UNIVERSITY OF VAASA

FACULTY OF BUSINESS STUDIES

DEPARTMENT OF ACCOUNTING AND FINANCE

Amir Mobasheri

DIRECTIONAL FORECASTING OF EURUSD EXCHANGE RATE USING CLASSIFICATION APPROACH

Master's Thesis in
Accounting and Finance
Line of Finance

VAASA 2016

Table of Content

List of Figures

List of Tables

UNIVERSITY OF VAASA
Faculty of Business Studies
Author: Amir Hossein Mobasheri
Topic of the Thesis DIRECTIONAL FORECASTING OF EURUSD EXCHANGE RATE USING CLASSIFICATION APPROACH

Supervisor: Professor Timo Rothovius
Degree: Master of Science in Economics and Business Administration
Department: Department of Accounting and Finance
Major Subject: Accounting and Finance
Line: Finance
Year of Entering the University 2014
Year of Completing the Thesis 2016 **Pages: 90**

ABSTRACT

This thesis study aims to use classification methods in forecasting EURUSD direction of change. A number of classifiers including logistic regression, knn, naïve bayes, and classification tree are used. The input variable universe is comprised of three major categories: currency pairs, interest rates and market indices (stock and commodity indices). All series are from 1.1.2004 to 8.2.2016. Two main types of models are constructed. First are models with fixed predictors that are based on ideas from literature. Second are models which select predictors at each step from a pool of predictors using an input selection algorithm. The input selection algorithms are MIM, MRMR, JMI and DISR originated from information theory field. In estimating the models two types of predictors are used: original form and discretized version. Models are estimated using both recursive and rolling window. Finally, the out-of-sample forecast is formally tested for statistical significance. Among the models built according to the combination of classifier, input selector, predictor and estimation scheme, a few models are found to be marginally significant, indicating the promising outlook of using more sophisticated methods.

Keywords: Directional forecasting, Forex, Exchange rate, Classification

1. Introduction

In this introduction section, at first the motivation for forecasting exchange rate is discussed, then an overview of the previous methods to forecast FX is described. Then, advantages of considering directional prediction is explained. Finally, factors influencing forecast performance and some open questions are mentioned.

A forecast of future movement of any financial asset including forex in itself is valuable information for speculators and international portfolio managers. In addition to speculation, according to (Bekaert and Hodric 2011) currency forecasts are used in the following contexts in companies:

- Quantifying foreign exchange risk.
- Setting prices for their products in foreign markets.
- Valuing foreign projects.
- Developing international operational strategies.

In another study (Rossi 2013) reports a few reasons exchange rates forecasts are useful for Central Banks and policy makers such as using forecast to project the consequences of particular policy measures

Regarding forecasting methods, the literature abounds with studies originating from different fields such as economics, statistics, signal processing and machine learning. Maybe it is not exaggerating to state that for any method of prediction there is a study using that method in forecasting currency movements. For example, (Yu, Wang and Lai 2007) reports around 45 studies that just used different types of Neural Network for predicting currencies. For studies rooted in economics and econometric methods, (Rossi 2013) conducted a literature review regarding papers in the last ten years.

Regardless of different method used for model building, the majority of the previous studies seek to find the expected value of the currency level $E_t(S_{t+h})$ or change $E_t(\Delta S_{t+h})$ in next *h* period, and compare it with some benchmark which is usually random walk or some AR process. There are various measures used to perform this comparison. MSFE (Mean Squared Forecast Error) or its root RMSFE (Root Mean Squared Forecast Error) is almost used in majority of studies (along with other distance based measures). Another measure is called "Directional Accuracy" (i.e. $sign(E_t(S_{t+h})-S_t)=sign(S_{t+h}- S_t)$) which calculates the proportion of forecasts that correctly predict the direction of

change ,and in contrast to RMSFE it is not affected by distance between the forecast and the actual realization.

