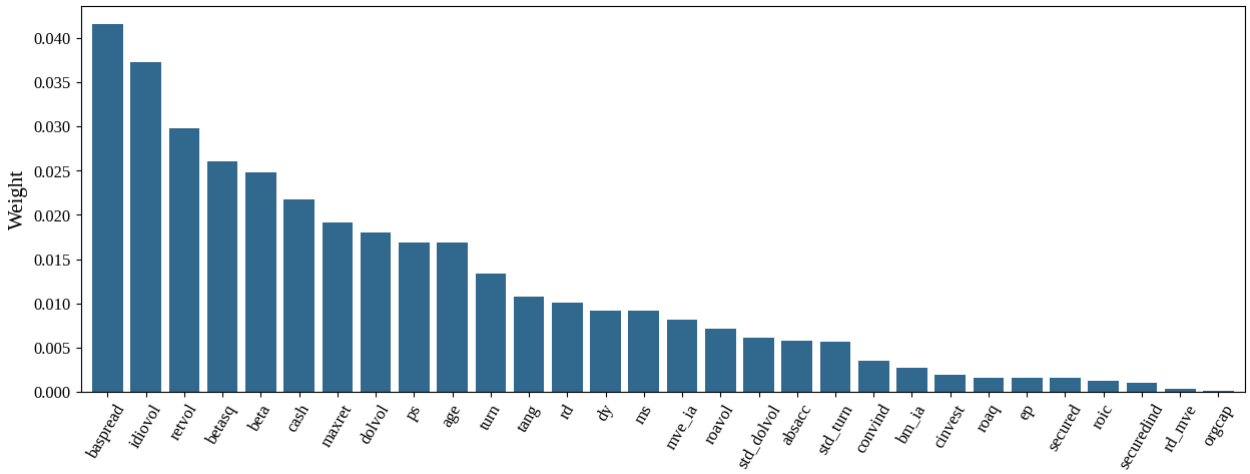


Figure 1: Coefficients of thirty features for the IVS levels

The figure displays a selection of thirty features that follow from an Elastic Net screening. The features are ranked based on their respective absolute coefficients within the Elastic Net model. The selection makes use of the following parameters in the Elastic Net regularisation given by equation (1): $\alpha = 0.5$ and $\rho = 0.0112$. This ensures a selection of thirty variables while the coefficients corresponding to the other variables are set to zero. The sample for the elastic net screening ranges from January 1996 to December 2011.

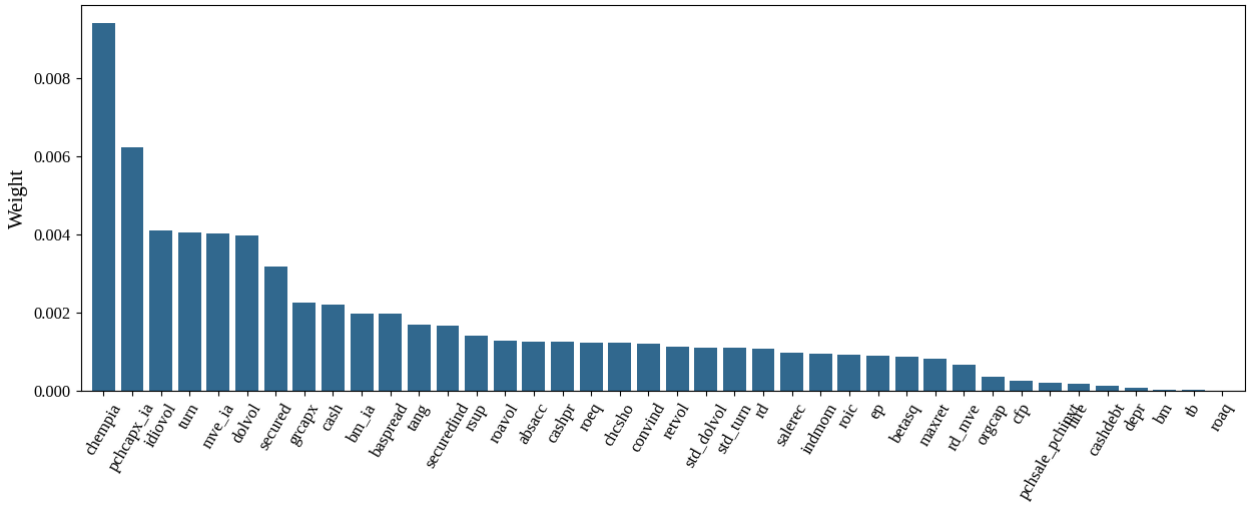


Figure 2: Coefficients of forty features for the IVS slopes

The figure displays a selection of forty features that follow from an Elastic Net screening. The features are ranked based on their respective absolute coefficients within the Elastic Net model. The selection makes use of the following parameters in the Elastic Net regularisation given by equation (1): $\alpha = 0.5$ and $\rho = 0.00155$. This ensures a selection of forty variables while the coefficients corresponding to the other variables are set to zero. The sample for the elastic net screening ranges from January 1996 to December 2011.

A remarkable difference is that the feature bid-ask spread (*baspread*) is the most important for IVS level prediction, in line with research on the IVS shape by Chen et al., 2023a, while it does not hold the same importance for the shape features slope and curvature. Similarly, the feature for industry-adjusted change in employees (*chempia*) is the most important for both level and slope while it is not included

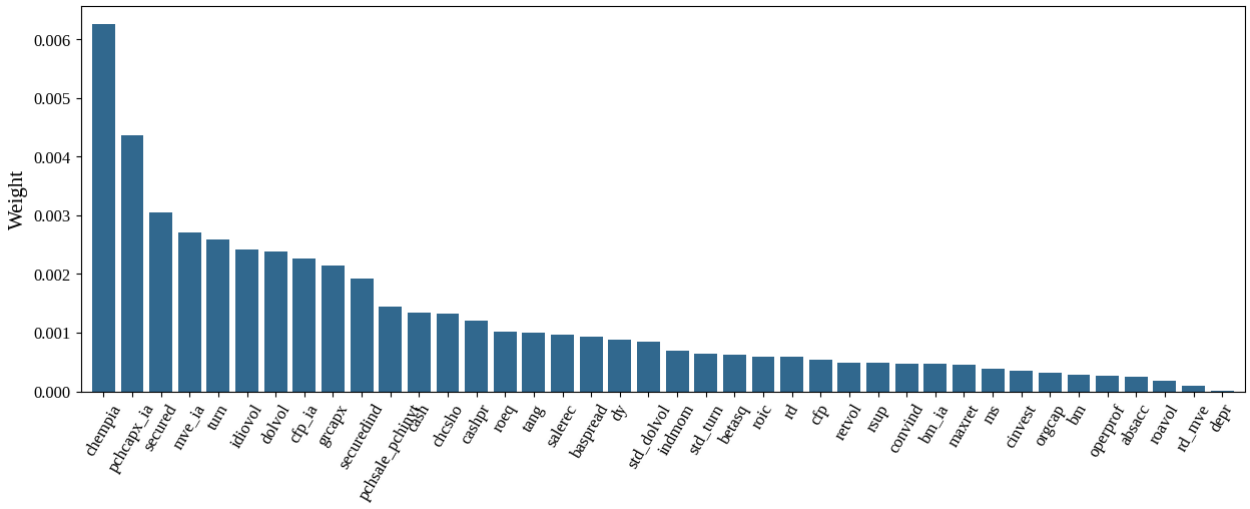


Figure 3: Coefficients of forty features for the IVS curve

The figure displays a selection of forty features that follow from an Elastic Net screening. The features are ranked based on their respective absolute coefficients within the Elastic Net model. The selection makes use of the following parameters in the Elastic Net regularisation given by equation (1): $\alpha = 0.5$ and $\rho = 0.00125$. This ensures a selection of forty variables while the coefficients corresponding to the other variables are set to zero. The sample for the elastic net screening ranges from January 1996 to December 2011.

in the selection for the IVS levels.

Notably, the features that show to be most important in the feature selection do not rank among the top variables in terms of missing values before the data imputation process. This is illustrated in figure 12 in Appendix B, which displays the proportion of missing values per feature. As a result, our data imputation process has limited impact on the IVS feature predictions.

5.2 Evaluating predictive performance

The main research question of this paper reads 'Can machine learning models improve the accuracy of option implied volatility forecasts compared to traditional econometric models?'. To answer this question, we conduct a comprehensive comparison of predictive performance among various machine learning methods, a simple Black-Scholes formulation, and a selection of linear benchmarks. Subsequently, we aim to shed light on the functioning of these models by examining the feature importance assigned to the variables by the best performing models.

5.2.1 Out-of-sample performance individual models

The predictive performance of the models is evaluated using the out-of-sample R^2_{OOS} test statistic. The corresponding values for the IV feature predictions are presented in Table 2 and visualised in Figures 4, 5 and 6, for the level, slope and curvature, respectively. Bold values in the table and figures indicate the

Table 2: Overview of R^2_{OOS} test statistics

	BS	OLS-All	OLS-4	OLS-30	Elnet	RF	Extra	Dart	XGBoost	NN
Level	0.000	**0.401***	**0.395***	**0.451***	**0.424***	**0.435***	**0.487***	**0.465***	**0.478***	**0.483***
Slope	0.000	0.251	0.213	0.205	0.226	0.240	0.272	0.270	0.269	0.269
Curvature	0.000	**0.245**	0.203	0.262	**0.213**	0.242	0.262	0.257	0.261	0.251

	EW1	EW2	EW3	Stacking 1	Stacking 2	Stacking 3
Level	**0.485***	**0.485***	**0.491***	**0.439***	**0.473***	**0.459***
Slope	0.264	0.271	0.274	0.261	0.258	0.235
Curvature	0.250	0.262	0.262	0.260	0.244	**0.248**

This table summarises the out-of-sample R^2_{OOS}, in percentages, for the out-of-sample IVS feature predictions. The bold values are significant at a 5% level and the values accompanied by an asteriks indicate the significance at a 1% level. The significance is based on the significance of a Diebold-Mariano comparative test statistic, comparing the models to a simple Black-Scholes benchmark, which follows from a two-sided t-test.

statistical significance of R^2_{OOS} at a 5% significance level. The statistical significance of the R^2 statistics is based on the significance of a Diebold-Mariano comparative test statistic, comparing the models to a simple Black-Scholes benchmark, which follows from a two-sided t-test. The exact significance levels for the Diebold-Mariano test statistics can be found in Table 7, 8 and 9 in Appendix F.2. In section 4.3, we propose three different model combinations for ensembles. Based on the results obtained from the individual models, we construct the ensembles. The first ensemble (EW1 & Stacking 1) includes three models with different architectures: a regularisation model (elastic net), the best performing machine learning model (extremely randomised trees) and a neural network. In general, for the prediction of the different IV features, the extremely randomised trees, the Dart algorithm and the XGBoost algorithm are found to consistently perform well. The second ensemble (EW2 & Stacking 2) incorporates these optimally performing models. For the third ensemble (EW1 & Stacking 1), the same selection is used in combination with a neural network. Subsequent tables and figures incorporate these proposed ensemble configurations.

From Figure 4 it follows that all models form a significant improvement over the Black-Scholes model at a 5% confidence level in predicting the IVS level. Within the category of machine learning models, extremely randomised regression trees (Extra), exhibit the best results for IV level prediction, followed by the neural network (NN) and the gradient boosting algorithms (Dart, XGBoost). The analysis of OLS regressions reveals a clear bias-variance trade-off in IV level prediction. Reducing the variance of the model by moving from 94 to 30 predictors improves the R^2_{OOS} test statistic by 5% (from *40.15%* to *45.15%*). However, further reducing the number of predictors from 30 to 4 leads to an increase in the bias, resulting in a decline of the R^2_{OOS} test statistic by approximately 6% (from *45.15%* to *39.52%*). For the 'correct' model an optimal middle ground has to be found. Since the machine learning models all yield substantial R^2_{OOS} values, the problem of predicting IV level is found to be a non-linear problem. The superior performance of the neural network indicates the complexity of the problem.

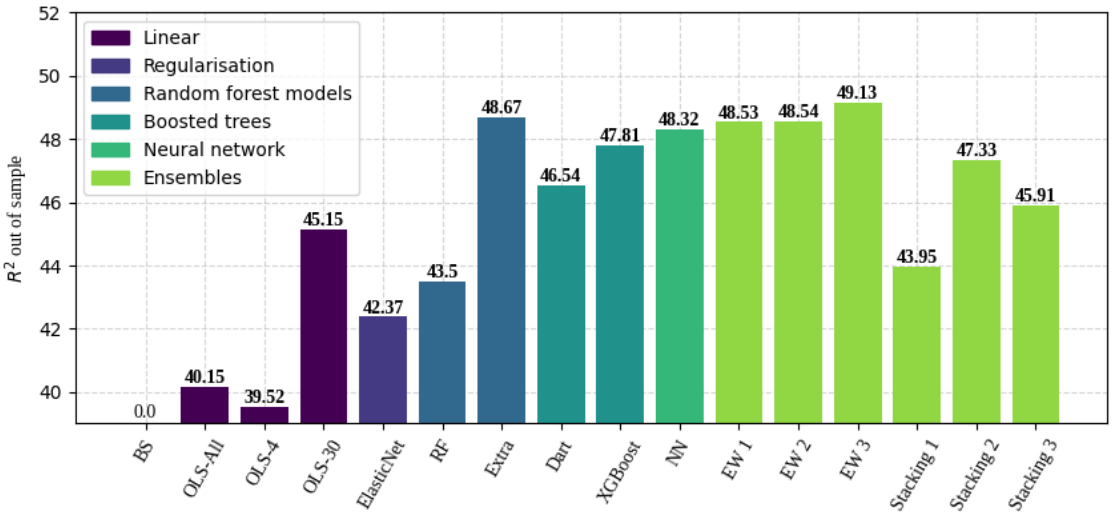


Figure 4: Out-of-sample IVS level prediction performance (percentage R^2_{OOS})

This figure displays the out-of-sample R^2_{OOS}, in percentages, for the out-of-sample level predictions computed by the models in the six model categories examined. Bold values indicate significance at a 5% confidence level. The significance is based on the significance of a Diebold-Mariano comparative test statistic, comparing the models to a simple Black-Scholes benchmark, which follows from a two-sided t-test.

It is noteworthy that the R^2_{OOS} test values are considerably higher in comparison to the values found in research by Bali et al., 2021a, who examine option price prediction. This is to be expected as predicting volatility is easier than predicting option prices. Nevertheless, the difference amongst the different categories and models is, on the other hand, quite similar. Bali et al., 2021a find the lowest values for regularisation techniques, and the highest for an ensemble of non-linear models, followed by the boosted regression trees and the random forest. Contrasting to our findings, their FFN does not outperform the forest models. The ensemble category, particularly an equally weighted ensemble consisting of Extra Trees, Dart, XGBoost, and a NN, yields the most accurate predictions. This promising insight is further explored in section 5.3. Figure 4 clearly demonstrates that the group of equally weighted ensembles outperforms the stacking methods.

