Can Machine Learning Improve Accuracy in Predicting the Implied Volatility Surface of American-Style Options for US Equities?

Alan Watters (657192aw)

ERASMUS UNIVERSITY ROTTERDAM
ERASMUS SCHOOL OF ECONOMICS
Master Thesis MSc Econometrics and Management Science (Quantitative Finance)

Abstract

In this paper we investigate whether machine learning techniques are an effective tool in improving the predictability of the implied volatility surface (IVS) beyond the accuracy of traditional parametric modelling. We examine the efficacy of a two-step methodology for predicting the IVS applied to individual American-style equity options. Our approach involves initially modelling the IVS parametrically. We then fit the model implied errors to a feedforward neural network and evaluate its ability to correct the errors. To assess the accuracy of the nonparametric corrections, we use a large dataset of the options of the US equities Amazon, JPMorgan and Microsoft. In our analysis, we apply the correction to three models, Black-Scholes, Ad-Hoc Black-Scholes and Carr and Wu. The purpose for the range of the models is to asses the versatility of this two-step framework and the importance of the inital parametric model within it. Our research has shown effectiveness of this framework for improving predictability of the IVS for equity options. We found that the accuracy of initial the parametric model used was important in determining the accuracy of the corrected model, the parametric models performed best for JPMorgan and worst for Amazon. This translated to the corrected models with JPMorgan again

yielding the best results and Amazon the worst. We found that as time-horizon increased the accuracy of our framework decreased. Lastly, inclusion of macro-economic features gave no improvement in prediction accuracy.

Key abbreviations: Black-Scholes (BS), Ad-Hoc Black-Scholes (AHBS), Carr and Wu (CW), Implied Volatility (IV), Feed-Forward Neural Network (FFNN), Implied Volatility Mean Squared Error (IVRMSE)

Supervisor:	dr. M. Grith
Second assessor:	dr. A. Pick
Submission date:	13th March 2024

Contents

1 Introduction

Options are financial instruments whose prices are determined by four key characteristics: price of underlying, strike price, time to maturity and future underlying volatility, the only unobservable parameter. The three observable parameters can vary greatly between option contracts which makes direct comparison by price difficult. Two options on the same underlying, will have the same future volatility. This gives volatility a crucial role in option pricing analysis as it offers a method of comparison between option contracts. The volatility in question is theoretically the future volatility of the underlying security up until expiry. In practice, as the future is yet to occur, we use (model) implied volatility (IV). IV, for each option, is found by calculating which volatility, given the observed values for the other characteristics, gives us the current option price given the Black-Scholes model. When implied volatility is measured across the entire cross-section of moneyness[1] and time to expiry we call it the implied volatility surface (IVS). Our analysis focuses on predicting the IVS over multiple time horizons[2]. A key assumption of the Black-Scholes model is that volatility is constant across time, moneyness and maturity. Since 1973 it has been accepted this assumption is imperfect. Particularly Dupire et al. (1994) found that the IVS of options exhibits a 'skew' or 'smile'. This finding has motivated many researchers to attempt to find improved option pricing models which means there is a plethora of parametric models available today. Improvements have been made in the parametric modelling of options but we are still yet to, and may never, find the perfect parametric model. I would argue this is because there are non-linear relationships at play. Our paper is testing the robustness of the framework first proposed by Almeida et al. (2022). The researchers found great success from their two-step framework for European-style index options. In our paper we will extend their framework to working with American-style equity options which in Bernales and Guidolin (2014) were found to exhibit predictability. First, we fit the parametric model to the observed implied volatilities. Then, we non-parametrically estimate the model-implied pricing error function.

For the non-parametric correction a neural network was the natural choice. Neural networks have been extensively used in asset pricing and have been proven to be more effective than other ML techniques as in Gu et al. (2020). We use out-of-sample prediction accuracy to compare across model, time-horizon and equity. We find strong evidence that the two-step framework is more effective in predicting the IVS than parametric modelling alone. We found adding macro-economic factors to the data further improves results, but under specific circumstances. In order to reap this benefit you must sacrifice re-fitting the NN each day, which yields the best results. Under the restriction of fitting a single NN to the entire dataset, adding macro-economic variables improves prediction accuracy. Our macro-economic analysis showed that the VIX index was by far the most important addi-

[1] Moneyness, m, of an option is defined as: $m = S/K$ where S is underlying price and K is strike price

[2] Four to be precise: Same-day, 1-day, 5-day and 20-day

tional feature. We found the least important features were the two interest rate proxies, the 3 Month treasury bill rate being slightly more important than the 10 Year treasury bond rate. The remainder of the paper is organised as follows. There is a brief review of related literature next. In Section 2 we discuss the data, handling, processing and practicalities. Section 3 details the modelling, both parametric and non-parametric. Section 4 discusses the results. Section 5 explores the of inclusions additional macroeconomic variables and Section 6 concludes the research.

1.1 Literature Review

This paper is built on the fundamentals of parametric options pricing, The Black-Scholes model first proposed by Black and Scholes (1973), offered the first closed-form solution for option prices. Despite being revolutionary there are some clear limitations of the BS model, namely the assumption volatility is constant across both strike price and time to expiry, giving a constant implied volatility surface. It is now widely recognised that the IV of options exhibits a volatility smile/skew Dupire et al. (1994), the earliest papers that found evidence of this phenomena did not formulate the results in such terms but instead described how Black Scholes pricing errors vary systematically with strike price or with time to expiry. For example, MacBeth and Merville (1979) reported that the Black-Scholes model undervalues in-the-money and overvalues out-of-the-money call options. Research from Rubinstein (1985) then tested the null hypothesis that implied volatility is constant across strike prices and yielded statistically significant results rejecting their null hypothesis, disproving the constant volatility assumption that is imperative to the Black-Scholes model. Rubinstein's most robust result is that for out-of-the-money calls implied volatility is systematically higher for options with shorter times to expiration. Later papers treated local volatility as a deterministic function of other parameters and applied a smoothing effect to volatility estimation. The Ad-Hoc Black-Scholes (AHBS) model, first proposed in Dumas et al. (1998), adapts the BS model over time to improve accuracy. It does this by incorporating the extent of volatility variation across time and moneyness using data from previous observations. ,

The BS model is not perfectly adapted for continuous data and in practice option prices are a set of continuous data. Papers from Heston (1993), Bates (2000), Duffie et al. (2000) and Andersen et al. (2015) have all attempted to account for this by implementing mechanisms that account for added risk associated with continuous data such as stochastic volatility and jump risk. The dynamics of volatility modelling are complicated. Luckily, in 2022 Christensen et al. (2021) analysed several machine learning techniques for volatility modelling with neural networks producing the most best results. Thus we will use a neural network for the non-parametric adjustment of the model errors in this paper.

In 2014, a paper from Bernales and Guidolin (2014) showed that individual equity options exhibit predictability in their IVS and eight years later Almeida et al. (2022) was re-

leased, which showed feed forward neural networks (FFNNs) were able to substantially improve the accuracy of IVS predictions for several option pricing models, for index options. The Almeida methodology coupled with the Bernales findings motivated this paper, an attempt to extend the Almeida methodology to be used for individual equity options. Our research is assessing the frameworks feasibility for predicting the IVS of the equities; Amazon, JPMorgan and Microsoft. If it is useful beyond index options it will have large ramifications for the broader financial industry and the approach to IVS forecasting for many options.

2 Data

For our research we selected 3 large-volume US equities; Amazon, the global e-commerce retailer with ticker AMZN, JPMorgan Chase & Co, the world's largest bank with ticker JPM and Microsoft, a technology giant the world's biggest company with ticker MSFT. Options for these three equities are traded at the Chicago Board Options Exchange (CBOE). Our dataset is all the qualifying options over a sample period of 2 years from January 1 2017 to December 31 2018. This data was obtained from OptionMetrics.

We followed standard practice when cleaning our data set. Firstly, following the industry standard, we removed all ITM[3] options and deal only with OTM[4]options in our analysis. Then we removed all option contracts with less than 10 or more than 240 trading days to expiry. Options with an excessively short time to expiry are hyper sensitive to market noise while options with an excessively long time to maturity are, to a point, much unaffected by macro-economic conditions compared to nearer term options. Because of this options with extreme time to maturity at either side of the spectrum contribute little to describing the overall IVS shape.

Next, similarly to Dumas et al. (1998) and Heston and Nandi (2000), we excluded contracts with extreme moneyness. Our classification of extreme moneyness is less than 0.70 or greater than 1.60. At extreme moneyness option IVs become uninformative for the overall IVS. In Almeida et al. (2022) their categorisation of extreme moneyness was $m < 0.8$ or $m > 1.4$. The Almeida paper is based on index options, We are dealing with individual equities which typically have a higher volatily than an index. Therefore, we felt it would be prudent to extend our categorisation of extreme moneyness. The remaining dataset consists of Put options with $m_{i,t} \in [1.00, 1.60)$ and Call options with $m_{i,t} \in (0.70, 1.00]$. Following Bakshi et al. (1997) and Goncalves and Guidolin (2006), we excluded option contracts with prices lower than 1.00. Options with excessively low prices are heavily impacted by the effects of price discreteness which can heavily skew their IVs. In all three models we are remaining in the implied volatility space this means in our model estimations we don't use the

[3]ITM means In The Money. An ITM option is an option with non-zero intrinsic value
[4]OTM means Out of The Money. An OTM option has no intrinsic value

option price. Consequently we don't need to include the interest rate or dividend rate as the purpose of both is to discount the option price. On any given day these will be the same for all options of a given ticker.