Two advantages in focusing on "Directional Accuracy" (DA) are discussed here. First, from statistical perspective, models that cannot beat random walk in RMSFE sense show predictive ability when considering the DA measure. For example, (Meese and Rogoff 1983) in their seminal article compared RMSFE and other distance based measures to random walk and find their model performance is not better than random walk. Later in other study (Cheung, Chinn and Pascual 2005) find that evaluating models by the direction-of-change criteria shows more empirical evidence that models can outperform the random walk. Besides, (Christoffersen and Diebold 2006) show that conditional mean dependence is not required for having sign dependence. This finding is especially important for financial markets famous for having weak mean dependence. Second, from an economic point of view, according to (Chung and Hong 2007) the directional predictability is more relevant to many financial applications such as utility-based measures and market timing.

(Rossi 2013) Conducted a comprehensive literature review regarding out-of-sample performance of previous studies mainly using economic models. She concluded predictability of exchange rates depends on the following:

- Choice of predictors.
- Forecast horizon.
- Sample period for evaluation.
- Model specification.
- Forecast evaluation method.

Other findings of the study was instabilities in the models' forecasting performance across models, predictors and time. These finding leads to two questions:

1. What are the reasons predictability change over time?
2. Is it possible to find a way to exploit instabilities to improve forecast?

These Issue of changing predictive performance of a model is also known to researchers in machine learning community and is coined as "concept drift". (Gama, et al. 2014) Conducted a survey regarding methods to deal with this issue.

One important question is in order regarding the forecasting of exchange rates in general. If the efficient market hypothesis is correct, doesn't it make the whole subject of this study pointless? This question is answered using the arguments put forward by (Rossi 2013) .

Rossi (2013:31) states,

> *It is important to note that the efficient market hypothesis does not imply that exchange rate changes should be unpredictable. That is, the Meese and Rogoff (1983) finding that the random walk provides the best prediction of exchange rates should not be interpreted as a validation of the efficient market hypothesis. The efficient market hypothesis means that bilateral exchange rate is the market's best guess of the relative, fundamental value of two currencies based on all available information at that time. The efficient market hypothesis does not mean that exchange rates (like any asset price) are unrelated to economic fundamentals, nor that exchange rates should fluctuate randomly around their past values.*

1.1. Purpose of the study

The purpose of this thesis is to investigate the usefulness of using classification approach in forecasting the direction-of-change in foreign exchange market specifically for EURUSD currency pair. In other words, instead of forecasting the expected value of the future exchange rate $E_t(S_{t+h})$ and use it to study the direction of forecast ($sign(E_t(S_{t+h})-S_t)$), in this study the response variable is constructed as a discrete variable i.e. $Y_t=sign(S_{t+h}- S_t)$ and is directly forecasted.

In line with the above mentioned purpose, several classification models are used. The reason for using several methods is because there is no best classification method. In other words, depending on the dataset, some methods perform better than others.

After generating various classification models, the study is extended in the direction of abovementioned purpose by examining if combining the output of several classifier would enhance the performance or not.

1.2. Structure of the thesis

The thesis is organized as follows. In brief, chapters 2-6 covers the theoretical subjects. Chapter 7 introduce previous studies and chapters 8-9 deals with the empirical part. At length, chapter 2 introduce a few theory regarding exchange rates. Chapter 3 covers a concise explanation of the classification methods used in this study. Chapter 4 extends on classification theory by describing the performance measures and how to test the significance of results. Chapter 5 describe the methods to combine the result of several classifiers. In chapter 6, after a concise theoretical part, the methods used to select variables from a larger pool of variables are introduced. Chapter 7 is a literate review. In chapter 8 the methodology is described and the result is reported in chapter 9.

2. Theories regarding exchange rate determination

In this section, a few theories regarding exchange rates are described. Generally, these theories are not supported by empirical findings. These models are not used directly in this thesis study but being aware of them provides valuable insight .Therefore, it is tried to maintain brevity in the presentation of these models.

