A Derivatives Trading Recommendation System: the Mid-Curve Calendar Spread Case

Adriano S. Koshiyama[1*], Nikan Firoozye[1] and Philip Treleaven[1]

[adriano.koshiyama.15, n.firoozye, p.treleaven]@ucl.ac.uk

∗ Corresponding author - +44 (0)74 2607-9265

[1]Department of Computer Science, University College London
Gower Street, London WC1E 6BT
United Kingdom - Tel: +44 (0) 20 7679 2000

Abstract

Derivative traders are usually required to scan through hundreds, even thousands of possible trades on a daily basis. Up to now, not a single solution is available to aid in their job. Hence, this work aims to develop a trading recommendation system, and apply this system to the so-called Mid-Curve Calendar Spread (MCCS). To suggest that such approach is feasible, we used a list of 35 different types of MCCS; a total of 11 predictive models; and 4 benchmark models. Our results suggest that linear regression with lasso regularisation compared favourably to other approaches from a predictive and interpretability perspective.

Keywords: Trading Recommendation System; Machine Learning; Derivatives

1 Introduction

Derivative traders are usually required to scan through hundreds, even thousands of possible trades on a daily basis. A concrete case is the so-called Mid-Curve Calendar Spread (MCCS), a derivatives package that involves selling an option on a forward-starting swap and buying an option on a spot-starting swap with longer expiration (Corb, 2012; Natenberg, 2014). In such a package, traders look for the historical carry and the breakeven width levels, metrics that can be easily inferred from the terminal or aged payoff profile of the MCCS, shown in several heatmaps made by the research team. After that, they rank the most prominent ones to offer a client or to proceed in some proprietary trading. In general, the straightforwardness and swiftness that the decisions are made is the main upside of this framework.

However, one might notice that the main downsides of such approach are: (i) substantial information on the underlying like sensitivities, implied volatility, etc. are usually not taken into account; (ii) using the previous example, high historical values for carry and breakeven widths are more necessary rather than sufficient conditions for a profitable MCCS trade, being such argument extensible to other trades as well; (iii) a trader can quickly judge if an individual trade is worthwhile to invest, but may take some time to find it; and (iv) after a given period, traders tends to only look at a small subset of possible trades (small area on the heatmap), rather than the all available selection. Hence, a systematic approach where more information at hand is crossed and aggregated to find good trading picks can be highly useful and undoubtedly increase the trader's productivity.

Therefore, the objective of this work is to develop a trading recommendation system that can aid derivatives traders in their day-to-day routine. Being more specific, our solution is based on the following pipeline: (i) on a certain trade date, we compute metrics and sensitivities related to MCCS; (ii) these metrics are feed in a model that can predict its expected return for a given holding period; and after repeating (i) and (ii) for all trades we (iii) rank the trades using some dominance criteria. Our final solution is a model-based heatmap with the attractiveness scores for each MCCS trade, which can be offered to the traders and salespeople on a daily basis.

In order to suggest that such approach is methodologically and computationally feasible, we started using a list of 35 different types of MCCS regarding expiration (3-60 months), forward (3-60 months) and swap tenure (1-8 years). For each MCCS we used 10 years of historical data, ranging weekly from Sep/06 to Sep/16. Usually, on each Wednesday of a week we computed the (a) carry and breakeven for an MCCS, (b) the package sensitivities (gamma, vega, theta, etc.), (c) its present value

for an at the money forward strike, and (d) its present value after holding this package for a given period – allowing us to compute the holding h-periods-ahead return. We carried out an exploratory analysis, showing the overall performance of such trades, highlighting the major factors that droves their performance over time.

After this exploratory step, we started the modelling stage by: (i) fitting a predictive model using as the input those previous metrics, and the subsequent annualized returns as the output (in this case, holding 1-year-ahead the trade); (ii) making predictions and comparing different modelling methodologies by a set of performance metrics and benchmarks. Being more precise, we used a total of 11 predictive models, ranging from simple linear regression to support vector regression and multi-layer perceptron. We also implemented 4 benchmark models, being two intimately linked with breakeven width and carry level commonly used by the traders to decide which MCCS looks more attractive to a selling pitch. After establishing a backtesting engine and setting performance metrics, our results suggest that in general Linear Regression with Lasso Regularization (Lasso Regression) compared favourably to other approaches from a predictive and interpretability perspective.

In this sense, we organised this work as follows: next section presents a literature review on existing approaches to return/price prediction/estimation in different areas and instruments, as well as a brief description on MCCS trades. The third section displays the dataset that comports the MCCS trades, showing how the information is computed and gathered, which variables are the input and outputs, and the main assumptions that are embedded in it. Then, we move to modelling strategy, highlighting the main models that are going to be used as candidates for the recommendation system, how they are tested and have their performance assessed. Finally, we exhibit the results and discussions, starting with an exploratory analysis of the dataset, and then moving towards the performance analysis of each model through different metrics and perspectives. We close this work with some concluding remarks and future directions for research.

2 Background

2.1 Related Works

2.1.1 Cash Instruments Strategies

Literature provides a growing body of evidence that price changes can be predicted, that is, in particular circumstances and periods securities violate the Efficient Market Hypothesis (Campbell, Lo, & MacKinlay, 1997; Malkiel, 2003). This hypothesis

states that price changes must be unforecastable if they are properly anticipated, that is if they fully incorporate the expectations and information of all participants. Therefore, if returns can be predicted based on an information set (e.g., historical prices, economic indicators, news, etc.), one can use this information to trade and generate profits that are unexpected given the risk level that the market participant is assuming.

In this sense, researchers have employed different modelling approaches and information sets to predict price changes across a range of assets. When we scan the literature for cash instruments (equities, bonds, foreign exchange, etc.) focused only in using past returns as the main source for prediction, we can find works that tap into Bayesian forecasting (X. Zhou, Nakajima, & West, 2014), Nonparametric Predictive Inference (Baker, Coolen-Maturi, & Coolen, 2017), Forecasting Combination (Elliott, Gargano, & Timmermann, 2013), Generalized Exponential Weighted Moving Average (Nakano, Takahashi, & Takahashi, 2017), Support Vector Machines (SVM) (Karathanasopoulos et al., 2016), Shallow and Deep Neural Networks (NN) architectures (Chong, Han, & Park, 2017; Deng, Bao, Kong, Ren, & Dai, 2017; Gerlein, McGinnity, Belatreche, & Coleman, 2016; T. Zhou et al., 2016), Random Forest and Gradient Boosting Trees (Krauss, Do, & Huck, 2017), and so forth. The list of proposed methodologies keeps growing, in which equities or indices appears as the dominant asset class to apply these algorithms. Collectively, they provide evidence that some forecastability over returns can be achieved by putting in place complex models with a suitable training scheme.

Nonetheless, the main criticism of such approaches lies on the limited information set that they are tapping into: past returns. Even though past returns is a valuable source of data for forecasting, backed by many references in the previous paragraph, for some authors they are treated as an another source of information, which in some cases plays a minor role for prediction. Approaches that uses mainstream and novel news sources (Reuters, The Wall Street Journal, Twitter, etc.) have provided evidence that adding such features can improve return prediction and aid in the design of a trading strategy across several asset classes (Andersen, Bollerslev, Diebold, & Vega, 2007; H. Chen, De, Hu, & Hwang, 2014; Y.-L. Chen & Gau, 2010; El Ouadghiri, Mignon, & Boitout, 2016; Feuerriegel & Prendinger, 2016). Taking an even larger information set, (Weng, Ahmed, & Megahed, 2017) searches for disparate data sources (Google, Wikipedia, and so on) to increase the knowledge base that an algorithm can tap into for making predictions. Their inference engine used three modelling techniques (decision trees, NN and SVM) and compared favourably to other reported results in the literature.

2.1.2 Derivatives Instruments Strategies

Contrasting with the emphasis that researchers in cash instruments put on return predictability, when we devote our attention to research in derivatives instruments (options, swaps, swaptions, etc.) it is clear that most of the effort is concentrated on pricing these contracts. Similarly, researchers have adopted different approaches and information sets to describe the price movements of these contracts properly. The traditional framework is via stochastic calculus, in which pricing is made under some market assumptions (frictionless market, no arbitrage, risk-neutral investor, etc.) as well as an assumed asset price dynamics over time (e.g., Geometric Brownian Motion) (Glasserman, 2013; Shreve, 2004). By discovering the fair price of a contract, trading strategies can be established to tap into any potential mispricing (Ehrman, 2006; Natenberg, 2014).

In parallel to the traditional framework, alternative ways of pricing and trading started to emanate relying on fewer assumptions and more data-driven. We can pinpoint approaches that use NN for option pricing and hedging with daily S&P 500 index daily call options (Gencay & Qi, 2001) as well as for real-time pricing and hedging options on currency futures of EUR/USD at tick level (von Spreckelsen, von Mettenheim, & Breitner, 2014). It is worth to mention other approaches in the derivatives realm, such that the prediction of pricing and hedging errors for equity-linked warrants with Gaussian Process models (Han & Lee, 2008), building machine learning models for predicting option prices over KOSPI 200 Index options (Park, Kim, & Lee, 2014) and a general study on forecasting non-negative option price distributions using Bayesian kernel methods (Park & Lee, 2012).

When we devote our attention to the asset type that this work is dedicated, interest rate swaptions, a similar pattern persists: most of the research is related to pricing and not to return prediction. Regarding pricing, the same tradition of relying on stochastic calculus techniques is followed (Brigo & Mercurio, 2007; Rebonato, McKay, & White, 2011). Regarding potential alternatives using more data-driven approaches as we saw with currency, indices and equities options, we can only mention the work of Souza et al. (Sousa, Esquível, & Gaspar, 2012) which calibrates the Vasicek interest rate model under the risk neutral measure by learning the model parameters using Gaussian processes for regression. Considering trading strategies and return prediction, we can find even less academic research, being perhaps most of the research residing inside the counterparts that exchange such products (banks, hedge funds, etc.). This shortage of published research might be linked with the absence of ready to use and publicly available datasets, similar to the ones found in cash products since these instruments are traded off-exchange.

Nonetheless, we can still find research that has some connection with our work, such by (Duyvesteyn & de Zwart, 2015) where the authors build trading strategies by long-short combinations of two at-the-money straddles for the four top swaption markets (USD, JPY, EUR and GBP), in search to harvest the volatility risk premium in such markets. Their research taps into a large dataset of at-the-money implied volatility quotes on the 10-year swap rate and 1 to 12-month swaption maturities between April 1996 and December 2011 to calculate the returns of the long-short straddle strategies. They find statistically significant returns for all markets and both delta–gamma and delta–vega neutral strategies. The main similarities of their approach with our work reside mainly in the asset type that they are trading as well as with the idea of using a set of indicators (mostly based on the shape of the volatility surface) to devise different trading strategies. However, regarding modelling strategy, there are clear differences between their approach (more theory-driven) and the methodology envisioned by this work (tilted to data-driven). See (Choi, Mueller, & Vedolin, 2017) for a recent review on similar approaches in the harvesting volatility risk premium.

2.1.3 Summary

Based on this review of existing approaches to return/price prediction/estimation in different areas and instruments, to the best of our knowledge, our work is the first attempt to apply computational statistics and machine learning techniques to build trading strategies in the context of interest rate swaptions. Our approach is not the only novel from a modelling perspective, but instead of trading the vanilla product (receiver/payer interest rate swaption), we prefer to focus on options strategies (calendar spreads, straddles, etc.) which in many cases is the package that is in practice traded. By thinking in terms of the package, in this case, a Mid-Curve Calendar Spread, rather than the individual constituents we unlock some features that can only be computed in this situation, like the carry at expiry, breakeven width and so on.

Therefore, we can train our models not only using past returns but also using sensitivities as well as information derived from the package payoff function. By portraying in this manner our investment strategy, we have a large information set that can substantially add information to aid forecasting returns. But as a counter-effect, this poses a new challenge on separating relevant features in a dynamic context. In this respect, the combination of temporal cross-validation, a diverse set of models and regularisation/feature selection can provide a robust framework for trading strategies backtesting and assessment. But before presenting such framework, next

section gives a brief view on Mid-Curve Calendar Spreads trades.

2.2 Mid-Curve Calendar Spreads

Mid-Curve Calendar Spread (MCCS) is a package involving short selling an option on a forward-starting swap and going long a longer-expiry swaption on the same underlying swap (Corb, 2012; Firoozye & Zheng, 2016). There is a counterpart with many similarities for equities – check in (Natenberg, 2014) for more information. Investors typically use MCCS to take a view on forwarding volatility. This comes from the fact that, conceptually, spot volatility can be decomposed into forward volatility and mid-curve volatility. Taking 10y10y[1] for example, Figures 1, 2 and 3 below illustrates examples of the reference time periods covered by different interest rate volatilities and their instruments. The red lines indicate the time over which interest rate volatility exposure is taken, and the grey line indicates the underlying forward swap rate.

- (Spot) Swaption: 10y10y Swaption – a plain vanilla swaption has its strike set at inception and the underlying swap starts on a sport basis from the option expiry date.

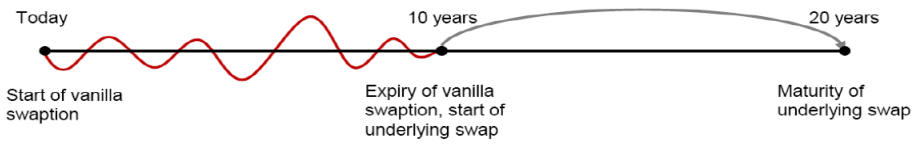

Figure 1: Spot Swaption example. Source: (Firoozye & Zhang, 2014a).

- Mid-curve Swaption: 5y mid-curve on 5y10y swap rate – the volatility of a forward-starting swaption, called mid-curve, whose strike is set at inception and but the underlying swap starts several years following the option expiry date.