An initial summary of our data is included in Table 1. Options with less than 60 days to expiry were defined as short-term and options with between 60 and 240 days to expiry were defined as long-term. This table shows a majority of our options are OTMC, ATM or OTMP. For both companies there is a relatively even overall number of options with both short and long maturities. Interestingly there are more ATM options with short maturities but for every other option type there are more with long maturities.

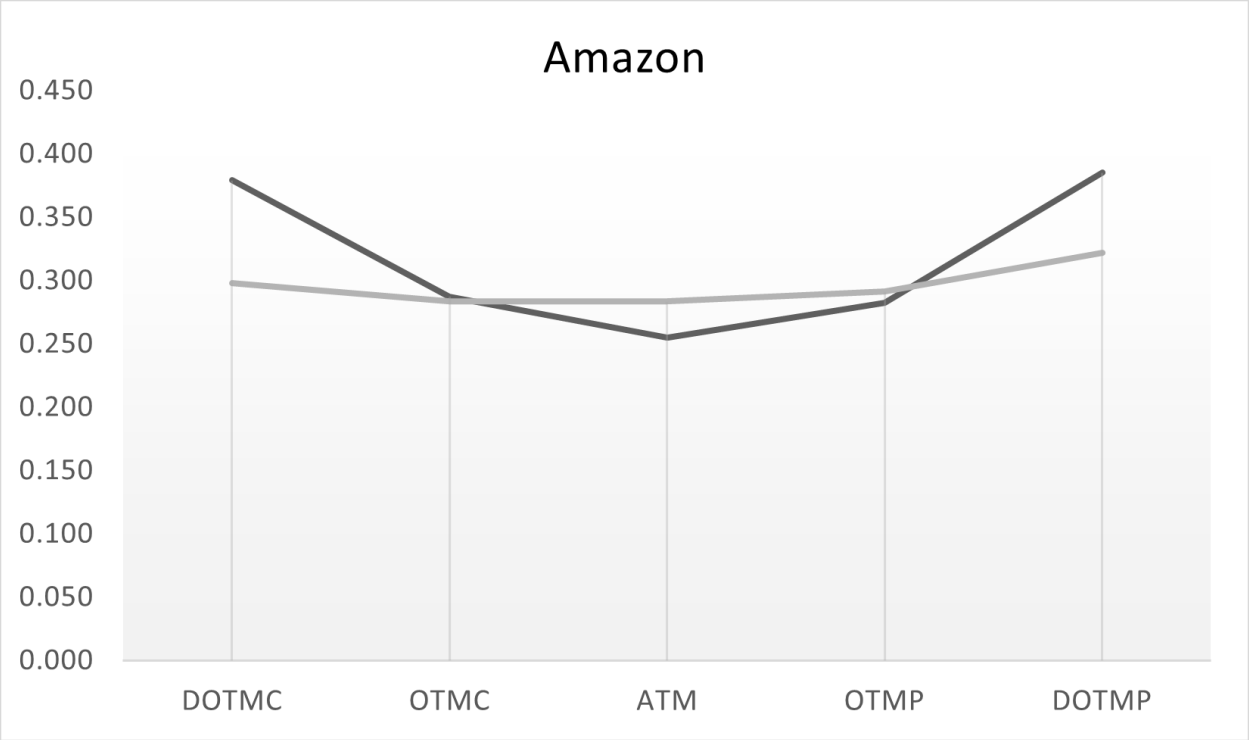
Table 1: Option Cross-Section

		Number		μ_{IV}		σ_{IV}	
Company	Moneyness/Expiry	Short	Long	Short	Long	Short	Long
Amazon	DOTMC: [0.7, 0.9)	1647	1396	37.92%	29.78%	8.35%	5.56%
	DOTMC: [0.9, 0.97)	5900	1399	28.73%	28.36%	9.37%	5.70%
	ATM: [0.97, 1.03)	10847	1627	25.50%	28.33%	8.98%	5.47%
	OTMP: [1.03, 1.1)	7159	917	28.26%	29.11%	8.85%	5.64%
	DOTMP: [1.1, 1.6]	2920	1682	38.52%	32.19%	11.48%	6.05%
	Total	28473	7021	28.92%	29.65%	9.25%	5.70%
JPMorgan	DOTMC: [0.7, 0.9)	1	255	30.27%	20.33%	0.00%	2.58%
	DOTMC: [0.9, 0.97)	352	1633	22.90%	19.16%	5.06%	2.12%
	ATM: [0.97, 1.03)	3855	1599	20.25%	19.64%	4.51%	2.24%
	OTMP: [1.03, 1.1)	495	1069	25.03%	21.70%	5.66%	2.47%
	DOTMP: [1.1, 1.6]	24	576	37.91%	25.09%	10.69%	3.26%
	Total	4727	5132	21.04%	20.56%	4.71%	2.38%
Microsoft	DOTMC: [0.7, 0.9)	10	662	37.28%	23.77%	5.71%	3.62%
	DOTMC: [0.9, 0.97)	929	2274	29.13%	21.49%	6.24%	3.56%
	ATM: [0.97, 1.03)	5144	2240	24.28%	21.40%	6.55%	3.66%
	OTMP: [1.03, 1.1)	825	1569	30.60%	23.18%	7.42%	3.78%
	DOTMP: [1.1, 1.6]	69	1006	40.85%	26.68%	10.08%	4.56%
	Total	6977	7751	25.85%	22.67%	6.64%	3.37%

In table Table 1 there is a clear pattern in the IVs of the options. ATM options have low IV and options with more extreme moneyness have higher IV. We plotted this in Figures 1-3 with the darker line being short-term options and the lighter line being long-term options. The graphs are showing average IV plotted on the y-axis and option type plotted on the x-axis. This graph supports the idea of a volatility skew/smile as there is a clear violation of the Black-Scholes model's constant volatility assumption. The skew/smile/smirk has been heavily researched by Dupire et al. (1994) and our data supports its existence. Notably the standard deviation of short term options is always considerably higher than their long

term equivalents [5], sometimes even greater than 3 times larger, as for JPMorgan DOTMP options. This observation is consistent with the idea that short term options are more sensitive to news that their long term options. Figure 1, Figure 2 and Figure 3 show mean IV plotted against option type for each of our three equities. The dark lines are short-term options and the light lines are long-term options. This figure is clear evidence for the existence of volatility skew with all six curves exhibiting some level of skew. For all three equities both the short and long curves have IV increasing as moneyness gets more extreme. [6]

Figure 1: Amazon Mean IV by Option Type



[5]Except for DOTMC JPM which has a single option therefore no standard deviation.
[6]Except $IV_{ATM} > IV_{OTMC}$ for both Amazon and JPMorgan.