2.1. Interest Rate Parity (IRP)

The relationship between exchange rates and interest rates can be stated in two settings with different assumptions. Uncovered Interest Rate Parity (UIRP) and Covered Interest Rate Parity (CIRP). These two are explained using the following notation

S_t: $Nominal\ exchange\ rate\ domestic\ per\ foriegn\ at\ time\ t.$

i_t: $Domestic\ interest\ rate$

i_t^*: $Foreign\ interest\ rate$

F_t: $h\ period\ ahead\ forward\ rate\ at\ time\ t.$

- Uncovered Interest Rate Parity (UIRP): According to UIRP, a domestic investor can invest $\frac{1}{S_t}$ units with the rate of return i_t^* between time *t* and time *t+h* using 1 unit of domestic currency. At the end of the period, the payoff of the investment in foreign currency unit will be $\frac{1}{S_t}(1+i_t^*)$. Now if this amount is converted back to domestic currency, with the exchange rate S_{t+h} , the investor owns $\frac{S_{t+h}}{S_t}(1+i_t^*)$, in absence of arbitrage, the expected value of this amount should be equal $(1+i_t)$ that is

$$\frac{E_t(S_{t+h})}{S_t}(1+i_t^*) = (1+i_t)$$

Can be rearranged to

(1) $$\frac{E_t(S_{t+h})}{S_t} = \frac{(1+i_t)}{(1+i_t^*)}$$

All the values that are known at time t, come out of expectation and only S_{t+h} which is random stays in expectation function.

- Covered Interest Rate Parity (CRIP): CIRP almost follows the same line of reasoning, except that the investor hedges its investment. Suppose the forward rate is F_t i.e. the price of 1 unit foreign currency h-period ahead at time t. Since investor knows at time t+h he/she will own $\frac{1}{S_t}(1+i_t^*)$ amount of foreign currency, he can sell this amount in advance at time t, $\frac{F_t}{S_t}(1+i_t^*)$ is the amount the investor own in domestic currency unit. In absence of arbitrage the following relationship holds

$$\frac{F_t}{S_t}(1+i_t^*) = (1+i_t)$$

Which can be rearranged to

(2) $$\frac{F_t}{S_t} = \frac{(1+i_t)}{(1+i_t^*)}$$

The content of interest rate parity is adapted from (Rossi 2013)

2.2. Purchasing Power Parity (PPP)

In simplest form, PPP states that common currency price of an identical basket of goods becomes equal:

(3) $$s_t + p_t^* = p_t$$

Where s_t, p_t are foreign exchange and price of a basket of goods in log level. And asterisk denote foreign. If q is defined as the domestic units of the domestic basket of goods required to purchase a single basket of foreign goods, then

(4) $$q_t = s_t - p_t + p_t^*$$

The price can be decomposed into several dimensions such as tradable and non-tradable. Tradable category can further divided into importable and exportable. If price is decomposed into tradable and non-tradable, then,

(5) $$P_t = \alpha P_t^N + (1-\alpha)P_t^T$$

Where N and T denotes non-tradable and tradable, respectively. Assuming the weight is the same, the expression for real exchange rate can be obtained as

(6) $$q_t = q_t^T + \alpha(\hat{p}_t^N - \hat{p}_t^T)$$

Where $q_t^T = s_t - p_t^T + p_t^{T*}$ and hat denotes price differential.

Consequently, the real exchange rate deviates from zero if either tradable prices differ, or the relative price of non-tradable versus tradable differs across countries.

Relative prices can be determined by demand side factors and/or supply side factors. For example, in the long run, rising preference for services, which are non-tradable, a may lead to a rise in relative price of non-tradables.

The content of this section is adapted form (James, Marsh and Sarno 2012).

2.3. Balassa-Samuelson Model

As mentioned in earlier in PPP section, by decomposing price into tradable and non-tradabele the real exchange rate can be written as

$$q_t = q_t^T + \alpha(\hat{p}_t^N - \hat{p}_t^T)$$

Where $q_t^T = s_t - p_t^T + p_t^{T*}$ and hat denotes price differential. Assuming perfect capital mobility, free inter-sectoral factor mobility and identical production functions, the following relation can be written

(7) $$p_t^N - p_t^T = a_t^T - a_t^N$$

Where a_t^T, a_t^N are TFP (Total Factor Productivity) levels in the traded and nontrade sectors, respectively. This relationship states that the relative price of traded goods moves one-for-one with the productivity differential. By combining this relationship with the expression for real exchange rate we obtain

(8) $$q_t = q_t^T + \alpha(\hat{a}_t^N - \hat{a}_t^T)$$

Hat denotes relative productivity. This relation shows relative exchange rate as a function of inter-country relative productivity differential. One implication of this model is that if PPP holds for tradable then $q_t^T = 0$, and the real exchange rate depends only on the productivity differential.