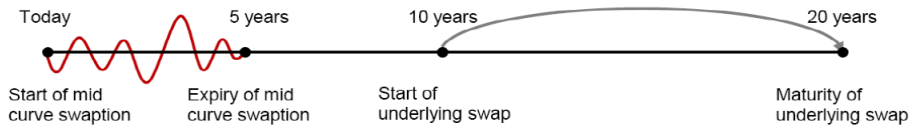

Figure 2: Mid-curve Swaption example. Source: (Firoozye & Zhang, 2014a).

[1]This notation is extensively used during this work. In this case, the first 10y means a spot swaption with 10 year of expiration, while the second 10y refers to the swap tenure.

- Forward-starting Swaption: 5y forward 5y10y option – 5y forward 5y10y volatility is the implied volatility of a forward strike swaption whose strike is only set after several years at the then At the Money Forward (ATMF) level, and the underlying swap starts on a spot basis from the option expiry date.

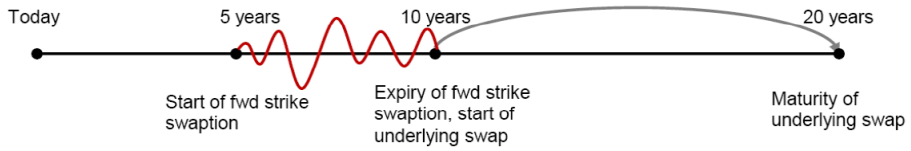

Figure 3: Forward-starting Swaption example. Source: (Firoozye & Zhang, 2014a).

In mathematical terms, the below relation holds for those volatility exposures of different time periods for the underlying 10y10y rate (Taleb, 1997):

$$10y \times \sigma^2_{10y10y} = 5y \times \sigma^2_{5y5y10ymid-curve} + 5y \times \sigma^2_{5yfwd5y10y} \tag{1}$$

With the above (assuming flat skews where volatility is the standard deviation), 5y fwd 5y10y volatility can be backed out if we know 10y10y swaption volatility and 5y5y10y mid-curve volatility. Therefore, a forward strike swaption, in nature a pure exposure to forward volatility, can be approximated by short selling a mid-curve and buying a plain vanilla swaption, which is an MCCS. Due to put-call parity, at the expiry date, an MCCS becomes either an out of the money payer or an out of the money receiver no matter it is payer or receiver at the inception.

In Figure 4 is presented the payoff profile for an EUR 1m1y2y[2].

We plot the payoff profiles for current volatility and up and down volatility scenarios, noting that the long vega position means that the payoff profile shifts up in a rising volatility environment and correspondingly shifts down in a falling vol environment. We calculate the (volatility adjusted) breakevens as being $0.41\% - 0.46\%$, giving little protection against selloffs. We note that forwards in a ± 1 volatility band leave them at $0.40\% - 0.48\%$, a range just marginally larger than our breakeven range (i.e., the trade should pay off just slightly less than 66% of the time).

In Figure 5, we show the performance of the EUR 1y2y forwards overlaid with breakevens during 2012-14. Our breakeven band has largely covered most of the extremes since the June ECB meeting, in which ECB President Draghi said that he does not plan on synchronising the ECB meetings with the Fed or any other central

[2]Short selling a 1m1y2y mid-curve swaption and going long a longer-expiry 13m2y spot swaption.

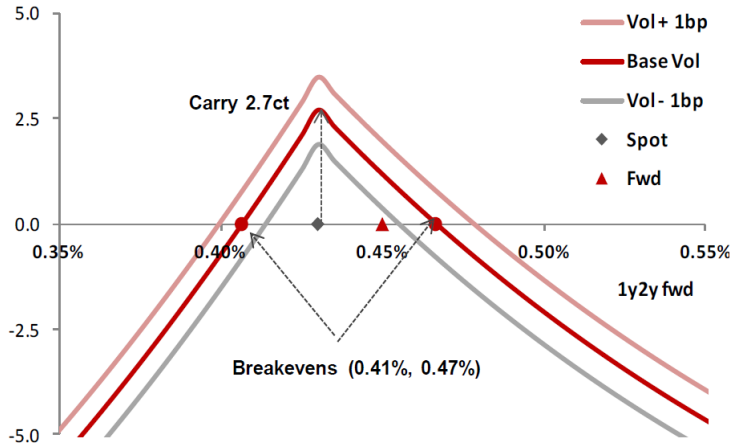

Figure 4: Payoff profile for an EUR 1m1y2y. Source: (Firoozye & Zhang, 2014b).

bank. This suggests a disconnection of EUR rates from their US counterparts, and the EUR curve is likely to be less affected by developments in US rates in coming months. However, in that occasion, our breakeven width should provide a decent buffer against either a sell-off or rally.

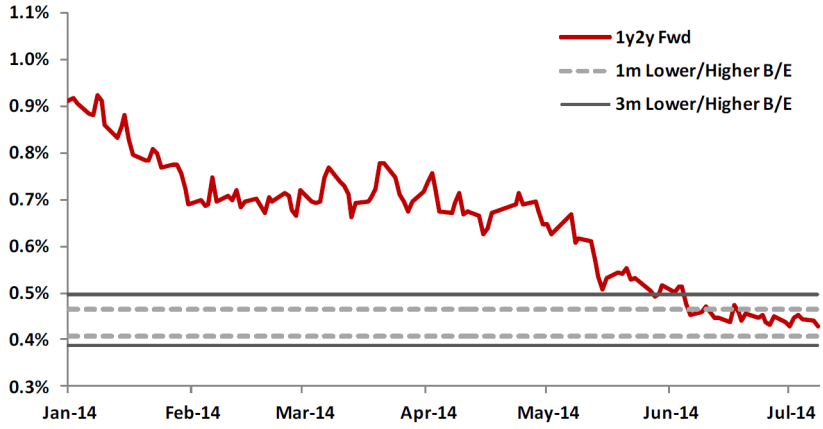

Figure 5: EUR 1y2y forward rate (%)a. Source: (Firoozye & Zhang, 2014b).

MCCS can result in what we think of as turbocharged carry, primarily because of the risks that they have (which fortunately can be balanced with the returns in a way which results in relatively attractive trades). The results are in vast contrast to more standard carry trades. For instance, cash or swaps typically have less risk and correspondingly less carry. Going long outright typically benefits from far less positive carry because although it performs poorly in a selloff, it also performs well in a rally. Based on these characteristics of MCCS, next section presents how we elaborated the methodology to build this trading recommendation system.

3 Methodology

In summary, our solution develops the following roadmap (also schematically describe in Figure 6):

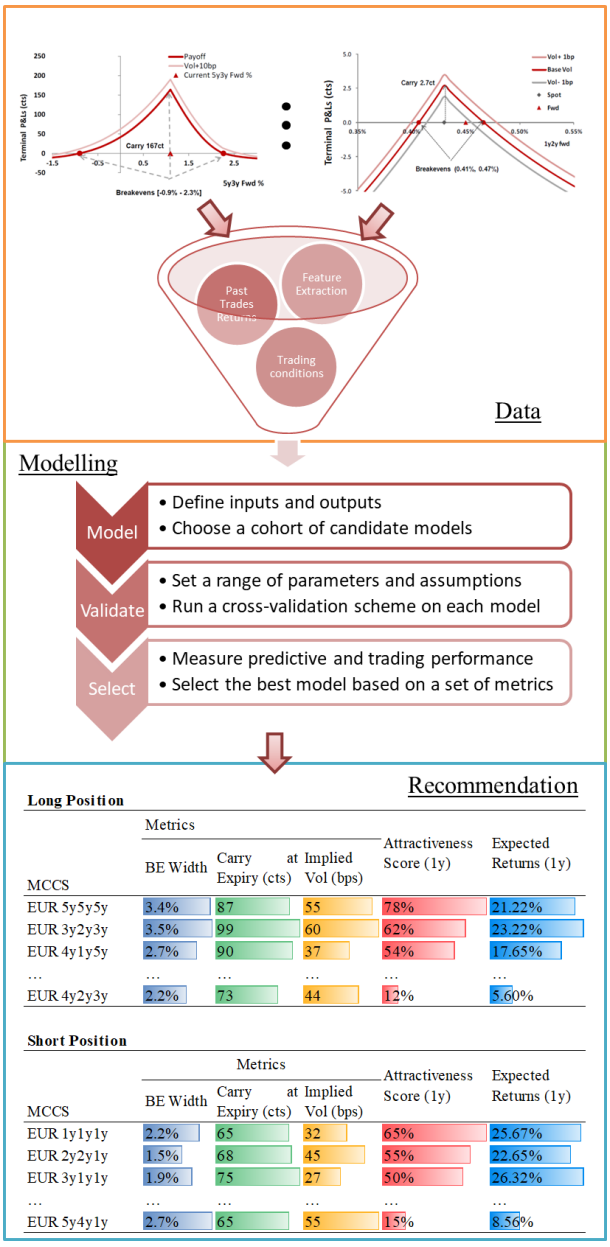

Figure 6: Flowchart describing the input-output schemes from the proposed trading recommendation system for MCCS trades.

1. **Data**: On a certain trade date, we **calculate metrics and sensitivities** related to an MCCS package;

2. **Modelling**: These metrics are **feed in a predictive model** that outputs its expected return for a given holding period (e.g., one year);

3. **Recommendation**: After repeating (i) and (ii) for all MCCS we (iii) rank them based on the expected returns using some criteria.

Following this outlined structure, the next three subsections describes in more details when and which MCCS trades were recorded (Sampling Scheme), which predictive models were trained and how they were assessed (Modelling Strategy) and how the long/short trading signal is computed for each MCCS (Recommendation System). Finally, last subsection presents which metrics were used to evaluate the recommendation system performance when a certain predictive model candidate is underpinning it.

3.1 Sampling Scheme

During our experiments, we opted to use the trades displayed in Table 1.

Although many other configurations are available in practice, these are the ones with longest historical data available, which is important when it is necessary to fit a predictive model. As it can be seen, all trades are in Euro, ranging from different expiries (1y-5y), forwards (1y-5y) and swap tenures (1y-5y and 8y).

For each configuration, at time t we agree with a counterpart to trade this package using the At the Money Forward (ATMF) rate as the strike, paying or receiving the present value PV_t. The PV_t is computed via SABR model (Rebonato et al., 2011), using information and parameters (e.g., spot, forward rates and rate-rate correlation) calibrated using market data on a daily basis. From the same model that computed the PV_t, we can also obtain other metrics and sensitivities as those displayed in Table 2.

Carry and BE Width are those obtained looking at the payoff profile at expiry. The Aged 1y Carry is produced by ageing the trade by one year (moving closer to the expiration) and estimate the payoff profile computing the carry. Theta, Vega and Gamma are the sensitivities of the instruments by a change in time, volatility and a wider range of underlying rate movements, respectively. These and the ATMF Implied Vol are backed by the SABR model too. Curve, Time and Volatility Carry are the amount of Aged 1y Carry that can be attributed to the changes in certain sensitivities from spot to forward, such as the Delta (Curve), Theta (Time) and Vega (Volatility). These can also be used as tools to understand which factors most influence the instrument value over time.

After computing all these metrics at time t, we hold the trade until $t + h$ where h can be two weeks, one month, one year, and so on, as long as $t + h$ is before or at expiration. In time $t + h$ we compute the PV_{t+h} of the same trade again, using

the new economic scenario available (e.g. rates, change in model parameters). By agreeing on buying back or selling the current trade for PV_{t+h} we can compute the Holding k-period Return of the trade started at time t by:

$$R_t^{(h)} = \frac{PV_{t+h} - PV_t}{PV_t} \tag{2}$$

In summary, Table 3 presents an example of information in a wide format that is available when we combine the data from time t and $t + h$.

Note that in the most contemporaneous period (close to T) we do not have the PV_{T+h} and so, we cannot compute $R_T^{(h)}$. Conversely, if we want to use lagged returns $R_{t-h-1}^{(k)}$ as explanatory forces for $R_t^{(h)}$, then in the beginning this information is also not available. Therefore, our dataset is trimmed at the beginning and the end mainly by the value of h. If h is small, such as two weeks or one month, the trimming is imperceptible and, therefore, may not affect the model fitting and validation. However, if h is large such as two or three years, this might reduce the samples available substantially, decreasing the range of models and cross-validation schemes that might be employed for this task. We depicted most of these details in Figure 7, highlighting the fact that we need to wait for almost two holding horizons to get at least one lagged return.

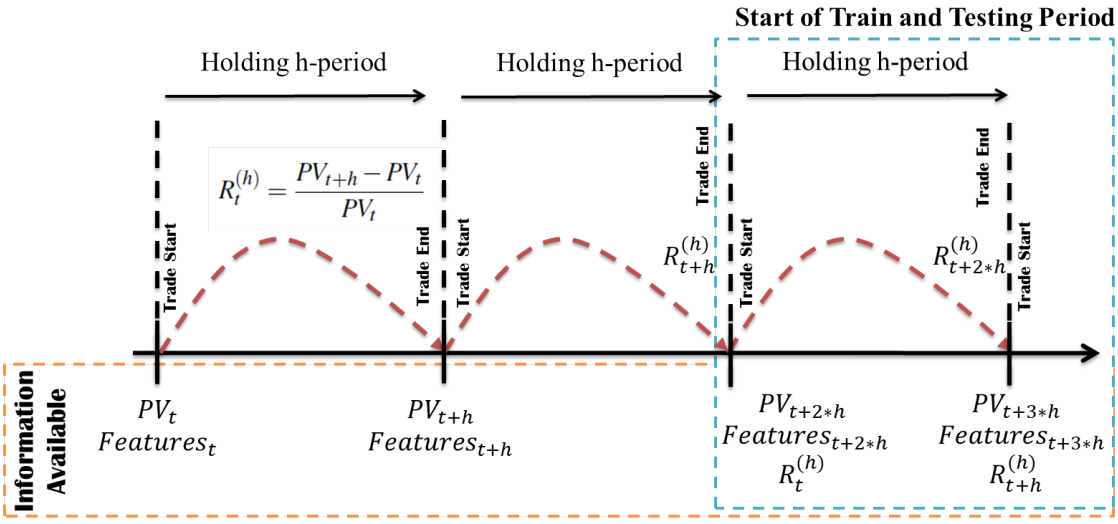

Figure 7: Visual description of the information presented in Table 3.