Contrary to the level feature predictions, none of the models produce slope predictions significantly different from a simple Black-Scholes model. Although all machine learning models exhibit positive R^2_{OOS} values, only the statistic for the OLS regression including all regressors is found to be close to significant, with a significance level of (7%). The R^2_{OOS} values for predicting the IV slope are substantially smaller compared to the values for the IV level predictions (from 39.52% to 49.13% versus 20.51% to 27.43%). This is in line with research on implied volatility shape features by Chen et al., 2023a, who also find smaller R^2_{OOS} for slope and curvature compared to level predictions. Notably, Figure 5 shows that the OLS regression, including all predictor variables, achieves the highest R^2_{OOS} at 25.12%. Thus, we conclude that the information contained in the large set of predictor variables is of great importance for optimal IV slope prediction.

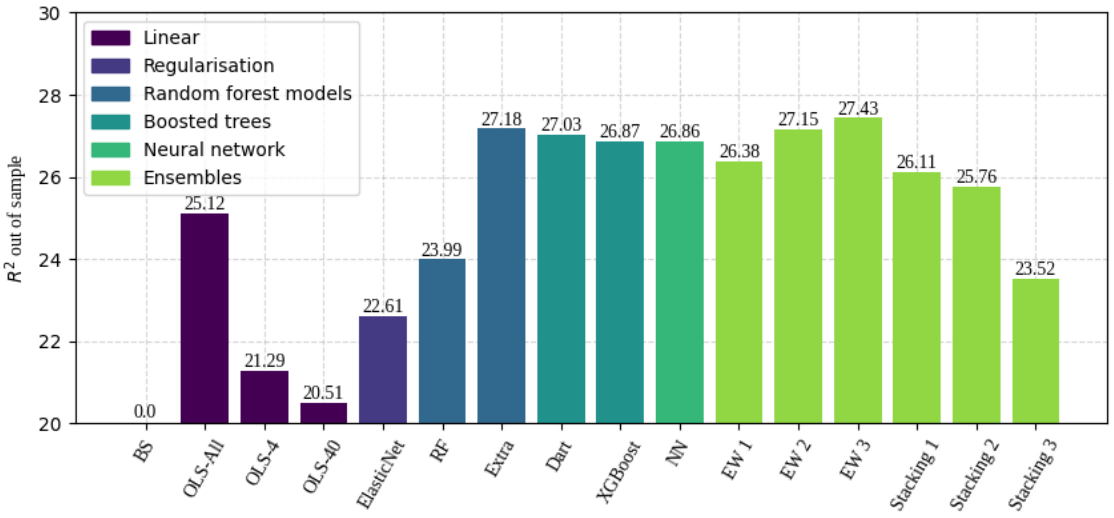


Figure 5: Out-of-sample IVS slope prediction performance (percentage R^2_{OOS})

This figure displays the out-of-sample R^2_{OOS}, in percentages, for the out-of-sample slope predictions computed by the models in the six model categories examined. Bold values indicate significance at a 5% confidence level. The significance is based on the significance of a Diebold-Mariano comparative test statistic, comparing the models to a simple Black-Scholes benchmark, which follows from a two-sided t-test.

Although the performance of machine learning models does not demonstrate a significant improvement over the BS model, Figure 5 exhibits a promising pattern regarding the predictive power of these models for the IV slope feature. While the results do not demonstrate a substantial advantage over the benchmark, there are indications that the machine learning models can capture meaningful patterns and trends in predicting the IV slope. Additionally, while the individual performance of the NN does not surpass that of the forest models, its inclusion in the ensemble results in higher predictive performance. The ensembles, especially the third ensemble, with the inclusion of NN, display high predictive performance (EW1: *26.38%*, EW2: *27.15%*, EW3: *27.43%*). This suggests that IV slope predictions are sensitive to an increase in diversity of model architecture, affecting the variation in handling predictive variables and the computation of predictions.

Figure 6 illustrates the predictive performance of the proposed models for IV curve prediction. Among the linear and machine learning models, the OLS model with 40 predictors stands out as a strong contender. Interestingly, unlike the IV slope predictions, variable selection yields favourable results compared to an OLS model including all predictor variables. Overall, it can be concluded that the IV curvature is challenging to predict accurately, as indicated by the relatively low R^2_{OOS} values, especially when compared to the values found for IV level and, to a lesser extent, slope predictions. From the figure, we observe that only the OLS-all, elastic net, and Stacking 3 methods exhibit significant outperformance of the BS model. Thus, apart from the elastic net algorithm, we reject the hypothesis of improved model performance of the machine learning models for IV curve prediction. Figure 6 displays a positive predictive performance

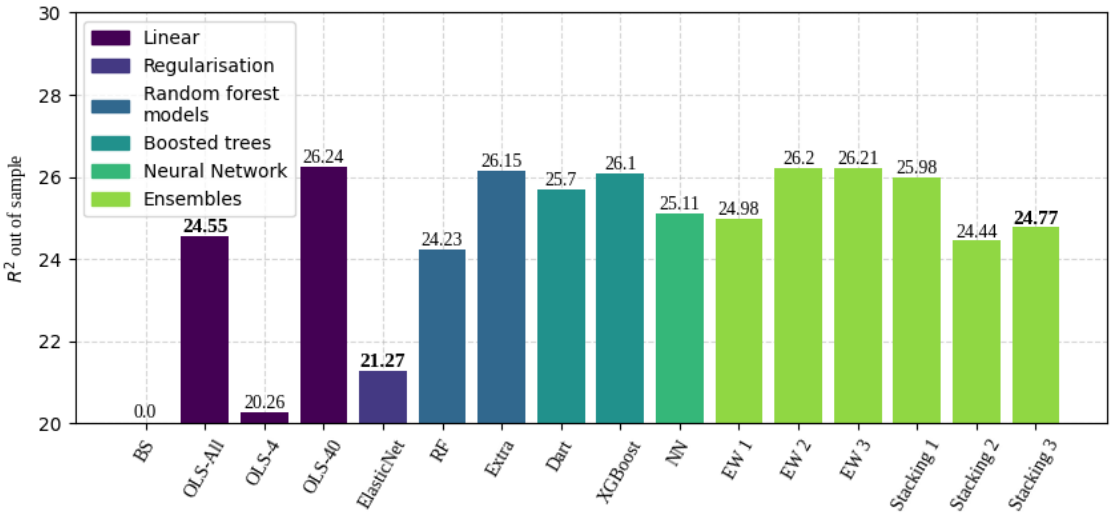


Figure 6: Out-of-sample IVS curve prediction performance (percentage R^2_{OOS})

This figure displays the out-of-sample R^2_{OOS}, in percentages, for the out-of-sample curve predictions computed by the models in the six model categories examined. Bold values indicate significance at a 5% confidence level. The significance is based on the significance of a Diebold-Mariano comparative test statistic, comparing the models to a simple Black-Scholes benchmark, which follows from a two-sided t-test.

for the extremely randomised forest and boosted regression trees, however, none of the R^2_{OOS} values are found to be statistically significant. The 'correct' model's relative simplicity is evident from the NN's underperformance (*25.11%*) compared to the linear and forest models (OLS-40: *26.24%,* Extra: *26.15%,* Dart: *25.70%,* XGBoost: *26.10%*). Including the NN in the equally weighted ensemble 3 (EW3) results in a negligible improvement in R^2_{OOS}, with values changing from *26.20%* to *26.21%*. This suggests that the NN does not provide additional information beyond what is already captured by the other models in the ensemble. This observation aligns with the relatively high R^2_{OOS} of the OLS-40 model (*26.24%*) and the statistically significant performance of the OLS-All model (*24.55%*), indicating a linear relation in the IV curve data, as opposed to complex non-linear behaviour.

5.2.2 Diebold-Mariano model comparison

In this section the significance of the notable differences between models and their relative importance is examined using a Diebold Mariano (DM) test statistic. An extensive overview of all models included in this research can be found in Appendix F.2. Table 7, 8 and 9, for comparison of, respectively, IV level, IV slope and IV curve predictions, includes the Diebold-Mariano test statistics, accompanied by their significance which follows from a two-sided t-test. In the tables, a positive value denotes superior performance of the column model over the row model.

From Table 7, denoting the DM test statistics for the IV level predictions, we observe that all of the

models significantly outperform the Black-Scholes model (as indicated by the bold font). However, only a few relative performance measures are found to be statistically significant. Specifically, the Extra Trees model (Extra), the first equally weighted ensemble (EW1), and the second stacking ensemble (Stacking 2) show significant outperformance compared to the OLS regression model that includes all predictors.

Notably, based on the sign of the DM values, it is evident that the Extra Trees model outperforms all other models. Moreover, the large number of positive DM values in the equally weighted 1 (EW1), equally weighted 2 (EW2) and equally weighted 3 (EW3) columns suggests a favourable performance of these ensembles.

Even though the DM statistics of the machine learning models look promising, it is important to note that none of the comparative performance measures between the linear models and machine learning models are statistically significant, except for the comparison between OLS-All and Extra Trees for level predictions.

Within the Tables 8 and 9, displaying the model comparison for IV slope and IV curve, even more DM statistics are found to be insignificant, indicating the rejection of superior model performance for machine learning models. Nevertheless, in the columns for the models Extra, Dart, XGBoost, and NN, positive values are observed for the performance comparisons with the BS benchmark and the linear regressions. This is no longer the case for the IV curve predictions, where OLS-40 is found to outperform all other machine learning models.

The overall findings of the comparative analysis of a variety of models for IV feature prediction are four-fold. Level predictions exhibit higher R^2_{OOS} values in comparison to slope and curve predictions, indicating that predicting the latter two features (slope and curve) is more challenging. Specifically, the R^2_{OOS} values for level range from *39.52%* to *49.13%*, for slope from *20.51%* to *27.43%*, and for curvature from *20.26%* to *26.24%*. These results align with Chen et al., 2023a, who also reported higher R^2_{OOS} values for level prediction compared to slope and curve. While models such as extremely randomised forest and boosted regression trees still show a small difference of around 2% compared to the best-performing linear model for slope prediction, for curve prediction, machine learning models do not outperform a simple linear regression. Secondly, the extremely randomised regression trees and gradient boosted models (Dart and XGBoost) emerged as the best-performing models. These findings are consistent with Bali et al., 2021a, who also reported the superiority of boosted models over other machine learning techniques like elastic net and random forest. For IV level and slope prediction, the neural network also proves to be a valuable addition to the predictions. Thirdly, the equally weighted ensemble consistently outperforms the ensembles constructed using a stacking procedure. This suggests that a simple equal-weighted combination of models yields better results than the more complex stacking-based ensembles. Lastly, while the comparative results yield favourable

DM-test statistics, indicating potential improvement of machine learning models over standard models in IV feature prediction, it is essential to emphasise that very few of these statistics are statistically significant. This lack of statistical significance makes it difficult to draw firm conclusions solely based on these values.

5.2.3 Feature Importance

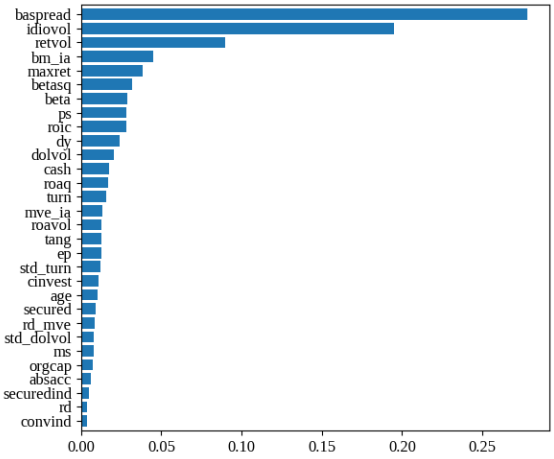
In the previous part we have found a selection of four models that have shown to consistently perform well: Extremely randomised trees, Neural neworks, Dropout additive regression trees and XGBoosted trees. In order to gain a deeper understanding of the model architecture and variable handling within the model, we compute the feature importance, in line with (Gu et al., 2020). Figures 7, 8 and 9 illustrate the features which are found to be most important for the best performing models ERF, Dart and XGBoost, in predicting IV level, slope and curvature. For an easier comparison of the feature importance between different models and between the three IVS shape features, we have constructed a heatmap of the importance scores which can be found in Figure 20 in Appendix F.1. The feature importance of a variable is determined by randomly permuting the values of the predictor while keeping the other variables unchanged. This process allows us to observe the resulting decrease in the R^2 test statistic, which measures the variable's contribution to the overall performance of the model.

The shape of the feature importance plots for IV level predictions shown in Figure 7 is strongly concave. For all four models, the importance is assigned to a few variables, while the other variables show to have little influence. Notably, the Dart and XGBoost plots portray a very similar feature importance order and distribution, which can be linked to their similar model architecture. The first heatmap in Figure 20 clearly highlights the dissimilarity of the partition of feature importance between the neural network and forest models. While the feature importance shape is similar, the NN assigns the largest importance to *betasq* and *beta*, as opposed to *baspread* and *idiovol*. These disparities are a consequence of the differences in feature selection within the algorithms.

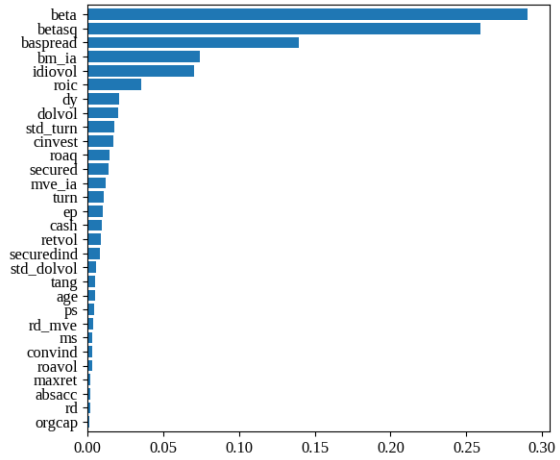
The most important feature found is *baspread*, which represents the Bid-ask spread of an option. The consistent identification of *baspread* as an important feature by all four models can be attributed to the economic relationship between IVS and stock liquidity. Stocks with low liquidity (high bid-ask spread) tend to exhibit higher IVS levels (Freire and Kleen, 2023). This is in agreement with findings by Bali et al., 2021a, who also identify *baspread* as one of the most important features, along with implied volatility itself, for predicting option returns. Similarly, Chen et al., 2023a find *baspread* to be amongst the top ten fundamentals in predicting IV levels.