Figure 2: JPMorgan Mean IV by Option Type

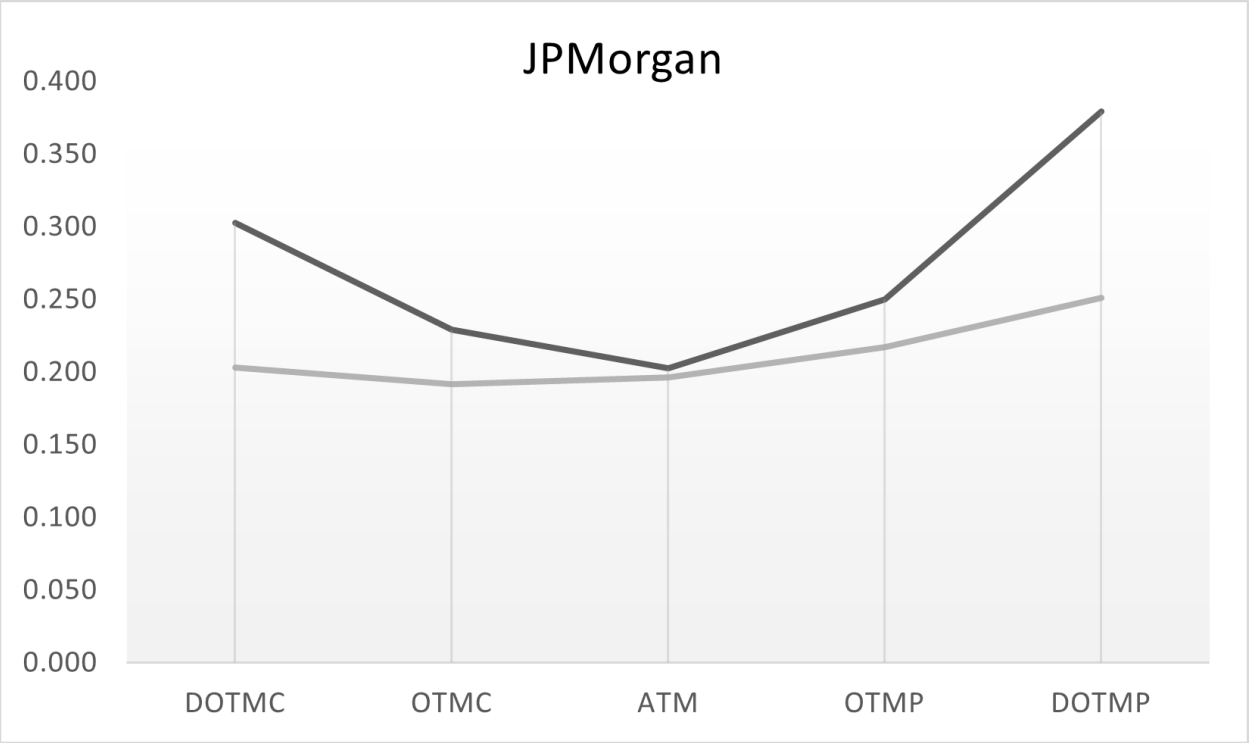
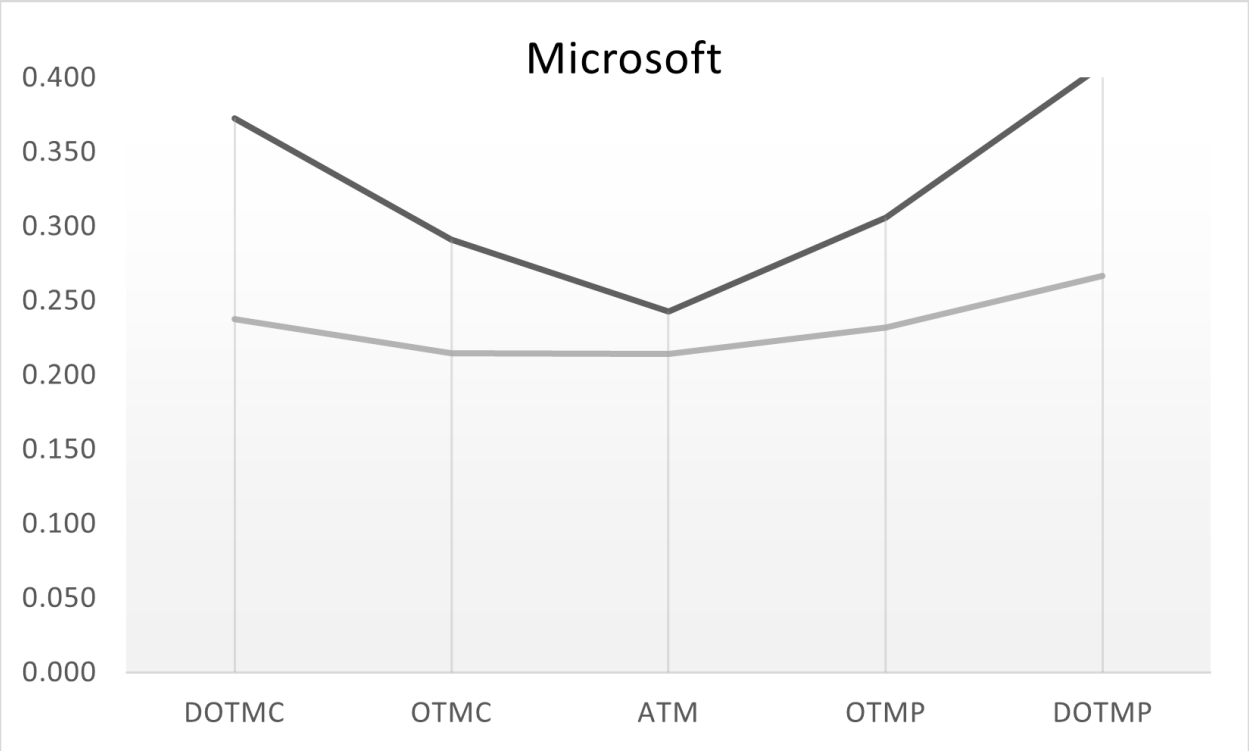


Figure 3: Microsoft Mean IV by Option Type



2.1 Practicalities

After cleaning the datasets we are left with 35,494, 4,771 and 14,728 options for Amazon, JPMorgan and Microsoft respectively. For our prediction exercise we split our dataset into training and testing. In order for both our testing and training data to have a full spectrum of moneyness and time to maturity in the sets we chose a divisor [7] such that any options with strike price divisible by said divisor were put into the training set and the rest in the testing set. Our goal was to have between 65 and 80 percent of our data allocated for training, our chosen divisors gave us 77:23, 65:35 and 69:31 train-test data splits for each of our equities.

When constructing our datasets for our four different prediction horizons we only included trading days that had at least 5 options in both the testing and training sets. The sample size for each prediction horizon is shown in table Table 2.

Table 2: Trading Days in Analysis

	Horizon			
Company	Same-day	1-day	5-day	20-day
Amazon	427	406	400	386
JPMorgan	159	152	147	133
Microsoft	301	279	269	246

3 Methodology

Our methodology intends to build on the research of Almeida et al. This paper found that a FFNN is effective in improving the predictive power of more traditionally used parametric option pricing models when using European-style S&P500 options. In 2014, research from from Bernales et al. found that American-style equity options exhibit predictability. In our paper, we want to combine these findings and investigate whether the two-step framework used by Almeida et al. is effective at capturing the predictability of American-style equity options. As our option data is made up of American-style options we don't have data for European option prices. This makes finding an IV estimate via reverse-calculating the model formulas tedious. Instead we opted to remain in the implied volatility space for all our models.

3.1 Parametric Option Pricing Models

In our research we used models that vary in complexity in hopes to yield initial parametric model results which vary in accuracy, this will allow us to analyse the effect of the initial models accuracy on the non-parametric models ability to correct errors.

[7] Our divisor was 2 for Amazon and 1 for JPMorgan and Microsoft

3.1.1 Black-Scholes

When pubished in 1973 the Black-Scholes Model was ground-breaking as it was the first formalised option pricing model which offered a closed-form solution.
The model expressed as a differential equation is as follows:

$$\frac{dS_t}{S_t} = \mu + \sigma dW_t, \tag{1}$$

The formula for the Put and Call prices both have five unknowns. [8] [9]

$$Call_{BS}(S_t, K, \tau, r, \sigma) = S_t\Phi^{10}(d_1) - Ke^{-r\tau}\Phi(d_2) \tag{2}$$

$$Put_{BS}(S_t, K, \tau, r, \sigma) = Ke^{-r\tau}\Phi(-d_1) - S_t\Phi(-d_2) + \tag{3}$$

$$d_1 = \frac{ln(S_t/K) + (r + \sigma^2/2)\tau}{\sigma\sqrt{\tau}} \tag{4}$$

$$d_2 = d_1 - \sigma\sqrt{\tau} \tag{5}$$

The BS model assumes a constant IVS across time, moneyness and time to expiry. Despite our research focus on American-style options, an option type which Black-Scholes wasn't specifically designed for, we felt it would provide valuable insight into the effectiveness of this framework to correct consistently inaccurate models.
As we are remaining in the implied volatility space, our Black-Scholes estimate was calculated by the below minimisation.

$$\hat{\sigma}_{BS} = \sigma : \min_{\sigma} \frac{1}{n}\sum_{i=1}^{n} [\sigma_{i,t} - \sigma]^2 \text{ with } \sigma > 0 \tag{6}$$

3.1.2 Ad-Hoc Black Scholes

A key consequence of the Black-Scholes model is that the IV is constant across moneyness and time to expiry. This is an idea that has been repeatedly refuted over the years, for example by Dupire et al. and Heston et al., even our own option panel in Table 1 seemingly refutes the assumption. Somewhat motivated by this failing, a paper from Dumas et al. was

[8] Stock Price (S), Strike Price (K), Risk Free Rate (r), Volatility (σ) and Time to Expiry (τ)
[9] $\tau = (T - t)$, T being the expiration date and t the current date

released developing the Ad-Hoc Black-Scholes (AHBS) model. The AHBS model is characterised by specifying implied volatility as a quadratic function of option moneyness and time to expiry. Importantly the AHBS model assumes a time-varying IV, meaning it is to be calculated for each day as well as for each moneyness and time to expiry.

$$\sigma_{i,t} = \beta_{0,t} + \beta_{1,t} m_{i,t} + \beta_{2,t} m_{i,t}^2 + \beta_{3,t} \tau_i + \beta_{4,t} \tau_i^2 + \beta_{5,t} m_{i,t} \tau_i + \epsilon_{i,t} \tag{7}$$

Using observed values for $\sigma_{i,t}$ to get parameter estimates via ordinary least squares (OLS). This is done by minimizing the IVMSE such that:

$$\min_{\beta} \frac{1}{n} \sum_{i=1}^{n} [\sigma_{i,t} - \sigma_{AHBS}(\beta_t, m_{i,t}, \tau_{i,t})]^2 \tag{8}$$

with

$$\sigma_{AHBS(\beta_t, m_{i,t}, \tau_{i,t})} = \beta_{0,t} + \beta_{1,t} m_i + \beta_{2,t} m_i^2 + \beta_{3,t} \tau_i + \beta_{4,t} \tau_i^2 + \beta_{5,t} m_i \tau_i \tag{9}$$

Given the estimated parameters $\hat{\beta}_t$, the IV predicted by the AHBS model for an option with moneyness, m and time to expiry, τ is: $\sigma_{AHBS}(\hat{\beta}_t, m, \tau)$ [11]

3.1.3 Carr and Wu Model

The Carr and Wu (CW) model was first proposed in Carr and Wu (2016) with the goal of better incorporating position management techniques, used by institutional investors, into option pricing. The Carr and Wu model starts with the near-term dynamics of the IVS and derives no-arbitrage constraints from its current shape. The Carr and Wu model, like the AHBS model, is recalculated on a daily basis. This means we can specify the near-term dynamics of the IVS while leaving its long-term variation unspecified and highlights the models 'semi-parametric' flavour. The theory specifies just enough dynamic structure to achieve a fully parametric characterisation of the IVS. The Carr and Wu model offers a computationally efficient framework and a straight-forward quadratic solution. An interesting and important feature of Eq. (10) is that the no-arbitrage constraint depends only on the current levels of the five dynamic processes (υ_t, π_t, λ_t, ω_t, ρ_t), but it does not depend directly on the exact dynamics of these processes. Thus, the dynamics of the five state variables are left unspecified. As a result, fitting the relation to observed implied volatility surfaces only involves extracting the current values of the five dynamic states, which we calculate by least squares on the training data set, but does not involve the estimation of any model parameters that govern their state dynamics.