Based on these procedures, metrics and observations, Table 4 express other details that we used during our experiments to generate the dataset.

Therefore, we gathered data from trades entered on a weekly basis from September 2006 to September 2016. These trades are struck ATMF, using the PV_t computed from the Middle Rate (in practice, some bid-ask spread would be imbued pro-

portional to the Vega). After holding for one year ($h = 1y$) the trade, we compute the arithmetical returns that are, therefore by definition, automatically annualised. These returns are gross, and so we need to take into account the transaction costs (hedging costs and fixed fees charged by the derivatives desk) as well as some future funding rate. These values are also outlined in Table 4, where the transaction costs of 0.75 as a fraction of Vega were chosen not only to taken into account the transaction cost, but also some potential bid-ask spread on the start/unwind of the trade. The 3-month London Interbank Overnight Rate (LIBOR) was chosen as the funding cost/benchmark rate to compute excess returns.

Using these assumptions, next subsection presents the modelling strategy that taps into this dataset to create the recommendation system for the MCCS trades.

3.2 Modelling Strategy

In relation to modelling, our general model is a system of uncoupled equations:

$$R_{t,1}^{(1y)} = f_1(features_{t,1}) + \varepsilon_{t,1} = \hat{R}_{t,1}^{(1y)} + \varepsilon_{t,1} \tag{3}$$

$$R_{t,2}^{(1y)} = f_2(features_{t,2}) + \varepsilon_{t,2} = \hat{R}_{t,2}^{(1y)} + \varepsilon_{t,2} \tag{4}$$

$$\dots$$

$$R_{t,n}^{(1y)} = f_n(features_{t,n}) + \varepsilon_{t,n} = \hat{R}_{t,n}^{(1y)} + \varepsilon_{t,n} \tag{5}$$

where for each MCCS trade ($i = 1, ..., n$) there is an i-th predictive model f_i that is feed with a set of pre-calculated features (BE Width, Carry, etc.) and returns an estimate of the holding 1y-period return $\hat{R}_{t,i}^{(1y)}$. As the model is an approximation, some noise/error is expected, and in the modelling aspect, this is expressed as the $\varepsilon_{t,i}$ component. After defining which variable is intended to be predicted, the remaining points are: which models are available to embody f_i and how the fitting, validation and selection of these models are going to be made.

About the first point, in the first rows of Table 5 we display the models that we used during our experiments. We should mention that these models are standard techniques commonly found in the computational statistics and machine learning literature, with their mathematical descriptions and usage can be consulted in the following references (Bishop, 2007; Duda, Hart, & Stork, 2000; Friedman, Hastie, & Tibshirani, 2001; Haykin, 2009).

The Abbreviation column of Table 5 presents shortened version for the usual name of models in the Model column. We mainly use these abbreviations for plots and tables during the results and discuss section. Table 5 Model column presents a plethora of models that this work has fitted for this prediction purpose: we started

from simple predictive models such as Classical Linear Regression, k-Nearest Neighbours and Classification and Regression Tree, towards those that can seamlessly exhibit nonlinear behaviours, like Random Forest, Kernel Ridge Regression, Multi-Layer Perceptron and Support Vector Regression. Some of these methods had their hyperparameters held constant across all experiments (Fixed Hyperparameters column), or because we wanted to apply a particular form of a method (RBF kernel, single hidden layer, etc.) or because during a warm-up phase we noticed that they did not affect substantially the results (hyperbolic tangent, increasing number of trees, etc.).

For certain models, the Cross-Validated Parameters column shows which hyperparameters were optimised before the prediction step. For instance, suppose the case of Ridge Regression and the need to define the regularisation value (λ) appropriately. Consider that we have a set of training pairs $(features_t, R_t^{(1y)})_{t=1}^{L}$[3] of size L, and for this sample we subset it in k-rolling-cross-validation (k-rolling-cv) folders (better explained later in this subsection). Then, we train and test using this scheme the Ridge Regression model with one of the predefined λ, say $\lambda = 10^0$. We compute some performance function on the test set (Mean Squared Error – MSE) and repeat this process for all λ values available. We use in the final model the λ that on average had the lowest MSE.

The process explained previously is repeated for all cross-validated parameters in each model. When a model needs to cross-validate a pair or more of hyperparameters, the procedure is to perform all the feasible combinations of them (e.g., KRR-RBF with $(\lambda, \gamma) = \{(1, 0.01), (0.1, 0.01), ..., (0.001, 100)\}$). Another relevant point is that this process is in fact made inside a bigger loop called nested resampling (Bischl, Mersmann, Trautmann, & Weihs, 2012), and that is the reason we have k and L inner and outer. Figure 8 present an example of nested resampling validation for the k-rolling-cv ($k_{outer} = k_{inner} = 1$, $L_{outer} = 3$ and $L_{inner} = 1$).

As it can be seen, before applying the model to the test set it needs first to be calibrated. This calibration is performed doing cross-validation within the training set, splitting in the same fashion the available dataset. After the model is well tuned, it is fully applied in the test set, and furthermore, the test set is incorporated into the training set, and the process is repeated. Converting this paradigm to our problem, consider the case of Ridge Regression again. Therefore, the validation process of this model is made by:

1. Divide all available $(features_t, R_t^{(1y)})_{t=1}^{T}$ data using the k-rolling-cv, with $k_{outer} = 1$ and $L_{outer} = 2$ years (Table 5);

[3]For the sake of brevity we dropped the subscript that refers to a particular trade (i).

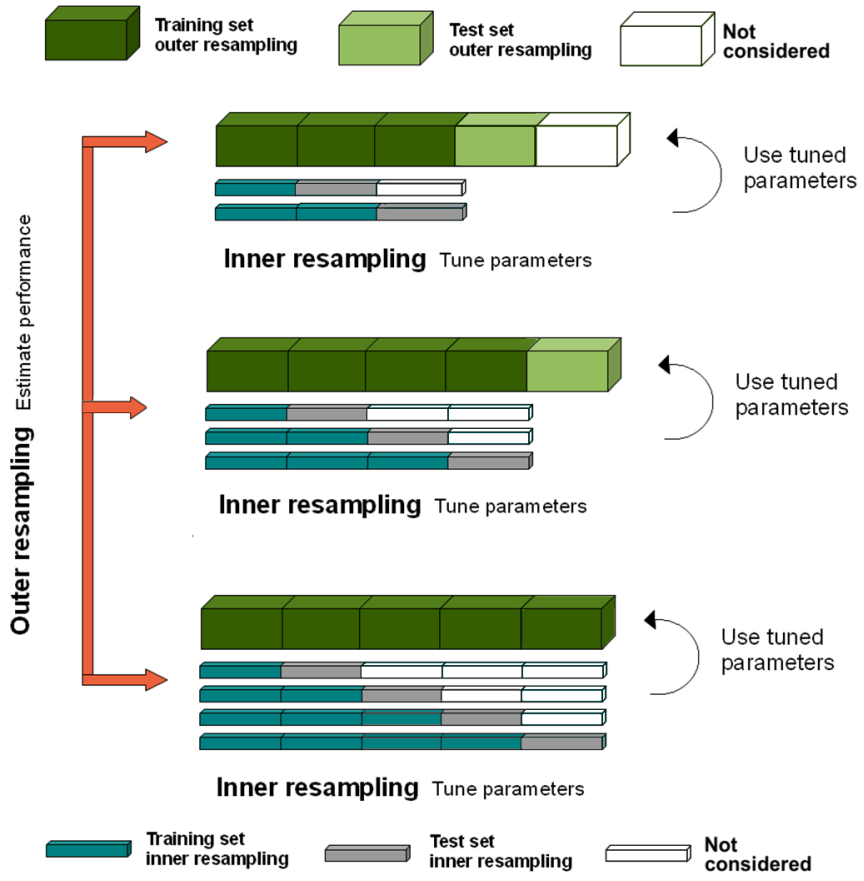

Figure 8: Nested resampling for k-rolling-cv case.

2. For each training subsequence $(features_t, R_t^{(1y)})_{t=1}^{L_{outer}}$, $(features_t, R_t^{(1y)})_{t=1}^{L_{outer}+1}$, ..., $(features_t, R_t^{(1y)})_{t=1}^{T}$, train a model and if it is necessary to calibrate hyperparameters, do:

   (a) Split the data using the k-rolling-cv, with $k_{inner} \approx (T_{train} - L_{inner})/5$ and $L_{inner} = 1$ year, where T_{train} is the number of elements in the training set;

   (b) Run the method for all folders using the pre-defined hyperparameters;

   (c) Return the hyperparameters that yielded the best results.

3. Apply each model to its assigned test subsequence $(features_t, R_t^{(1y)})_{t=k_{outer}}$, $(features_t, R_t^{(1y)})_{t=k_{outer}+1}$, ..., $(features_t, R_t^{(1y)})_{t=T}$ and compute the performance metrics in the test set.

In repeating this process for all models, we want to assure that all are adequately calibrated as well as hedged against any overfitting. We used performance metrics

not only related to predictive power, but also to profits and losses that a certain model generated overtime. These are going to be introduced in the last subsection.

We fitted usual benchmarks found in the literature for regression and forecasting modelling: the Average and Naive models (Friedman et al., 2001; Hyndman, Koehler, Ord, & Snyder, 2008). We also implemented the benchmarks that traders use to assess whether a particular MCCS is worth to be pitched or traded: BE Width and Carry at Expiry. We replicated the way traders look to these features, by computing z-scores[4] based on average and standard deviation on rolling window of size equal to 1 year. The signal for going long/short (S_t) is given by a thumb rule with a simple rationale: if a certain metric has a z-score above or equal to ± 3, the trader goes fully long (+)/short(-) in the trade, since it is a very extreme event. Otherwise, it reduces the leverage on it, until it below one standard deviation of distance from the rolling average.

We removed any missing data, and clipped extremes values, mainly in returns above the 95% percentiles (in our case it can be due to some numerical problems, or some extreme scenarios related to 2008-2009 financial crisis period). Next subsection presents the final component of our roadmap: recommendation system.

3.3 Recommendation System

The recommendation of a certain trade can be made solely on some normalised version of the expected return for holding 1y-period the i-th trade ($\hat{R}_{t,i}^{(1y)}$). Given that each model will be providing individual forecasts for each MCCS and after that their performance will be assessed locally and globally, a more suitable manner to proceed would be to assign a credit based on the tracking record of a model to predict a particular MCCS trade. Hence, we will be weighted up or down a signal not only based on the magnitude of a model prediction but also by its quality. Then, consider as $\hat{R}_{t,i}^{(1y)}$ the expected return for holding 1y-period the i-th trade. Now, define the new signal function $S_{t,i}$ by:

$$S_{t,i} = \frac{\hat{R}_{t,i}^{(1y)} \times \max(Rho_{\hat{R}_{t,i}^{(1y)}, R_{t,i}^{(1y)}}, 0)}{\max(|\hat{R}_{t,i}^{(1y)} \times \max(Rho_{\hat{R}_{t,i}^{(1y)}, R_{t,i}^{(1y)}}, 0)|, ..., |\hat{R}_{t-h,i}^{(1y)} \times \max(Rho_{\hat{R}_{t-h,i}^{(1y)}, R_{t-h,i}^{(1y)}}|, 0)} \in [-1, 1] \quad (6)$$

where the strength of the i-th long/short signal is given by its expected return, scaled by the maximum weighted return that a long/short position on the same

[4] a z-score is defined by: $Z - score = \frac{X-\mu}{\sigma}$ where X represent the actual value of a certain variable, μ and σ the average and standard deviation of X in a period.

trade (that is why the returns are in absolute terms) was expected to yield in the previous h-period (in this case 1 year). Therefore, the trade with the maximum weighted return in absolute terms will have $|S_{t,i}| = 1$ as well as those close to zero will yield $S_{t,i} \approx 0$. The weight/credit of a certain prediction is based on the historical adherence between the actual and predicted values – Pearson correlation coefficient, outlined in the next subsection.

The caveat of this approach in practice is that a trade can have a small annualised return (say 3%), but if it is historically the largest in absolute terms as well as given by an accurate model, then it will yield $S_{t,i} = 1$. To avoid that in practice, and also by to link this methodology to Figure 6, the column attractiveness score is represented by $S_{t,i}$ whereas expected return by $\hat{R}_{t,i}^{(1y)}$. By providing both information, the trader can check how much he should go long/short as well as knows how much is expected for that particular trade. If the $S_{t,i}$ is high, but $\hat{R}_{t,i}^{(1y)}$ is below his target, the trader can opt to not pitch/enter in this trade.

3.4 Evaluation Metrics

Below we outline two types of metrics: one that focuses on the predictive performance that the model provided, and other four that are based on the profit/loss that its application harvested during the backtest. Set by $R_t^{(S)} = R_t^{(1y)} \times S_t(\hat{R}_t^{(1y)})$ the strategy return (combination of the realized/observed excess returns and the signal – function of a model prediction), we can compute the following metrics:

- Pearson Correlation Coefficient (Rho): it is a dimensionless measure of the linear dependence between the actual and predicted values:

$$Rho = \frac{Cov[R_t^{(1y)}, \hat{R}_t^{(1y)}]}{\sqrt{Var[R_t^{(1y)}]Var[\hat{R}_t^{(1y)}]}} \tag{7}$$

  where Cov and Var are the covariance and variance operators. It ranges from $[-1, +1]$, with -1 representing a perfect inverse linear association, and $+1$ the opposite. In our case, we benefit more when Rho is close to $+1$. In the context of linear models, a higher predictive power is a necessary condition for profitable trades (see (Acar & Satchell, 2002)), hence by minimising the predictive error we are somewhat trailing a path for profits maximisation, albeit such causation is not very clear since this is not a sufficient condition.