Figure 7: Feature importance by model for IV level predictions

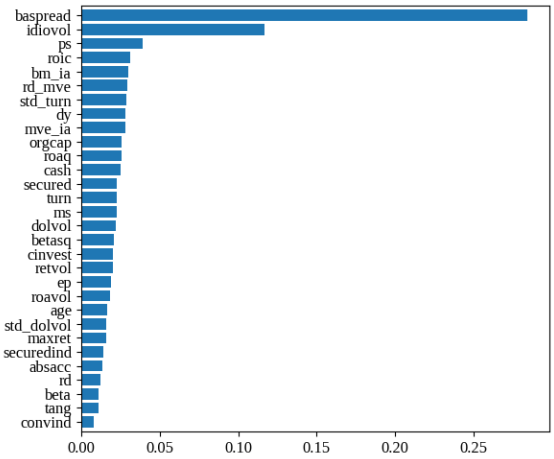
(a) Extra Trees feature importance



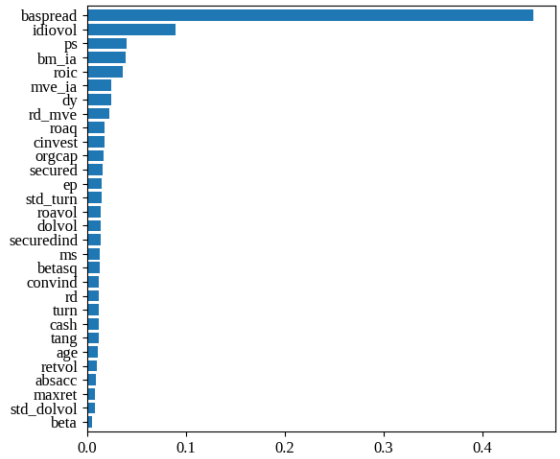
(b) NN feature importance



(c) Dart feature importance



(d) XGBoost feature importance



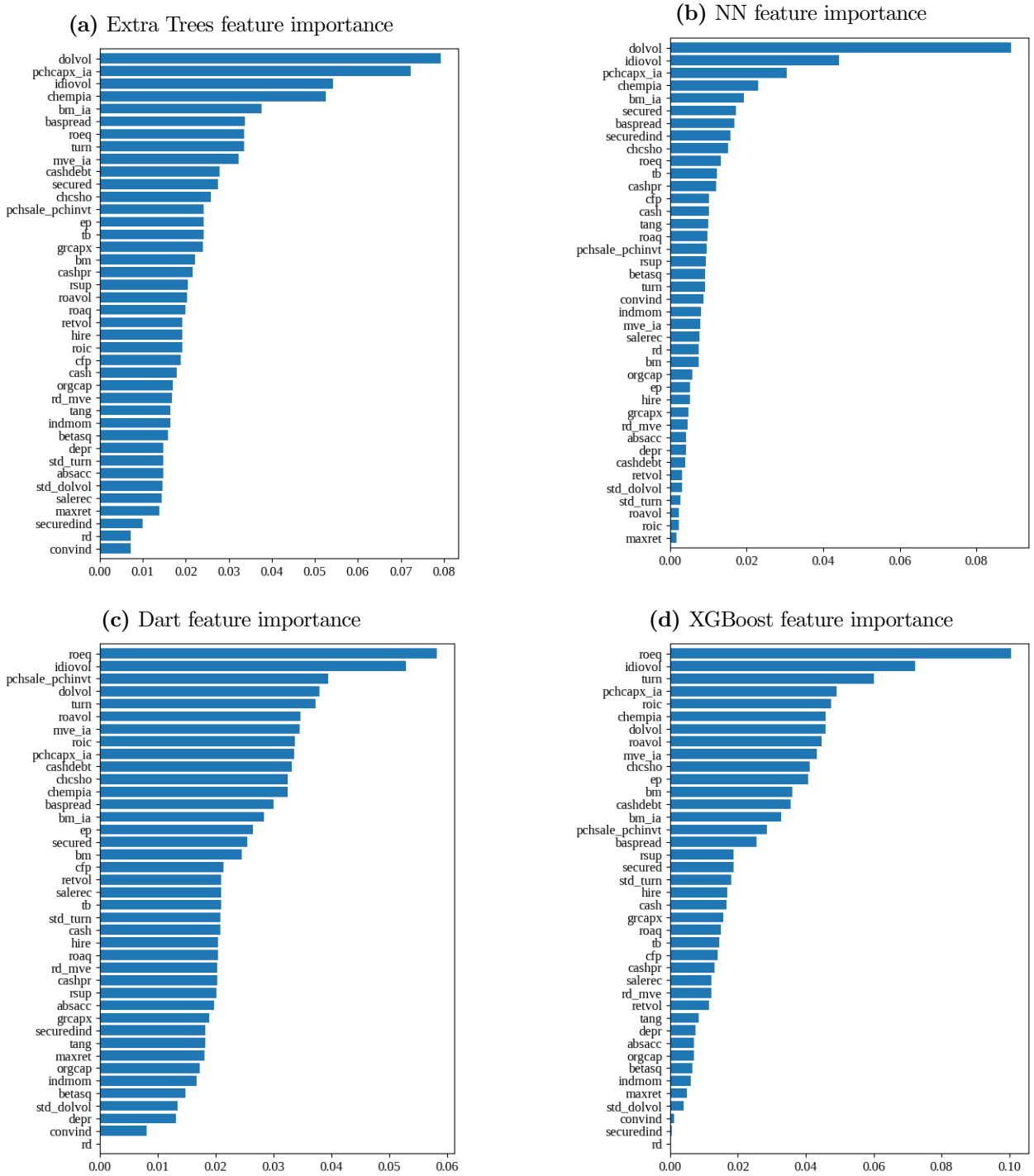
These figures display the variable importance of 30 features incorporated in the IV level predictions for the top four best performing models: Extra, Dart, NN, and XGBoost. The variable importance for each of the models is normalised to sum to one.

Upon analysing the feature importance for IV slope predictions depicted in Figure 8, we observe that the forest models exhibit similar feature importance scores, whereas the NN model displays noticeable differences. The NN still assigns large importance to a few variables while the other models have a more evenly distributed variable importance. A possible explanation of this can be the L2 regularisation terms that are added to the neural network, as mentioned in the model formulation in section 4.2.5. These regularisation terms minimise the impact of some variables by shrinking some of the weights towards zero.

All of the models agree on the importance of *idiovol* in predicting IV slopes. *Idiovol* denotes the idiosyncratic return volatility. The *idiovol* is the component of return volatility that is specific to that particular stock. It is to be expected that this variable is linked to the IV slope and contributes to an increased IVS when *idiovol* is large, and the corresponding stock return, therefore, volatile (Freire

and Kleen, 2023). To substantiate this finding, Chen et al., 2023a also include *idiovol* in their top ten fundamental variables for IV slope predictions.

Figure 8: Feature importance by model for IV slope predictions



These figures display the variable importance of 40 features incorporated in the IV slope predictions for the top four best performing models: Extra, Dart, NN, and XGBoost. The variable importance for each of the models is normalised to sum to one.

Other variables that are included in the top variables for IV slope prediction are *roeq* (return on equity), and *pchcapx_ia* (industry % change in capital expenditures). Similar to the IV level predictions, we observe a comparable ranking between the boosting algorithms (Dart and XGBoost) for the feature importance of IV slope predictions, as shown in Figure 20. Notably, both models assign a high importance to *roic* (return on invested capital), *roavol* (earnings volatility), *roeq* (return on equity) and *chcsho* (change in shares outstanding), while these variables do not rank as high in the feature importance of the other two models. This observation highlights the influence of model configuration on the processing of variable information the computation of predictions.

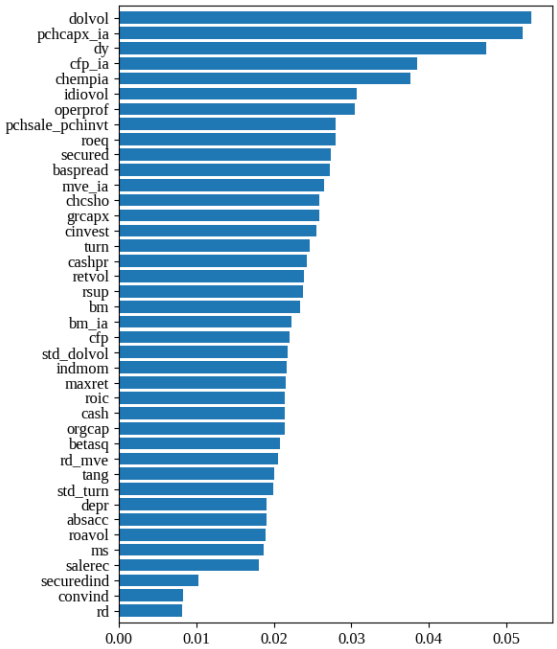
Lastly, we examine the feature importance of the IV curve predictions, displayed in figure 9. The feature importance rankings for IV curve predictions exhibit a similar pattern to the rankings for IV slope predictions. Once again, the NN model assigns high importance to a few variables, while the other models distribute importance more evenly among the variables. The variables that show to be most important for the IV curve predictions are similar to the IV slope prediction variables, *dolvol* (dollar trading volume) and *pchcapx_ia* (industry adjusted % change in capital expenditures) for the Extra Trees and NN, and *roeq* (return on equity) and *roic* (return on invested capital). These variable contribute most to the prediction of the IVS 'smile'.

A noteworthy variable that appears in the feature importance for IV curve prediction and consistently ranks among the top five variables for every model is *dy* (dividend to price). *Dy* is a valuation characteristic and has been previously identified by Freire and Kleen, 2023 as a factor that can account for cross-sectional differences in the implied volatility surface (IVS). Chen et al., 2023a also find *dy* to be of importance for IVS prediction, contrasting to our research they find *dy* to be amongst the top ten fundamentals for IV slope prediction as opposed to IV curvature.

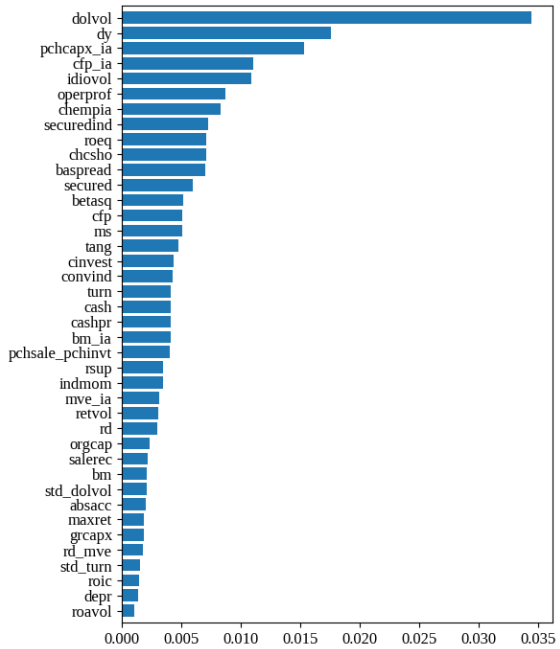
Overall, the variable importance analysis offers valuable insights into the feature importance for the different IVS prediction problems. It highlights the differences and similarities in the variable selection for the best-performing machine learning models. Lastly, this information can help to identify the most important features for future modelling, potentially leading to a reduction in the number of features used without compromising predictive power.

Figure 9: Feature importance by model for IV curve predictions

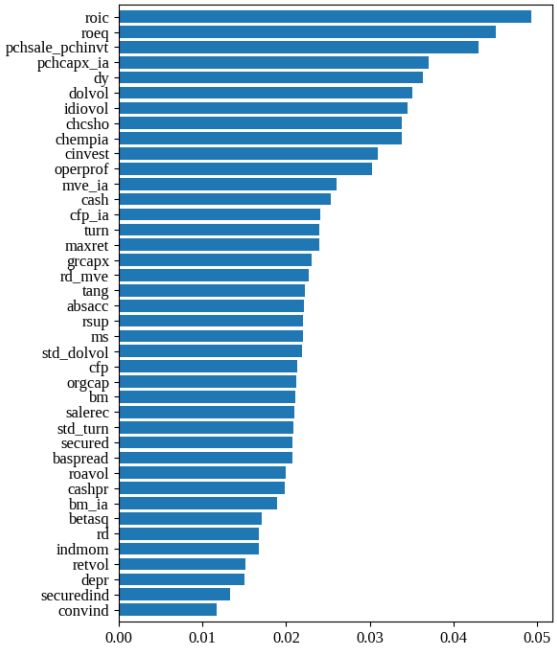
(a) Extra Trees feature importance



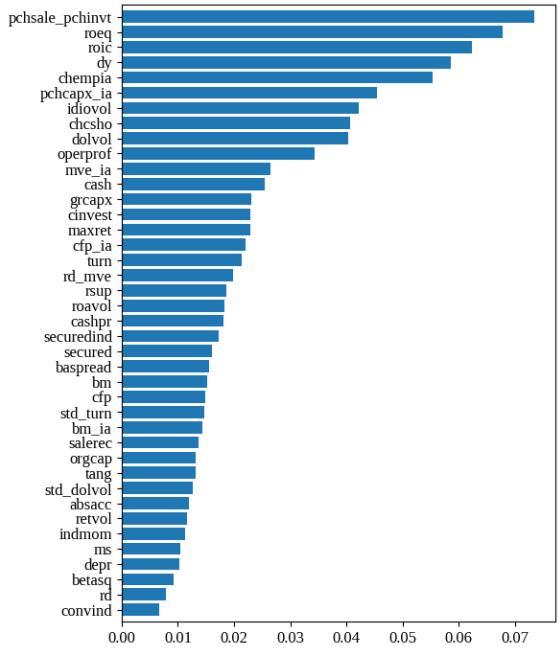
(b) NN feature importance



(c) Dart feature importance



(d) XGBoost feature importance



These figures display the variable importance of 40 features incorporated in the IV curve predictions for the top four best performing models: Extra, Dart, NN, and XGBoost. The variable importance for each of the models is normalised to sum to one.

5.3 Performance of Ensembles

After having analysed the predictive performance of individual models and examined the variables influencing the predictions of IV features, our focus now shifts to addressing the second research question: 'Can an optimised combination of models yield better results in predicting implied volatility features compared to individual models?' As described in section 4.3, in this paper we propose three ensemble configurations. The first ensemble consists of three different categories of models, a regularisation technique (Elnet), a forest model (Extra Trees) and a feedforward neural network (NN). The second ensemble builds upon the individual predictive performance of the proposed machine learning models and combines the best perfoming models in an ensemble (Extra Trees, Dart, XGBoost). The third ensemble is similar to the second one but with the inclusion of an additional neural network. This addition aims to expand the ensemble's capability to capture more complex relationships in the data. Furthermore, for all three ensembles, the models are both combined using a stacking procedure, as well as an equally weighted combination.

Figure 10 visualises the course of R^2_{OOS} values over the out-of-sample period (2012 - 2021), for all three ensembles. Each of the prediction problems is displayed in a row of three subfigures. The first row of subfigures displays the R^2_{OOS} values for the IV level predictions, the second row is for the IV slope predictions and the third row is for the IV curvature. The column of subfigures each show one of the three ensembles. Within each subfigure both the equally weighted and the stacking ensembles are presented.