The content of this section is adapted from (James, Marsh and Sarno 2012).

2.4. Monetary model with flexible prices

In this model, it is assumed that PPP holds continuously. Money-demand functions in the two countries are expressed as

(9) $$m_t^d - p_t = \phi y_t - \lambda i_t$$

Where m is the log nominal money stock. y is log income, I is the short-term interest rate, and the d superscripts indicate “demand”. For simplicity, it is assumed money demand parameters are the same across two countries. Rearranging and assuming money supply equals money demand, and imposing PPP results in

$$s_t = (m_t - m_t^*) - \phi(y_t - y^*) + \lambda(i_t - i_t^*) \quad (10)$$

This model is sometimes termed monetarist.

The model predicts that increasing interest rate differential causes the domestic currency to depreciate. This result makes sense when PPP holds both in short run and long run because according to Fisher relation, positive interest rate differentials arise from inflation differentials. Consequently, the faster a currency loses value against a basket of goods, the faster it loses value against another currency. The reason is that domestic and foreign prices are linked through PPP.

The material of this section was adapted from (James, Marsh and Sarno 2012).

2.5. Portfolio balance model

Traditional portfolio balance models include a measure of stock balance between domestic assets and foreign assets held by domestic investors. The model is

$$s_t = \beta_0 + \beta_1(i_t - i_t^* - E_t(s_{t+1} - s_t)) + b_t - b_t^* \quad (11)$$

Where $b_t$the stock of domestic assets is held by domestic investors and b_t^* is the stock of foreign assets held by domestic investors. Since $E_t(s_{t+1} - s_t)$ is unobservable, it is approximated by zero. Various measures have been used in literature as proxy for balance such as cumulated trade balance differentials, cumulated current account balance differentials, and government debt.

This equation implies if the amount of foreign assets held by domestic investors rise, the exchange rates will fall i.e. domestic currency appreciate.

The content of this section is adapted from (Rossi 2013). A more detail presentation can be found in (James, Marsh and Sarno 2012).

2.6. Supply and demand analysis

(Mishkin, Matthews and Giuliodori 2013) Explain the reason for exchange rate movement in a supply and demand analysis framework. To simplify the analysis it is assumed that the amount of domestic assets is fixed which means the supply curve is vertical at a given quantity and does not shift. Therefore, under this assumption, it is enough to consider only the factors that shift the demand curve for domestic assets. Figure **1** depicts this setting and shows the result of a hypothetical shift in demand.

- ***Important Note**: Unlike previous section, in this section exchange rate is the amount of foreign currency per unit of domestic currency, i.e. it shows the price of one unit domestic currency in foreign currency unit. Therefore, if exchange rates **rise**, it means domestic currency **appreciates**.*

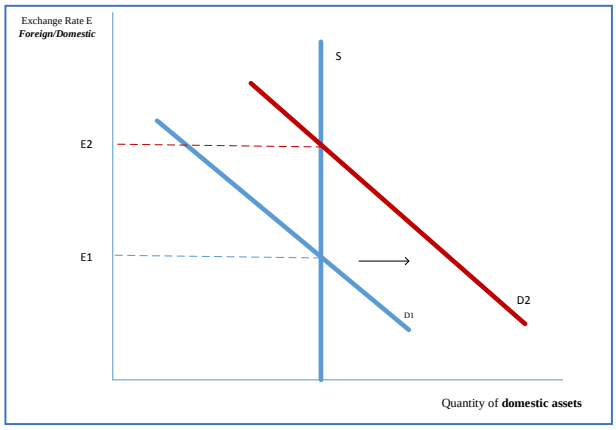
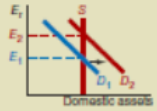
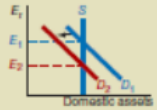
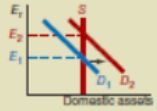
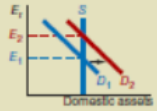
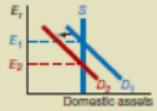
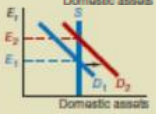
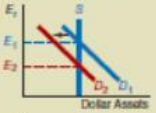
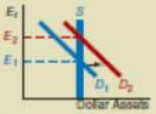


Figure 1: Shift in demand.