- Average Return (Avg Return): is the arithmetic average of the strategy re-

turns:

$$\bar{R}^{(S)} = \frac{\sum_{t=1}^{T} R_t^{(S)}}{T} \tag{8}$$

- Standard Deviation: is the estimator of the dispersion around the strategy average returns (a risk measure in certain sense):

$$\sigma_{R^{(S)}} = \sqrt{\frac{\sum_{t=1}^{T} (R_t^{(S)} - \bar{R}^{(S)})^2}{T}} \tag{9}$$

- Information Ratio: is the average annualized return of a strategy earned in excess of a particular benchmark per unit of risk (measured in terms of standard deviation):

$$IR = \frac{\bar{R}^{(S)} - \bar{B}}{\sigma_{R^{(S)}}} \tag{10}$$

  where $\bar{B}$ is the average return of the benchmark (e.g., treasury bond, equity index). In our case, it was already set to the 3-month LIBOR rate (Table 4). It should be mentioned that Information Ratio makes each strategy performance comparable: since we are adjusting average returns by the risk assumed for each strategy, it removes the leverage component that is magnifying/shrinking the returns provided by a certain strategy. In analogy, it can be viewed as standardising a random variable by stripping it from its location and scaling factor.

- VIX Quintiles: consider an asset V that has a sequence of returns $R_1^{(V)}, ..., R_T^{(V)}$. Let $I_{\tau_1 < R_t^{(V)} < \tau_2}$ represents an indicator function that outputs 1 if $\tau_1 < R_t^{(V)} < \tau_2$, otherwise 0 – with τ_1 and τ_2 denoting thresholds set by the user. In our case τ_1 and τ_2 represents quantiles from the Volatility Index (VIX) returns (Whaley, 2009). This index is an approximation of the market volatility and it is widely used to find tail events and measure robustness of certain models during its rising (when returns are high) and falling (when returns are low) periods.

  Hence, we can compute the performance of an asset or strategy conditional to market circumstances of high/low volatility. In this sense, consider the conditional average return during a certain regime of an asset V by:

$$\bar{R}^{(S)}_{\tau_i < R_{(V)} < \tau_{i+1}} = \frac{\sum_{t=1}^{T} R_t^{(S)}}{T} \tag{11}$$

  In our case we adopt V as VIX, and $\tau_1 = Q(R_{(V)}, 0.0)$, $\tau_2 = Q(R_{(V)}, 0.2)$,

$\tau_3 = Q(R_{(V)}, 0.4)$, $\tau_4 = Q(R_{(V)}, 0.6)$, $\tau_5 = Q(R_{(V)}, 0.8)$ and $\tau_6 = Q(R_{(V)}, 1.0)$, where $Q(R_{(V)}, x)$ is the x quantile of $R_{(V)}$, that is, $Q(R_{(V)}, x)$ is the number that is exactly above $x\%$ of $R_{(V)}$ values. We will refer to this five buckets – (τ_1, τ_2), (τ_2, τ_3), (τ_3, τ_4), (τ_4, τ_5), (τ_5, τ_6) – as the quintiles. In this occasion, τ_1 and τ_2 represents falling periods of VIX, in opposition to τ_5 and τ_6 that represents rising periods of VIX levels.

Based on the methodology developed in this section, next one presents the results and discussions of this work.

4 Results and Discussions

This section presents the results and discussions on MCCS trades. We start with an exploratory analysis of the MCCS trades dataset, mainly focusing on each MCCS returns, its relationship with some features (Carry, BE Widths, etc.) and responses to rises and falling periods of VIX. Then, using the methodology as mentioned earlier, we check the results of each model performance per MCCS. These suggest that Lasso Regression performed significantly well from different metrics and perspectives. Based on that, we close this section delving into Lasso Regression results, highlighting its overall and per MCCS returns, the relevance of each feature on predicting a certain MCCS trade, and a particular analysis on the EUR 4y5y5y MCCS trade.

4.1 Exploratory Analysis

Figure 9 exhibits boxplots with the annualised returns after holding for 1 year the trade[5], aggregated by trade type (Expiry, Forward and Swap Tenor) from September 2006 to September 2015.

Some patterns can be spotted in these boxplots: (i) the short expiries (mainly the 1y) tends to have negative median returns, whilst the long expiries (in major the 5y) reflects an opposite pattern with some yielding median returns close to 10%; (ii) when we increase the tenure (forward and swap) the returns tend to be more concentrated around the median, while for those with short tenure the spread tends to be greater; and (iii) although the others follow approximately a similar shape, a particular range of a package, the Xy5y5y, presents a clear upward sloping when the expiration (X) increases as well as an increase in the spreads around

[5]With the aim to improve our exposition, along the text we refer to this phrase as Holding 1y returns or just as returns.

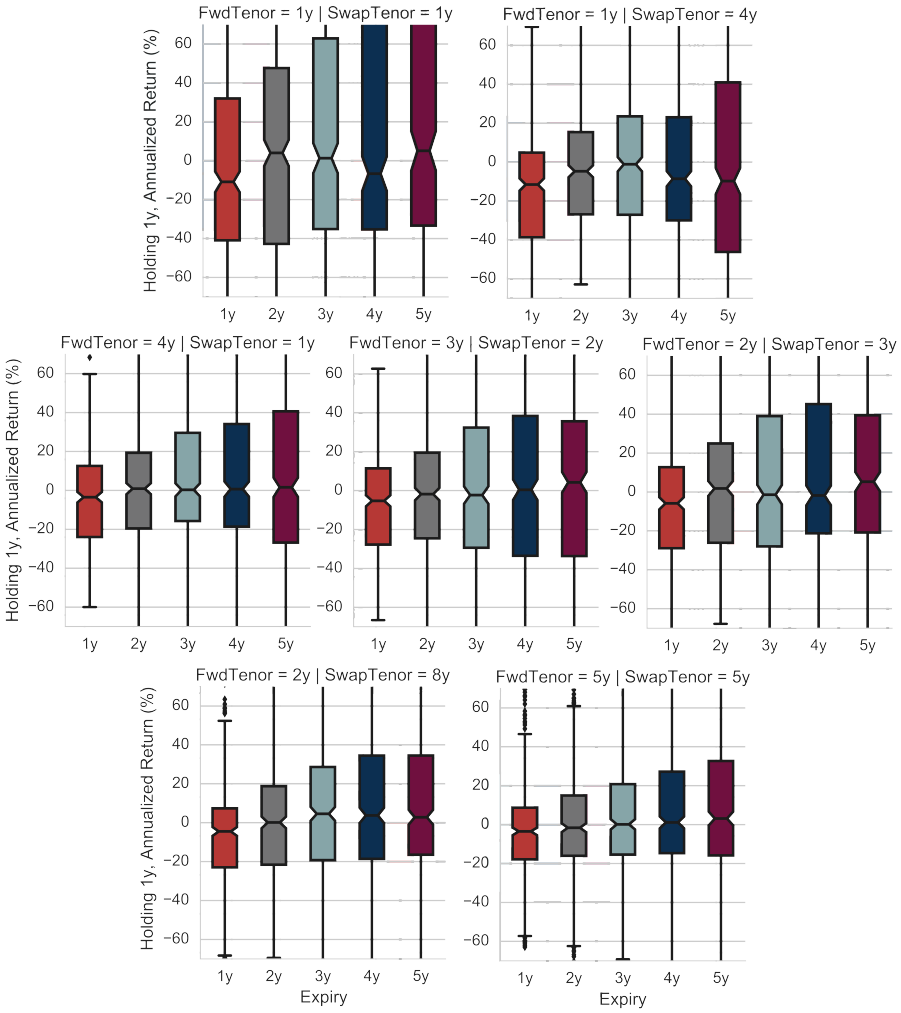

Figure 9: Boxplot from Sep/2006-Sep/2015 of annualised returns from holding during 1 year the instrument. All y-axes were fixed between 60% and -60% annualized return to facilitate their visualisations and comparisons.

the median. Drilling-down in this particular range, Figure 10 shows the Holding 1y returns obtained during the trading period, as well as the evolution of the BE Width and Carry at Expiry.

Concerning returns, before 2009 these trades did not see quite promising from a long positioning perspective, and it should be pointed out that during the bulk of the Financial Crisis (2008) these trades performed badly (from a long-only standpoint). In 2009, when the turmoil started to reduce, a long position in any of the expiries for the Xy5y5y MCCS yielded high positive returns. Potentially, this was due to the surge in forward volatility for long tenures, which caused the wide breakeven and great carry that we can observe during that year. After 2009, the returns tended to decrease, with a drawdown in 2014 probably caused by the lowest carry levels since 2008. In 2015 the returns were stabilised around the origin, with historically small

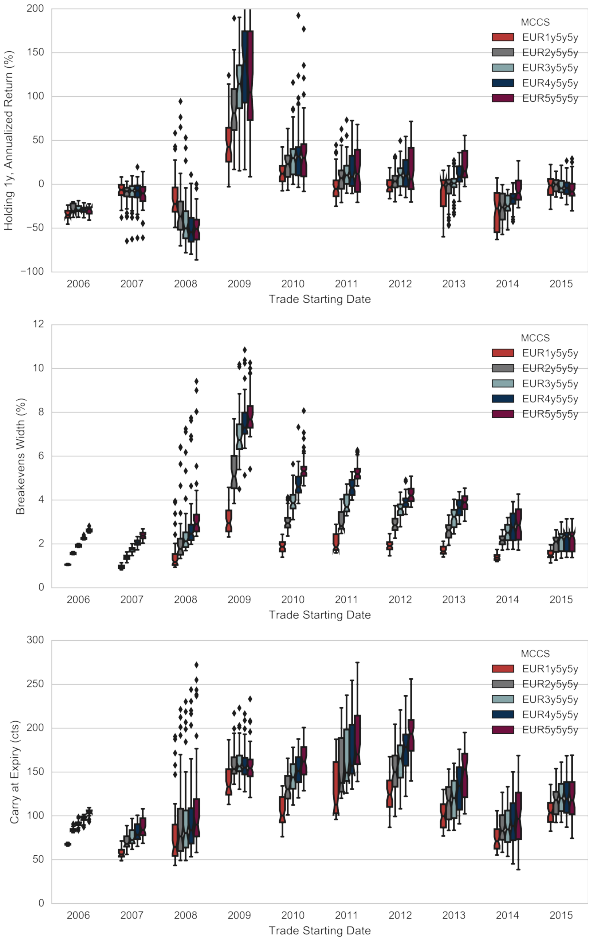

Figure 10: Boxplots of holding 1y annualised returns, BE Width and Carry at Expiry for different expirations (X) of Euro 5 year forward and swap tenure (EUR Xy5y5y).

breakeven levels, but a higher carry compared to 2014.

Overall there is a necessary, yet no sufficient relationship between returns and breakeven and carry values: high returns are linked to a surge in breakeven and carry levels, but the opposite is not necessarily true (2011 is a good example for this particular package). Therefore, there might be other variables that are explaining this variation but are not taken into account in this analysis. Figure 11 present the all period correlation matrix between the Holding 1y returns of each MCCS with different metrics and sensitivities extracted from it.

As expected, breakeven and carry levels have a positive correlation, meaning that high returns are associated with increases in carry and largely in BE Width levels; implied vol, conversely, is negatively related with MCCS returns, meaning that bigger eventual swings in the rates will tend to damage any yield that can be harvested by assuming long positions. Interesting is that usually, the traders receive reports based on carry and breakeven levels, as well as some view in the implied vol.

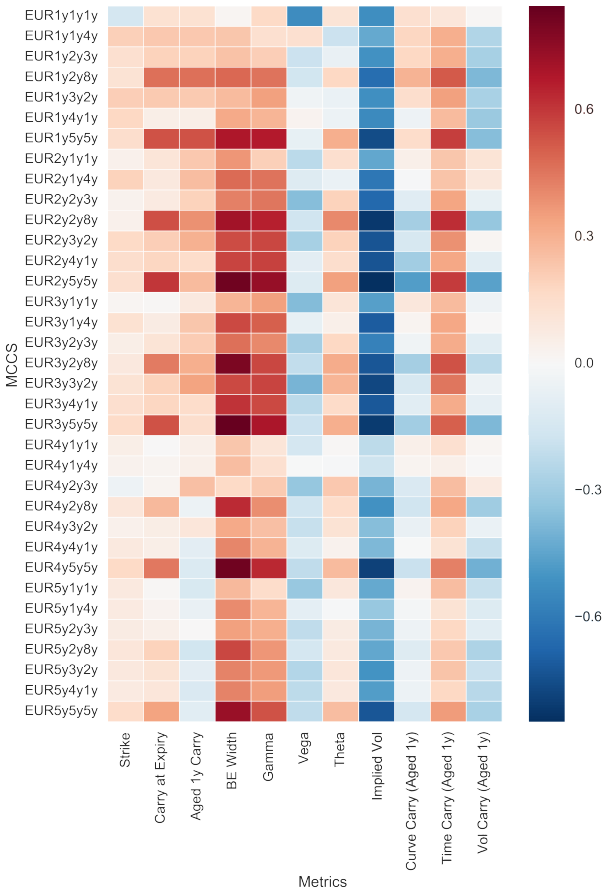

Figure 11: Heatmap with the pairwise correlation coefficient between features and returns.