In all figures we can clearly see the effect of the COVID-19 crisis on the IV feature predictions from the dip in predictive performance over 2019 and 2020. The pandemic has interfered with the established relations between variables and IV features, resulting in a negative impact on their predictability. However, following the crisis period, the R^2_{OOS} values in all figures show a tendency to revert to their former levels. An additional observation from Figure 10, consistent across all figures, is that the stacking procedure (indicated by the light green graph) produces less favourable predictions compared to an equally weighted ensemble (represented by the dark green line). The stacking method relies on historical data for model selection, making it vulnerable to breakdowns when faced with unpredictable events like the COVID-19 crisis. In such cases, it would be more advantageous to average out prediction errors. The stacking method is better suited for relatively consistent data, whereas for data sensitive to unpredictable events, the equally weighted ensemble outperforms the stacking procedures for (nearly) every out-of-sample year, across all models and IV features. The predictive performance for IV slope and curvature exhibits an increase as we progress over the out of sample data, which is expected due to the increasing training sample size available for model training. Particularly, after the COVID-19 crisis, there is an evident increase of around 5% for IV slope and around 2.5% for IV curve compared to the previous levels before the crisis. However, for

level predictions, the predictive performance over 2021 does not exhibit an increase compared to the values before the COVID-19 crisis. This can be attributed to the fact that towards the end of the characteristics dataset, 40% of the values were missing (Figure 12). Additionally the dataset only contained information for a selection of stocks resulting in empty rows after the linking procedure (Figure 13). Despite using a cross-sectional median procedure to account for missing data, the set contains less valuable information for the last few years. As a result, the predictive performance for level predictions, we have found to gain the most from information in the predictor variables, is more sensitive to the information shortage.

We examine the first column of figures for the first ensemble. While the individual models (represented by gray lines) exhibit fluctuations and considerable deviations from each other, the ensembles often surpass their R^2_{OOS} levels and yield an optimal outcome. This indicates a significant improvement in the predictive performance of IV features by employing ensemble methods.

In the second ensemble, the individual model performances and those of the ensembles appear to be closely aligned. Apart from the COVID-19 years, the Stacking 2 ensemble consistently underperforms when compared to the individual models for IV slope and curve prediction. Remarkably, it even exhibits lower R^2_{OOS} values than the Stacking 1 model, which incorporates a broader but less qualified range of base learners. Therefore, we reject the hypothesis that a combination of optimal performing models used as base learners for a stacking ensemble leads to an improvement in the predictions over individual models.

The inclusion of a NN in the third equally weighted ensemble results in a marginal improvement in performance for level predictions (compared to ensemble 2), but does not influence the ensemble performance to a great extent. In terms of slope and curve predictions, particularly between the years 2016 to 2019, the NN makes a valuable contribution to improving the predictions; Compared to ensemble 2, the graphs lie notably higher and, moreover, tower above the graphs of individual models. Remarkably, the NN itself does not perform exceptionally well; it is the averaging of prediction mistakes that leads to visible predictive improvements.

Overall, the ensembles yield satisfactory results, outperforming the individual models. This merging of model architectures proves to be particularly effective when the base learners' individual performances are strong. Notably, the addition of an NN to the ensemble leads to noticeable improvements in predictive performance, particularly for slope and curvature predictions.

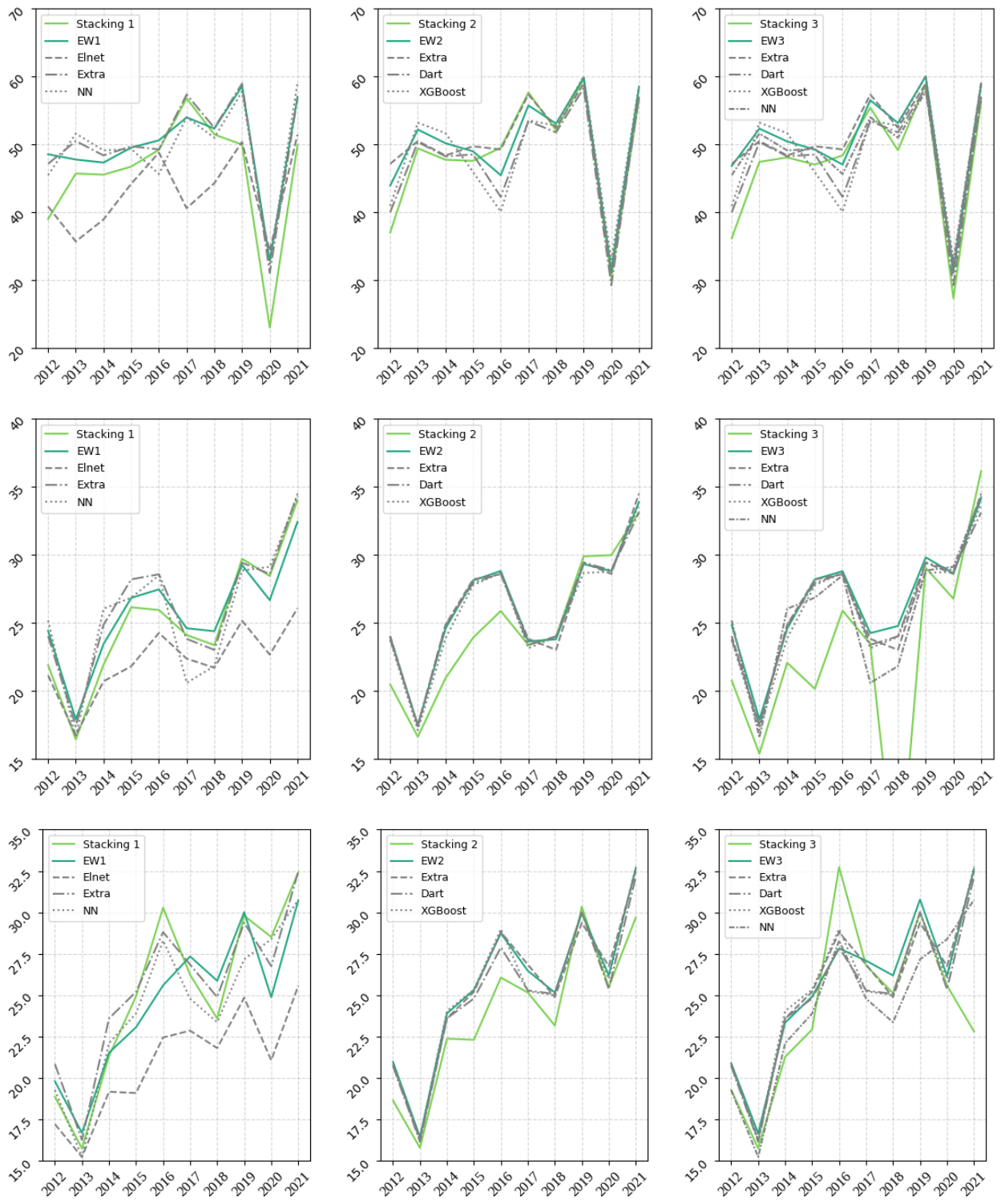


Figure 10: Out-of-sample IVS feature prediction performance (percentage R^2_{OOS}) for different ensembles

This figure displays the out-of-sample R^2_{OOS}, in percentages for ensembles, computed in two different ways, and the models from which they are constructed. The rows of subfigures portray the predictive performance for each feature (f.t.t.b Level, Slope, and Curve), and every column of subfigures shows a different type of ensemble (f.l.t.r. Ensemble 1, Ensemble 2, and Ensemble 3). The OOS sample ranges from January 2012 to December 2021.

6 Conclusion

The primary objective of this study was to investigate the predictability of option implied volatility features, namely level, slope, and curvature, using machine learning models. Additionally, the study explored whether combining models in ensembles would enhance the predictive performance of individual models. One of the key contributions of this research is the comprehensive comparison of a wide range of models in the relatively unexplored domain of option implied volatility of equity options. The analysis encompasses a traditional Black-Scholes model, several linear models (including Ordinary Least Squares with various numbers of predictors), and various machine learning models (Elastic Net, Random Forest, Extremely Randomised Forest, Dropout Additive Regression Trees, Extreme Gradient Boosted Trees, and Neural Network). Furthermore, the study proposes three ensemble combinations, all created through both a stacking and an equally weighted weighing approach.

The evaluation of the predictive performance of the models and ensembles for IV features, is divided into three parts: a comparative analysis of the model performances, feature importance analysis and examination of model ensembles. First, the performance of the individual machine learning methods is compared to the Black-Scholes benchmark and the linear models. The best performing models are Extremely randomised trees for the IV level ($R^2_{OOS} = \textit{0.487}$), Extremely randomised trees for the IV slope ($R^2_{OOS} = \textit{0.272}$), and an Ordinary least squares regression with 40 predictors for the IV curvature predictions ($R^2_{OOS} = \textit{0.262}$). The machine learning models significantly outperform the Black-Scholes model only for IV level prediction. However, the results form a promising starting point for further predictive improvements. The observed range of R^2_{OOS} for curve and slope features is considerably lower. This indicates that predicting the latter two features (slope and curve) is more challenging (level: *39.52%* to *49.13%*, slope: *20.51%* to *27.43%*, curve: *20.26%* to *26.24%*). Moreover, while satisfactory results are produced by the neural network for IV level and slope, this does not translate to the curve predictions. Since the neural network is the model with the highest potential complexity, this suggests that this complex structure does not generalise well to the IV curve. This leads to the conclusion that the correct model for IV curve predictions does not need to capture complex interactions and a simple OLS model is preferred. Furthermore, from this analysis, it follows that the extremely randomised regression trees and gradient boosted models (Dart and XGBoost) and neural network emerged as the best-performing models.

Secondly, the analysis feature importance assigned by the best performing models provides valuable insights into the differences and similarities in the use of information for IV feature predictions. For IV level prediction, *baspread* is consistently selected as one of the most important features by all models.

We also observe differences in variable importance between forest models and the neural network due to their different model configurations. Regarding slope predictions, all models agree on the importance of *idiolvol*. Lastly, for curve predictions, *dy* is identified as the most important feature. In terms of the variable selection the boosting methods display a similar ranking in feature importance. The IV level predictions are primarily influenced by a few variables, while all models, except for the elastic net, show a more evenly distributed variable importance for the IV slope and curvature predictions.

Lastly, the course of the predictive performance of ensembles is analysed. In general the ensembles produce satisfactory outcomes, demonstrating superior performance compared to the individual models for IV level ($R^2_{OOS} = 0.491$) and IV slope ($R^2_{OOS} = 0.274$). For curve predictions the best performing OLS-40 model is not outperformed by the ensembles. The integration of model architectures proves highly effective, especially when the base learners' individual performances are strong. Notably, the inclusion of an NN in the equally weighted ensemble results in noticeable enhancements in predictive performance, particularly for slope and curvature predictions. This implies that incorporating diverse model configurations adds significant value to model ensembles.

All in all, the primary conclusion is that machine learning models offer satisfactory enhancements for predicting IV shape features over traditional models. This discovery encourages further exploration into the implementation of machine learning models in the realm of IVS prediction. Another key finding is that predictions can be enhanced by integrating a simple equally weighted ensemble, as it consistently outperforms the individual models and yields satisfactory results even when the base learners may not be optimally suited for the predictions.

7 Discussion

In the discussion section, we address several limitations of our study and propose future research directions. Notably, our IV feature prediction conclusions are subject to a critical remark. The construction of features in our research relies on a straightforward approximation using formulations proposed by Chen et al., 2023a. Consequently, the conclusions drawn from our constructed features may not accurately reflect the true IV level, slope, or curve, as the models might not capture them precisely. To improve results, it is advisable to pursue more accurate estimation of IV features, such as obtaining values through polynomial fitting. This approach is expected to yield better outcomes, considering its closer relationship with the predictor variables.

Another limitation of our research, is the fact that the dataset obtained from Gu et al., 2020 contains a large percentage of missing values in the last two years, as evident in Figure 12. Moreover, the updating of the dataset appears to be limited after 2008, resulting in a lack of information for many stocks for which option IV features are predicted. Figure 13 illustrates the missing values after linking options to firm characteristics, indicating the absence of predictor variables for the stocks corersponding to the incorporated options. After 2008, the number of missing values is no longer constant, as depicted in Figure 14. It is observed that the number of options gradually increases. The decline in the number of options after 2011 originates from new options which are not traded for over 10 years and, therefore, not included in the dataset. With the rise of new options the number of missing values increases, indicating a lack of variable information available for newly incorporated options. To enhance results, we recommend improving the existing dataset by incorporating information on new options and addressing missing values. This may involve retrieving the missing data from a different source or finding related variables to include, leading to a more complete and informative variable dataset. Such improvements in predictors are expected to significantly enhance prediction accuracy.

Despite the seemingly favourable results from the machine learning models, their performance improvements over the linear models are not statistically significant in most cases. This suggests that the differences observed in the performance measures may be due to random fluctuations or noise rather than genuine superiority of the machine learning models. Therefore, while the machine learning models show potential based on their DM statistics, further analysis and evaluation are needed to establish their superiority over the linear models with more robust statistical tests. For this research we make use of options with a time to maturity of 30 days. To substantiate our findings it is advised to repeat the methods for a range of maturities or on a selection of the IV dataset. By sorting the observations on, for example, firm size, we are able to check whether the models produce consistent results.

Due to limited computational resources and time constraints, we employed static hyperparameter tuning for some machine learning models, potentially hindering the models from reaching their full potential. To overcome this limitation, we propose the adoption of dynamic hyperparameter tuning, selecting hyperparameters based on the training data for each iteration. Furthermore, the models' performance could be enhanced by employing dynamic feature selection. Currently, a fixed set of predictors is chosen at the initial loop and used throughout all iterations. However, since the importance of features can vary over the training sample, selecting a new set of features at each training iteration is likely to improve model performance.

This paper examines only three proposed model combinations and two weight assigning formulations. As the equally weighted model ensemble consistently outperforms individual models, further research into ensemble construction is expected to yield even better results. Potential advancements could involve adopting a revised weighing scheme based on characteristics in the input data, assigning larger weights to models that have demonstrated improved predictive power due to certain data changes, while reducing the weight of models that do not respond to such information. Enhanced predictive power can also be achieved through different model combinations. Additionally, an extension on ensembles could involve developing a switching model that selects the appropriate model based on observed data states. This could be combined with a classification model to detect significant periods, such as recession periods.

Finally, we recommend exploring the practical applications of our findings in option pricing, investment strategies, and portfolio formation.

References

Ackerer, D., Tagasovska, N., & Vatter, T. (2020). Deep smoothing of the implied volatility surface. *Advances in Neural Information Processing Systems*, *33*, 11552–11563.

Almeida, C., Fan, J., Freire, G., & Tang, F. (2022). Can a machine correct option pricing models? *Journal of Business & Economic Statistics*, 1–14.

Amilon, H. (2003). A neural network versus black–scholes: A comparison of pricing and hedging performances. *Journal of Forecasting*, *22*(4), 317–335.

Audrino, F., & Colangelo, D. (2010). Semi-parametric forecasts of the implied volatility surface using regression trees. *Statistics and Computing*, *20*(4), 421–434.

Audrino, F., & Knaus, S. D. (2016). Lassoing the har model: A model selection perspective on realized volatility dynamics. *Econometric Reviews*, *35*(8-10), 1485–1521.

Audrino, F., Sigrist, F., & Ballinari, D. (2020). The impact of sentiment and attention measures on stock market volatility. *International Journal of Forecasting*, *36*(2), 334–357.

Avramov, D., Cheng, S., & Metzker, L. (2023). Machine learning vs. economic restrictions: Evidence from stock return predictability. *Management Science*, *69*(5), 2587–2619.

Bakshi, G., Cao, C., & Chen, Z. (1997). Empirical performance of alternative option pricing models. *The Journal of finance*, *52*(5), 2003–2049.

Bali, T. G., Beckmeyer, H., Moerke, M., & Weigert, F. (2021a). Option return predictability with machine learning and big data. *The Review of Financial Studies (forthcoming)*.