Figure **2** summarizes the supply demand analysis. In the figure the symbol † means relative to other countries. For example, if relative risk of domestic asset to foreign asset increases ceteris paribus, then the demand curve for domestic asset shift to the left, implying reduction in demand for any given level of exchange rates. As a result, the equilibrium exchange rates will fall or equivalently, the domestic currency depreciate.

Factors that shift the demand curve for domestic assets and affect the exchange rate

Factor	Change in factor	Change in quantity demanded of domestic assets at each exchange rate	Response of exchange rate*, E	
Foreign wealth	↑	↑	↑	
Risk†	↑	↓	↓	
Liquidity†	↑	↑	↑	
Domestic interest rate, i^D	↑	↑	↑	
Foreign interest rate, i^F	↑	↓	↓	
Expected domestic price level†	↑	↓	↓	
Expected trade barriers†	↑	↑	↑	
Expected import demand	↑	↓	↓	
Expected export demand	↑	↑	↑	
Expected productivity†	↑	↑	↑	

*Units of foreign currency per unit of domestic currency: ↑ indicates domestic currency appreciation; ↓, depreciation.

†Relative to other countries.

Note: Only increases (↑) in the factors are shown; the effects of decreases in the variables on the exchange rate are the opposite of those indicated in the 'Response' column.

Figure 2: Summary of supply demand analysis.

3. Brief background of classifiers

The materials that will follow in this section are mainly based on (Hastie and Tibshirani 2013) and (Kuncheva 2014) .

Classification methods are used when the response variable is qualitative or categorical. Figure **3** displays different elements in a classifier. A combination of a number of classifier is called an ensemble.

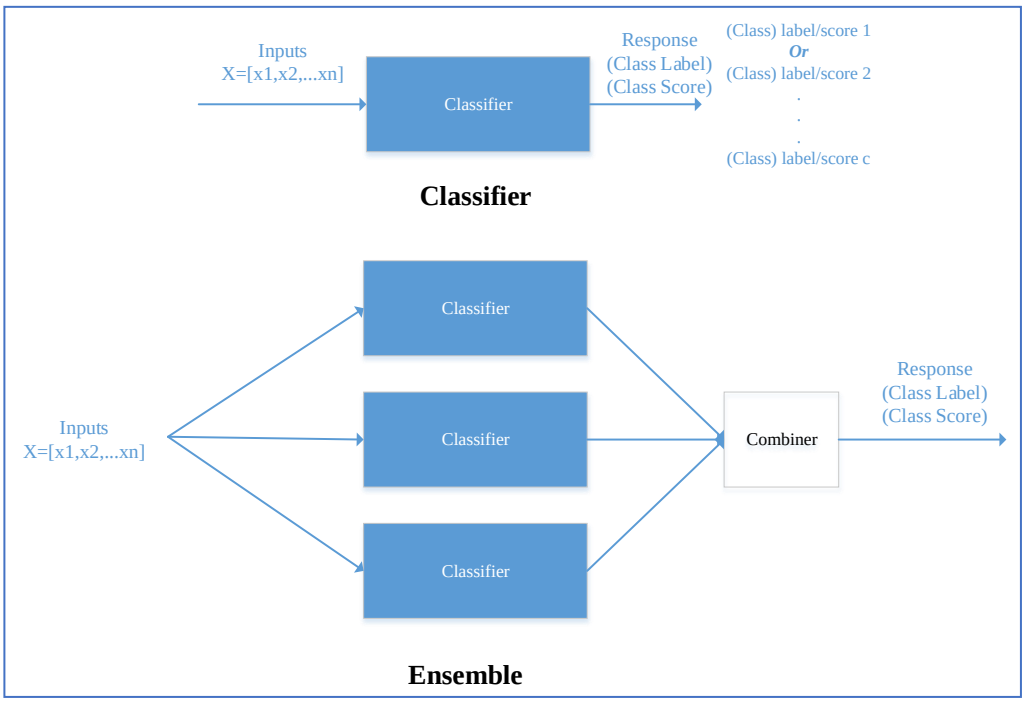


Figure 3: Elements of classifier and ensemble.

