Implied Volatility Forecasting using Neural Networks

Master Thesis Financial Economics

Author:
Jeroen Alexander Stam
Student ID: 535267
Erasmus University Rotterdam
Erasmus School of Economics

Supervisor:
Dr. Jens Soerlie Kværner
Tilburg University
Second Assessor:
Jan Lemmen

Date of Submission: September 22, 2023

Abstract

I examine the efficacy of neural networks in forecasting implied volatility, utilizing firm characteristics data from Gu et al. (2020). I use end-of-the-month data from at-the-money US call options with maturities of 30- and 91-days over the sample period 1996 to 2021. In line with my expectations, neural networks outperform linear regressions in forecasting implied volatility using firm characteristics. Specifically, utilizing neural networks yields out-of-sample R-squared increases of 25 and 10 percentage points for the 30- and 91-day maturities, relative to OLS. Diebold-Mariano statistics show neural networks outperform linear regression achieving higher accuracy in 87.63% and 86.07% of the assessed options for the respective maturities. Option maturity does not play a major role in IV predictability.

Contents

I. Introduction

I examine the effectiveness of neural network (NN) models compared to linear regression (OLS) in predicting implied volatility (IV) for US options with 1-month (30d) and 3-months (91d) maturity using the firm characteristics of Gu et al. (2020). Also, I study the impact of option maturity on the relative forecasting performance of NN and OLS. Lastly, I analyse variable importance of the forecasting models to uncover which variables have the most relevance for forecasting IV. The data structure and methodology of this research are largely based on Crego et al. (2023). The economic motivation for this research is to lay the foundation of IV forecasting using NN, which could be used to make profits as an investor. In Appendix A, I present a simplified example on how one can make profit by forecasting IV using the Black-Scholes formula[1]. The scientific goal of this thesis is to find what factors influence implied volatility and which model, linear versus nonlinear, is best capable of using these factors to predict IV. This has been researched for asset pricing using realized returns (Gu et al., 2020) or yields (Crego et al., 2023). However, it has not been discovered what the implications are for IV. I expect results similar to Crego et al. (2023), because IV is closely related to their yields estimator. Therefore, I expect NN to outperform OLS in forecasting IV and anticipate liquidity-related characteristics to be important IV predictors.

Asset prices are dynamic and are affected by many other factors aside from cash flows and performance of underlying companies. Macroeconomic (Hondroyiannis & Papapetrou, 2001) and geopolitical stability (Balcilar et al., 2018), investor sentiment (Baker & Wurgler, 2007) and other factors play important roles in determining riskiness of an asset. A parameter that is often used to quantify the riskiness of an asset is implied volatility. Implied volatility is an estimate for future volatility of an underlying asset, based on the option contract prices on that asset. IV has been popularized by models such as the Black-Scholes model (Black

[1] In this scenario, a European call option is listed for \$14.70. The relevant factors are as follows: spot price = \$110, strike price = \$100, risk-free rate = 3%, and maturity = 1 year, resulting in an implied volatility of 15%. If one correctly anticipates that tomorrow the IV will increase to 16%, a financial gain of \$0.30 can be achieved by purchasing the call option at its current price.

& Scholes, 1973) and the Rubinstein Implied Binomial Tree (Rubinstein, 1994). These models show the relation between IV and option prices, which allows for prediction of option prices changes, by predicting future IV. IV is forward looking, i.e., it reflects the investors expectations of future price changes of the asset. Cao et al. (2020) show that the put-call implied volatility spread can be used to effectively predict equity premiums, stating that the predictive power can be traced back to the forward-looking information underlying the implied volatility spread, capturing general market sentiment. Similarly, the VIX, the Chicago Board Options Exchange's Volatility Index, is a measure of the stock market's expectation of future volatility based on S&P500 index options. Its ability to capture the market expectations makes implied volatility a valid predictor for asset prices. Hence, in general, the predictive power of IV for asset prices stems from its ability to capture market expectations.

A vast body of literature has been dedicated to forecasting asset prices using factors. Factor models such as the Capital Asset Pricing Model of Sharpe (1964) and the FF-5 of Fama and French (2015) are regarded as cornerstones of modern-day asset pricing. These models perform relatively well, but are limited to establishing linear relations between asset prices and factors. This is where Machine Learning models can improve on the standard methods. Machine learning (ML) is a relatively new approach to academic research. Its relevance is found in its effectiveness in deciphering, especially nonlinear, relations between dependent and independent variables. Gu et al. (2020) assess the task of asset pricing using multiple ML models using a vast amount of predictors. The authors find that using ML models yields large economics gains, in some cases doubling the return obtained using linear models. Gu et al. (2020) and other studies such as Christensen et al. (2021) accredit ML's superior performance, compared to linear models, to the ability to capture nonlinear relations. However, the literature regarding whether nonlinear is better than linear analysis in predicting volatility instead of realized returns is not conclusive. For example, Vortelinos (2017) finds that a Heterogeneous Auto Regressive model slightly outperforms nonlinear models. Vortelinos attributes the outperformance to the property of persistence of volatility. However, Hosker et al. (2018) show that the recurrent neural networks (RNN) and long short-term memory (LSTM) provide improved volatility forecasts compared to linear regression, principal components analysis and ARIMA methods for 1-month option contracts of the VIX. Vrontos et al. (2021) discovered that the utilization of machine learning techniques for forecasting implied volatility can surpass conventional econometric models and model selection methods. This superiority of machine learning models holds true in both statistical and economic evaluation contexts. In the same vein, Malliaris and Salchenberger (1996) show that neural networks are more accurate at forecasting implied volatility compared to historic forecasts.

The central paper on which this research is based is Crego et al. (2023). In their paper, the authors examine the usage of multiple models on predicting an option-based yield estimator derived from Martin and Wagner (2019), using the characteristics of Gu et al. (2020) and macro data of Welch and Goyal (2008). IV closely relates to yield estimator of Martin and Wagner (2019), as IV is a key ingredient in their yield calculation. Hence, I expect the data structure and methodology of Crego et al. (2023) to also be compatible with IV research. Crego et al. (2023) find that their best performing model is NN3, which considerably outperforms OLS at predicting option yields. Considering the similarities in research design, I expect to obtain similar results to Crego et al. (2023), meaning neural networks outperforming OLS in forecasting IV. Using this, I construct the following hypothesis

H1: Neural network outperforms linear regression in forecasting implied volatility based on firm characteristics

I also aim to examine the effect of option maturity on the potential outperformance of NN compared to OLS. The closer the option is to its expiration, the less opportunities there are for an investor to make profit on it, as less volatility can affect its price. As the trading price of the option will eventually converge to the strike price, the option price tends to become less volatile. A longer time until expiration allows for more changes of IV, which is why I expect the implied volatility to be more complex to predict further from expiration. I anticipate the ML model to be better suited for this complexity. Therefore, I predict the difference in forecasting performance of the models to be larger for options with longer maturities. I will not formally test this using a hypothesis as that is beyond the scope of this thesis. Combining the aforementioned hypothesis with the maturity analysis, the Research Question reads:

RQ: To what extent does neural network outperform linear regression in forecasting implied volatility based on firm characteristics, and how does option maturity affect this relative performance?