Bali, T. G., Goyal, A., Huang, D., Jiang, F., & Wen, Q. (2020). Predicting corporate bond returns: Merton meets machine learning. *Georgetown McDonough School of Business Research Paper*, (3686164), 20–110.

Bali, T. G., Goyal, A., Huang, D., Jiang, F., & Wen, Q. (2021b). Different strokes: Return predictability across stocks and bonds with machine learning and big data. *Swiss Finance Institute, Research Paper Series*, (20-110).

Bali, T. G., Hu, J., & Murray, S. (2019). Option implied volatility, skewness, and kurtosis and the cross-section of expected stock returns. *Georgetown McDonough School of Business Research Paper*.

Ball, C. A., & Torous, W. N. (1985). On jumps in common stock prices and their impact on call option pricing. *The Journal of Finance*, *40*(1), 155–173.

Ballings, M., Van den Poel, D., Hespeels, N., & Gryp, R. (2015). Evaluating multiple classifiers for stock

price direction prediction. *Expert systems with Applications*, *42*(20), 7046–7056.

Basak, S., Kar, S., Saha, S., Khaidem, L., & Dey, S. R. (2019). Predicting the direction of stock market prices using tree-based classifiers. *The North American Journal of Economics and Finance*, *47*, 552–567.

Bates, D. S. (1996). 20 testing option pricing models. *Handbook of statistics*, *14*, 567–611.

Beckers, S. (1980). The constant elasticity of variance model and its implications for option pricing. *the Journal of Finance*, *35*(3), 661–673.

Bennell, J., & Sutcliffe, C. (2004). Black–scholes versus artificial neural networks in pricing ftse 100 options. *Intelligent Systems in Accounting, Finance & Management: International Journal*, *12*(4), 243–260.

Bianchi, D., Büchner, M., & Tamoni, A. (2021). Bond risk premiums with machine learning. *The Review of Financial Studies*, *34*(2), 1046–1089.

Black, F., & Scholes, M. (1973). The pricing of options and corporate liabilities. *Journal of political economy*, *81*(3), 637–654.

Breiman, L. (2001). Random forests. *Machine learning*, *45*, 5–32.

Brownlee, J. (2020). *Data preparation for machine learning: Data cleaning, feature selection, and data transforms in python*. Machine Learning Mastery.

Bryzgalova, S., Pelger, M., & Zhu, J. (2020). Forest through the trees: Building cross-sections of stock returns. *Available at SSRN 3493458*.

Bucci, A. (2020). Realized volatility forecasting with neural networks. *Journal of Financial Econometrics*, *18*(3), 502–531.

Campbell, J. Y., & Thompson, S. B. (2008). Predicting excess stock returns out of sample: Can anything beat the historical average? *The Review of Financial Studies*, *21*(4), 1509–1531.

Caporin, M., & Poli, F. (2017). Building news measures from textual data and an application to volatility forecasting. *Econometrics*, *5*(3), 35.

Carr, P., & Wu, L. (2009). Variance risk premiums. *The Review of Financial Studies*, *22*(3), 1311–1341.

Carr, P., & Wu, L. (2016). Analyzing volatility risk and risk premium in option contracts: A new theory. *Journal of Financial Economics*, *120*(1), 1–20.

Carr, P., Wu, L., & Zhang, Z. (2019). Using machine learning to predict realized variance. *arXiv preprint arXiv:1909.10035*.

Chandrashekar, G., & Sahin, F. (2014). A survey on feature selection methods. *Computers & Electrical Engineering*, *40*(1), 16–28.

Chang, B. Y., Christoffersen, P., & Jacobs, K. (2013). Market skewness risk and the cross section of stock returns. *Journal of Financial Economics*, *107*(1), 46–68.

Chang, B.-Y., Christoffersen, P., Jacobs, K., & Vainberg, G. (2012). Option-implied measures of equity risk. *Review of Finance*, *16*(2), 385–428.

Chen, D., Guo, B., & Zhou, G. (2023a). Firm fundamentals and the cross-section of implied volatility shapes. *Journal of Financial Markets*, *63*, 100771.

Chen, L., Pelger, M., & Zhu, J. (2023b). Deep learning in asset pricing. *Management Science*.

Chinco, A., Clark-Joseph, A. D., & Ye, M. (2019). Sparse signals in the cross-section of returns. *The Journal of Finance*, *74*(1), 449–492.

Chowdhury, R., Mahdy, M., Alam, T. N., Al Quaderi, G. D., & Rahman, M. A. (2020). Predicting the stock price of frontier markets using machine learning and modified black–scholes option pricing model. *Physica A: Statistical Mechanics and its Applications*, *555*, 124444.

Christensen, K., Siggaard, M., & Veliyev, B. (2021). A machine learning approach to volatility forecasting. *Available at SSRN*.

Christoffersen, P., Fournier, M., & Jacobs, K. (2018). The factor structure in equity options. *The Review of Financial Studies*, *31*(2), 595–637.

Clements, A., & Vasnev, A. L. (2021). Forecast combination puzzle in the har model. *Available at SSRN 3875026*.

Corsi, F. (2009). A simple approximate long-memory model of realized volatility. *Journal of Financial Econometrics*, *7*(2), 174–196.

Cox, J. C., & Ross, S. A. (1976). The valuation of options for alternative stochastic processes. *Journal of financial economics*, *3*(1-2), 145–166.

Das, S. P., & Padhy, S. (2017). A new hybrid parametric and machine learning model with homogeneity hint for european-style index option pricing. *Neural Computing and Applications*, *28*, 4061–4077.

De Spiegeleer, J., Madan, D. B., Reyners, S., & Schoutens, W. (2018). Machine learning for quantitative finance: Fast derivative pricing, hedging and fitting. *Quantitative Finance*, *18*(10), 1635–1643.

Dennis, P., & Mayhew, S. (2002). Risk-neutral skewness: Evidence from stock options. *Journal of Financial and Quantitative Analysis*, *37*(3), 471–493.

Diavatopoulos, D., Doran, J. S., Fodor, A., & Peterson, D. R. (2012). The information content of implied skewness and kurtosis changes prior to earnings announcements for stock and option returns. *Journal of Banking & Finance*, *36*(3), 786–802.

Diebold, F. X., & Mariano, R. S. (2002). Comparing predictive accuracy. *Journal of Business & economic*

statistics, *20*(1), 134–144.

Dietterich, T. G. (2000). Ensemble methods in machine learning. *Multiple Classifier Systems: First International Workshop, MCS 2000 Cagliari, Italy, June 21–23, 2000 Proceedings 1*, 1–15.

Donaldson, R. G., & Kamstra, M. (1997). An artificial neural network-garch model for international stock return volatility. *Journal of Empirical Finance*, *4*(1), 17–46.

Dörries, J. (2021). Decomposed higher-moment risk premiums and market return predictability. *Available at SSRN 3784496*.

Dugas, C., Bengio, Y., Bélisle, F., Nadeau, C., & Garcia, R. (2009). Incorporating functional knowledge in neural networks. *Journal of Machine Learning Research*, *10*(6).

Dumas, B., Fleming, J., & Whaley, R. E. (1998). Implied volatility functions: Empirical tests. *The Journal of Finance*, *53*(6), 2059–2106.

Elyasiani, E., Gambarelli, L., & Muzzioli, S. (2020). Moment risk premia and the cross-section of stock returns in the european stock market. *Journal of Banking & Finance*, *111*, 105732.

Ewing, J. A. (2010). *Comparison of implied volatility approximations using "nearest-to-the-money" option premiums* (Doctoral dissertation). Clemson University.

Félix, L., Kräussl, R., & Stork, P. (2020). Implied volatility sentiment: A tale of two tails. *Quantitative Finance*, *20*(5), 823–849.

Fernandes, M., Medeiros, M. C., & Scharth, M. (2014). Modeling and predicting the cboe market volatility index. *Journal of Banking & Finance*, *40*, 1–10.

Freire, G., & Kleen, O. (2023). Equity options and firm characteristics. *Available at SSRN 4342597*.

Friedman, J. H. (2001). Greedy function approximation: A gradient boosting machine. *Annals of statistics*, 1189–1232.

Galindo, J., & Tamayo, P. (2000). Credit risk assessment using statistical and machine learning: Basic methodology and risk modeling applications. *Computational economics*, *15*, 107–143.

Ganaie, M. A., Hu, M., Malik, A., Tanveer, M., & Suganthan, P. (2022). Ensemble deep learning: A review. *Engineering Applications of Artificial Intelligence*, *115*, 105151.

Garcia, R., & Gençay, R. (2000). Pricing and hedging derivative securities with neural networks and a homogeneity hint. *Journal of Econometrics*, *94*(1-2), 93–115.

Genre, V., Kenny, G., Meyler, A., & Timmermann, A. (2013). Combining expert forecasts: Can anything beat the simple average? *International Journal of Forecasting*, *29*(1), 108–121.

Geske, R. L., & Zhou, Y. (2009). *Capital structure effects on prices of firm stock options: Tests using implied market values of corporate debt*. SSRN.

Geurts, P., Ernst, D., & Wehenkel, L. (2006). Extremely randomized trees. *Machine learning*, *63*, 3–42.

Ghosh, I., & Sanyal, M. K. (2021). Introspecting predictability of market fear in indian context during covid-19 pandemic: An integrated approach of applied predictive modelling and explainable ai. *International Journal of Information Management Data Insights*, *1*(2), 100039.

Giglio, S., Kelly, B., & Xiu, D. (2022). Factor models, machine learning, and asset pricing. *Annual Review of Financial Economics*, *14*, 337–368.

Goncalves, S., & Guidolin, M. (2006). Predictable dynamics in the s&p 500 index options implied volatility surface. *The Journal of Business*, *79*(3), 1591–1635.

Green, J., Hand, J. R., & Zhang, X. F. (2017). The characteristics that provide independent information about average us monthly stock returns. *The Review of Financial Studies*, *30*(12), 4389–4436.

Gu, S., Kelly, B., & Xiu, D. (2020). Empirical asset pricing via machine learning. *The Review of Financial Studies*, *33*(5), 2223–2273.

Guyon, I., & Elisseeff, A. (2003). An introduction to variable and feature selection. *Journal of machine learning research*, *3*(Mar), 1157–1182.

Han, Y., He, A., Rapach, D., & Zhou, G. (2022). Expected stock returns and firm characteristics: E-enet, assessment, and implications. *Assessment, and Implications (August 28, 2022)*.

He, X., Feng, G., Wang, J., & Wu, C. (2021). Predicting individual corporate bond returns. *Available at SSRN 4374213*.

Heston, S. L. (1993). A closed-form solution for options with stochastic volatility with applications to bond and currency options. *The review of financial studies*, *6*(2), 327–343.

Hillebrand, E., & Medeiros, M. C. (2010). The benefits of bagging for forecast models of realized volatility. *Econometric Reviews*, *29*(5-6), 571–593.

Hull, J., & White, A. (1987). The pricing of options on assets with stochastic volatilities. *The journal of finance*, *42*(2), 281–300.

Hull, J., & White, A. (2017). Optimal delta hedging for options. *Journal of Banking & Finance*, *82*, 180–190.

Hung, C., & Chen, J.-H. (2009). A selective ensemble based on expected probabilities for bankruptcy prediction. *Expert systems with applications*, *36*(3), 5297–5303.

Hutchinson, J. M., Lo, A. W., & Poggio, T. (1994). A nonparametric approach to pricing and hedging derivative securities via learning networks. *The journal of Finance*, *49*(3), 851–889.

Isengildina-Massa, O., Curtis Jr, C. E., Bridges, W., & Nian, M. (2007). *Accuracy of implied volatility approximations using" nearest-to-the-money" option premiums* (tech. rep.).

Itkin, A. (2015). To sigmoid-based functional description of the volatility smile. *The North American Journal of Economics and Finance*, *31*, 264–291.

Kaeck, A. (2018). Variance-of-variance risk premium. *Review of Finance*, *22*(4), 1549–1579.

Kim, H. Y., & Won, C. H. (2018). Forecasting the volatility of stock price index: A hybrid model integrating lstm with multiple garch-type models. *Expert Systems with Applications*, *103*, 25–37.

Kingma, D. P., & Ba, J. (2014). Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*.

Krauss, C., Do, X. A., & Huck, N. (2017). Deep neural networks, gradient-boosted trees, random forests: Statistical arbitrage on the s&p 500. *European Journal of Operational Research*, *259*(2), 689–702.

Langlois, H. (2020). Measuring skewness premia. *Journal of Financial Economics*, *135*(2), 399–424.

Lee, S., Lee, J., Shim, D., & Jeon, M. (2007). Binary particle swarm optimization for black-scholes option pricing. *Knowledge-Based Intelligent Information and Engineering Systems: 11th International Conference, KES 2007, XVII Italian Workshop on Neural Networks, Vietri sul Mare, Italy, September 12-14, 2007. Proceedings, Part I 11*, 85–92.

Li, H., Cao, Y., Li, S., Zhao, J., & Sun, Y. (2020). Xgboost model and its application to personal credit evaluation. *IEEE Intelligent Systems*, *35*(3), 52–61.

Li, S. (2005). A new formula for computing implied volatility. *Applied mathematics and computation*, *170*(1), 611–625.

Liu, D., Liang, Y., Zhang, L., Lung, P., & Ullah, R. (2021). Implied volatility forecast and option trading strategy. *International Review of Economics & Finance*, *71*, 943–954.

Liu, J., Zhang, S., & Fan, H. (2022). A two-stage hybrid credit risk prediction model based on xgboost and graph-based deep neural network. *Expert Systems with Applications*, *195*, 116624.

Liu, S., Oosterlee, C. W., & Bohte, S. M. (2019). Pricing options and computing implied volatilities using neural networks. *Risks*, *7*(1), 16.

Luong, C., & Dokuchaev, N. (2018). Forecasting of realised volatility with the random forests algorithm. *Journal of Risk and Financial Management*, *11*(4), 61.

Ma, Y., Han, R., & Wang, W. (2021). Portfolio optimization with return prediction using deep learning and machine learning. *Expert Systems with Applications*, *165*, 113973.

Malliaris, M., & Salchenberger, L. (1996). Using neural networks to forecast the s&p 100 implied volatility. *Neurocomputing*, *10*(2), 183–195.

Mayhew, S. (1995). Implied volatility. *Financial Analysts Journal*, *51*(4), 8–20.

Merton, R. C. (1973). Theory of rational option pricing. *The Bell Journal of economics and management science*, 141–183.

Mittnik, S., Robinzonov, N., & Spindler, M. (2015). Stock market volatility: Identifying major drivers and the nature of their impact. *Journal of banking & Finance*, *58*, 1–14.

Muzzioli, S. (2010). Option-based forecasts of volatility: An empirical study in the dax-index options market. *The European Journal of Finance*, *16*(6), 561–586.

Nagel, S. (2021). *Machine learning in asset pricing* (Vol. 1). Princeton University Press.

Pan, J. (2002). The jump-risk premia implicit in options: Evidence from an integrated time-series study. *Journal of financial economics*, *63*(1), 3–50.