Since classifiers are used in different fields, various names are used for the same concept. Table **1** is adapted from (Kuncheva 2014) showing various terminology used interchangeably. The left column is the usage in this study.

Table 1: Various terminology used interchangeably.

This study	Synonyms in other discipline.
Feature, Input	Attribute.
Classifier	Hypothesis, Learner, Inducer, Generalizer, Expert.
Object, Observation	Example, Instance, Case, Data point.
Ensemble	Team, Pool, Committee, Meta learner.

In the above representation, a classifier can be viewed as a function that assigns a class label to an object or observation. In fact, if there are c class labels, any set of *c* functions can be used to construct a classifier. These functions are called *discriminant functions* each will yield a result called *score* and the object (observation) is assigned to a class with highest score (ties are assigned randomly). Figure 4 depicts the classifier as a set of discrimination functions.

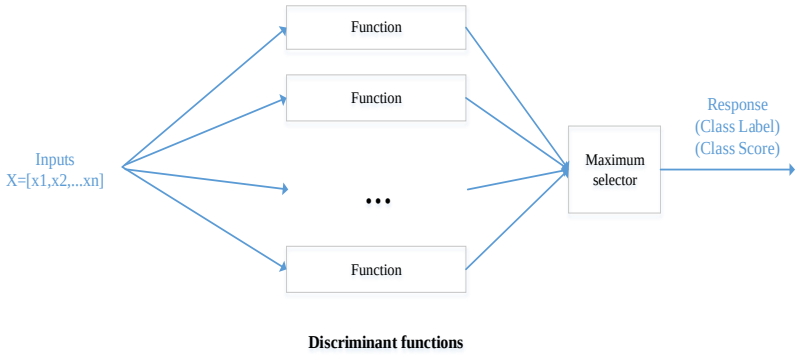


Figure 4: Classifier as a set of discrimination functions.

3.1. Classifiers

There are many classifiers that can be used to predict the class label. In this study the following classifiers are considered: naïve classifier, logistic regression, knn (k-nearest neighbor), naïve bayes, classification tree and linear regression.

- Naïve classifier

Perhaps this is the simplest classifier that can be used. It is the prior or unconditional probability of a class label. To estimate it using sample the following formula can be used:

$$\hat{P}(Y = i) = \frac{N_i}{N}\ , \qquad i = 1,2,..c \tag{12}$$

Where N_i is the number of observation is labeled as i, N is the total sample size and c is the number of classes. The classifier will assign label to class label with the highest probability. For a two-class problem:

$$\pi_1 = P(Y = 1) = \frac{N_1}{N}\ ;\ \ and\ \ \pi_0 = P(Y = 0) = 1 - \pi_1$$

- Logistic regression

Logistic regression is introduced in several ways in major text books. One simple way explained in (Hastie and Tibshirani 2013), starts by pointing out if linear regression is used to estimate the binary response variable it is possible to produce probabilities outside the range of [0,1]. Consequently, the probabilities are modeled using a function that produces values between 0 and 1. One such function is the *logistic function*,

(13) $$p(X) = P(Y = 1|X) = \frac{\exp(\beta_0 + X\beta)}{1 + \exp(\beta_0 + X\beta)}$$

To estimate the parameters (coefficients), the method of maximum likelihood is used. Figure **5** displays three logistic functions with different values for coefficient. It can be seen this function can get various shapes.