II. Data

In this analysis, I attempt to follow the data collection methodology from Crego et al. (2023). I obtain implied volatility data of options with 30- and 91-days maturity from the volatility surface of OptionMetrics. I limit the sample to end-of-the-month observations, i.e., the last trading day in each month. I use Center for Research in Security Prices (CRSP) to obtain stock price data of the underlying assets. Next, I limit my sample to at-the-money options by monthly filtering on the option contracts that are closest to the actual asset prices. I use the database of Gu et al. (2020) to obtain the data on 94 firm characteristics. I normalize the predictor data and use monthly median values to replace missing data. Throughout this thesis, these predictors will be referred to as the GKX firm characteristics. A list of these firm characteristics with descriptions can be found in the Table B1 (Appendix B). To be able to optimize the NN model, I split all my data into a training, validation and a testing section. Following Crego et al. (2023), I use 1996-2003 as the training data, 2004-2007 as the validation data and the remaining years, 2008-2021, as the testing set. Lastly, I exclude options with fewer than 12 consecutive monthly observations in the test set, as an imbalanced panel introduces additional complexity to the forecasting process. The final data set stretches from January 1996 to December 2021. The descriptive statistics for implied volatility after aforementioned adjustments are presented in the following table:

Table 1: Desciptive Statistics: Implied Volatility

	30d			**91d**		
	Train	Validation	Test	Train	Validation	Test
Mean	0.4987	0.4234	0.4805	0.4891	0.4170	0.4568
Std	0.2433	0.2274	0.2937	0.2341	0.2163	0.2665
Med	0.4429	0.3728	0.4049	0.4342	0.3708	0.3906
Obs	108993	123738	543975	108988	123737	543959

Mean, standard deviation, median and amount of observations of implied volatility. Train is the period 1996-2003, validation is 2004-2007 and test is 2008-2021. Values, except observations, are rounded to the fourth decimal.

III. Methodology

I utilize the train-validate-test approach to optimize the neural network. In this process, the training set is utilized to estimate coefficients for each predictor variable, while the validation set assesses model performance beyond the training data. This facilitates model refinement, enhancing out-of-sample forecasting accuracy. Subsequently, the testing set evaluates the actual forecasting proficiency of the model. However, for the OLS model, the validation set holds no utility since this model lacks tuning parameters (hyperparameters) for refinement. Integrating the validation period into the training and/or testing sets for OLS would result in imbalanced performance comparisons. Consequently, the linear regression model excludes the validation period to ensure a fair evaluation. Both models share the same structure: the implied volatility serves as the dependent variable, while the feature variables from the database by Gu et al. (2020) act as the independent variables.

III.I Linear Regression

The first forecasting model is the pooled linear regression. Initially, I perform a regression on the entire training dataset to obtain the coefficients that capture the relationship between each GKX feature and implied volatility. This regression takes the following form:

$$\boldsymbol{IV} = \boldsymbol{\alpha} + \boldsymbol{X}^{train}\hat{\boldsymbol{\beta}} + \boldsymbol{\varepsilon} \tag{1}$$

Here, $\boldsymbol{IV}$ represents a $N^{train} \times 1$ vector of implied volatilities, $\boldsymbol{\alpha}$ denotes a $N^{train} \times 1$ vector of constants, $\boldsymbol{X}^{train}$ indicates a $N^{train} \times 94$ matrix of the 94 firm characteristics from the Gu et al. (2020) of the training period, $\hat{\boldsymbol{\beta}}$ represents a 94×1 vector of estimated coefficients of the firm characteristics, and $\boldsymbol{\varepsilon}$ denotes a $N^{train} \times 1$ vector of estimation errors. N^{train} represents the amount of observations in the training set. For forecasting the IV for the test set, I utilize the coefficients obtained from the training set's regression, resulting in the following equation:

$$\widehat{\boldsymbol{IV}} = \boldsymbol{X}^{test}\hat{\boldsymbol{\beta}} \tag{2}$$

Here, $\widehat{\boldsymbol{IV}}$ represents a $N^{test} \times 1$ vector of implied volatilities, $\boldsymbol{X}^{test}$ indicates a $N^{test} \times 94$ matrix of the 94 firm characteristics from the Gu et al. (2020) of the training period and $\hat{\boldsymbol{\beta}}$ represents the 94×1 vector of coefficients estimated in Eq. (1). N^{test} represents the amount of observations in the test set.

III.II Neural Network

The more advanced model of this research is the feed-forward neural network. This artificial intelligence model feeds the raw data from the *input layer* through one or multiple *hidden layers* of neurons that interact with each other and non-linearly transform the raw input data. These transformations are then linearly aggregated in the final *output layer*. Neural networks have a vast amount of computational flexibility, which is both an asset as well as a liability. It allows for complex nonlinear function estimation, but has the drawback of having low transparency and being highly parameterized.

The key objective in Machine Learning is to create a model that performs well on data which it has not been trained on, achieving a low *regularisation* error. A proper model achieves a sufficiently small training error as well as a sufficiently small gap between training and test error. This ties into the challenge of optimizing a model's capacity, i.e., its ability to fit a wide variety of functions. Ideally, the ML model's capacity closely matches the true complexity of its task. An insufficient capacity leads to *underfitting*, meaning the model does not learn enough from the training data to be effective on the test set. More formally, the training error is not sufficiently low. Underfitting is associated with a higher bias of the estimator. The bias of the estimator is defined as:

$$Bias(\hat{\boldsymbol{\theta}}_m) = \mathbb{E}[\hat{\boldsymbol{\theta}}_m] - \boldsymbol{\theta}, \tag{3}$$

where $\hat{\boldsymbol{\theta}}_m$ is the estimator and $\boldsymbol{\theta}$ is the true underlying value. In contrast, a too high capacity leads to *overfitting*. This means the model is too tailored to the training data, such that it becomes ineffective when used on the test set. More formally, the gap between training and test error is too large. Overfitting is associated with a higher variance of the estimator. The variance of the estimator is written as: $Var(\hat{\theta})$. Finding the optimal model capacity means optimizing the objective function. From the multiple possible objective functions, I use the MSE objective function as this is also the objective function for OLS. MSE is defined as:

$$MSE = \mathbb{E}[(\hat{\theta}_m - \theta)^2] = Bias(\hat{\theta}_m)^2 + Var(\hat{\theta}_m) \tag{4}$$

The model optimizes its capability to forecast out-of-sample by minimizing the MSE equation on the validation set. This ensures, given the data and hyperparameters, the model finds the optimal trade-off between training error and the gap between training and test error. The MSE is calculated monthly and averaged to obtain a single value.