Park, H., Kim, N., & Lee, J. (2014). Parametric models and non-parametric machine learning models for predicting option prices: Empirical comparison study over kospi 200 index options. *Expert Systems with Applications*, *41*(11), 5227–5237.

Pasupulety, U., Anees, A. A., Anmol, S., & Mohan, B. R. (2019). Predicting stock prices using ensemble learning and sentiment analysis. *2019 IEEE Second International Conference on Artificial Intelligence and Knowledge Engineering (AIKE)*, 215–222.

Patel, J., Shah, S., Thakkar, P., & Kotecha, K. (2015). Predicting stock and stock price index movement using trend deterministic data preparation and machine learning techniques. *Expert systems with applications*, *42*(1), 259–268.

Polamuri, S. R., Srinivas, K., & Mohan, A. K. (2019). Stock market prices prediction using random forest and extra tree regression. *Int. J. Recent Technol. Eng*, *8*(1), 1224–1228.

Poon, S.-H., & Granger, C. W. J. (2003). Forecasting volatility in financial markets: A review. *Journal of economic literature*, *41*(2), 478–539.

Qin, Q., Wang, Q.-G., Li, J., & Ge, S. S. (2013). Linear and nonlinear trading models with gradient boosted random forests and application to singapore stock market.

Rahimikia, E., & Poon, S.-H. (2020). Machine learning for realised volatility forecasting. *Available at SSRN*, *3707796*.

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

Rapach, D. E., Strauss, J. K., & Zhou, G. (2013). International stock return predictability: What is the role of the united states? *The Journal of Finance*, *68*(4), 1633–1662.

Sadorsky, P. (2022). Using machine learning to predict clean energy stock prices: How important are market volatility and economic policy uncertainty? *Journal of Climate Finance*, 100002.

Sasaki, H. (2016). The skewness risk premium in equilibrium and stock return predictability. *Annals of Finance*, *12*(1), 95–133.

Sylvester Walusala, W., Rimiru, R., & Otieno, C. (2017). A hybrid machine learning approach for credit scoring using pca and logistic regression. *International Journal of Computer (IJC)*, *27*(1), 84–102.

Tang, Y., Song, Z., Zhu, Y., Yuan, H., Hou, M., Ji, J., Tang, C., & Li, J. (2022). A survey on machine learning models for financial time series forecasting. *Neurocomputing*, *512*, 363–380.

Varian, H. R. (2014). Big data: New tricks for econometrics. *Journal of Economic Perspectives*, *28*(2), 3–28.

Vinayak, R. K., & Gilad-Bachrach, R. (2015). Dart: Dropouts meet multiple additive regression trees. *Artificial Intelligence and Statistics*, 489–497.

Vrontos, S. D., Galakis, J., & Vrontos, I. D. (2021). Implied volatility directional forecasting: A machine learning approach. *Quantitative Finance*, *21*(10), 1687–1706.

Wang, C.-P., Lin, S.-H., Huang, H.-H., & Wu, P.-C. (2012). Using neural network for forecasting txo price under different volatility models. *Expert Systems with Applications*, *39*(5), 5025–5032.

Wang, W., Li, W., Zhang, N., & Liu, K. (2020a). Portfolio formation with preselection using deep learning from long-term financial data. *Expert Systems with Applications*, *143*, 113042.

Wang, Y., Zhang, Y., Lu, Y., & Yu, X. (2020b). A comparative assessment of credit risk model based on machine learning——a case study of bank loan data. *Procedia Computer Science*, *174*, 141–149.

Witten, I. H., & Frank, E. (2002). Data mining: Practical machine learning tools and techniques with java implementations. *Acm Sigmod Record*, *31*(1), 76–77.

Yang, Y., Zheng, Y., & Hospedales, T. (2017). Gated neural networks for option pricing: Rationality by design. *Proceedings of the AAAI conference on artificial intelligence*, *31*(1).

Ye, Z. J., & Schuller, B. W. (2021). Capturing dynamics of post-earnings-announcement drift using a genetic algorithm-optimized xgboost. *Expert Systems with Applications*, *177*, 114892.

Zaffaroni, P., & Zhou, G. (2022). Asset pricing: Cross-section predictability. *Available at SSRN 4111428*.

Zeng, Y., & Klabjan, D. (2019). Online adaptive machine learning based algorithm for implied volatility surface modeling. *Knowledge-Based Systems*, *163*, 376–391.

Zhang, Q., Liu, J., Tian, D., & Yue, H. (2021). Application of stacking ensemble learning in option implied volatility. *2021 2nd International Conference on Artificial Intelligence and Computer Engineering (ICAICE)*, 623–627.

Zheng, Y. (2017). *Machine learning and option implied information* (Doctoral dissertation). Imperial College London.

Zou, H., & Hastie, T. (2005). Regularization and variable selection via the elastic net. *Journal of the royal statistical society: series B (statistical methodology)*, *67*(2), 301–320.

Zulfiqar, N., & Gulzar, S. (2021). Implied volatility estimation of bitcoin options and the stylized facts of option pricing. *Financial Innovation*, *7*, 1–30.

Appendices

A Details on stock characteristics

The stock characteristics dataset is provided by Gu et al., 2020, who have incorporated several lags in the data, mimicking real-world delay. This lagging ensures that the most recent data available at the time of prediction is used, to avoid incorporating future information into our analysis. The release for most of the characteristics in their dataset is delayed. To mitigate any forward-looking bias Gu et al., 2020 incorporate lag periods. Monthly characteristics are considered up to the end of the previous month (month t), quarterly data is lagged by at least 4 months (end $t-4$), and annual data is lagged by at least 6 months (end $t-6$).

Table 3 gives the details on some of the most important characteristics used in this research. Details on the other 94 characteristics can be found in Gu et al., 2020.

B Data transformation

Table 11 shows the distribution of the number of missing values in the dataset provided by Gu et al., 2020. Figure 12, shows for which part of the missing values each variable account for.

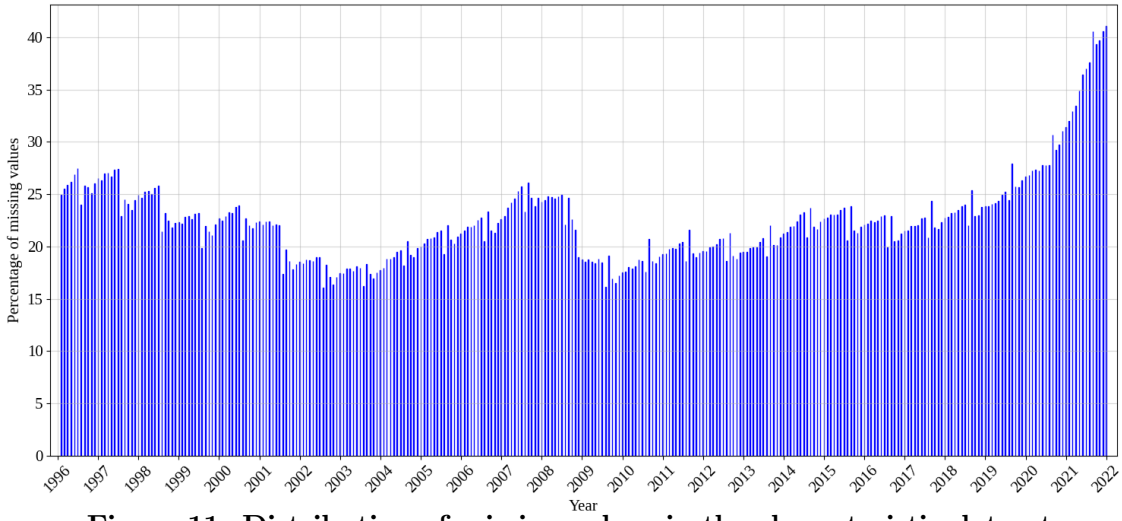


Figure 11: Distribution of missing values in the characteristic dataset

This figure displays the missing values in percentages in the stock characteristics provided by Gu et al., 2020, over an interval from 1996 to 2021.

After transforming the data by linking the option implied volatility data from *WRDS* to the corresponding stock characteristics, missing data emerge. This is due to the fact that the dataset of stock characteristics does not have information for every option present in the IVS dataset. The rise of missing values through the linking procedure is visualised in Figure 13

Table 3: Details of stock characteristics

Nr.	Name	Description
1	absacc	Absolute accruals
2	age	# years since first Compustat coverage
3	baspread	Bid-ask spread
4	beta	Beta
5	bm	Book-to-market
6	bm_ia	Industry-adjusted book to market
7	cash	Cash holdings
8	cashdebt	Cash flow to debt
9	cashpr	Cash productivity
10	cfp	Cash flow to price ratio
11	cfp_ia	Industry-adjusted cash flow to price ratio
12	chcsho	Change in shares outstanding
13	chempia	Industry-adjusted change in employees
14	cinvest	Corporate investment
15	convind	Convertible debt indicator
16	depr	Depreciation / PP&E
17	dolvol	Dollar trading volume
18	dy	Dividend to price
19	ep	Earnings to price
20	grcapx	Growth in capital expenditures
21	hire	Employee growth rate
22	idiovol	Idiosyncratic return volatility
23	indmom	Industry momentum
24	ms	Financial statement score
25	mve_ia	Industry-adjusted size
26	orgcap	Organizational capital
27	pchcapx_ia	Industry adjusted % change in capital expenditures
28	pchsale_pchinvt	% change in sales - % change in inventory
29	ps	Financial statements score
30	rd	R&D increase
31	rd_mve	R&D to market capitalization
32	retvol	Return volatility
33	roaq	Return on assets
34	roavol	Earnings volatility
35	roeq	Return on equity
36	roic	Return on invested capital
37	rsup	Revenue surprise
38	salerec	Sales to receivables
39	secured	Secured debt
40	securedind	Secured debt indicator
41	std_dolvol	Volatility of liquidity (dollar trading volume)
42	std_turn	Volatility of liquidity (share turnover)
43	tang	Debt capacity/firm tangibility
44	tb	Tax income to book income
45	turn	Share turnover

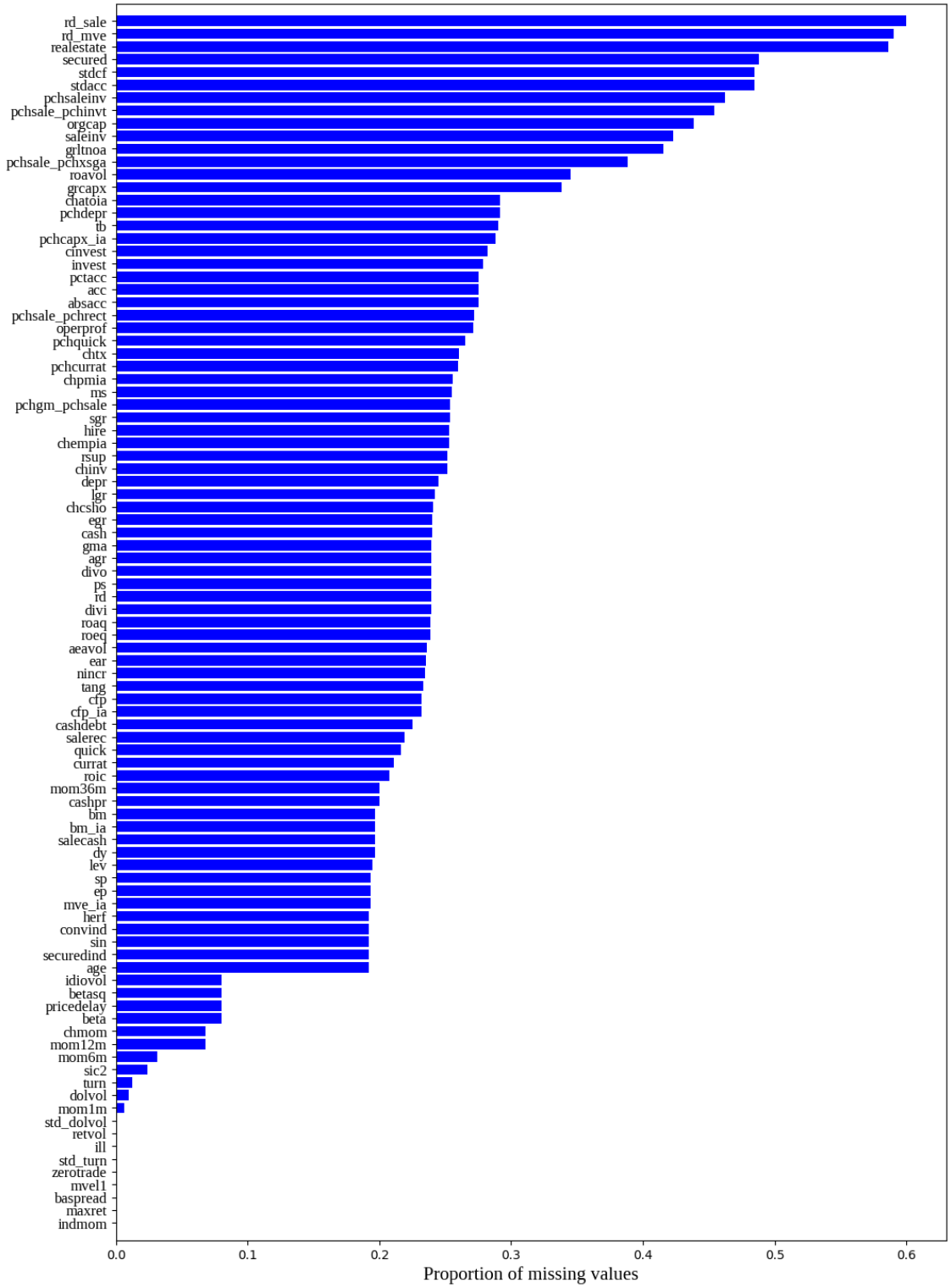


Figure 12: Portion of missing values per characteristic

This figure displays proportion of the total number of missing values accounted for by each characteristic in the dataset provided by Gu et al., 2020, over an interval from 1996 to 2021.