The Carr and Wu model assumes that the stock price is log-normally distributed with

[11] With the constraint that $\sigma_{AHBS}(\hat{\beta}_t, m, \tau) > 0 \forall t$.

time-varying volatility and has 5 time-varying parameters: (υ, π, λ, ω, ρ). The models differential equations are defined as:

$$\frac{dS_t}{S_t} = \sqrt{\upsilon_t} dW_t \tag{10}$$

$$\frac{d\sigma_t(K,\tau)}{\sigma_t(K,\tau)} = e^{-\lambda_t \tau}(\pi_t dt + w_t dZ_t) \tag{11}$$

Given that we are going to be calculating daily estimates using the Carr and Wu model, it is sufficient to assume these parameters on each day are constants. The Carr and Wu model's five dynamic processes have some constraints[12] which ensure feasibility of the model. For each day the value of our 5 parameters, $\beta_t = (\upsilon_t, \pi_t, \lambda_t, \rho_t, \omega_t)$ are found by minimizing Equation 12 across all the options in the training dataset for day t[13].

$$\hat{\beta}_t = (\hat{\upsilon}_t, \hat{\pi}_t, \hat{\lambda}_t, \hat{\rho}_t, \hat{\omega}_t) = {}_{arg_{\beta_t}\, min} \sum_{i=1}^{n} [\frac{1}{4} e^{-2\lambda_t \tau_i} \omega_t^2 \tau_i^2 \sigma_i^4 + (1 - 2e^{-\lambda_t \tau_i} \pi_t \tau_i - e^{-\lambda_t \tau} \omega_t \rho_t \sqrt{\upsilon_t} \tau_i) \sigma_i^2 - (\upsilon_t + 2e^{-\lambda_t \tau_i} \omega_t \rho_t \sqrt{\upsilon_t} \kappa_i + e^{-2\lambda_t \tau_i} \omega_t^2 \kappa_i^2)] \tag{12}$$

Each $\hat{\sigma}_{CW}$ is then obtained by solving Equation 13.

$$\hat{\sigma}_{CW\,i,t} = \sigma : [\frac{1}{4} e^{-2\hat{\lambda}_t \tau_i} \hat{\omega}_t^2 \tau_i^2 \sigma^4 + (1 - 2e^{-\hat{\lambda}_t \tau_i} \hat{\pi}_t \tau_i - e^{-\hat{\lambda}_t \tau} \hat{\omega}_t \hat{\rho}_t \sqrt{\hat{\upsilon}_t} \tau) \sigma^2 - (\hat{\upsilon}_t + 2e^{-\hat{\lambda}_t \tau_i} \hat{\omega}_t \hat{\rho}_t \sqrt{\hat{\upsilon}_t} \kappa_i + e^{-2\hat{\lambda}_{t\tau_{i,t}}} \hat{\omega}_t^2 \kappa_i^2)]^2 = 0 \text{ with } \sigma \in \mathbb{R}^+ \tag{13}$$

3.2 Non-Parametric Correction

In this section we will discuss the second part of our framework in more detail. For the nonparametric correction of the our parametric model, as in Almeida, we are opting to use a feed forward neural network. After fitting a parametric model to the data we then fit the model-implied error to a FFNN.
The model implied error, of model p, is defined as:

$$\epsilon_p(m_{i,t}, \tau_{i,t}) = \sigma(m_{i,t}, \tau_{i,t}) - \sigma_p(m_{i,t}, \tau_{i,t}) \tag{14}$$

There are several types of non-parametric model we could fit the model implied errors to, the most promising and heavily researched is the FFNN. For a given model p the FFNN

[12] Carr and Wu model constraints: $\upsilon > 0$, $\pi \in \mathbb{R}$, $\lambda > 0$, $\rho \in [-1, 1]$, $\omega > 0$
[13] $\kappa_i = S_i / K_i$

works by minimizing the following objective function:

$$\min_{f}(\frac{1}{n}\sum_{i=1}^{n}[\hat{\epsilon}_p(m_{i,t},\tau_i) - f(m_{i,t},\tau_i)]^2) \tag{15}$$

Over an appropriate space of functions f.
The function $\hat{f}(m,\tau)$ is the one which best approximates the pricing errors of the parametric model. Our new estimate is:

$$\hat{\sigma}_{NN}(m,\tau) = \sigma_p(m,\tau) + \hat{f}(m,\tau) \tag{16}$$

3.2.1 Feed Forward Neural Network

Feed forward neural networks consist of an input layer of explanatory variables, in our case the model implied errors, and multiple intermediate hidden layers. The hidden layers are responsible for the nonlinear transformations. The number of hidden layers in a FFNN is the known as the depth and the number of nodes in a hidden layer is referred to as its width. The depth and width are both important factors in model effectiveness, we trialled different architectures and decided to use 3 hidden layers, with our initial layer having 128 input nodes as this yielded the best results for a subset of our data, details of this analysis are found in the appendix. We used the step property to determine the width of our remaining hidden layers and hence the widths are 64 and 32. Finally, there is the output layer which gives the predicted value.
The generalised formulas for a FFNN architecture are Equation 17 and Equation 18 .
Starting from $\mathbf{z_0} = \mathbf{x_{i,t}} = \epsilon_p(m_{i,t},\tau_{i,t}) \in \mathbb{R}$ we iterate using Equation 17:

$$\underset{d_l\times 1}{\boldsymbol{z}_l} = \mathring{h}(\underset{d_l\times d_{l-1}}{\boldsymbol{A}_{l-1}} \times \underset{d_{l-1}\times d_1}{\boldsymbol{z_l}} + \underset{d_l\times 1}{\boldsymbol{b}_{l-1}}), \text{ for } l = 1,...,L \tag{17}$$

$$f(\mathbf{x}_{i,t}) = \underset{1\times d_L}{\boldsymbol{A}_L}\ \underset{d_L\times 1}{\mathbf{z}_L} + \underset{1\times 1}{\mathbf{b}_L} \tag{18}$$

L is the number of hidden layers with each layer l containing d_l output neurons. $\boldsymbol{A}_{l-1}$ is the weight matrix, $\mathbf{b}_{l-1}$ is the intercept vector and $\mathring{h} : \mathbb{R}^{d_l} \to \mathbb{R}^{d_l}$ applies the activation function h(·) to each element of the vector $\boldsymbol{A}_{l-1}\mathbf{z}_{l-1} + \mathbf{b}_{l-1}$. We used the Rectified Linear Unit (ReLU) activation function defined as: $h(x) = \max(0,x)$.
The formula for our particular architecture is found in Equation 19.

$$f(\underset{1\times1}{\boldsymbol{x}_{i,t}}) = \underset{1\times32}{\boldsymbol{A}_3}\,\mathring{h}\Big[\underset{32\times64}{\boldsymbol{A}_2}\,\mathring{h}\Big(\underset{64\times128}{\boldsymbol{A}_1}\,\mathring{h}\Big[\underset{128\times1}{\boldsymbol{A}_0}\times\underset{1\times1}{\boldsymbol{z_0}}+\underset{128\times1}{\boldsymbol{b_0}}\Big]+\underset{64\times1}{\boldsymbol{b_1}}\Big)+\underset{32\times1}{\boldsymbol{b_2}}\Big]+\underset{1\times1}{\boldsymbol{b_3}} \tag{19}$$

4 Prediction of the Implied Volatility Surface

In this section, we conduct prediction exercises in the option cross-section based on a day-by-day model estimation, which is common practice in the option pricing literature. In fact, the AHBS and CW models are designed to fit the implied volatility surface on a given day. While the daily re-estimation approach is theoretically inconsistent with the Black and Scholes model which assumes that parameters are constant over time, these models are often also implemented on a day-by-day basis as in Bakshi et al. (1997), we chose to implement the BS model on a day-to-day basis to allow for comparison between all three models. Our model parameters change on a daily basis but our input data lacks a time-varying signal, we decided it is logical to isolate each day of predictions and apply our non-parametric correction to each day individually . The main performance metric for our analysis is the out-of-sample Implied Volatility Root Mean Square Error (IVRMSE). The IVRMSE is computed by aggregating prediction errors across the testing samples. We find this error metric particularly valuable because it is expressed in terms of implied volatility, which is easily interpretable while also facilitating comparisons scross equities and models. Additionally, it aligns well with the modeling process and our analysis of implied volatility surface. [14]

We investigate the performance of 6 total models. Consisting of our 3 parametric option pricing models BS, AHBS and CW and our 3 non-parametrically corrected models BS+NN3[15], AHBS+NN3, CW+NN3. Our main goal is to answer whether neural networks can successfully correct and improve upon the parametric models by learning their pricing error surface. Secondarily we want to determine which parametric model performs best pre and post non-parametric correction. For each equity we split the datasets by day, we then split the daily data into a training and testing set. We fit each model to the training set and tested the model on the appropriate testing set. For same-day prediction we used the training and testing datasets of the same day and for 20-day ahead prediction we used the training set on day X and the testing set on day X+20. The FFNN is applied to the model-implied errors of the training set and then the output model is tested on the corresponding testing set.