Model Parameters

To optimize the ML model's capacity for out-of-sample forecasting, hyperparameters are used to tune the model such that the MSE of validation set is minimised. Neural networks have a complicated structure and require a number of hyperparameters. The first hyperparameter is the amount of neurons in each hidden layer. To find the optimal hidden layer structure and learning rate (LR), I use a grid search on the 30d validation data. To save computational expense and time, it is assumed that the optimal parameters found with the 30d data are also optimal for the 91d data. I examine the performance of $\{32, 16, 8\}$ and $\{16, 8\}$ where each element represents the amount of neurons per hidden layer. Within each hidden layer, neurons receive input and perform mathematical operations, such as assigning weights, nonlinearly transforming the data. The output is then passed onto the next layer. Learning rate (LR) is the hyperparameter that determines how much the model adjusts to estimated error each time the model is updated. The grid search includes LR $\in \{0.1, 0.01, 0.001\}$. Following modern NN default recommendation, I use rectified linear unit (ReLu) as the activation function of my hidden layers (Jarrett et al., 2009). ReLu activation function, defined as $g(z) = max\{0, z\}$, has the benefits of being nonlinear, yet preserving many desired properties of linear functions. These desired properties include easy optimisation and good generalisation. Following Crego et al. (2023), I use the Adam optimisation algorithm with 100 epochs. The L1 penalty is the hyperparameter that assigns a penalty based on the amount of variables, used to prevent inclusion of too many variables. To determine the optimal L1 penalty, I use cross-validation. Unlike OLS, neural networks exhibit variability in forecasting outcomes. I use a randomly selected seed to ensure reproducibility and consistency of the results.

III.III Forecast Evaluation

To evaluate the goodness-of-fit of a model, a commonly used measure is the R-squared (R^2). R-squared measures the proportion of variance of the dependent variable that is explained by the model of independent variables. When evaluating a forecasting model, it is common to calculate the R-squared based on the forecasts of the test sample, as this shows how well the model performs on not previously-shown data compared to the simple mean. This is referred to as the out-of-sample R-squared, R^2_{oos}. The R^2_{oos} is calculated as follows:

$$R^2_{\text{oos}} = 1 - \frac{\sum_{i=1}^{n}(y_{\text{test},i} - \hat{y}_{\text{test},i})^2}{\sum_{i=1}^{n}(y_{\text{test},i} - \bar{y}_{\text{test}})^2} \tag{5}$$

where y represents the actual IV, $\hat{y}$ represents the model's estimation of IV and $\bar{y}$ represents the simple average actual IV. Given a correctly constructed model, a higher R^2 means a more accurate model. R^2 has an upper limit of 1 and is only negative if the model's forecast is a worse approximate of the true value than the simple average. The R^2 is calculated monthly to capture how much of the cross-sectional variation can be explained using the forecasts. This also allows for assessing how the models' accuracy develops over time. Finally, for each model I take the average of the monthly R^2s to obtain a single evaluation metric.

Finally, I partition the generated multi-asset Panel data containing NN and OLS forecasts into distinct single-asset time series. For every individual asset, I compute forecast errors for both NNs and OLS, followed by conducting the Diebold-Mariano (DM) test.

$$\text{DM} = \sqrt{n} \times \frac{\tilde{d}}{\sigma_d} \tag{6}$$

$$\tilde{d} = \frac{\sum_{i=1}^{n} (\hat{y}_i^{NN} - y_i) - (\hat{y}_i^{OLS} - y_i)}{n} = \frac{\sum_{i=1}^{n} \hat{y}_i^{NN} - \hat{y}_i^{OLS}}{n} \tag{7}$$

where $\tilde{d}$ represents the average difference between NN forecast error and OLS forecast error, σ_d represents the standard deviation of the difference between NN forecast error and OLS forecast error and n represents the amount of forecasts. I use DM such that a test statistic larger than the critical value indicates NN outperforms OLS. Conversely, a test statistic lower than the negative critical value indicates OLS outperforms NN. Using two-sided testing and a 95% confidence level, the critical value is approximately ±1.96. Finally, I take the average of the DM test statistics and observe what percentage of test statistics exceeds the critical value.

III.IV Variable Importance

In contrast to OLS, neural networks are infamously nontransparent, meaning it is hard to determine the contribution of each variable to the model. Following Crego et al. (2023), I use the leave-one-covariate-out (LOCO) from Lei et al. (2018) on the training set to obtain the variable importance for the neural network. First, I calculate the Spearman

Rank Correlation ρ for the entire model, i.e., using all F = 94 firm characteristics. Then, the model is trained F = 94 times, each iteration dropping feature f = 1,2,...,F from the full model. The variable importance of feature f is then calculated as the absolute change of Spearman Rank Correlation, $|\Delta\rho|$, when dropping f from to the full model.

IV. Results

The first objective of this research is to obtain the optimal neural network hyperparameters. The hyperparameters I kept adjustable are the hidden layer structure, learning rate and L1 penalty. Using the LASSO model, I obtained optimal L1 penalties of approximately 0.00032 and 0.00027 for the 30-day and 91-day maturities, respectively. Using the grid search, I obtained the optimal configuration for the other hyperparameters. In Table 2, I present the performance measures of out-of-sample forecasting for all configurations of the grid search. The results are from the validation data of the 30-day maturity options.

Table 2: Grid Search

	LR=0.1		**LR=0.01**		**LR=0.001**	
Layers	$\boldsymbol{R^2}$	$\boldsymbol{MSE}$	$\boldsymbol{R^2}$	$\boldsymbol{MSE}$	$\boldsymbol{R^2}$	$\boldsymbol{MSE}$
$\boldsymbol{\{32, 16, 8\}}$	-0.2042	0.0623	-0.5692	0.0811	0.1175	0.0456
$\boldsymbol{\{16, 8\}}$	-0.2041	0.0623	0.6168	0.0198	-1.5860	0.1337

Table presents the out-of-sample R^2 and Mean Squared Error. LR represents Learning Rates and Layers represents the layer structure of the Neural Network. Values are rounded to the fourth decimal.

From this table, one can see that the configuration LR = 0.01 and $\{16, 8\}$ performs best, both achieving the highest R-squared, as well as lowest Mean Squared Error. LR = 0.1 was unable to effectively produce different forecasts, i.e, it produced a single value for nearly the entire panel of forecasts. This is likely the result of overfitting; the model is too tailored to the training data which causes it to become insensitive to new inputs. For LR = 0.1, changing the hidden layer structure hardly affected the performance, as can be seen from the minimal alteration in both R^2 and MSE. The models using learning rates of 0.1 are both unable to achieve positive R-squared values, meaning their forecasts are worse approximates than the simple average. The same holds for the configuration LR = 0.001 and $\{32, 16, 8\}$, which has the worst performance of all configurations considered. Increasing the complexity of the hidden layer structure did vastly improve the forecasting capabilities, even obtaining a positive R^2. In contrast, the LR = 0.01 model vastly improved when reducing the hidden layer complexity, obtaining the best R^2 and MSE of all configuration.