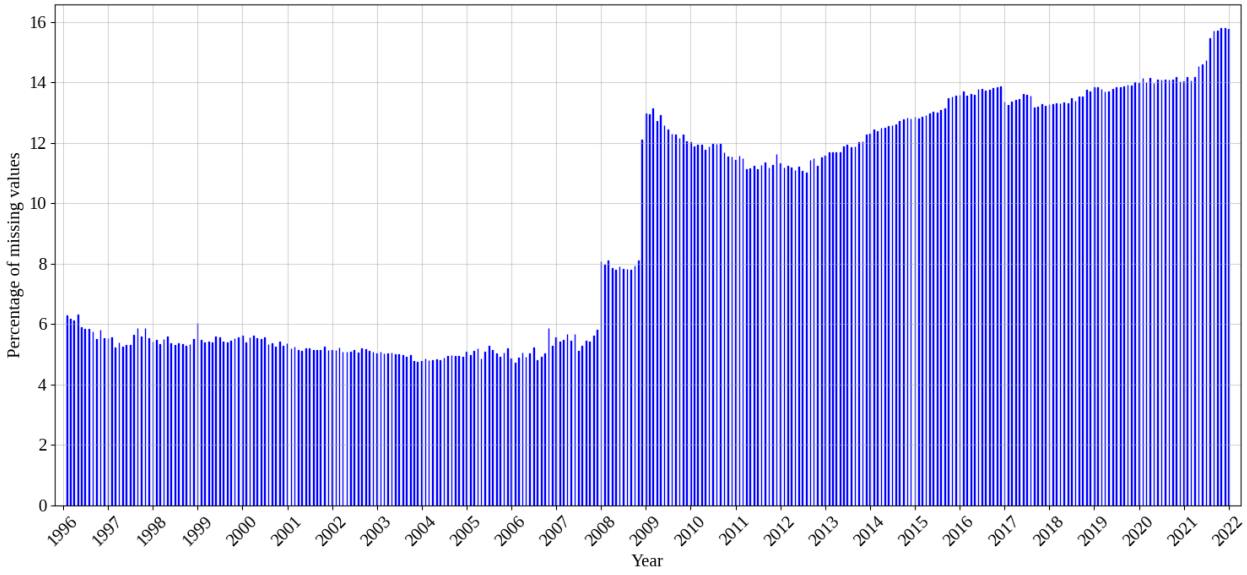


Figure 13: Percentage of missing values after linking

This figure displays the missing values in percentages in the stock characteristics that emerge when linking the characteristics dataset to the IVS data, over an interval from 1996 to 2021

To obtain a liquid dataset we only incorporate stocks for which the options are traded for more than ten years. The number of stocks for which IVS data is given, differs per month. This distribution is given in figure 14.

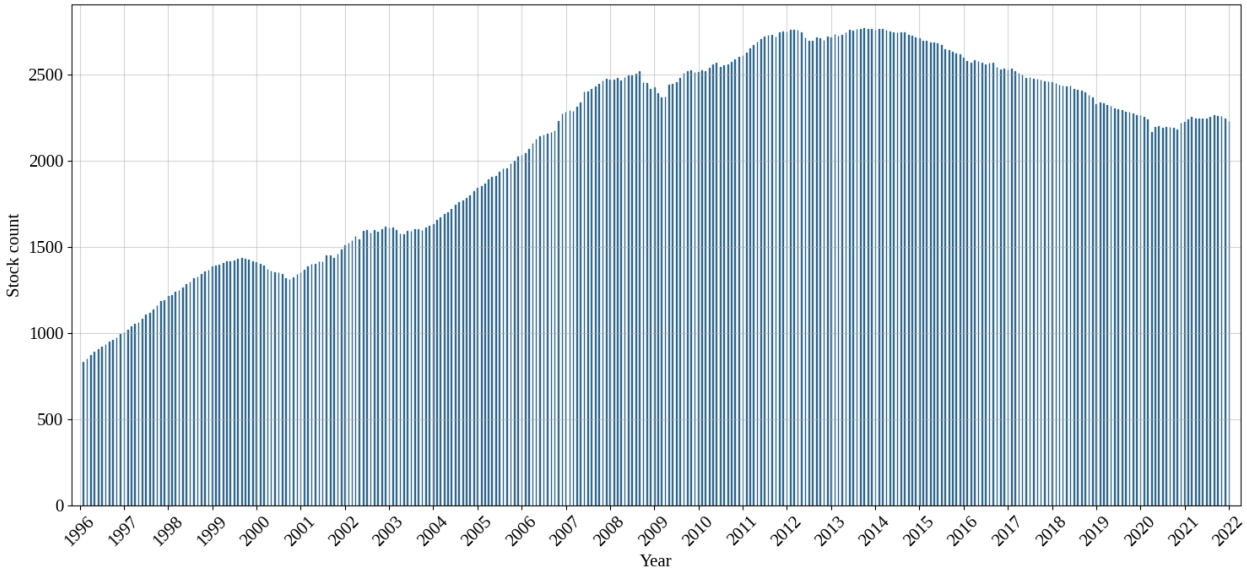


Figure 14: Number of stocks per month

This figure displays the number of stocks per month in the final dataset for which options are traded and IVS data is given, over an interval from 1996 to 2021

Based on in-the-money, at-the-money and out-of-the-money options, three different IVS shape features: level, slope and curvature, are computed. The distribution of these features is given in Figure 15.

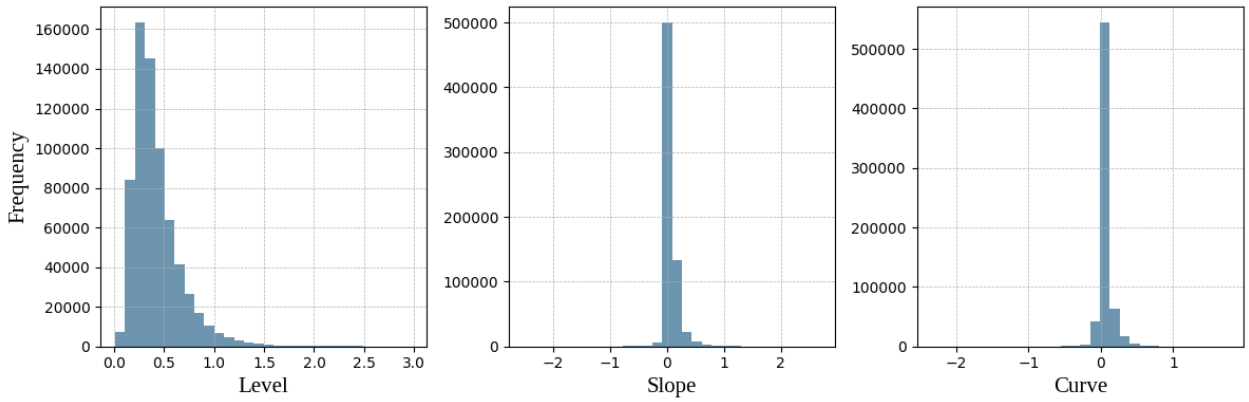


Figure 15: Distribution of IVS features

This figure displays the distribution of the values for (from left to right) the IVS level, slope and curvature, over an interval from 1996 to 2021

C List of common abbreviations

IVS	Implied Volatility Surface
IV	Implied Volatility
ITM	In-The-Money
ATM	At-The-Money
OTM	Out-The-Money
BS	Black-Scholes
OLS	Ordinary Least Squares
Elnet	Elastic Net
RF	Random Forest
Extra Trees	Extremely Randomised Trees
GBM	Gradient Boosting Model
Dart	Dropout Additive Regression Trees
XGBoost	Extreme Gradient Boosting
NN	Neural Network
EW	Equally Weighted ensemble

D Model tuning

D.1 Sample splitting

Figure 16 visualises the rolling window approach used for tuning the models and computing out of sample predictions.

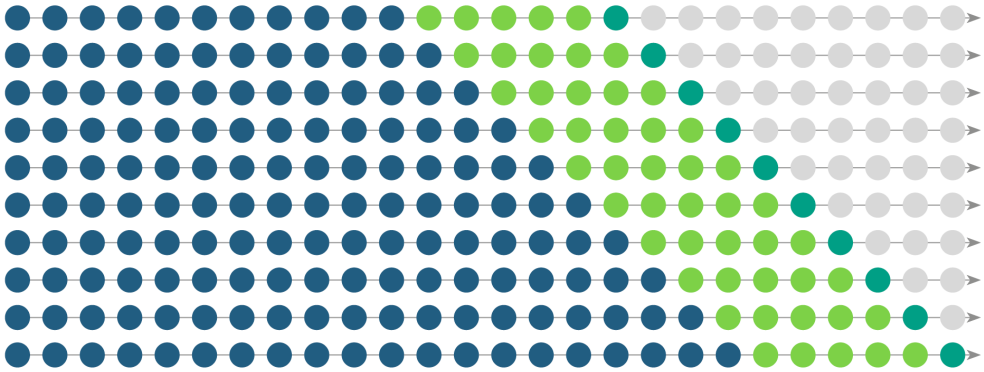


Figure 16: Expanding window training procedure

This figure illustrates the expanding rolling window strategy. The sample starts with a training sample of 11 years (1996 - 2006), a validation sample of 5 years (2007 - 2011) and a out-of-sample test set of 1 year (2021). After one year, the training sample is extended by a year and the models are retrained, while the sizes of the other samples remain constant. In the figure the dot represent one year.

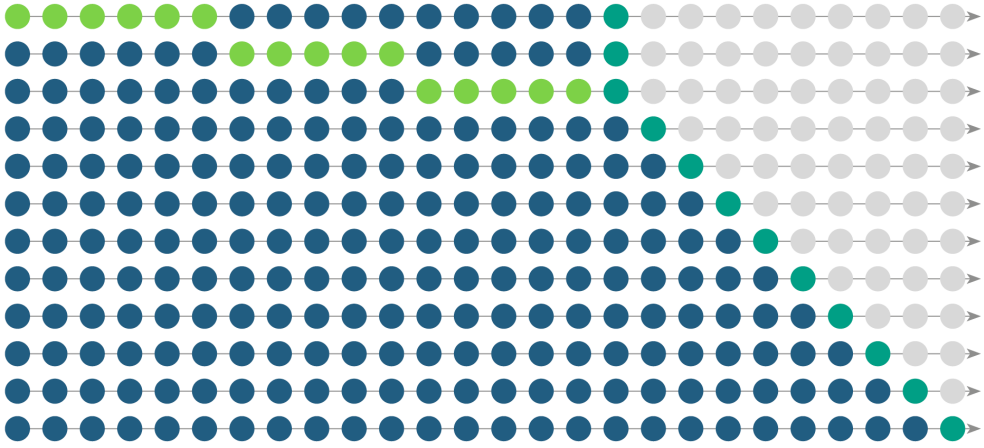


Figure 17: Static cross validation training procedure

This figure illustrates a static training strategy. The sample starts with a training and validation sample of 16 years on which a three fold crossvalidation is performed. After the first year, the training sample is extended by a year and the models are retrained using the hyperparameters found in the first loop.

D.2 Hyperparameter tuning

We make use of a dynamic tuning method, which entails that the hyperparameter tuning is performed for every loop. To examine the different combinations of hyperparameters, a model is fit on the training set $\mathcal{T}_1$ and a validation score is obtained from the predictions for the validation set $\mathcal{T}_2$. The parameter combination with the best score is used in the final model. Due to the computation time of the Random Forest model and the GBM models, a static tuning method is employed for the hyperparameter tuning of these methods. When using this approach, hyperparameter tuning is only performed once, after which the optimal combination of hyperparameters is used for each of the models in the following loops. A three fold crossvalidation is performed on the initial training and validation set. Consequently, the optimal combination of hyperparameters is determined and used for the remaining 9 loops.

The optimal hyperparameters for the various machine learning models are found by means of the aforementioned methods. Table 4 contains the grid of parameters for every model for the IV level prediction, Table 5 contains the grid of parameters for every model for the IV slope prediction and Table 6 contains the grid of parameters for every model for the IV curve prediction. In bold are the chosen parameters based on the static hyperparameter selection method.

Table 4: Hyperparameter grid for level predictions

	Elnet	**RF**	**Extra Trees**
Alpha	**0.01**, 0.05, 0.1, 0.15	-	-
L1_ratio	0.1	**0.01**, 1e-11, 1e-14, 1e-18	-
Max depth	-	3, 4, **5**, 6	6, 8, 9, 12, 14, **16**
Max features	-	5, **10**, 15	13, 17, **20**
Number of estimators	-	200, 300, **400**, 500	**100**, 200, 300, 400, 600

	Dart	**XGBoost**	**NN**
Max depth	3, 4, **5**, 6, 9	7, 9, **11**, 13	-
Number of estimators	200, **300**, 400, 500, 600	200, 400, 500, 600, **700**	-
Rate drop	**0.1**, 0.01	-	-
Learning rate	-	0.1, **0.01**, 0.001	-
Units	-	-	64, **32**, 16
Regularisation	-	-	1e-2, **1e-4**, **1e-6**

Table 5: Hyperparameter grid for slope predictions

	Elnet	RF	Extra Trees
Alpha	0.01, 0.05, 0.1, 0.15, **1e-7**	-	-
L1_ratio	0.7, 0.8, **0.9**	-	-
Max depth	-	**3**, 4, 5, 6	6, 7, 8, 9, 12, **13**, 14, 16
Max features	-	5, 10, **15**, 20, 30	13, 15, 17, **20**, 30
Number of estimators	-	200, 300, 400, 500, **600**, 700	100, 200, 300, 400, **500**, 600

	Dart	XGBoost	NN
Max depth	3, **4**, 5, 6, 9	**3**, 5, 7, 9, 11, 13	-
Number of estimators	200, 300, **400**, 500, 600	200, 400, 500, 600, **700**	-
Rate drop	**0.1**, 0.01	-	-
Learning rate	-	0.1, **0.01**, 0.001	-
Units	-	-	64, **32**, 16
Regularisation	-	-	1e-2, **1e-4**, **1e-6**

Table 6: Hyperparameter grid for curve predictions

	Elnet	RF	Extra Trees
Alpha	**1e-15**, 1e-17, 1e-19	-	-
L1_ratio	0.2, 0.5, **0.7**	-	-
Max depth	-	3, 4, **5**, 6	6, 7, 8, 9, 12, 13, 14, **16**
Max features	-	5, 10, 15, **20**, 30	13, **15**, 17, 20
Number of estimators	-	200, 300, 400, **500**, 600, 700	100, 200, 300, 400, **500**, 600

	Dart	XGBoost	NN
Max depth	3, 4, **5**, 6, 9	3, **5**, 7, 9, 11, 13	-
Number of estimators	200, **300**, 400, 500, 600	200, 400, **500**, 600, 700	-
Rate drop	**0.1**, 0.01	-	-
Learning rate	-	0.1, **0.01**, 0.001	-
Units	-	-	64, **32**, 16
Regularisation	-	-	1e-2, **1e-4**, **1e-6**

E Schematic model formulations

Figures 18 and 19 portray, respectively, a visualisation of branching out in a regression tree and a 3-layer neural network (with layers 1, l+1 and final layer L).

Figure 18: Random Forest example

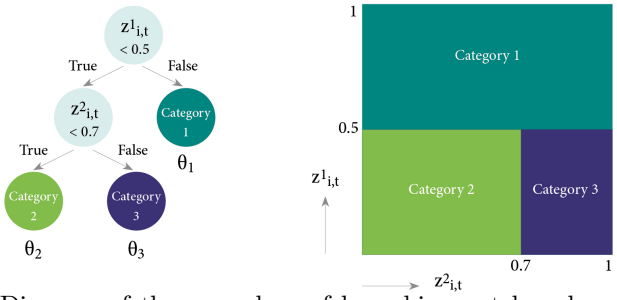


Diagram of the procedure of branching out based on different features, hence partitioning data points. The terminal nodes sort data points into categories 1, 2 and 3. Their corresponding constant θ_k is used for the forecasts.