[14] Our IVRMSEs are a daily average of all eligible test days

[15] NN3 refers to a feed-forward neural network of depth 3.

4.1 Results

4.2 Predictions in 4 Time Horizons

In Table 2 we compare the IVRMSE of predictions for the three models aswell as the NN3 corrected models. In 100% of instances the NN3 correction improved the IVRMSE compared to the base pricing model. The best performing model for each company at each time horizon is highlighted in bold.

Table 3: Data Summary

		AMZN		JPM		MSFT		Mean	
Horizon	Model	No NN	NN3	No NN	NN3	No NN	NN3	No NN	NN3
Same-day	BS	3.49%	0.96%	1.71%	**0.49%**	2.46%	**0.69%**	2.55%	**0.71%**
	AHBS	1.81%	**0.64%**	1.75%	0.54%	6.87%	1.87%	3.48%	1.02%
	CW	4.49%	1.05%	2.28%	0.55%	4.79%	1.19%	3.85%	0.93%
1-day	BS	4.01%	**1.49%**	1.84%	**0.79%**	2.76%	**1.08%**	2.87%	**1.12%**
	AHBS	4.02%	1.61%	4.18%	1.66%	6.04%	2.02%	4.75%	1.76%
	CW	5.12%	1.70%	2.40%	0.86%	5.08%	1.82%	4.20%	1.46%
5-day	BS	4.96%	2.54%	2.07%	**1.04%**	3.38%	**1.59%**	3.47%	**1.72%**
	AHBS	5.10%	**2.57%**	6.90%	3.46%	4.47%	2.21%	5.49%	2.75%
	CW	6.87%	2.91%	2.69%	1.14%	5.68%	2.29%	5.08%	2.11%
20-day	BS	7.65%	**4.08%**	2.27%	**1.27%**	4.20%	**2.41%**	4.71%	**2.59%**
	AHBS	12.63%	6.89%	3.52%	2.01%	6.08%	3.29%	7.41%	4.06%
	CW	10.24%	5.01%	2.92%	1.41%	6.86%	3.21%	6.67%	3.21%

For most combinations of equity and model, increasing the prediction time horizon increases the IVRMSE. This is logical as the training data upon which the models are estimated becomes further removed from the test data as time horizon increases. The BS+NN3 is the clear best-performing model followed by CW+NN3 and then AHBS+NN3. Interestingly BS is also the best-performing base model which suggests initial parametric model performance is influential in the FFNN's ability to correct. CW and AHBS perform quite comparitively while CW+NN3 substantially outperforms AHBS+NN3, suggesting the dynamics of CW are better suited to the two-step framework. The AHBS model is particularly bad at predicting for JPMorgam which is interesting as it performs similarly to CW for Microsoft and even outperforms CW slightly for Amazon. The relative difference in IVRMSE of Microsoft for AHBS+NN3 and CW+NN3 is substantially lower than for the corresponding uncorrected models. This shows the flexibility of the framework to provide similar but not identical results for different parametric models

4.3 Prediction Accuracy

In Figure 4, Figure 5 and Figure 6 we are measuring the performance of 1-day ahead predictions for our models on a single day and comparing them before and after the non-parametric correction. For this section we wanted a day on which all three equities had at least 10 options in the test sample, we ultimately chose 17/01/2017 because our three equities also had a similar amount of options which makes for more interesting comparison.

In both the right and left panels we are plotting option moneyness [16] on the x-axis against IV % error [17] on the y-axis. In the left hand (LH) panel, the blue circles are the prediction errors of the parametric model and the red circles are the prediction errors of the two-step framework, that is after the parametric model errors have been fit to a FFNN. In the right hand (RH) panel, the black circles show the improvement[18] in IV % error for each observation and the dotted red lines show the the mean improvement for each model. Across all three option models, the NN adjustment contributes a significant improvement in the IV% error [19], ranging from 4.5% improvement up to 9.8%. The Black-Scholes model benefits the most from the non-parametric correction but not by a significant degree compared to Ad-Hoc Black-Scholes and Carr and Wu.

Firstly we have Figure 4 which analyses the different models for Amazon. If we look at the LH graphs we can see from the blue circles that the CW model performed best with BS and AHBS performing similarly to one another. CW had no predictions with greater than 20% error while both other models had several. After fitting the neural network, the CW model is still the best performing, however the CW has the smallest average correction at 3.6% compared to 5.1% and 4.8% for BS and AHBS respectively. This highlights the importance of the initial model accuracy when using this framework.

Now referencing Figure 5, from the LH panel, the blue circles show us that the three parametric models perform similarly with most predictions having an IV % error between 10% and 25%. Now looking at the RH panel, with the neural network correction the BS model clearly outperforms the other models and AHBS is the worst performing, BS benefits the most from the non-parametric correction with an average corrections of 9.8%. AHBS and CW have average corrections of 4.1% and 6.5% for. AHBS performs particularly poorly on the options with moneyness of less than 1.0, giving significant negative improvements for two of these observations.

In Figure 6 we are analysing Microsoft. In this instance the CW model is the best initial model with all predictions having a percentage error less than 15%. BS and AHBS perform similarly on average but the BS predictions are of lower volatility. The improvement percentages for all three models are similar with 7.8%, 5.2% and 5.7% this results in CW also

[16] $moneyness = StockPrice/StrikePrice$

[17] IV % error is defined as: $(\frac{|IV - IV_{model}|}{IV})100$

[18] $improvement = (IV\%Error)_{model} - (IV\%Error)_{NN}$

[19] We opted to use IV % error in this section due to isolating individual observations for analysis

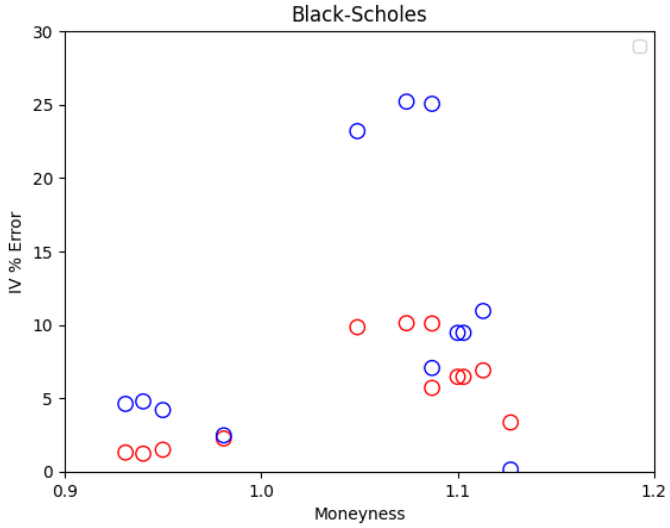
having the best results after the neural network correction, again highlighting the importance of the initial model.

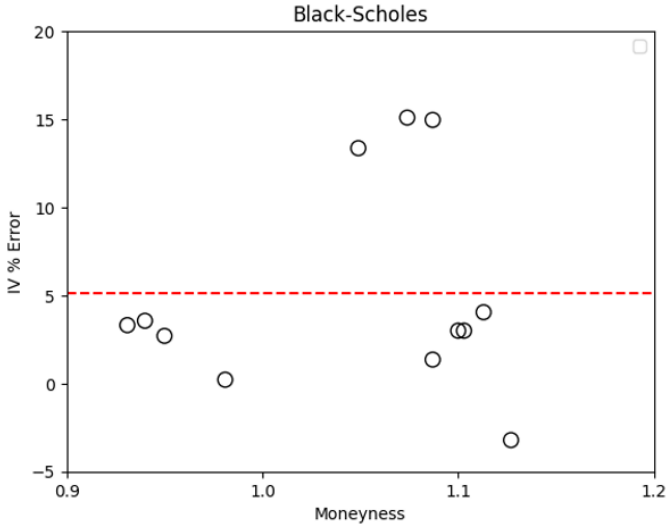
It is worth noting that these results represent a single day from a two-year data set and are included simply to illustrate the framework more clearly, to draw concrete conclusions more robust analysis in this way is required. On 01/17/2017 across our three equities, the CW model performed best followed by the BS and then the AHBS. Interestingly, for JPMorgan the AHBS was actually the best model, lacking the extreme anomalies present in both Amazon and Microsoft for this model. The above graphs highlight the importance of the initial model in this framework, with the best performing model in the parametric approach translating to the best performing model in the two-step approach for all three equities. This can be seen from the correction percentages, which are relatively similar for all three models for each of our equities. The results for Microsoft are more similar to Amazon than JPMorgan. From the LH panel we can see that CW performs best followed by BS and then AHBS, with the latter two performing similarly. As seen in the RH panel the degree of correction is very similar for all three models with 6.4%, 4.5% and 4.6% respectively. The BS correction for Microsoft is the most accurate of all our examples, with all predictions having an IV % error below 1% after correcting.