The optimal configuration obtained using 30-day data, $\{16, 8\}$ and LR = 0.01, was used to implied volatility forecasts for both maturities. To show how the models' accuracy develops over time, I grouped the IV forecasts by month and calculated the out-of-sample R^2 for both OLS and NN at both maturities. The result of this analysis is presented in Figure 1.

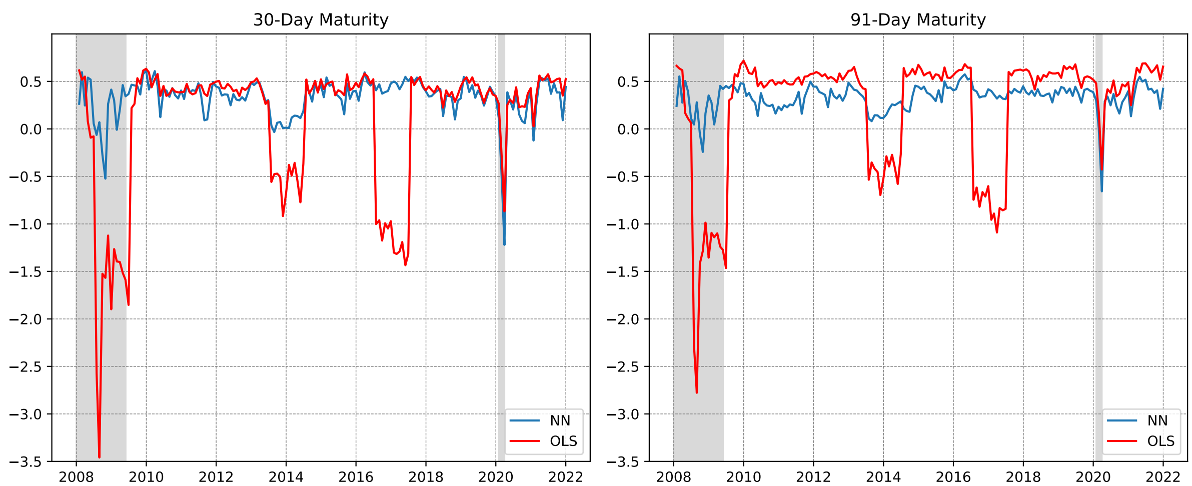


Figure 1: Monthly R^2

Y-axis represents monthly out-of-sample R^2 and the x-axis represents the period 2008-2022. Grey bars indicate periods of economic recessions according to National Bureau of Economic Research (01/2008 - 06/2009 and 02/2020 - 04/2020).

Both models show decreased R^2 in the periods of economic recession at both maturities. NN does not experience as severe accuracy drops as the OLS models, except for the 2020 recession. Out of all forecasts, the R^2 of the 30-day maturity OLS experiences the largest decreases.

Next, I averaged the monthly values to obtain a single R^2 per model. Similarly, I calculated and averaged the monthly MSEs. The results are presented in Table 3.

Table 3: Neural Networks versus OLS

	30d		**91d**	
	R^2	MSE	R^2	MSE
OLS	0.0699	0.0678	0.2260	0.0473
NN	0.3274	0.0521	0.3302	0.0417

Table presents the out-of-sample R^2 and Mean Squared Error. Values are rounded to the fourth decimal.

Looking at R^2, the neural network model performs better than the linear regression at forecasting implied volatility using the GKX firm characteristics. Note that for both models are more accurate at forecasting the 91d maturity. Using neural networks instead of OLS yields an increase in out-of-sample R-squared of more than 25 and 10 percentage points for the 30- and 91-day maturities, respectively. NN also obtains slightly lower MSE than OLS at both maturities.

To formally test the forecasting ability difference between NN and OLS, I used the Diebold-Mariano test on each asset separately. The average DM test statistic and the percentage of cases where NN is significantly better than OLS ($DM^*\%$) are shown in Table 4:

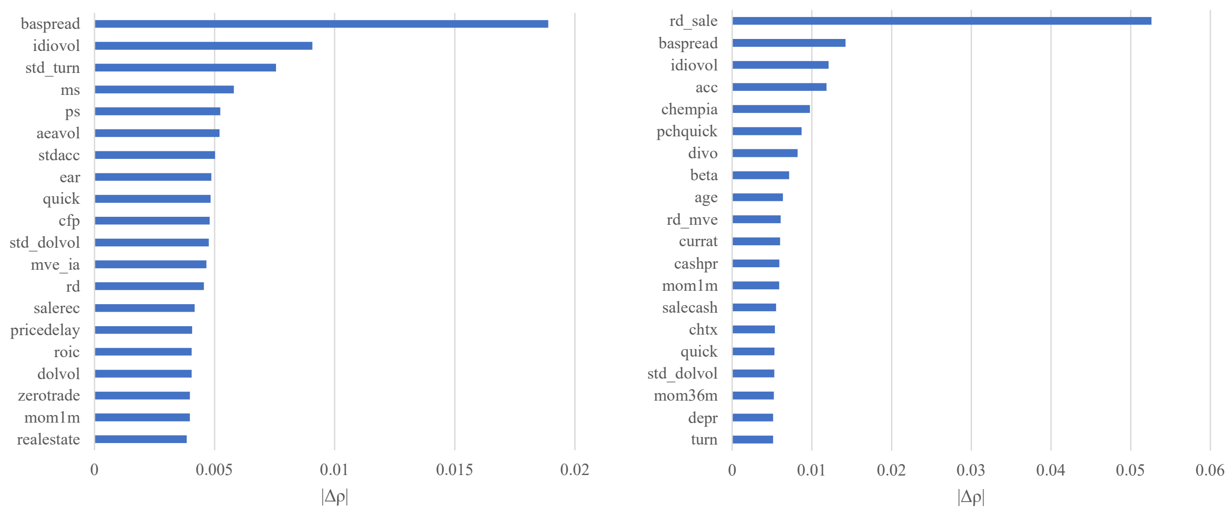
Table 4: Diebold-Mariano Test

	$AverageDM$	$DM^*\%$
$30d$	17.47	87.63
$91d$	11.42	86.07

Values are rounded to the second decimal. $DM^*\%$ shows the amount of cases for which the DM statistic is larger than critical value 1.96, meaning NN has significantly lower forecast errors than OLS

Table 4 presents the main result of the paper. Both maturities have average DM test statistics much larger than 1.96, 30d having the largest of the two. A DM test statistic that exceeds the critical value of 1.96 indicates that NN has significantly lower forecasting errors than OLS. NN performs better than OLS in more than 86% of cases at forecasting IV using GKX firm characteristics at both maturities. Despite the difference in average DM statistic, there is not a large difference in NN outperformance between maturities.