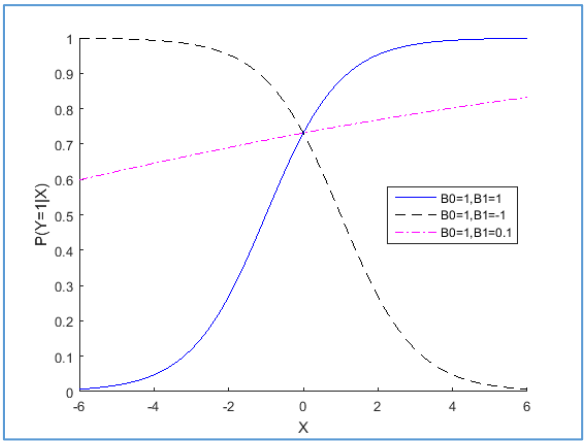


Figure 5: Logistic function.

Another approach to introduce logistic regression as explained by (Greene 2008) is as follows. Let y^* be an *unobserved* variable such that

(14) $$y^* = x'\beta + \epsilon \; ; x' \; is \; a \; row \; vector$$

The variable y is the *observed* variable such that

$$\begin{cases} y = 1 & \text{if} y^* \geq 0 \\ y = 0 & \text{if} y^* < 0 \end{cases}$$

Now,

$$P(y = 1|x) = p(y^* \geq 0|x) = p(x'\beta + \epsilon \geq 0|x) = p(\epsilon \geq -x'\beta|x)$$

If the distribution of ϵ is symmetric like logistic or normal distribution then,

$$p(\epsilon \geq -x'\beta|x) = p(\epsilon < x'\beta|x) = F(x'\beta)$$

Where F is the cumulative distribution of ϵ . In case the logistic distribution is chosen for ϵ then it leads to the *logistic* regression. The choice of normal distribution will lead to the *probit* regression.

Finally, another way to introduce logistic regression is in the context of "*Generalized Linear Model*" which we do not deal with it. In brief, the representation of this approach incorporate several regression model such as linear, logistic, probit and several other models as special case.

- K-nearest neighbors or Knn

The Knn classifier uses the idea that, objects in the same class are more similar to each other than the objects from another class. As explained in (Hastie and Tibshirani 2013) , to classify an observation with features x_0 , Knn first finds *K* nearest neighbor of this observation, then assign the class label with highest number to the new observation. Figure **6** depicts this mechanism more clearly. In the figure, there are two classes, red and blue. In order to classify the new observation using Knn with *K=3* neighbors, at first, three nearest neighbors of the observation are identified. Then, the number of each class member is calculated to find the class which has the highest number of members in the neighborhood (here, 2 blue and 1 red) .Finally the class label with greatest number is assigned to the new observation which here is " blue".

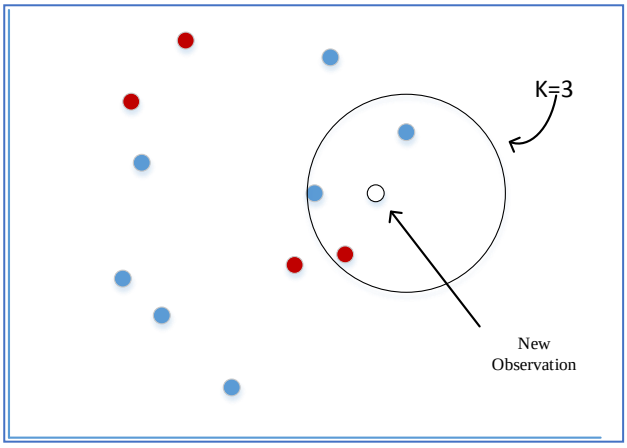


Figure 6: K-nearest neighbor classifier.

In fact, the abovementioned procedure is equivalent to stating the class is assigned based on maximum estimated probability of each class in the neighborhood of the new observation. In general, if N_0 is the set containing the K neighbors of the observation with features x_0 for a problem with c class label we have,

$$P(Y = j|X = x_0) = \frac{1}{K} \sum_{i \in N_0} I(y_i = j) \;\; ; j = 1 \dots c \tag{15}$$

and the predicted class will be the class with highest probability.