However, as we can see, other metrics might be worthwhile to look and as well tap into when building a prediction model, like Gamma and Time Carry, although it is not clear from a linear sense whether they will add something new (e.g., Gamma exhibits the same pattern as BE width).

Figure 12 displays for each MCCS its average returns during different VIX regimes.

We can trace some patterns from Figure 12. Regardless of forward and swap tenure, as expiration increases a more solid relationship between expected returns and VIX levels appear, being turbulent periods met with losses and vice-versa. Also, the EUR ZyXyXy with Z $\geq$ 2 years tended to behave approximately similar as S&P 500 behave with VIX (Whaley, 2009), but with the shorter expiries being more insensitive. About this last point, we still do not have a clear explanation of why such pattern appeared, being our guess the current predominant period of a negative correlation between stock markets and bond markets. Hence, as strong swings happen in S&P 500 it tends to affect some parts of the yield curve, largely the 2y and 10y maturity because investors see government treasuries as safe heavens

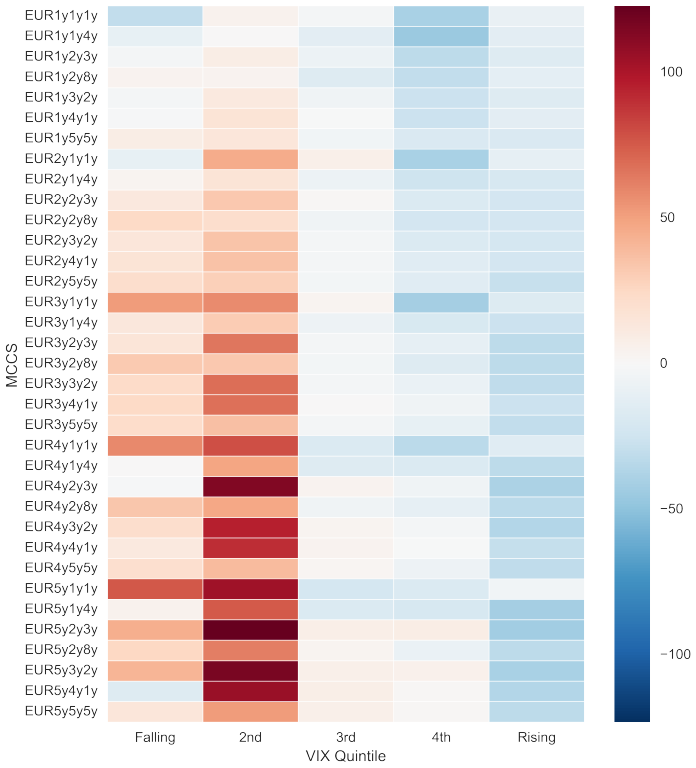

Figure 12: Heatmap with MCCS average returns during different VIX quintiles.

for stressful moments – and vice-versa. That is our guess of why the EUR 1yXyXy is less affected rather than the EUR 5yXyXy.

After outlining this exploratory analysis, next subsection head towards the recommendation system results.

4.2 Recommendation Systems Results

Figure 13 presents the Rho measure in the test set of each different model for all the MCCS trades. We can spot different patterns: (i) the predictive baselines (Naive and Mean Prediction) did not perform that well when compared to other predictive models; (ii) generally any model tested did not fared well for the 1y expires when compared to their longer counterparts; (iii) from the linear regression family, Lasso Regression followed by Ridge Regression are the ones that performed better; and (iv) the tree-based approaches performed more or less aligned, with good results for Random Forest and Grad Boost Reg; and (v) MLP fared well for a particular range of trades (e.g., EUR Xy1y1y), mainly when compared to other models in special Lasso Regression. Overall, all models exhibited reasonable predictive skills for a set of trades, with some complementarity between them.

However, a high predictive power is more a necessary condition rather than a

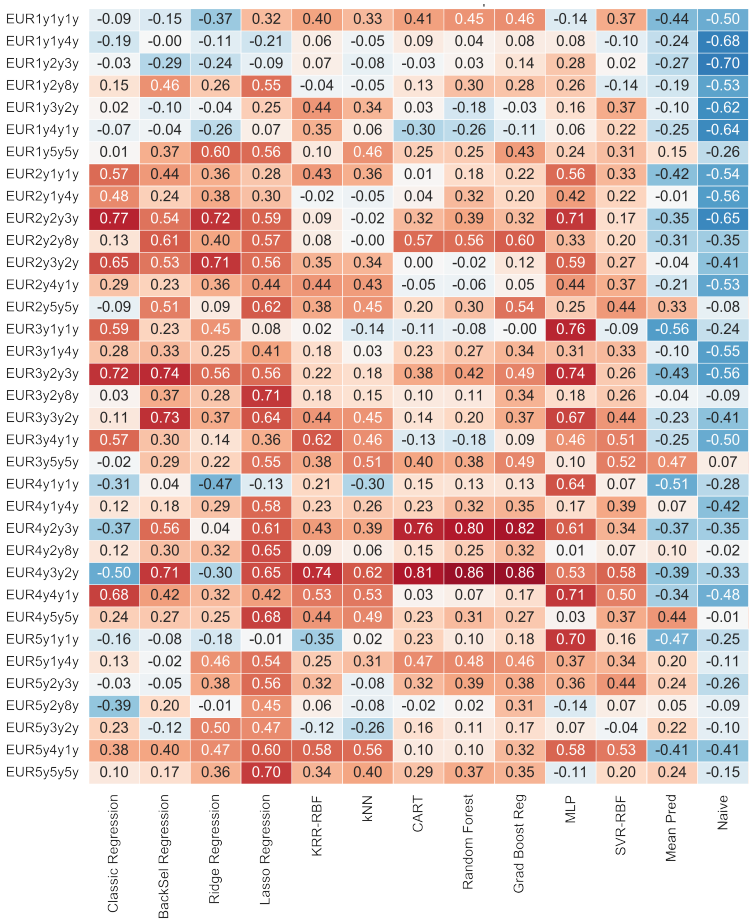

Figure 13: Heatmap with Rho values during the test set.

sufficient for higher return. In Figure 14 we display a heatmap with the results of all models for each trade regarding Average Return (%). Similarly, different remarks can be made over the global picture: (i) the Naive and Mean Pred models underperformed as well, but the traders benchmarks did perform reasonably well, surpassing the predictive models in many occasions; (ii) the linear regression models kept approximately their positions when we swapped from Rho to Average Return metric, with a clear edge for Lasso Regression; (iii) there was a clear inversion for the nonlinear models, especially the tree-based approaches; and (iv) MLP still fares well for the trades in the EUR Xy1y1y range, but did not repeat a more stable across other trades.

We believe a plausible explanation for the stability of linear models and degradation from the nonlinear is perhaps due to the objective function we are targeting when these models are trained and selected. Minimizing squared error is a necessary condition for profitable directional trades with linear models (Acar & Satchell, 2002), but it says nothing about nonlinear models (like the tree-based family, MLP,

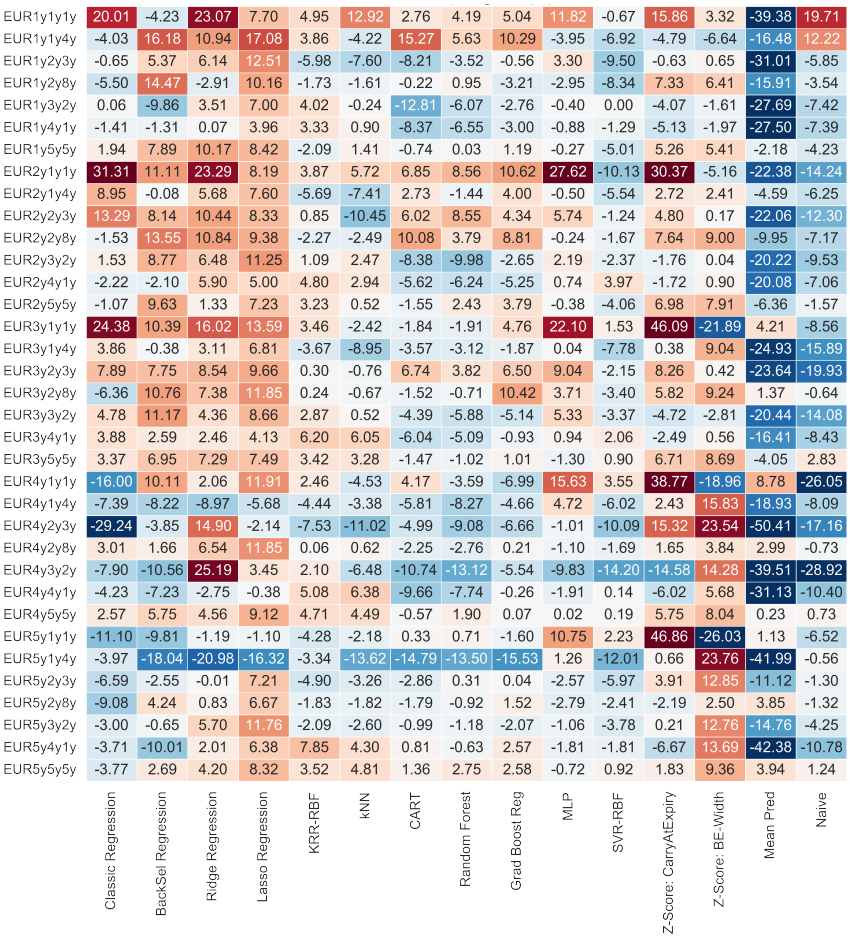

Figure 14: Heatmap with the historical Average Return (%) during test set.

KRR-RBF and SVR-RBF). Our results provide some evidence that such conjecture is eventually untrue, hence demanding changes in the objective function for nonlinear models.

When we take into account the variability seen in the stream of returns generated by the recommendation system, we may encounter a different picture. In this sense, Figure 15 shows a heatmap with the Information Ratio for all the available combinations of models and trades.

In general, the models approximately kept their advantages, but now all are standing on a similar scale. Based on these Information Ratio results, Table 6 presents a statistical analysis using the average ranks[6], Friedman test and Holm posthoc procedure (Derrac, García, Molina, & Herrera, 2011).

[6]When we rank the models for a single MCCS, it means that we sort all them in such way that the best performer is in the first place (receive value equal to 1), the second best is positioned in the second rank (receive value equal to 2), and so on. We can repeat this process for all trades and compute metrics, such as the average rank (e.g., 1.35 means that a particular model was placed mostly near to the first place).

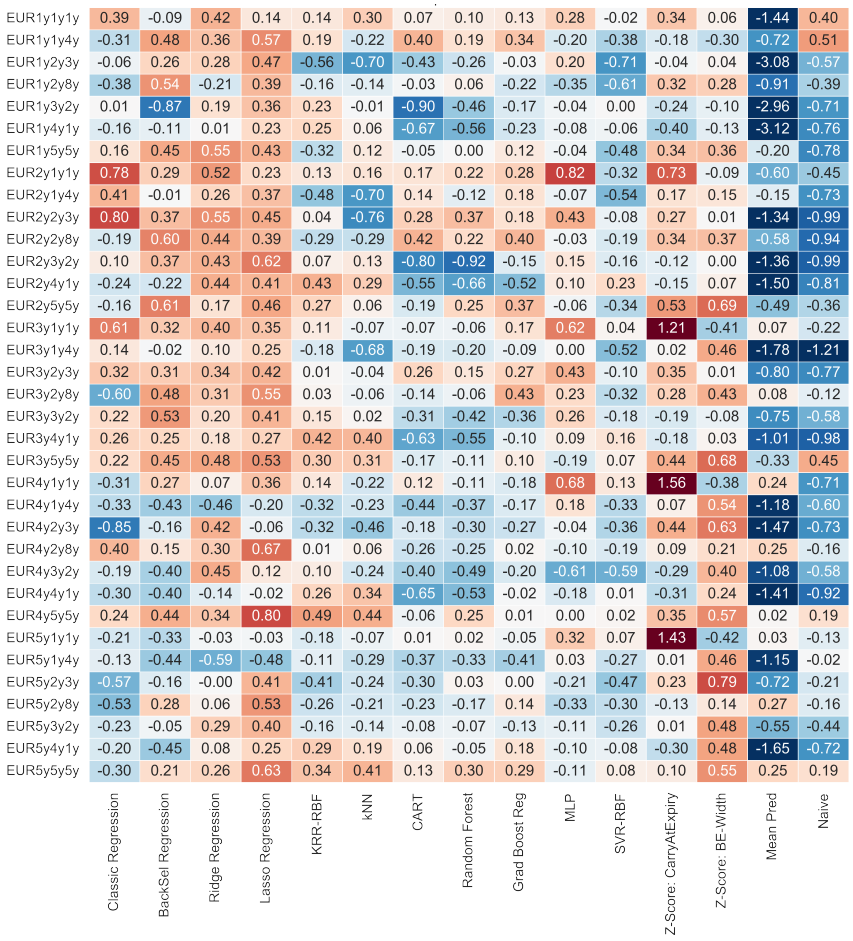

Figure 15: Heatmap with the Information Ratio during the test set.

When we look at the average rank, Lasso Regression was the top positioned (3.23) while Mean Pred remained most of the time as the worst choice (12.86). The trader's benchmarks performed pretty well, being placed in the third and fourth places. When we compare whether such result fared by Lasso Regression was substantially different from Ridge Regression (4.86), we arrive with a Z-score equal to 1.63 and a p-value of 0.0517. If we set our initial significance level as 0.05 and correct using the Holm procedure (last column) we can assert that Lasso did not perform significantly different from Ridge Regression, but way better than the other models. Therefore, Lasso Regression is capturing some information beyond that is being spanned by the trader's benchmarks, as well as beating almost all other predictive models for this particular task.