Lastly, the results of the variable importance analysis are shown in Figure 1. Figure 1.a and 1.b show the 20 most important variables for the 30d and 91d training sets, respectively.



(a) 30d

(b) 91d

Figure 2: Variable Importance

This figure shows of the variable importance analysis using leave-one-covariate-out on Spearman rank correlation on the training set (1996-2003). This figure shows the variable importance of the 20 most important variables for the 30d and 91d maturity options. The x-axis variable, $|\Delta\rho|$, represents the absolute change in Spearman rank correlation from the full model, when the y-axis variable feature is excluded.

Figure 2.a shows that, for the 30d maturity, the three most important variables are: bid-ask spread, idiosyncratic return volatility and standard deviation of asset turnover. Figure 2.b shows that at the 91d maturity the three most important variables are: research and development-to-sales ratio, bid-ask spread, idiosyncratic return volatility.

V. Discussion

In this section, I discuss the results of this thesis. The first result to discuss are the results from Figure 1. Both models show decreased R^2 coinciding with the periods of economic recession at both maturities. Notice that NNs does not experience as large R^2 drops as OLS. The economic turbulence of a recession is likely causing an increase in complexity of implied volatility predicting, leading to both forecast models being less accurate. OLS suffers a total of four major accuracy drops, two of which do not occur during to economic recessions, between 2013-2015 and 2016-2018. In an attempt to explain these drops, I examined the trends of implied volatility and implied volatility changes. I expected that these drops might be caused by rapid changes in IV trends, making the prediction task more complicated. Looking at Figure C1 and C2 (Appendix C), IV trend spikes can be observed coinciding with the economic recession periods. However, no noticeable trend spikes occur around the accuracy drops between 2013-2015 and 2016-2018. Thus, I conclude that these R^2 drops are likely not the result of rapid IV trend changes. What actually did cause these accuracy drops remains unexplained.

Next is the comparison of forecast evaluation measures of NN and OLS and the results of the Diebold-Mariano test. From Table 3, one can see that, based on the evaluation measures presented, NN outperforms OLS for both maturities. OLS achieves better R^2 and MSE on the longer maturity options, likely due to having less intense accuracy drops as seen in Figure 1. NN's R^2 and MSE also both improve from 30- to 91-day maturity, however not by as much as is the case for OLS. The difference in R^2 is smaller for the 91-day maturity. This initially appears to contrast my expectation that the difference in forecasting performance of the models are larger for options with longer maturities. However, the results presented in Table 4 show that this is not necessarily the case, as the amount of cases in which NN is more accurate than OLS do not differ much between maturities (from 87.6% to 86.1%). Figure 1 shows that OLS often obtains a higher R^2 than NN, but the large accuracy drops is what deteriorates OLS' overall performance. NN is not consistently more accurate than OLS, but NN does not experience such accuracy drops leading to the majority of assets being better approximated by NN than OLS. It would be interesting to examine the relative performance between the forecasting methods when excluding recession periods.

In Table 4, I also present the average DM statistics. Both maturities have average DM much greater than the 1.96 critical value. The 30d has a larger average DM than the 91d maturity, however as discussed earlier, this does not lead to a large difference in percentual NN outperformance. The mean of DM statistics is included as it is an easy-to-understand metric. It also shows that significance of NN's outperformance is not highly affected by critical value choice. In retrospect, the median DM might also have been a relevant addition, as the median is more robust against outliers than the mean.

Next is the variable importance analysis, shown in Figure 2. The first thing to note is that the most important variable differs across the maturities, 30d having bid-ask spread (*baspread*) and 91d having research & development to sales ratio (*rd_sale*) as most important variable. Bid-ask spread, as well as idiosyncratic return volatility (*idiovol*), is amongst the three most important variables for both maturities. This could indicate that these two variables are important for IV forecasting using NN regardless of option maturity. To formally investigate this, future research should include more option maturities. std_turn is third most important for the 30d maturity. It is also worthy to note that the x-axes of the subfigures differ in magnitude. For 30d, *baspread* leads to a Spearman rank correlation change of 0.019, whereas *rd_sale* leads to a change of 0.053 for 91d, more than twice as much. Despite being highly contributing for 91d, research and development to sales is not included in the top 20 for 30d. Similarly, std_turn is in the top 3 for the 30d maturity but not included in the top 20 for 91d. An explanation for this could lie in the update frequency of variables. std_turn is updated monthly and important to short, 30-day maturity options, while *rd_sale* is annually updated and important for the longer, 91-day maturity options. Considering this, there might be a correlation between the significance of certain variables for short maturity options and the frequency of their updates, where short-term (30d) options tend to have more important variables updated more frequently. Conversely, for longer (91d) maturity options, it would imply crucial variables undergo less frequent updates. Looking at the update frequency of the top 20 variables of 30d, 8 are monthly, 4 are quarterly and 8 are annually. For 91d, 7 are monthly, 1 is quarterly and 12 are annually. 91d indeed has more annually updated variables than 30d, but this is not enough evidence to make any claims.

It is also relevant to discuss whether the variable importance results are in line with prior literature. Bid-ask spread denotes the difference between the lowest asking price and the highest bid for an asset and is a measure of liquidity. Standard deviation of asset turnover is a measure of asset liquidity volatility. Chou et al. (2011) show that an increase in option liquidity corresponds with an increase in the level of the implied volatility curve. This positive relation between liquidity and IV might explain why these variables are important for IV estimation. My results are in line with Crego et al. (2023) as they find that liquidity-related characteristics explain most of the variation in expected returns. Furthermore, Crego et al. (2023) show that for their predicted yield analysis for the S&P500 assets, bid-ask spread is the most import variable and asset turnover, which closely relates to the standard deviation of asset turnover, is the fifth most important variable. Next is the variable idiosyncratic return volatility. This variable shows the volatility associated with a specific individual asset. The importance of this variable is in line with Crego et al. (2023), as their variable importance shows that for predicted yield, idiosyncratic return volatility is the second most important variable for the S&P500 stocks. Lastly, research and development to sales ratio is a measure of a firm's level of investment in innovation activities. This variable is not amongst the top 5 most important variables for predicted yields in the S&P500 sample, but is the sixth most contributing variable for the total testing set in Crego et al. (2023). This total testing set does however contain non-optionable stocks which are not included in my dataset. Summarizing, the three most contributing variables of both 30d and 91d maturities are in line with the results of Crego et al. (2023), further confirming the likeness of the IV and predicted yields

Lastly, I discuss the implications of the results on the hypothesis and research question of this thesis. Table 4 shows NN outperforms in more than 86% of the cases. Based on these results, I conclude that the results are in favour of the hypothesis that neural network outperforms linear regression in forecasting implied volatility based on firm characteristics. This also answers the first section of the research question. When changing the option maturity from 30d to 91d, OLS' metrics improve more than NN's, but the percentage of cases in which NN is superior only drops from 87.6% to 86.1%. Even though this is not a formal test, I conclude that option maturity does not appear to have a large effect on the relative performance between NN and OLS. This is in contrast with my expectation, which is likely caused the expectation was poorly founded.