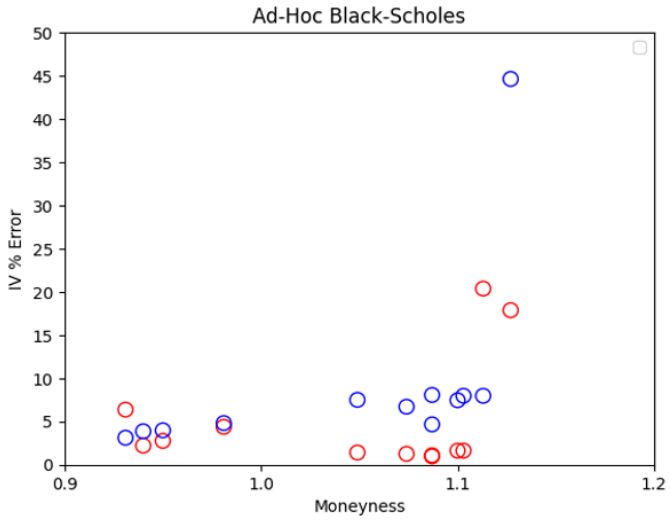
Similarly to for JPMorgan, the neural network correction performed poorly for the options with a moneyness less than 1.0, with an average improvement of -8%. Each model had some large anomalies with a % error greater than 25%

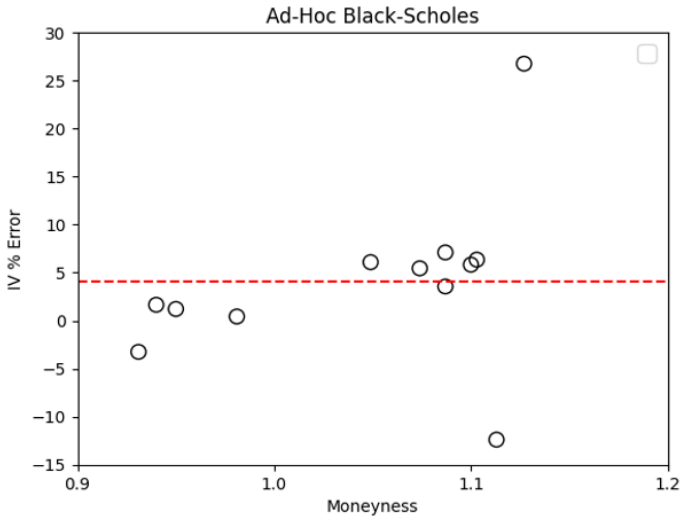
In summary, it is clear from the figures that on an individual day the equities have quite different shapes.

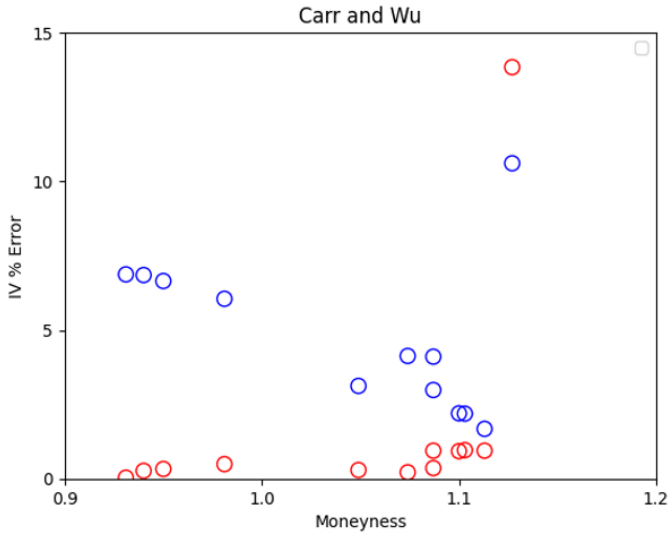
Figure 4: Prediction Errors of Models and NN Improvements: **Amazon 17/01/2017**











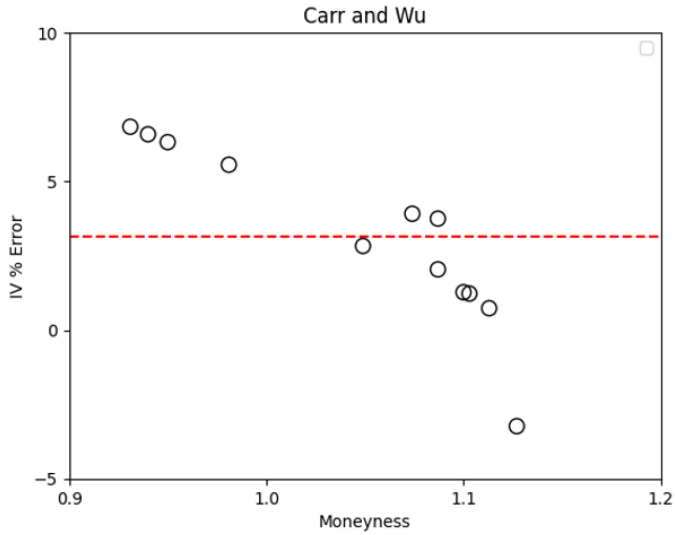


Figure 5: Prediction Errors of Models and NN Improvements: **JPMorgan 17/01/2017**

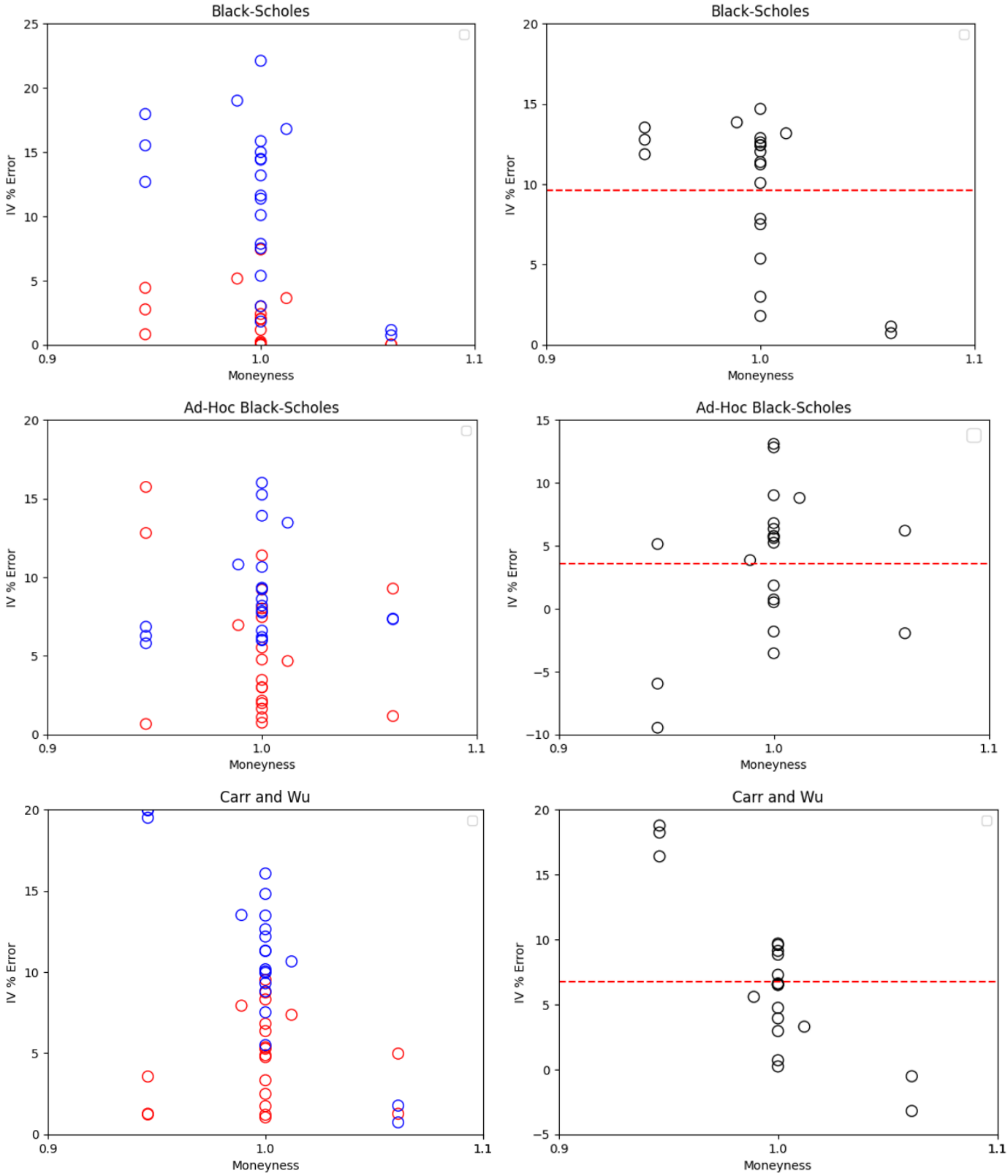
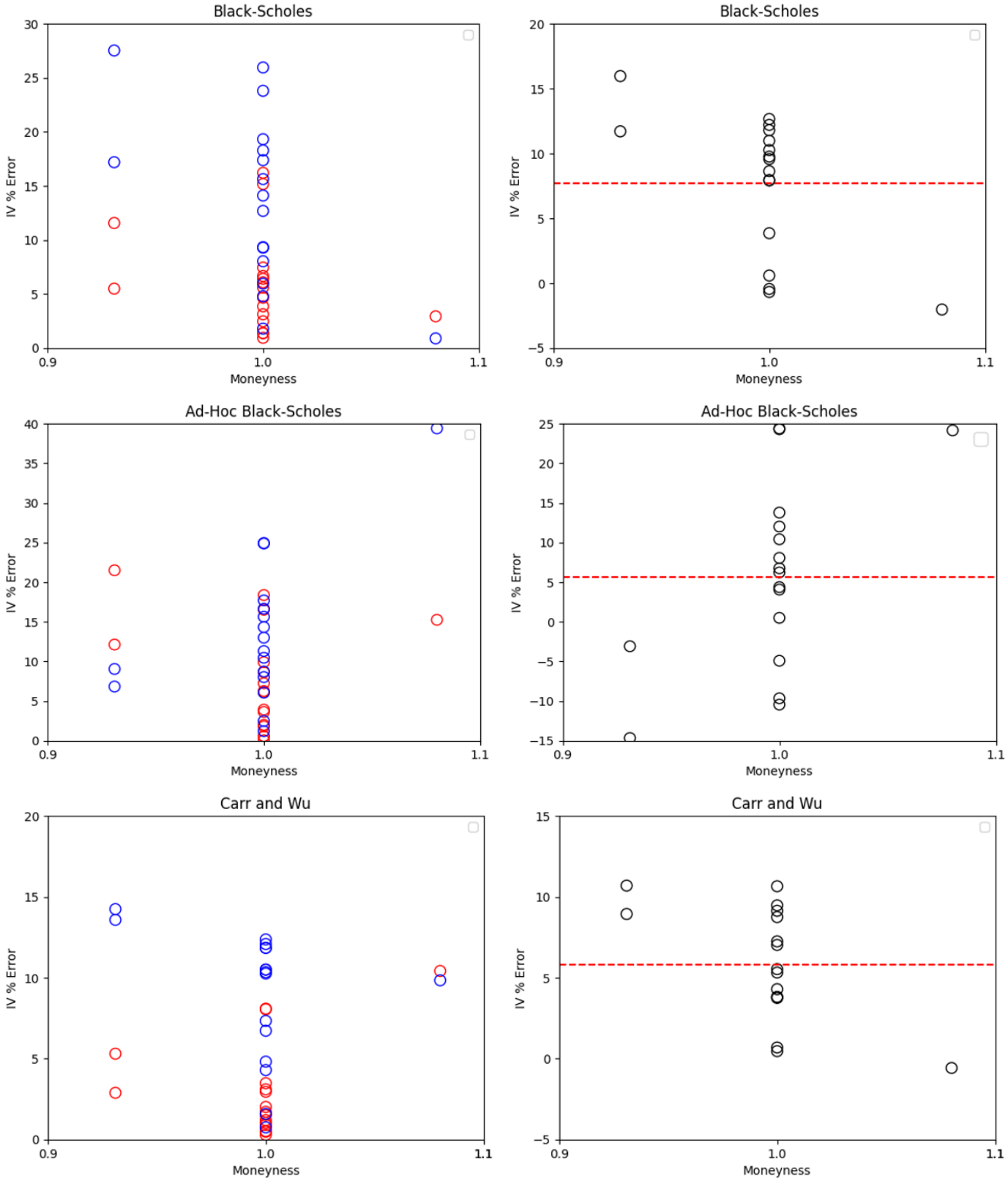