Figure 16 highlights how each model fared during different VIX regimes. They mostly exhibited an opposite profile from the ones found out for each MCCS (Figure 12), probably because predominantly the methods are taking short positions in the MCCS. This observation is evident in Lasso, Ridge and Backward Selection

Regression where these models performed well during rising periods of VIX levels. Some exceptions lie on MLP and BE Width: the former faring slightly well regardless of the VIX regime and the later performing well during jumping periods of VIX, losing only when its changes are small.

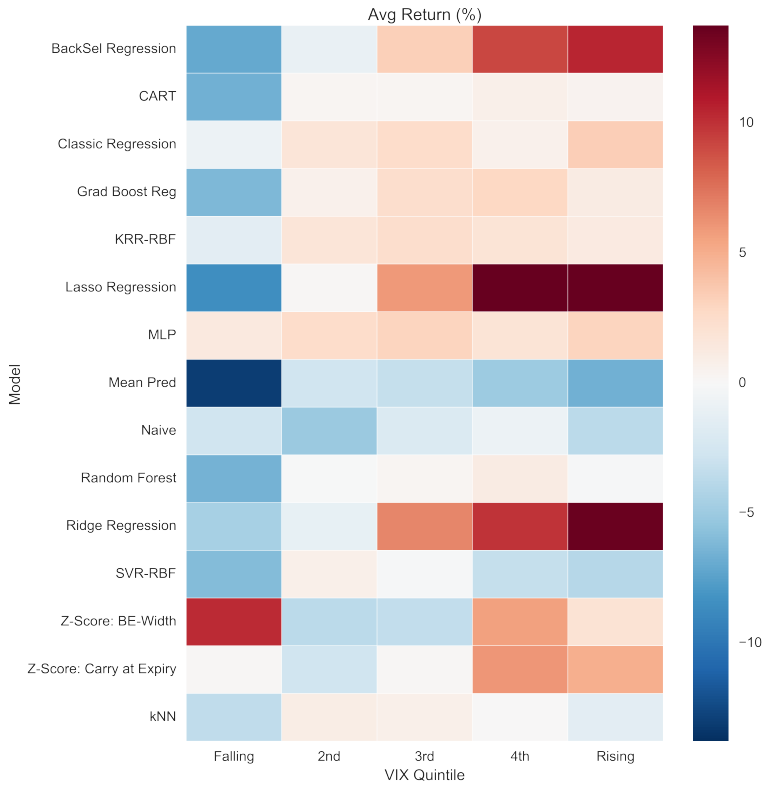

Figure 16: Heatmap with each recommendation system Average Returns (%) during different VIX regimes.

Our throughout analysis suggests that Lasso Regression seems to provide in general the best results across a range of metrics and criteria. In this sense, next subsection ends the results and discussion section by delving into Lasso Regression results. We start by showing its results across all MCCS, simulating the case where it was the sole member of the trading recommendation system. Then, we analysed its results on a particular trade, the EUR 4y5y5y since this trade has yielded half-half positive and negative returns (see Figure 9). Hence, this means that a predictive model will need to capture some signal to profit from it, rather than always going long or short in the case of 5 years and 1 years.

4.3 Delving into Lasso Regression Results

We start by looking at the aggregated returns harvested by Lasso Regression during its test phase. These results are consolidated in Figure 17, where: (i) the top plot shows the average return with standard deviations obtained across all MCCS per trading week; (ii) the middle reveal histograms, where the left one represents the returns obtained across all MCCS regardless of the trading date, while the right ones displayed the same data but conditioned per position; and (iii) finally the bottom image presents the trading success rate for each long/short position suggested by the model (left), with the break down by long/short position displayed as well (right). To clarify, trading success in this context means being long/short when the returns of trade were positive/negative regardless of its magnitude.

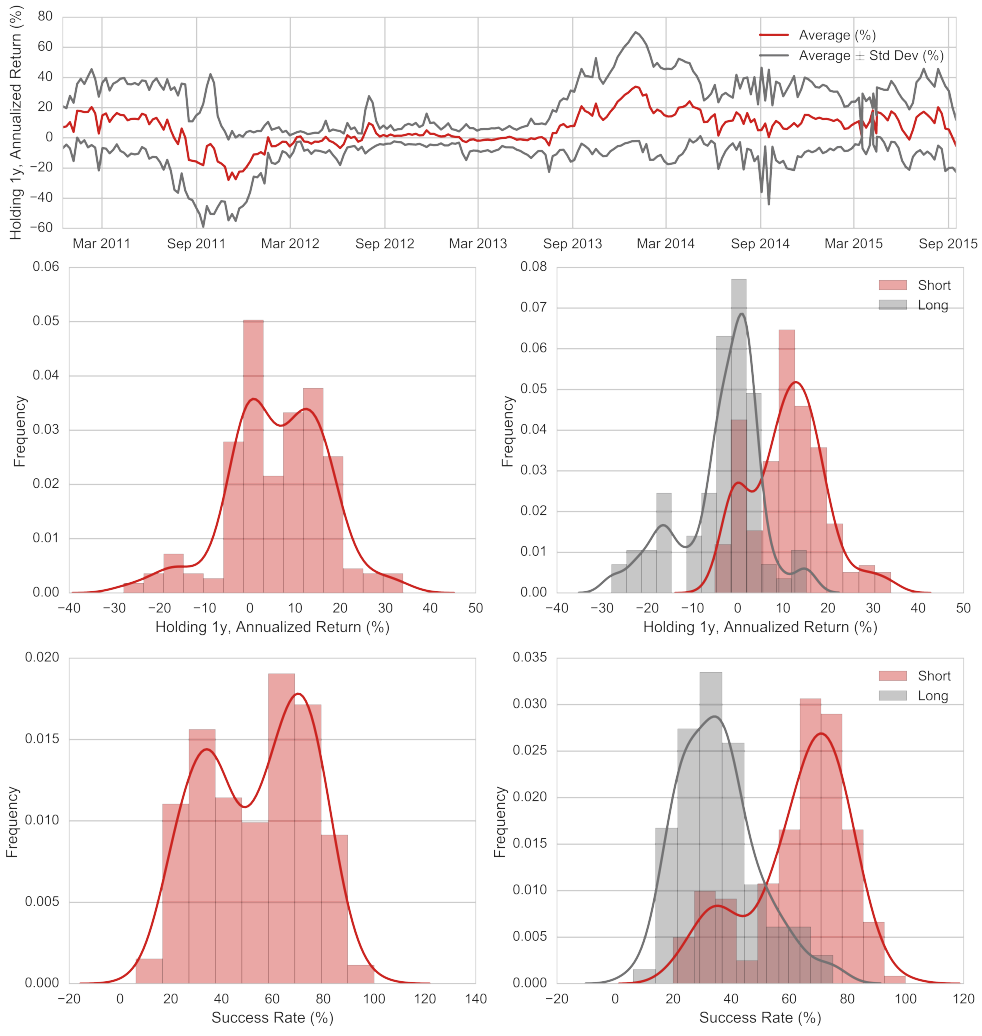

Figure 17: Aggregated returns over the period and histogram of aggregated returns and success rate from trades suggested using Lasso Regression.

In respect to the top image, we can see that Lasso Regression started well during

the first year but suffered a drawdown in the second to third year. This period was marked by higher volatility, mainly due to the final developments of the Euro Crisis period (2010-2012). However, from the third year onwards the average returns always scored positive values, usually ranging from 10% to 20% in average. Such performance can be seen stamped on the middle left histogram, where the bulk of returns lies above zero, and not only that but concentrated close to 15%. This performance was largely generated by Lasso Regression suggesting short positions (middle right), while the long positions were not so successful. Such pattern can be better seen in the histogram located at the bottom, where a bimodal distribution for trading recommendation success rate is depicted.

Probably the verified outperformance coming from taking short positions in the MCCS is linked with betting against the volatility/variance risk-premium trade (Choi et al., 2017). Roughly this strategy harvest the premium paid by a counterpart for the insurance on large swings in the market (almost the same as selling a put for equities options). Since in general, the market tends to remain range-bounded, the investor shorting the trade can repurchase it later for a smaller premium, profiting from this differential. Lasso Regression did dynamically the opposite and profited from it, largely because in this last 5-6 years was populated of higher volatility periods and tail events.

Figure 18 uses boxplots as a visualisation tool to decompose the aggregated results shown before, by informing per trade the returns obtained from using the Lasso Regression trading recommendation system.

Overall, some patterns can be spotted from the boxplots: (i) in general the medians are located above zero, meaning that more than half of the trades tended to yield positive returns; (ii) the upward sloping profile visualized in Figure 9 is not presented here, meaning that Lasso Regression is exploring long/short position regardless of the most frequent outcome for each MCCS tenure; (iii) Lasso Regression tended to perform well for EUR 3yXyXy trades and since these tended to be historically a challenging pick (medians are centred to zero in these trades), it means that the model is actually capturing some signal from the data and not naively guessing long/short positions; and (iv) the returns distribution, mostly with higher forward and swap tenure (second and third row), tended to be right-skewed – because the third quartile is far from the median, whereas the first is squeezed towards the median. This last fact denounces that Lasso Regression frequently generates small negative outcomes, and dangerous scenarios are not as likely, which tends to be a desired property for quantitative strategies in general.

Figure 19 help us to analyse which features are being most significant by Lasso

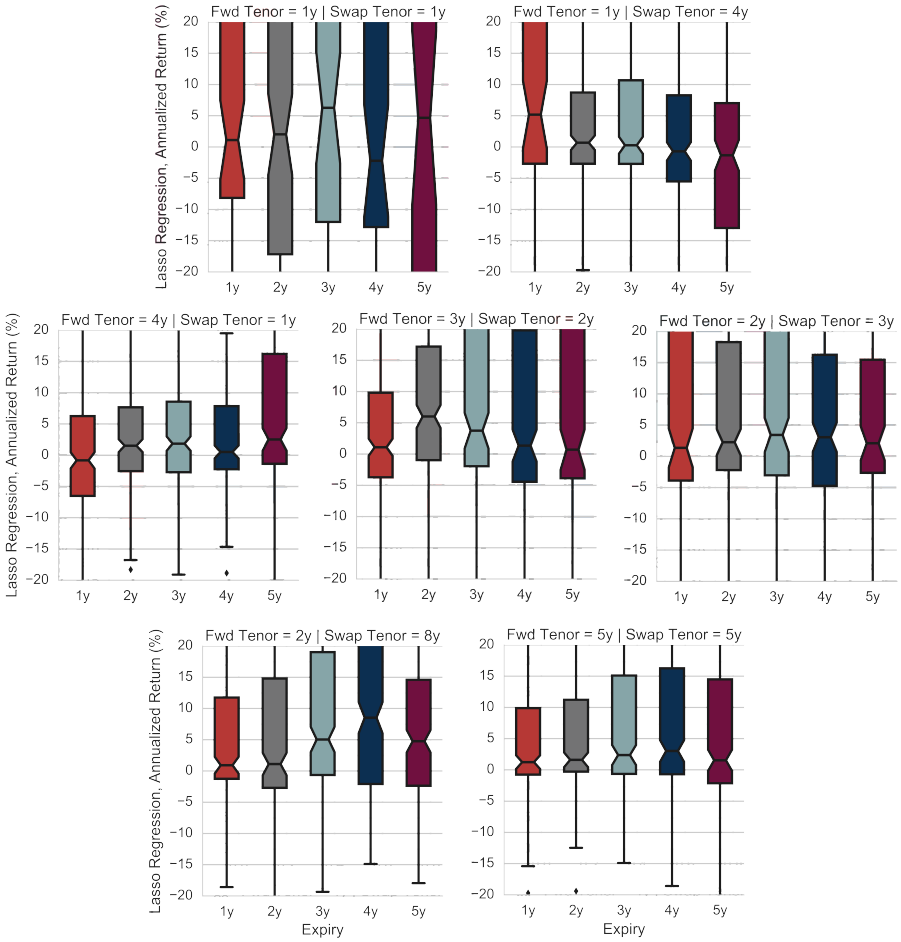

Figure 18: Boxplot with the returns obtained from Lasso Regression for each instrument. All y-axes were fixed between 0.3 and -0.3 (30% and -30% annualised return) to facilitate their visualisations and comparisons.

Regression for each particular trade.

Each cell corresponds to a normalised t-stats[7] from the model coefficients built in the last step from k-rolling-cv. Implied Vol was the most significant feature pointed out by Lasso Regression, is negatively related with the MCCS returns. Other important features were the BE Width – slightly positively correlated with returns – and the Carry at Expiry – strangely negatively related, but probably due to the depressed levels of carry that has been seen in the last batch of data (check Figure 10 for the EUR Xy5y5y case). Lasso Regression promoted in general very sparse models, being the other features playing specific roles for some trades like Time Carry for short-dated trades. Finally, the lagged returns were just relevant

[7]By normalised t-stats we mean dividing each coefficient t-stat by the sum of the absolute values of all t-stats in the model. The result is a number between -1 and +1, indicating the significant magnitude in comparison to other variables, as well as the direction in which it affects the model predictions. We multiplied it by a one hundred just to work on a more convenient scale of -100% and 100%.