VI. Application

In this section, I show how one can use the NN to form an actual trading strategy. Using the Black-Scholes formula (Appendix A) one can derive that, ceteris paribus, an increase (decrease) in implied volatility is associated with an increase (decrease) in option price. This means that investors should hold call options that are expected to rise in IV to profit off the associated expected call price increase. Conversely, one should write call options that are expected to drop in IV to profit off the associated expected call price decrease. The first step to form an actual trading strategy is to use NN to forecast IV as done in this research. Next, one could compute the monthly IV changes for all options and sort the options into deciles based on their IV change levels. Decile 10 contains the so-called *winner* options with the highest positive (or least negative) expected changes in IV, while decile 1 contains the so-called *loser* options with the highest negative (or least positive) expected changes in IV. Next, similar to a long-short portfolio for stocks, one can construct a hold-write portfolio for options, holding call options from the winner decile (10) and writing call options from the loser decile (1). As the contents of the deciles are subject to change each month, the hold-write portfolio has to be monthly-rebalanced. This is just one example of many possible trading strategies that include IV forecasting using NN.

VII. Conclusion

I examine the relative performance of neural networks and linear regressions at the task of forecasting implied volatility using firm characteristics. I use end-of-the-month data from at-the-money US call options with 30-day and 91-day maturities and 94 firm characteristics of Gu et al. (2020), from 1996 until 2021. Using a grid search, I find that the best performing NN model is able to produce higher out-of-sample R-squared and lower mean squared error values than OLS for both maturities. Using the Diebold-Mariano test, NN is the more accurate forecast model for more than 86% of the options. The maturity of the options does not appear to play a major role in the Diebold-Mariano analysis. NN's outperformance is likely attributable to that fact that its performance is more stable in periods in which OLS experiences sharp accuracy decreases. The most important drivers in explaining cross-sectional variation in IV are bid-ask spread and R&D-to-sales ratio, for option maturities 30 and 91, showing from neural network variable importance analysis

References

Baker, M., & Wurgler, J. (2007). Investor sentiment in the stock market. *Journal of economic perspectives*, *21*(2), 129–151.

Balcilar, M., Bonato, M., Demirer, R., & Gupta, R. (2018). Geopolitical risks and stock market dynamics of the brics. *Economic Systems*, *42*(2), 295–306.

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

Cao, C., Simin, T., & Xiao, H. (2020). Predicting the equity premium with the implied volatility spread. *Journal of Financial Markets*, *51*, 100531.

Chou, R. K., Chung, S.-L., Hsiao, Y.-J., & Wang, Y.-H. (2011). The impact of liquidity on option prices. *Journal of Futures Markets*, *31*(12), 1116–1141.

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

Crego, J. A., Soerlie Kvaerner, J., & Stam, M. (2023). Machine learning and expected returns. *Available at SSRN 4345646.*

Fama, E. F., & French, K. R. (2015). A five-factor asset pricing model. *Journal of financial economics*, *116*(1), 1–22.

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

Hondroyiannis, G., & Papapetrou, E. (2001). Macroeconomic influences on the stock market. *Journal of economics and finance*, *25*(1), 33–49.

Hosker, J., Djurdjevic, S., Nguyen, H., & Slater, R. (2018). Improving vix futures forecasts using machine learning methods. *SMU Data Science Review*, *1*(4), 6.

Jarrett, K., Kavukcuoglu, K., Ranzato, M., & LeCun, Y. (2009). What is the best multi-stage architecture for object recognition? *2009 IEEE 12th international conference on computer vision*, 2146–2153.

Lei, J., G'Sell, M., Rinaldo, A., Tibshirani, R. J., & Wasserman, L. (2018). Distribution-free predictive inference for regression. *Journal of the American Statistical Association*, *113*(523), 1094–1111.

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

Martin, I. W., & Wagner, C. (2019). What is the expected return on a stock? *The Journal of Finance*, *74*(4), 1887–1929.

Rubinstein, M. (1994). Implied binomial trees. *The journal of finance*, *49*(3), 771–818.

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

Vortelinos, D. I. (2017). Forecasting realized volatility: Har against principal components combining, neural networks and garch. *Research in international business and finance*, *39*, 824–839.

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

Welch, I., & Goyal, A. (2008). A comprehensive look at the empirical performance of equity premium prediction. *The Review of Financial Studies*, *21*(4), 1455–1508.

Appendix

Appendix A; IV Profitability Example

The Black-Scholes formula is defined as:

$$C = S_t \cdot \mathcal{N}(d_1) - K \cdot e^{-rt} \cdot \mathcal{N}(d_2) \tag{8}$$

$$P = K \cdot e^{-rt} \cdot \mathcal{N}(-d_2) - S_t \cdot \mathcal{N}(-d_1) \tag{9}$$

$$d_1 = \frac{\ln\left(\frac{S_t}{K}\right) + \left(r + \frac{\sigma^2}{2}\right) t}{\sigma\sqrt{t}} \tag{10}$$

$$d_2 = d_1 - \sigma\sqrt{t} \tag{11}$$

, where C and P represent the call and put option prices, respectively. S_t is the spot price, K is the strike price, r is the risk-free rate, t is the time to maturity, and σ is the volatility of the underlying asset. $\mathcal{N}(\cdot)$ denotes the cumulative distribution function of the standard normal distribution. Consider a European call option with $S_t = 110$, $k = 100$, $r = 3\%$, $t = 1$ year or $t_d = 365$ days. The call option is listed for \$14.70. Solving for these values in equations (9), (11) and (12) results in $\sigma = 15\%$. Suppose one forecasts that tomorrow the underlying asset's volatility will be $\sigma = 16\%$. Using the same equations (and $t_d = 364$), one can obtain that the call option should be worth \$15.00. Ceteris paribus, the call option is currently undervalued, meaning one theoretically gains \$0.30 by buying this call option. This example illustrates how to profit of the option Greek *vega*, meaning the price sensitivity to changes in implied volatility. This example is purely illustrative and is not meant to be realistic.