Figure 6: Prediction Errors of Models and NN Improvements: **Microsoft 17/01/2017**



5 Analysis with Additional Macroeconomic Factors

In this section we added four macroeconomic factors to our dataset. The VIX index (VIX) which is a measure of market volatility. The inflation index (INF) which is a measure of US inflation and both the 3-month treasury bill rate (3M) and 10-year treasury note rate (10Y) which are proxies for the interest rate representing the short and long end of the yield curve respectively. All four factors were obtained from the Centre for Research in Security Prices (CRSP) database. We assessed the effect of the inclusion of the additional macroeconomic factors on same-day and 1-day ahead predictions for the two best performing models from the previous sections, the Black-Scholes and Carr and Wu models.

For this section of our paper we chose to analyse the Black-Scholes and Carr and Wu models. The former because it is the benchmark model for option pricing while also performing the best in the previous section and the latter because, in the previous section, our NN found more success in correcting it than AHBS. We used the same-day and 1-day ahead time horizons to test the inclusion of macroeconomic factors. Our four factors don't vary intra-day therefore an option will give the data of the form:
$x_{i,t} = (m_{i,t}, \tau_{i,t}, VIX_t, INF_t, 3M_t, 10Y_t)$. This means on a given day all options share the same value for each of the four macro-factors. If we use our approach from the previous section, of fitting a neural-network to each day it would be equivalent to including four constants in our dataset. To capture the time-varying nature of these factors we decided to train our NN on one single large dataset. In Table 4 we have the IVRMSEs for each of our three equities and two chosen models, BS and CW, across the two time horizons. The NN3 column results are using the same methodology as the NN3+F columns, one non-parametric fitting on the entire dataset.

Table 4: Comparison of 3 Levels of Model Complexity

		Black-Scholes			Carr and Wu		
Horizon	Ticker	No NN	NN3	NN3 + F	No NN	NN3	NN3 + F
Same-day	Amazon	3.71%	1.87%	1.61%	3.90%	1.75%	**1.60%**
	JPMorgan	1.80%	0.90%	**0.80%**	2.52%	1.41%	1.16%
	Mircrosoft	2.85%	1.50%	**1.19%**	4.02%	2.41%	1.88%
1-day	Amazon	4.55%	2.01%	**1.67%**	4.80%	2.07%	1.79%
	JPMorgan	1.97%	1.05%	**0.89%**	2.75%	1.51%	1.22%
	Mircrosoft	3.38%	1.89%	1.62%	4.43%	1.94%	**1.43%**

Across both time-horizons in our analysis, a NN3+F corrected model was the best performing for all three companies. This suggests including the macro-economic factors can aid the ability of the FFNN to correct a parametric model. The BS+NN3+F slightly outperforms CW+NN3+F but to a lesser degree compared to when the macro-economic featurs are not included. These results have an important caveat, the nature of these factors means we cannot apply the correction to each day independently, as all four factors don't change on an

intra-daily basis, and would not impact the correction. Therefore, depite the clear improvement the factors provide, using them requires us to treat our data suboptimally and means we actually obtain a higher IVRMSE than if we had not used them. It can simultaneously be said that the factors are impactful at improving the correction but also that they require a less effective handling of the data. The inclusion of the macro-economic features changes the architecutre of our neural network. For NN3+F the architecture becomes as seen in Equation 20.

$$\underset{1\times1}{f(\boldsymbol{x}_{i,t})} = \underset{1\times32}{\boldsymbol{A_3}}\,\mathring{h}\Big[\underset{32\times64}{\boldsymbol{A_2}}\,\mathring{h}\Big(\underset{64\times128}{\boldsymbol{A_1}}\,\mathring{h}\Big[\underset{128\times5}{\boldsymbol{A_0}}\times\underset{5\times1}{\boldsymbol{z_0}}+\underset{128\times1}{\boldsymbol{b_0}}\Big]+\underset{64\times1}{\boldsymbol{b_1}}\Big)+\underset{32\times1}{\boldsymbol{b_2}}\Big]+\underset{1\times1}{\boldsymbol{b_3}} \tag{20}$$

with $\mathbf{z_0} = \mathbf{x_{i,t}} = [\epsilon_p(m_{i,t}, \tau_{i,t}), VIX_t, Inflation_t, 3M_t, 10Y_t] \in \mathbb{R}^5$

5.1 Feature Importance

To identify which covariates are most important when modelling the implied volatility surface and correcting our option pricing models, we use a simple notion of feature importance as used in Gu et al. (2020). We defined the importance of feature j as the increase in the IVRMSE arising from setting all values of feature j to zero, measured as a percentage. We then normalised all the features of our model such that the total feature importance adds to 100%. Feature importance should be high for covariates that help predict IVS.

Both models closely agree on feature importance, determining model error and VIX are the clear two most important features and that 3M and 10Y are rather unimportant. Model error as expected is the most important feature with 45.6% and 47.8% importance for Black-Scholes and Carr and Wu respectively. Combined with VIX they are responsible for close to 75% of variation in both models. The largest difference between the two models is that VIX is 4.5% more important in Carr and Wu than in Black-Scholes (27.7% vs 23.2%). The low feature importance of 3M and 10Y is consistent wit the fact they are highly correlated with one another.