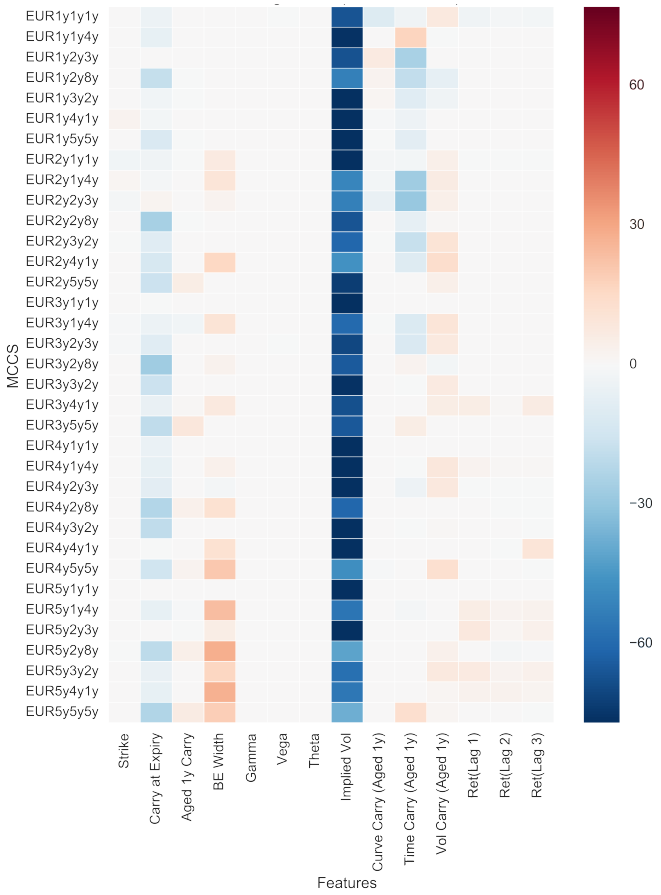

Figure 19: Heatmap with the Feature Significance (%) obtained from normalised t-stats of Lasso Regression coefficients.

for few trades, and perhaps could be omitted for certain trades in the future to guarantee a broader dataset.

We close this section showing results of Lasso Regression for a specific trade: EUR 4y5y5y – Figure 20.

Starting from the top, it can be seen that Lasso Regression was able to predict reasonably well the observed returns after 1 year holding this trade. It is not perceived any over/underestimation of values, with the mirror-shape format indicating a good fit. This observation is reflected in the middle image, where the long/short positions track well the observed returns of the EUR 4y5y5y. The main mistakes are due to events that were not incorporated into the model and perhaps were also unforecastable: (i) March to April 2012, possibly due to the Greek Debt Restructuring Agreement; (ii) May to November 2013, linked with the Taper Tantrum event in the US and its effect on Europe rates.

Although such events have influenced in the strategies return, last plot shows that the trading success of Lasso Regression has attained a historical Area Under

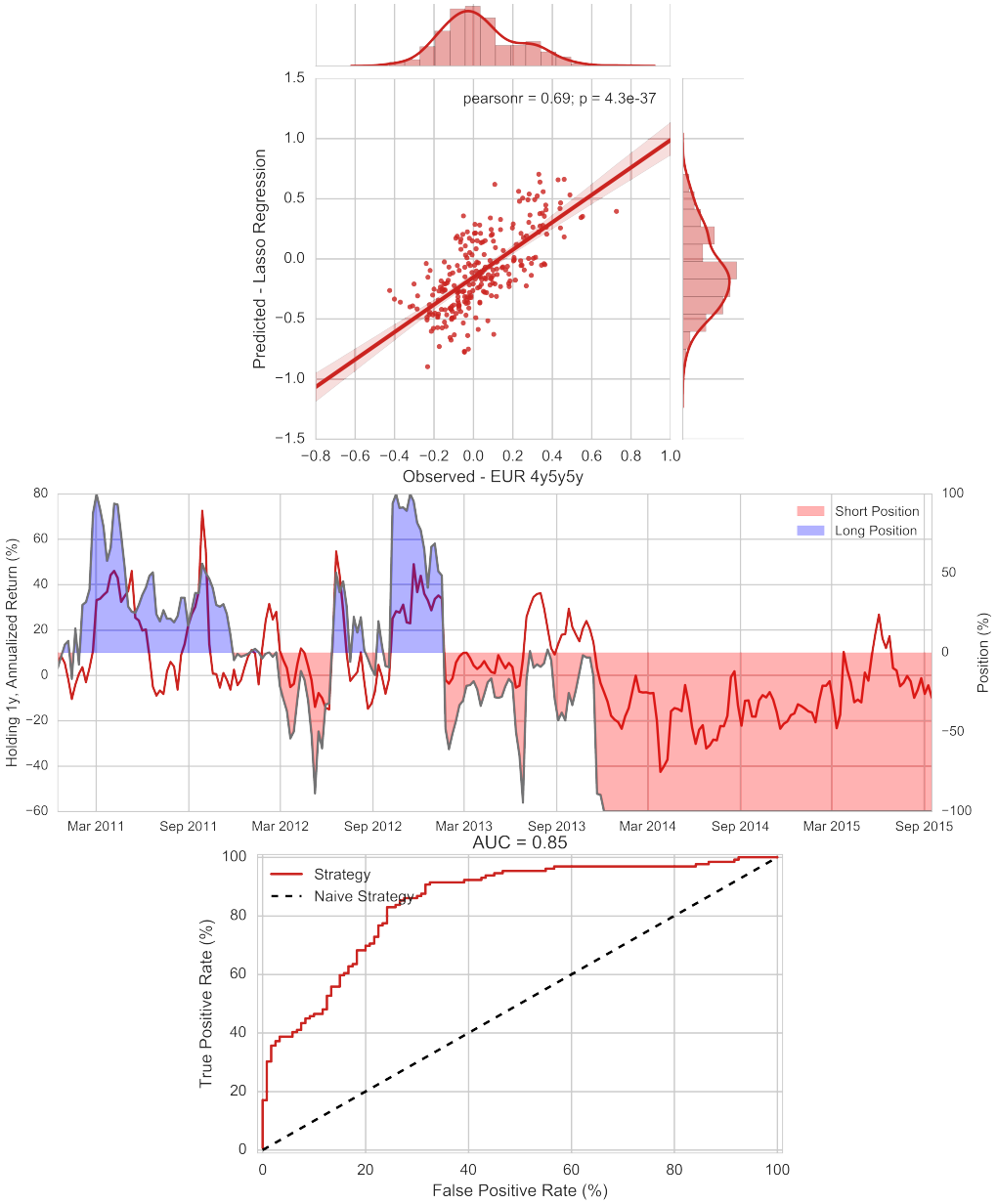

Figure 20: Lasso Regression results for EUR 4y5y5y: predicted versus observed values, returns and long/short signals over the period and receiving operating characteristic curve based on the success rate.

the Curve (AUC) of 0.85, with the capacity to control the false alarms to 5% and still recommend trades accurately 40% of the time. This is a good indicator since in the onset of the recommendation system it is better to reduce the chances of suggesting a bad trade than missing a good pick.

Next section draws some conclusions from our work as well as outline new directions for further research.

5 Conclusions

This work proposed a trading recommendation system for Mid-Curve Calendar Spread Trades (MCCS). We started with the main motivations, highlighting the bottlenecks that the derivatives traders face in their day-to-day routines. To tackle these issues, we propose a recommendation system that could analyse and rank a set of fixed income derivatives trades. Our first experiment is designing and applying this method for Mid-Curve Calendar Spread trades. However, before delving into the experimental setting, we first looked for current solutions available in the scientific literature. In this sense, to the best our knowledge we could not find anything suitable for interest rate swaptions strategies, though we can tap into works from the returns prediction literature for the modelling strategy.

Therefore, we started the methodology by showing the dataset: it comprised of 35 MCCS trades, ranging from September 2006 to September 2016, with different expirations, forward and swap tenures. For each particular trade, we described how the sampling of inputs (metrics, sensitivities and lagged returns) and outputs (returns from unwinding the trade after one year of its start) were computed on a weekly basis. Then, we displayed the modelling strategy by highlighting the models that were trained as well as which hyperparameters were investigated during the nested resampling step. Before entering the results section, we presented the backtesting setting with the performance measures used to compare different methodologies.

Our results section started from an exploratory analysis on the available trades. We look into their all period returns, relating them with Carry and BE width elements over time, computing their correlation with metrics and sensitivities as well as their performance during certain VIX regimes. We spotted many patterns, such as a clear upward sloping in the median returns when the expiration grows, with a proportional increase in the inter-quartile range. Also, the observation that these trades returns are driven by not only Carry, BE width and ATMF Implied Volatility (metrics that are often checked by the traders), but by other elements like Gamma and Time-Carry can give us an edge on separating good and bad deals. Finally, that regardless of forward and swap tenure, as expiration increases a more solid relationship between expected returns and VIX levels appear, being turbulent periods met with losses and vice-versa – with the exception coming from the shorter expiries.

Most models provided results better than the modelling benchmarks (Mean and Naive), yet very few were able to outperform the trader's benchmarks. Our results suggested that linear models with shrinkage procedures (e.g., Ridge and Lasso) tended to perform better than their nonlinear counterparts (like Kernel Ridge Regression, SVR and MLP). Also, regarding interpretability, they tend to be easier to

convey to the traders, since most are versed in linear models. Reasons linked with the nature of the dataset, sampling size and the limited choice for the hyperparameters to reduce the computational burden might have biased our results towards this direction, is not so clear which of these played a major/minor role. Further investigations and enhancements will provide a more clear picture on this topic.

When we delved into Lasso Regression results, we found out that this model wielded some interesting features like: (i) it learned a type of volatility buying/selling strategy without being programmed to do so; (ii) its returns distribution across all MCCS tended to be right-skewed, meaning that we are more hedged towards dangerous scenarios with greater chances of upsides; (iii) it matched traders view on selecting good trades, but adding some dynamic view on it since Carry at Expiry is now negatively linked with returns, rather than the original view from the traders; and (iv) it can control reasonably well the false positive rate without sacrificing much the discovery and suggestion of good trades. Since the system will be received with caution by traders and salespeople, we can in the short-term increase the threshold to recommend a trade without risking so much to lose profitable opportunities (like our argument for EUR 4y5y5y case).

We believe that Lasso Regression will be our choice for a first version of the trading recommendation system, with future developments giving space to different models and mixed approaches. Also, we foresee a multitude of directions for future works, with some highlighted below:

- Increasing the sampling frequency: we used for this experiment data from weekly trades. Although this turned the modelling process faster, we had an undersampling close to a fifth. This, of course, reduced the power of our backtest as well as limited the range of models that can be applied. We expect to start using daily trades data, but we do believe that it will only add more data between the gaps because of swaps rates, in general, tends to move quite slowly when compared to equities prices.

- Models and Ensemble methods: since linear models with shrinkage performed well, we believe that using some other methods of the same family like Elastic Net or LARS (Friedman et al., 2001) can provide results that outperforms Lasso Regression. Also, we perceive a negative association between traders benchmarks (BE width especially) and model-based strategies (Lasso Regression and MLP, for instance). This evidence can be seen on Figures 15 and 16.

- Signal generating function: a better analysis on the signal generating function

should be performed, attempting to find out potential alternatives for equation 6. Although it worked properly, we believe that some extra effort can overall improve models results and the nested resampling framework can help us find that out.

- Use a more application affine objective function to train/select the models: even though minimising prediction error is a necessary condition for profitable trades it is not a sufficient. Investing in a classification approach with a weighted objective function might be a potential further direction, with the upside of not requiring a signal function and with the downside of not being precise on the expected return of a trade.

- Regarding future applications, we expect to expand this work for US dollar trades as well as incorporating new models and ensemble approaches. Another viable alternative is to replicate such methodology to similar derivative trades but in the equities space (calendar-spreads, butterflies, etc.).

Acknowledgements

Adriano Soares Koshiyama wants to acknowledge the funding for its PhD studies provided by the Brazilian Research Council (CNPq) through the Science Without Borders program. Also, the authors would like to thanks, Guillaume Andrieux, Tomoya Horiuchi, Gerald Rushton, Tam Rajendran, and Anthony Morris for all the comments and support during this research.

References

Acar, E., & Satchell, S. (2002). *Advanced trading rules*. Butterworth-Heinemann.

Andersen, T. G., Bollerslev, T., Diebold, F. X., & Vega, C. (2007). Real-time price discovery in global stock, bond and foreign exchange markets. *Journal of International Economics*, *73*(2), 251 - 277. doi: https://doi.org/10.1016/j.jinteco.2007.02.004

Baker, R. M., Coolen-Maturi, T., & Coolen, F. P. A. (2017). Nonparametric predictive inference for stock returns. *Journal of Applied Statistics*, *44*(8), 1333-1349. doi: 10.1080/02664763.2016.1204429

Bischl, B., Mersmann, O., Trautmann, H., & Weihs, C. (2012). Resampling methods for meta-model validation with recommendations for evolutionary computation. *Evolutionary Computation*, *20*(2), 249–275.

Bishop, C. (2007). *Pattern recognition and machine learning*. Springer, New York.

Brigo, D., & Mercurio, F. (2007). *Interest rate models-theory and practice: with smile, inflation and credit*. Springer Science & Business Media.

Campbell, J. Y., Lo, A. W.-C., & MacKinlay, A. C. (1997). *The econometrics of financial markets*. Princeton University press.

Chen, H., De, P., Hu, Y. J., & Hwang, B.-H. (2014). Wisdom of crowds: The value of stock opinions transmitted through social media. *The Review of Financial Studies*, *27*(5), 1367. doi: 10.1093/rfs/hhu001

Chen, Y.-L., & Gau, Y.-F. (2010). News announcements and price discovery in foreign exchange spot and futures markets. *Journal of Banking and Finance*, *34*(7), 1628 - 1636. doi: https://doi.org/10.1016/j.jbankfin.2010.03.009

Choi, H., Mueller, P., & Vedolin, A. (2017). Bond variance risk premiums. *Review of Finance*, *21*(3), 987–1022.

Chong, E., Han, C., & Park, F. C. (2017). Deep learning networks for stock market analysis and prediction: Methodology, data representations, and case studies. *Expert Systems with Applications*, *83*, 187 - 205. doi: https://doi.org/10.1016/j.eswa.2017.04.030

Corb, H. (2012). *Interest rate swaps and other derivatives*. Columbia University Press.