Appendix B; GKX Firm Characteristics

Table B1: Details of the Firm Characteristics

No.	Acronym	Firm characteristic	Paper's author(s)	Year, Journal	Data Source	Frequency
1	absacc	Absolute accruals	Bandyopadhyay, Huang	2010, WP	Compustat	Annual
2	acc	Working capital accruals	Sloan	1996, TAR	Compustat	Annual
3	aeavol	Abnormal earnings	Lerman, Livnat	2007, WP	Compustat+CRSP	Quarterly
4	age	# years since first	Jiang, Lee	2005, RAS	Compustat	Annual
5	agr	Asset growth	Cooper, Gulen	2008, JF	Compustat	Annual
6	baspread	Bid-ask spread	Amihud	1989, JF	CRSP	Monthly
7	beta	Beta	Fama	1973, JPE	CRSP	Monthly
8	betasq	Beta squared	Fama	1973, JPE	CRSP	Monthly
9	bm	Book-to-market	Rosenberg, Reid	1985, JPM	Compustat+CRSP	Annual
10	bm_ia	Industry-adjusted	Asness, Porter	2000, WP	Compustat	Annual
11	cash	Cash holdings	Palazzo	2012, JFE	Compustat	Quarterly
12	cashdebt	Cash flow to debt	Ou	1989, JAE	Compustat	Annual
13	cashpr	Cash productivity	Chandrashekar	2009, WP	Compustat	Annual
14	cfp	Cash flow to price ratio	Desai, Rajgopal	2004, TAR	Compustat	Annual
15	cfp_ia	Industry-adjusted	Asness, Porter	2000, WP	Compustat	Annual
16	chatoia	Industry-adjusted	Soliman	2008, TAR	Compustat	Annual
17	chcsho	Change in shares	Pontiff	2008, JF	Compustat	Annual
18	chempia	Industry-adjusted	Asness, Porter	1994, WP	Compustat	Annual
19	chinv	Change in inventory	Thomas	2002, RAS	Compustat	Annual
20	chmom	Change in 6-month	Gettleman	2006, WP	CRSP	Monthly
21	chpmia	Industry-adjusted	Soliman	2008, TAR	Compustat	Annual
22	chtx	Change in tax expense	Thomas	2011, JAR	Compustat	Quarterly
23	cinvest	Corporate investment	Titman, Wei	2004, JFQA	Compustat	Quarterly
24	convind	Convertible debt	Valta	2016, JFQA	Compustat	Annual
25	currat	Current ratio	Ou	1989, JAE	Compustat	Annual
26	depr	Depreciation / PP&E	Holthausen	1992, JAE	Compustat	Annual
27	divi	Dividend initiation	Michaely, Thaler	1995, JF	Compustat	Annual
28	divo	Dividend omission	Michaely, Thaler	1995, JF	Compustat	Annual
29	dolvol	Dollar trading volume	Chordia, Subrahmanyam	2001, JFE	CRSP	Monthly
30	dy	Dividend to price	Litzenberger	1982, JF	Compustat	Annual
31	ear	Earnings announcement	Kishore, Brandt	2008, WP	Compustat+CRSP	Quarterly

Table B1 (Continued): Details of the Firm Characteristics

No.	Acronym	Firm characteristic	Paper's author(s)	Year, Journal	Data Source	Frequency
32	egr	Growth in common	Richardson, Sloan	2005, JAE	Compustat	Annual
33	ep	Earnings to price	Basu	1977, JF	Compustat	Annual
34	gma	Gross profitability	Novy-Marx	2013, JFE	Compustat	Annual
35	grCAPX	Growth in capital	Anderson	2006, JF	Compustat	Annual
36	grltnoa	Growth in long term	Fairfield, Whisenant	2003, TAR	Compustat	Annual
37	herf	Industry sales	Hou, Robinson	2006, JF	Compustat	Annual
38	hire	Employee growth rate	Bazdresch, Belo	2014, JPE	Compustat	Annual
39	idiovol	Idiosyncratic return volatility	Ali, Hwang	2003, JFE	CRSP	Monthly
40	ill	Illiquidity	Amihud	2002, JFM	CRSP	Monthly
41	indmom	Industry momentum	Moskowitz	1999, JF	CRSP	Monthly
42	invest	Capital expenditures	Chen	2010, JF	Compustat	Annual
43	lev	Leverage	Bhandari	1988, JF	Compustat	Annual
44	lgr	Growth in long-term	Richardson, Sloan	2005, JAE	Compustat	Annual
45	maxret	Maximum daily return	Bali, Cakici	2011, JFE	CRSP	Monthly
46	mom12m	12-month momentum	Jegadeesh	1990, JF	CRSP	Monthly
47	mom1m	1-month momentum	Jegadeesh	1993, JF	CRSP	Monthly
48	mom36m	36-month momentum	Jegadeesh	1993, JF	CRSP	Monthly
49	mom6m	6-month momentum	Jegadeesh	1993, JF	CRSP	Monthly
50	ms	Financial statement	Mohanram	2005, RAS	Compustat	Quarterly
51	mvel1	Size	Banz	1981, JFE	CRSP	Monthly
52	mve_ia	Industry-adjusted size	Asness, Porter	2000, WP	Compustat	Annual
53	nincr	Number of earnings	Barth, Elliott	1999, JAR	Compustat	Quarterly
54	operprof	Operating	Fama, French	2015, JFE	Compustat	Annual
55	orgcap	Organizational	Eisfeldt, Papanikolaou	2013, JF	Compustat	Annual
56	pchcapx_ia	Industry adjusted	Abarbanell, Bushee	1998, TAR	Compustat	Annual
57	pchcurrat	% change in current ratio	Ou	1989, JAE	Compustat	Annual
58	pchdepr	% change in depreciation	Holthausen, Larcker	1992, JAE	Compustat	Annual
59	pchgm_pchsale	% change in gross margin	Abarbanell, Bushee	1998, TAR	Compustat	Annual
60	pchquick	% change in quick ratio	Ou	1989, JAE	Compustat	Annual
61	pchsale_pchinvt	% change in sales	Abarbanell, Bushee	1998, TAR	Compustat	Annual
62	pchsale_pchrect	% change in sales	Abarbanell, Bushee	1998, TAR	Compustat	Annual