Figure 19: Diagram of a neural network

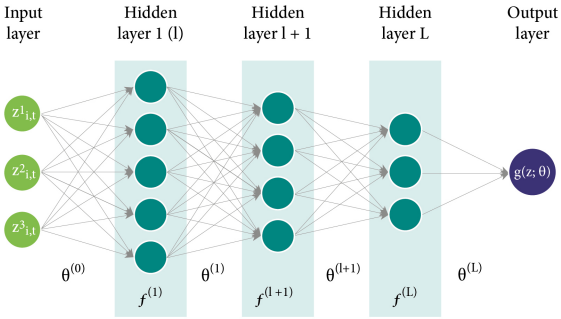
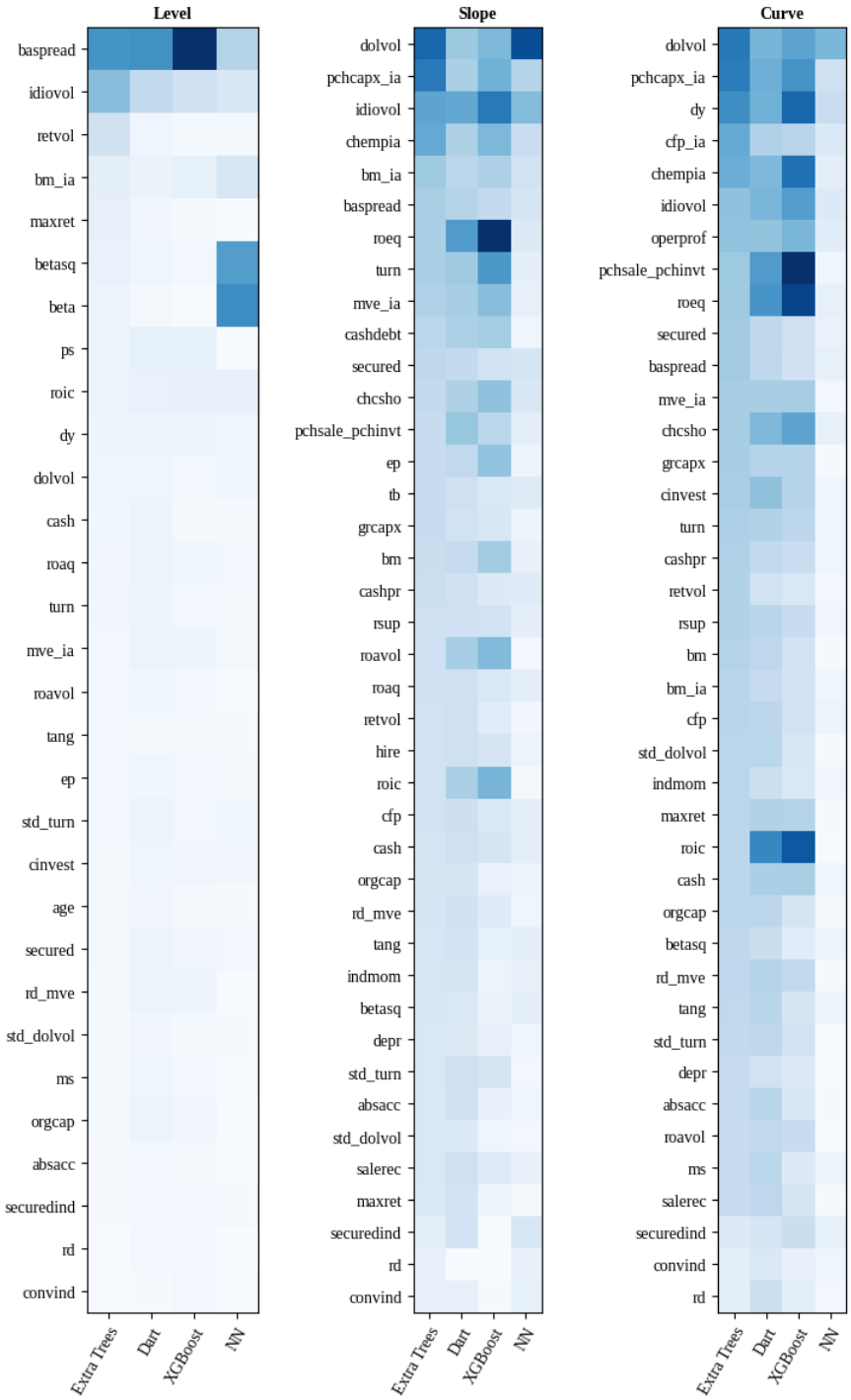


Illustration of the configuration of a feed-forward neural network. This specific network has three hidden layers with corresponding constants $\theta^{(l)}$, which produces forecast $g(z;\theta)$.

F Extensive results

F.1 Feature importance

Figure 20: Heatmap of feature importances for the models: Extra, NN, Dart and XGBoost, for the IV shape features level, slope and curvature



This figure displays a side-by-side comparison of the feature importance scores assigned to the variables within the training proces of the proposed models: Extra, NN, Dart and XGBoost, for the IV shape features level, slope and curvature. The importances are obtained over a sample from January 1996 to December 2011.

F.2 Diebold Mariano test-statistic

Table 7: Diebold-Mariano test statistics for IVS level predictions with corresponding p-values

	OLS-4	OLS-All	OLS-30	Elnet	RF	Extra	Dart	XGBoost	NN	EW 1	EW 2	EW 3	Stacking 1	Stacking 2	Stacking 3
BS	**4.899***(0.00)	**3.613***(0.01)	**5.257***(0.00)	**2.982**(0.02)	**5.978***(0.00)	**4.604***(0.00)	**4.075***(0.00)	**3.562***(0.01)	**3.940***(0.00)	**4.173***(0.00)	**4.152***(0.00)	**4.336***(0.00)	**5.911***(0.00)	**4.069***(0.00)	**4.219***(0.00)
OLS-4		-0.051(0.96)	0.826(0.43)	0.257(0.80)	0.807(0.44)	1.248(0.24)	0.851(0.42)	0.754(0.47)	0.944(0.37)	1.093(0.30)	1.043(0.32)	1.147(0.28)	1.163(0.27)	1.061(0.32)	0.993(0.35)
OLS-All			1.574(0.15)	0.620(0.55)	0.841(0.42)	**2.549**(0.03)	1.368(0.20)	1.242(0.25)	1.727(0.12)	**2.574**(0.03)	1.868(0.09)	2.049(0.07)	0.659(0.53)	**2.219**(0.05)	1.561(0.15)
OLS-30				-0.393(0.70)	-0.717(0.49)	1.606(0.14)	0.389(0.71)	0.473(0.65)	0.790(0.45)	1.089(0.30)	0.925(0.38)	1.217(0.25)	-0.281(0.78)	0.611(0.56)	0.250(0.81)
Elnet					0.095(0.93)	1.194(0.26)	0.667(0.52)	0.823(0.43)	1.208(0.26)	1.663(0.13)	1.104(0.30)	1.221(0.25)	0.119(0.91)	0.884(0.40)	0.548(0.60)
RF						1.346(0.21)	0.617(0.55)	0.589(0.57)	0.820(0.43)	1.026(0.33)	0.960(0.36)	1.135(0.29)	0.156(0.88)	0.875(0.40)	0.656(0.53)
Extra							-0.941(0.37)	-0.170(0.87)	-0.129(0.90)	-0.051(0.96)	-0.047(0.96)	0.321(0.76)	-0.870(0.41)	-0.519(0.62)	-1.106(0.30)
Dart								0.475(0.65)	0.887(0.40)	0.688(0.51)	**2.220**(0.05)	1.975(0.08)	-0.429(0.68)	0.304(0.77)	-0.234(0.82)
XGBoost									0.163(0.87)	0.151(0.88)	0.260(0.80)	0.403(0.70)	-0.451(0.66)	-0.107(0.92)	-0.379(0.71)
NN										0.103(0.92)	0.163(0.87)	0.619(0.55)	-0.600(0.56)	-0.267(0.80)	-0.634(0.54)
EW 1											-0.005(1.00)	0.269(0.79)	-0.697(0.50)	-0.395(0.70)	-0.782(0.45)
EW 2												0.812(0.44)	-0.693(0.51)	-0.416(0.69)	-0.864(0.41)
EW 3													-0.808(0.44)	-0.589(0.57)	-1.043(0.32)
Stacking 1														0.632(0.54)	0.446(0.67)
Stacking 2															-1.091(0.30)

This table presents the Diebold Mariano test statistics for all models based on the predictions made for the IVS level. The significance of the DM-statistics is given between brackets. Positive numbers indicate that the column model outperforms the row model. The bold values are significant at a 5% level and the values accompanied by an asterisk indicate the significance at a 1% level.

Table 8: Diebold-Mariano test statistics for IVS slope predictions with corresponding p-values

	OLS-4	OLS-All	OLS-40	Elnet	RF	Extra	Dart	XGBoost	NN	EW 1	EW 2	EW 3	Stacking 1	Stacking 2	Stacking 3
BS	1.785(0.11)	2.025(0.07)	1.884(0.09)	1.966(0.08)	1.720(0.12)	1.686(0.13)	1.725(0.12)	1.688(0.13)	1.637(0.14)	1.810(0.10)	1.704(0.12)	1.737(0.12)	1.602(0.14)	1.541(0.16)	1.294(0.23)
OLS-4		-1.018(0.34)	-1.135(0.29)	-0.873(0.41)	-0.471(0.65)	0.978(0.35)	1.098(0.30)	0.899(0.39)	0.699(0.50)	1.036(0.33)	1.045(0.32)	1.207(0.26)	0.399(0.70)	0.225(0.83)	-0.257(0.80)
OLS-All			-0.551(0.59)	1.284(0.23)	0.780(0.46)	1.033(0.33)	1.099(0.30)	1.024(0.33)	0.911(0.39)	1.241(0.25)	1.069(0.31)	1.153(0.28)	0.840(0.42)	0.725(0.49)	0.281(0.79)
OLS-40				1.350(0.21)	1.037(0.33)	1.141(0.28)	1.228(0.25)	1.152(0.28)	1.001(0.34)	1.402(0.19)	1.191(0.26)	1.283(0.23)	0.986(0.35)	0.880(0.40)	0.355(0.73)
Elnet					0.559(0.59)	0.966(0.36)	1.044(0.32)	0.954(0.37)	0.825(0.43)	1.211(0.26)	1.009(0.34)	1.111(0.30)	0.741(0.48)	0.613(0.56)	0.141(0.89)
RF						1.072(0.31)	1.204(0.26)	1.094(0.30)	0.815(0.44)	1.620(0.14)	1.162(0.28)	1.398(0.20)	0.860(0.41)	0.638(0.54)	-0.060(0.95)
Extra							-0.215(0.83)	-0.651(0.53)	-0.336(0.74)	-0.467(0.65)	-0.081(0.94)	0.418(0.69)	-0.715(0.49)	-0.512(0.62)	-0.800(0.44)
Dart								-0.322(0.75)	-0.126(0.90)	-0.509(0.62)	0.312(0.76)	0.865(0.41)	-0.606(0.56)	-0.468(0.65)	-0.690(0.51)
XGBoost									-0.015(0.99)	-0.355(0.73)	1.428(0.19)	2.032(0.07)	-0.594(0.57)	-0.431(0.68)	-0.687(0.51)
NN										-0.211(0.84)	0.246(0.81)	0.409(0.69)	-0.352(0.73)	-0.344(0.74)	-0.827(0.43)
EW 1											0.523(0.61)	0.834(0.43)	-0.081(0.94)	-0.155(0.88)	-0.479(0.64)
EW 2												0.814(0.44)	-0.758(0.47)	-0.528(0.61)	-0.750(0.47)
EW 3													-1.114(0.29)	-0.677(0.52)	-0.778(0.46)
Stacking 1														-0.238(0.82)	-0.528(0.61)
Stacking 2															-0.442(0.67)

This table presents the Diebold Mariano test statistics for all models based on the predictions made for the IVS slope. The significance of the DM-statistics is given between brackets. Positive numbers indicate that the column model outperforms the row model. The bold values are significant at a 5% level and the values accompanied by an asterisk indicate the significance at a 1% level.

Table 9: Diebold-Mariano test statistics for IVS curvature predictions with corresponding p-values

	OLS-4	OLS-All	OLS-40	Elnet	RF	Extra	Dart	XGBoost	NN	EW 1	EW 2	EW 3	Stacking 1	Stacking 2	Stacking 3
BS	2.112(0.06)	**2.344**(0.04)	1.985(0.08)	**2.273**(0.05)	2.150(0.06)	2.125(0.06)	2.166(0.06)	2.141(0.06)	1.985(0.08)	2.170(0.06)	2.154(0.06)	2.153(0.06)	1.936(0.08)	2.099(0.07)	**2.256**(0.05)
OLS-4		-1.375(0.20)	0.701(0.50)	-1.292(0.23)	-0.147(0.89)	1.191(0.26)	0.811(0.44)	1.004(0.34)	0.334(0.75)	0.290(0.78)	1.219(0.25)	1.193(0.26)	0.718(0.49)	-0.055(0.96)	0.042(0.97)
OLS-All			1.172(0.27)	1.086(0.31)	1.398(0.20)	1.508(0.17)	1.556(0.15)	1.508(0.17)	1.112(0.30)	1.592(0.15)	1.575(0.15)	1.620(0.14)	1.143(0.28)	1.289(0.23)	1.184(0.27)
OLS-40				-1.118(0.29)	-0.509(0.62)	-0.063(0.95)	-0.266(0.80)	-0.102(0.92)	-0.770(0.46)	-0.381(0.71)	-0.034(0.97)	-0.014(0.99)	-0.157(0.88)	-0.673(0.52)	-0.329(0.75)
Elnet					1.408(0.19)	1.547(0.16)	1.638(0.14)	1.552(0.16)	1.069(0.31)	1.709(0.12)	1.637(0.14)	1.721(0.12)	1.1(0.30)	1.318(0.22)	0.887(0.40)
RF						0.780(0.46)	0.703(0.50)	0.736(0.48)	0.309(0.76)	0.987(0.35)	0.841(0.42)	1.136(0.29)	0.542(0.6)	0.139(0.89)	0.105(0.92)
Extra							-0.651(0.53)	-0.138(0.89)	-1.092(0.30)	-0.661(0.53)	0.151(0.88)	0.086(0.93)	-0.072(0.94)	-1.225(0.25)	-0.340(0.74)
Dart								0.800(0.44)	-0.374(0.72)	-0.494(0.63)	1.129(0.29)	0.821(0.43)	0.18(0.86)	-1.153(0.28)	-0.237(0.82)
XGBoost									-0.788(0.45)	-0.595(0.57)	0.378(0.71)	0.15(0.88)	-0.022(0.98)	-1.164(0.27)	-0.32(0.76)
NN										-0.065(0.95)	0.914(0.38)	0.731(0.48)	0.877(0.4)	-0.412(0.69)	-0.102(0.92)
EW 1											0.736(0.48)	1.153(0.28)	0.383(0.71)	-0.49(0.64)	-0.078(0.94)
EW 2												0.038(0.97)	-0.089(0.93)	-1.355(0.21)	-0.353(0.73)
EW 3													-0.093(0.93)	-1.756(0.11)	-0.357(0.73)
Stacking 1														-0.697(0.5)	-0.291(0.78)
Stacking 2															0.053(0.96)

This table presents the Diebold Mariano test statistics for all models based on the predictions made for the IVS curvature. The significance of the DM-statistics is given between brackets. Positive numbers indicate that the column model outperforms the row model. The bold values are significant at a 5% level and the values accompanied by an asterisk indicate the significance at a 1% level.