Deng, Y., Bao, F., Kong, Y., Ren, Z., & Dai, Q. (2017). Deep direct reinforcement learning for financial signal representation and trading. *IEEE Transactions on Neural Networks and Learning Systems*, *28*(3), 653-664. doi: 10.1109/TNNLS.2016.2522401

Derrac, J., García, S., Molina, D., & Herrera, F. (2011). A practical tutorial on the use of nonparametric statistical tests as a methodology for comparing evolutionary and swarm intelligence algorithms. *Swarm and Evolutionary Computation*, *1*(1), 3–18.

Duda, R. O., Hart, P. E., & Stork, D. G. (2000). *Pattern classification (2nd edition)*. Wiley-Interscience.

Duyvesteyn, J., & de Zwart, G. (2015). Riding the swaption curve. *Journal of Banking Finance*, *59*, 57 - 75. doi: https://doi.org/10.1016/j.jbankfin.2015.05.012

Ehrman, D. S. (2006). *The handbook of pairs trading: strategies using equities, options, and futures* (Vol. 240). John Wiley & Sons.

Elliott, G., Gargano, A., & Timmermann, A. (2013). Complete subset regressions. *Journal of Econometrics*, *177*(2), 357 - 373. doi: https://doi.org/10.1016/j.jeconom.2013.04.017

El Ouadghiri, I., Mignon, V., & Boitout, N. (2016). On the impact of macroeconomic news surprises on treasury-bond returns. *Annals of Finance*, *12*(1), 29–53. doi: 10.1007/s10436-015-0271-3

Feuerriegel, S., & Prendinger, H. (2016). News-based trading strategies. *Decision Support Systems*, *90*, 65 - 74. doi: https://doi.org/10.1016/j.dss.2016.06.020

Firoozye, N., & Zhang, Q. (2014a). Turbo carry zooms ahead: A performance update of the turbo-carry trades. *Nomura International plc, Nomura Research*.

Firoozye, N., & Zhang, Q. (2014b). Usd short-term front-end turbo carry usd 1m1y2y trades: Short horizon for sizeable carry. *Nomura International plc, Nomura Research*.

Firoozye, N., & Zheng, X. (2016). Market update: Forward vol and midcurve calendar spreads in usd and eur recent levels and carry and trades of note. *Nomura International plc, Nomura Research*.

Friedman, J., Hastie, T., & Tibshirani, R. (2001). *The elements of statistical learning* (Vol. 1). Springer series in statistics Springer, Berlin.

Gencay, R., & Qi, M. (2001). Pricing and hedging derivative securities with neural networks: Bayesian regularization, early stopping, and bagging. *IEEE Transactions on Neural Networks*, *12*(4), 726-734. doi: 10.1109/72.935086

Gerlein, E. A., McGinnity, M., Belatreche, A., & Coleman, S. (2016). Evaluating machine learning classification for financial trading: An empirical approach. *Expert Systems with Applications*, *54*, 193 - 207. doi: https://doi.org/10.1016/j.eswa.2016.01.018

Glasserman, P. (2013). *Monte carlo methods in financial engineering* (Vol. 53). Springer Science & Business Media.

Han, G.-S., & Lee, J. (2008). Prediction of pricing and hedging errors for equity linked warrants with gaussian process models. *Expert Systems with Applications*, *35*(1–2), 515 - 523. doi: https://doi.org/10.1016/j.eswa.2007.07.041

Haykin, S. S. (2009). *Neural networks and learning machines* (Vol. 3). Pearson.

Hyndman, R., Koehler, A. B., Ord, J. K., & Snyder, R. D. (2008). *Forecasting with exponential smoothing: the state space approach*. Springer Science & Business Media.

Karathanasopoulos, A., Theofilatos, K. A., Sermpinis, G., Dunis, C., Mitra, S., & Stasinakis, C. (2016). Stock market prediction using evolutionary support vector machines: an application to the ase20 index. *The European Journal of Finance*, *22*(12), 1145-1163. doi: 10.1080/1351847X.2015.1040167

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

of Operational Research, *259*(2), 689 - 702. doi: https://doi.org/10.1016/j.ejor.2016.10.031

Malkiel, B. G. (2003). The efficient market hypothesis and its critics. *The Journal of Economic Perspectives*, *17*(1), 59–82.

Nakano, M., Takahashi, A., & Takahashi, S. (2017). Generalized exponential moving average (ema) model with particle filtering and anomaly detection. *Expert Systems with Applications*, *73*, 187 - 200. doi: https://doi.org/10.1016/j.eswa.2016.12.034

Natenberg, S. (2014). *Option volatility and pricing: advanced trading strategies and techniques*. McGraw Hill Professional.

Park, H., Kim, N., & Lee, J. (2014). Parametric models and non-parametric machine learning models for predicting option prices: Empirical comparison study over {KOSPI} 200 index options. *Expert Systems with Applications*, *41*(11), 5227 - 5237. doi: https://doi.org/10.1016/j.eswa.2014.01.032

Park, H., & Lee, J. (2012). Forecasting nonnegative option price distributions using bayesian kernel methods. *Expert Systems with Applications*, *39*(18), 13243 - 13252. doi: https://doi.org/10.1016/j.eswa.2012.05.077

Rebonato, R., McKay, K., & White, R. (2011). *The sabr/libor market model: Pricing, calibration and hedging for complex interest-rate derivatives*. John Wiley & Sons.

Shreve, S. E. (2004). *Stochastic calculus for finance ii: Continuous-time models* (Vol. 11). Springer Science & Business Media.

Sousa, J. B., Esquível, M. L., & Gaspar, R. M. (2012). Machine learning vasicek model calibration with gaussian processes. *Communications in Statistics - Simulation and Computation*, *41*(6), 776-786. doi: 10.1080/03610918.2012.625324

Taleb, N. (1997). *Dynamic hedging: managing vanilla and exotic options* (Vol. 64). John Wiley & Sons.

von Spreckelsen, C., von Mettenheim, H.-J., & Breitner, M. H. (2014). Real-time pricing and hedging of options on currency futures with artificial neural networks. *Journal of Forecasting*, *33*(6), 419–432. doi: 10.1002/for.2311

Weng, B., Ahmed, M. A., & Megahed, F. M. (2017). Stock market one-day ahead movement prediction using disparate data sources. *Expert Systems with Applications*, *79*, 153 - 163. doi: https://doi.org/10.1016/j.eswa.2017.02.041

Whaley, R. E. (2009). Understanding the vix. *The Journal of Portfolio Management*, *35*(3), 98–105.

Zhou, T., Gao, S., Wang, J., Chu, C., Todo, Y., & Tang, Z. (2016). Financial time

series prediction using a dendritic neuron model. *Knowledge-Based Systems*, *105*, 214 - 224. doi: https://doi.org/10.1016/j.knosys.2016.05.031

Zhou, X., Nakajima, J., & West, M. (2014). Bayesian forecasting and portfolio decisions using dynamic dependent sparse factor models. *International Journal of Forecasting*, *30*(4), 963-980.

Tables

Table 1: Configuration of the MCCS trades used. Remember that the package involves selling an option on a forward-starting swap (using the Expiration and Forward tenures) and buying a long expiration option (swap tenure) on a spot-starting swap.

Currency	Expiry	Forward	Swap	Currency	Expiry	Forward	Swap
EUR	1y	1y	1y	EUR	3y	3y	2y
EUR	1y	1y	4y	EUR	3y	4y	1y
EUR	1y	2y	3y	EUR	3y	5y	5y
EUR	1y	2y	8y	EUR	4y	1y	1y
EUR	1y	3y	2y	EUR	4y	1y	4y
EUR	1y	4y	1y	EUR	4y	2y	3y
EUR	1y	5y	5y	EUR	4y	2y	8y
EUR	2y	1y	1y	EUR	4y	3y	2y
EUR	2y	1y	4y	EUR	4y	4y	1y
EUR	2y	2y	3y	EUR	4y	5y	5y
EUR	2y	2y	8y	EUR	5y	1y	1y
EUR	2y	3y	2y	EUR	5y	1y	4y
EUR	2y	4y	1y	EUR	5y	2y	3y
EUR	2y	5y	5y	EUR	5y	2y	8y
EUR	3y	1y	1y	EUR	5y	3y	2y
EUR	3y	1y	4y	EUR	5y	4y	1y
EUR	3y	2y	3y	EUR	5y	5y	5y
EUR	3y	2y	8y				

Table 2: Metrics and sensitivities computed for each available package at time t.

Features	
PV	Strike
Carry at Expiry (Carry)	Breakeven Width (BE Width)
Aged 1y Carry	Theta
ATMF Implied Volatility (Implied Vol)	Gamma
Vega	Curve Carry (Aged 1y)
Time Carry (Aged 1y)	Volatility Carry (Aged 1y) (Vol Carry)

Table 3: Example of information available at time t and $t+h$ for the MCCS. A more schematic representation can be found in Figure 7.

Instant (t)	PV_t	$R_t^{(h)}$	$R_{t-h-1}^{(h)}$	Strike	Features
1	300	0.3	-	3.4	...
2	320	0.2	-	3.2	...
...	...	...	...	...	...
$t+h-1$	250	-0.1	-	1.8	...
$t+h$	260	-0.05	-	1.9	...
$t+h+1$	270	-0.2	0.2	2.0	...
...	...	...	...	...	...
T	250	-	0.1	2.2	...

Table 4: Details used to generate the MCCS trade dataset.

Detail	Value
Period	September 2006 to September 2016
Holding Period (h)	1 year
Trade Frequency	Weekly (usually on Wednesday)
Strike	At the Money Forward (ATMF)
Lagged data ($h-p$)	$p = 1, 2$ and 3 lagged returns
Assumption	Characteristic
Bid-Ask	Middle Rate
Transaction Costs	Entry and Unwind = $0.75 \times Vega_t$
Funding Rate	Libor 3 month rate

Table 5: Parameters used to model the MCCS trade dataset.

Abbreviation	**Model**	**Fixed Hyperparameters**	**Cross-Validated Hyperparameters**
Classic Regression	Classical Linear Regression	None	None
BackSel Regression	Stepwise Regression	Backward Selection	None
Ridge Regression	Ridge Regression	None	$\lambda = \{10^0, 10^{-1}, 10^{-1}, 10^{-2}, 10^{-3}\}$
Lasso Regression	Lasso Regression	None	$\lambda = \{10^0, 10^{-1}, 10^{-1}, 10^{-2}, 10^{-3}\}$
KRR-RBF	Kernel Ridge Regression	Radial-Basis Function kernel	$\lambda = \{10^0, 10^{-1}, 10^{-1}, 10^{-2}, 10^{-3}\}$ and $\gamma = \{10^{-2}, 10^{-1}, 10^0, 10^1, 10^2\}$
kNN	k-Nearest Neighbours	Euclidean Distance	$k = \{3, 5, 7, 9\}$
CART	Classification and Regression Tree	MSE Function	Max depth $= \{2, 3, 5, 7\}$
Random Forest	Random Forest	Number of trees = 333 and Max depth = 5	None
Grad Boost Reg	Gradient Boosting Tree	Number of trees = 333 and Max depth = 5	Learning Rate: $\{0.1, 0.3, 0.5\}$
MLP	Multi-Layer Perceptron	Single hidden layer with hyperbolic tangent as transfer function	$\lambda = \{10^0, 10^{-1}, 10^{-1}, 10^{-2}, 10^{-3}\}$ and $nn = \{5, 7, 12\}$
SVR-RBF	Support Vector Regression	Radial-Basis Function kernel	$C = \{10^0, 10^1, 10^2, 10^3\}$ and $\gamma = \{10^{-2}, 10^{-1}, 10^0, 10^1, 10^2\}$
Abbreviation	**Baseline Model**	**Parameter**	
Mean Pred	Average Prediction		
Naive	Naive Model		
Z-Score: BE-Width	BE Width feature	Rolling window of size = 1 year	$S_t = \lfloor_{-1}((Z > 1) * Z)/(3)\rceil^{+1}$
Z-Score: CarryAtExpiry	Carry at Expiry feature	Rolling window of size = 1 year	$S_t = \lfloor_{-1}((Z > 1) * Z)/(3)\rceil^{+1}$
	Other Parameters	**Values**	
	Warm-up Period (L)	$L_{outer} = 2$ years	$L_{inner} = 1$ year
	k-rolling-cv	$k_{outer} = 1$ week for outer	$k_{inner} \approx (T_{train} - L_{inner})/5$ for inner
	Outlier Treatment	Winsorizing	
	Winsorizing Quantiles	0.01 and 0.95	
	Missing Data Treatment	Remove	

Table 6: Average ranks, Friedman and Holm post-hoc statistical tests and analysis for Information Ratio.

Model	Average Rank	Z-score	p-value	Holm Correction
Mean Pred	12.86	9.63	**<0.0001**	0.0036
Naive	11.89	8.66	**<0.0001**	0.0038
SVR-RBF	10.60	7.37	**<0.0001**	0.0042
CART	10.11	6.89	**<0.0001**	0.0045
Random Forest	9.31	6.09	**<0.0001**	0.0050
kNN	8.23	5.00	**<0.0001**	0.0056
Classic Regression	8.17	4.94	**<0.0001**	0.0063
Grad Boost Reg	7.69	4.46	**<0.0001**	0.0071
KRR-RBF	7.49	4.26	**<0.0001**	0.0083
MLP	7.20	3.97	**<0.0001**	0.0100
BackSel Regression	6.37	3.14	**0.0008**	0.0125
Z-Score: CarryAtExpiry	6.09	2.86	**0.0021**	0.0167
Z-Score: BE-Width	5.91	2.69	**0.0036**	0.0250
Ridge Regression	4.86	1.63	0.0517	0.0500
Lasso Regression	**3.23**	-	-	-
Friedman Chi-Square	117.2203175	**<0.0001**		