Table B1 (Continued): Details of the Firm Characteristics

No.	Acronym	Firm characteristic	Paper's author(s)	Year, Journal	Data Source	Frequency
63	pchsale_pchxsga	% change in sales	Abarbanell, Bushee	1998, TAR	Compustat	Annual
64	pchsaleinv	% change sales-to-inventory	Ou	1989, JAE	Compustat	Annual
65	pctacc	Percent accruals	Hafzalla, Lundholm	2011, TAR	Compustat	Annual
66	pricedelay	Price delay	Hou, Moskowitz	2005, RFS	CRSP	Monthly
67	ps	Financial statements	Piotroski	2000, JAR	Compustat	Annual
68	quick	Quick ratio	Ou	1989, JAE	Compustat	Annual
69	rd	R&D increase	Eberhart, Maxwell	2004, JF	Compustat	Annual
70	rd_mve	R&D to market	Guo, Lev	2006, JBFA	Compustat	Annual
71	rd_sale	R&D to sales	Guo, Lev	2006, JBFA	Compustat	Annual
72	realestate	Real estate holdings	Tuzel	2010, RFS	Compustat	Annual
73	retvol	Return volatility	Ang, Hodrick	2006, JF	CRSP	Monthly
74	roaq	Return on assets	Balakrishnan, Bartov	2010, JAE	Compustat	Quarterly
75	roavol	Earnings volatility	Francis, LaFond	2004, TAR	Compustat	Quarterly
76	roeq	Return on equity	Hou, Xue	2015, RFS	Compustat	Quarterly
77	roic	Return on invested	Brown, Rowe	2007, WP	Compustat	Annual
78	rsup	Revenue surprise	Kama	2009, JBFA	Compustat	Quarterly
79	salecash	Sales to cash	Ou	1989, JAE	Compustat	Annual
80	saleinv	Sales to inventory	Ou	1989, JAE	Compustat	Annual
81	salerec	Sales to receivables	Ou	1989, JAE	Compustat	Annual
82	secured	Secured debt	Valta	2016, JFQA	Compustat	Annual
83	securedind	Secured debt indicator	Valta	2016, JFQA	Compustat	Annual
84	sgr	Sales growth	Lakonishok, Shleifer	1994, JF	Compustat	Annual
85	sin	Sin stocks	Hong, Kacperczyk	2009, JFE	Compustat	Annual
86	sp	Sales to price	Barbee, Mukherji	1996, FAJ	Compustat	Annual
87	std_dolvol	Volatility of liquidity	Chordia, Subrahmanyam	2001, JFE	CRSP	Monthly
88	std_turn	Volatility of liquidity	Chordia, Subrahmanyam	2001, JFE	CRSP	Monthly
89	stdacc	Accrual volatility	Bandyopadhyay, Huang	2010, WP	Compustat	Quarterly
90	stdcf	Cash flow volatility	Huang	2009, JEF	Compustat	Quarterly
91	tang	Debt capacity/firm	Almeida, Campello	2007, RFS	Compustat	Annual
92	tb	Tax income to book	Lev, Nissim	2004, TAR	Compustat	Annual
93	turn	Share turnover	Datar, Naik	1998, JFM	CRSP	Monthly
94	zerotrade	Zero trading days	Liu	2006, JFE	CRSP	Monthly

Source: Gu et al. (2020), supplementary material

Appendix C; Monthly IV Graphs

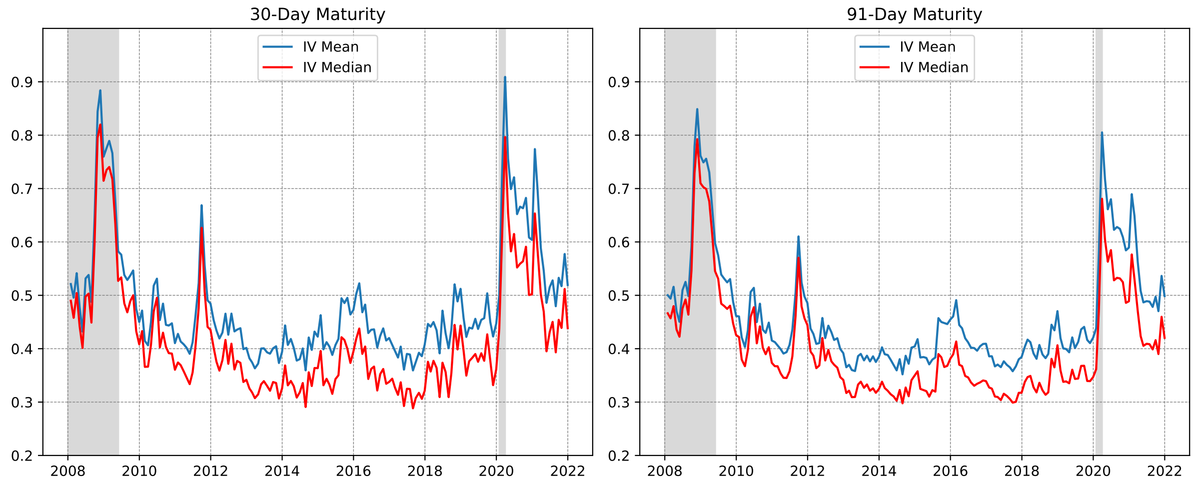


Figure C1: Monthly Mean and Median Implied Volatility

Y-axis represents monthly implied volatility level and the x-axis represents the period 2008-2022. Grey bars indicate periods of economic recessions according to National Bureau of Economic Research (01/2008 - 06/2009 and 02/2020 - 04/2020).

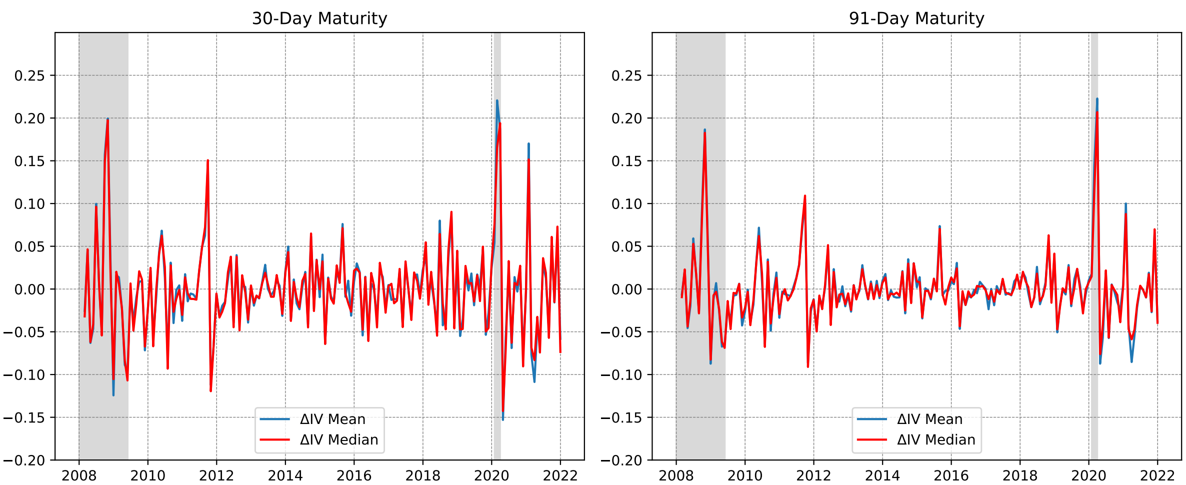


Figure C2: Monthly Mean and Median Implied Volatility change

Y-axis represents monthly change in implied volatility compared to the prior month and the x-axis represents the period 2008-2022. Grey bars indicate periods of economic recessions according to National Bureau of Economic Research (01/2008 - 06/2009 and 02/2020 - 04/2020).