Figure 8: Carr and Wu model+NN3+F Feature Importance

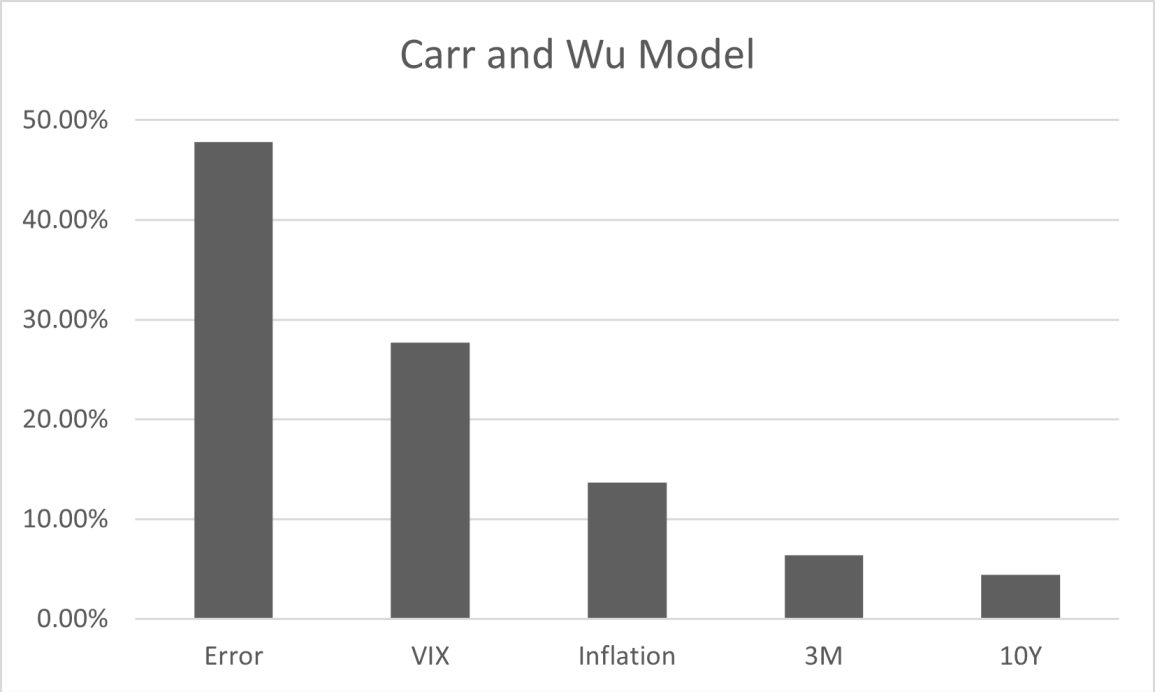
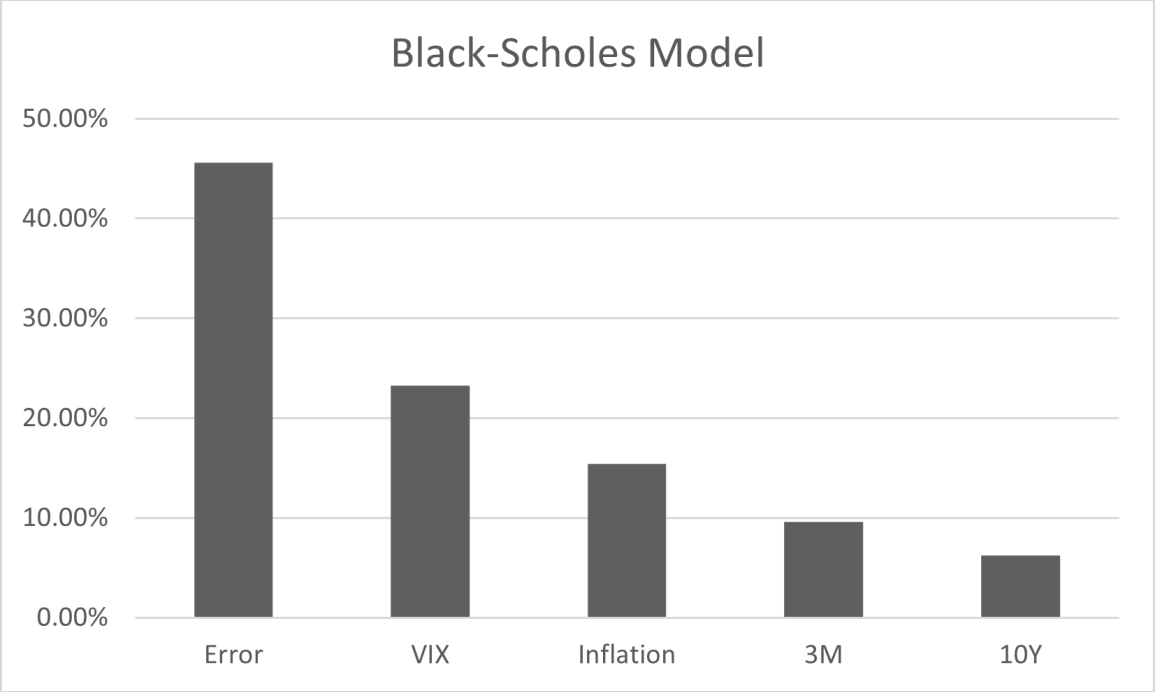


Figure 7: Black-Scholes model+NN3+F Feature Importance



6 Conclusion

Our study builds upon the research conducted by Almeida et al., wherein we assess the effectiveness of their two-step framework on individual US equities. Specifically, we analyzed three highly liquid stocks representing diverse sectors: Amazon in e-commerce, JPMorgan Chase & Co in finance, and Microsoft in technology. To begin, we constructed a parametric model based on the observed implied volatility surface. Subsequently, we applied this model to fit the pricing errors to a feedforward neural network. Our evaluation encompassed various models, including Black-Scholes, Ad-Hoc Black-Scholes, and Carr and Wu, selected based on their closed form solution for volatility, as we focused on American-style options. Our primary performance metric was out-of-sample implied volatility root mean square error (IVRMSE), enabling comparison across models, prediction horizons, and equities. Notably, we found that the corrected models consistently outperformed the parametric model, often significantly. Additionally, our analysis indicates the broader applicability of the framework to individual equities, especially for large financial institutions engaged in regular options trading for risk management purposes. For instance, institutions seeking to mitigate downside equity risk through put options can make more accurate assessments using this framework compared to traditional parametric modeling. However, our findings, while significant, fall short of industry-defining. Specifically, we observed less impressive results for individual equities compared to the S&P 500. Our results suggest that the neural network's ability to enhance accuracy may be contingent upon the number of eligible options, limiting the framework's effectiveness to larger companies. The inclusion of macroeconomic factors does improve performance but necessitates a sub-optimal data setup, as the time-varying nature of these factors is not fully captured in the daily neural network refitting process. We determined that the sacrifice in setup did not justify the marginal improvement in performance, leading us to recommend excluding macro factors and refitting the neural network daily. Our analysis revealed that relying solely on model-implied errors and daily refitting of the neural network yields the best results. Furthermore, we underscored the importance of the initial parametric model in determining the framework's effectiveness, with superior initial models yielding better results post-correction.

Upon comparison with the Almeida paper, we identified a shared emphasis on the significance of the parametric model for predicting corrected models' accuracy. While the Almeida paper highlighted the universal approximation feature of neural networks, our results indicate a weaker manifestation of this feature, likely attributable to variations in daily option quantities. In summary, our findings align with the Almeida paper but are comparatively less remarkable. Both studies observed an increase in IVRMSE with an extended time horizon.

This paper lays the groundwork for several potential avenues for future exploration. Initial indications suggest its applicability may be restricted to large companies with numerous

daily eligible options. It would be fascinating to explore how the framework performs with companies that have fewer eligible options. Moreover, examining the framework's viability in live trading scenarios would offer valuable insights into its practical viability.

References

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

Andersen, T. G., Fusari, N., & Todorov, V. (2015). *The pricing of short-term market risk: Evidence from weekly options* (tech. rep.). National Bureau of Economic Research.

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

Bates, D. S. (2000). Post-'87 crash fears in the s&p 500 futures option market. *Journal of econometrics*, *94*(1-2), 181–238.

Bernales, A., & Guidolin, M. (2014). Can we forecast the implied volatility surface dynamics of equity options? predictability and economic value tests. *Journal of Banking & Finance*, *46*, 326–342.

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

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.

Christensen, K., Siggaard, M., & Veliyev, B. (2021). *A machine learning approach to volatility forecasting* (tech. rep.). Department of Economics and Business Economics, Aarhus University.

Duffie, D., Pan, J., & Singleton, K. (2000). Transform analysis and asset pricing for affine jump-diffusions. *Econometrica*, *68*(6), 1343–1376.

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

Dupire, B., et al. (1994). Pricing with a smile. *Risk*, *7*(1), 18–20.

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.

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

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.

Heston, S. L., & Nandi, S. (2000). A closed-form garch option valuation model. *The review of financial studies*, *13*(3), 585–625.

MacBeth, J. D., & Merville, L. J. (1979). An empirical examination of the black-scholes call option pricing model. *The journal of finance*, *34*(5), 1173–1186.

Rubinstein, M. (1985). Nonparametric tests of alternative option pricing models using all reported trades and quotes on the 30 most active cboe option classes from august 23, 1976 through august 31, 1978. *The Journal of Finance*, *40*(2), 455–480.

The results in (**Optimal Architecture**) are for JPMorgan with same-day time horizon with our chosen architecuture in bold.

Table 5: Assessment of Optimal FFNN Architecture by IVRMSE

		Number of Hidden Layers				
No NN	Top Layer Size	1	2	3	4	5
1.71%	32	0.90%	0.74%	0.56%	0.61%	0.67%
1.71%	64	0.85%	0.70%	0.53%	0.51%	0.65%
1.71%	128	0.84%	0.67%	**0.49%**	0.54%	0.62%
1.71%	256	0.86%	0.69%	0.51%	0.56%	0.69%