- Naïve Bayes

We follow (Kuncheva 2014) and adapt its explanation regarding Naïve Bayes classifier. This classifier makes use of the Bayes formula with the assumption

that observation given the class label are independent of each other. In fact, this is the reason that this classifier is called "Naïve". Let $x = [x_1, x_2, \dots, x_n]^T$ be the feature vector and $\omega_1, \dots \omega_c$ are c class labels. Minimum classification error is guaranteed if the class with the largest posterior probability, $P(\omega_i|x)$ is chosen. According to Bayes formula we have,

$$P(\omega_i|x) = \frac{P(\omega_i)f(x|\omega_i)}{\sum_j P(\omega_j)f(x|\omega_j)}, \qquad i = 1, \dots, c \tag{16}$$

Here $f(x|\omega_i)$ is the conditional probability distribution of *x* given the class label is ω_i, and $P(\omega_i)$ is the prior probability of ω_i.Using the conditional independence assumption i.e. $[x_1, x_2, \dots, x_n]$ are independent given the class label, the joint distribution of x given the class label can be written as:

$$f(x|\omega_i) = f(x_1|\omega_i)f(x_2|\omega_i) \dots f(x_n|\omega_i), \quad i = 1, \dots, c.$$

Now the conditional distribution for each feature given the class label can be estimated separately using some method such as kernel density, and plugged in the Bayes formula to calculate the posterior probability. There are various ways to estimate this probability distribution. One parametric method is to assume normal distribution for each feature given the class and fit a normal distribution. To use a nonparametric approach, each feature can be first discretized, and for each feature the distribution estimated using the discrete values. (Kuncheva 2014).

One surprising characteristic of Naïve Bayes classifier is that, it works well although the independence assumption is clearly wrong in most practical situation (Kuncheva 2014).

- Classification tree

Classification tree classifies observation by partitioning the feature space into regions and assign the same label to all observations in the same region. Figure **7** presents a particular partition of the feature space with two feature x_1 and x_2 . Figure **8** displays this region in a tree-like fashion.

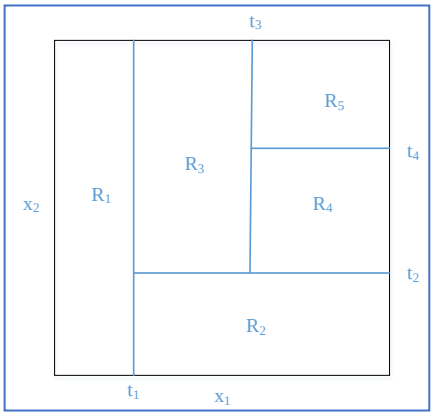


Figure 7: Portioned region.

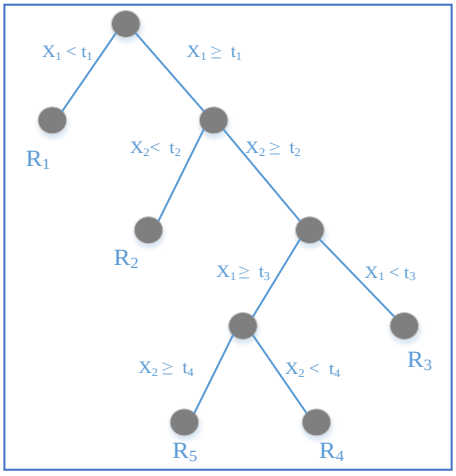


Figure 8: Tree classifier.

To classify a new observation the following rules are used:

$$if\ x_1 < t_1\ then\ R_1$$
$$elseif(\ x_1 \geq t_1\ and\ x_2 < t2\ )\ then\ R_2$$
$$elseif\ (x_1 \geq t_1\ and\ x_2 \geq t_2 and\ x1 \geq t_3\ and\ x_2 \geq t_4)\ then\ R_5$$
and so on

After finding the region in which the new observation belongs, the label with the maximum probability in that region is assigned to the new observation.

To grow a tree, the algorithm calculates the result of splitting each feature based on some criteria. Then, the feature with the best result is selected to be included in the tree. Therefore tree growing algorithm is a greedy algorithm. To perform the splitting there are several criteria that can be used such as Gini Index and cross-entropy.

(17) $$Gini\ Index = \sum_{k=1}^{c} \hat{p}_{jk}\left(1 - \hat{p}_{jk}\right)$$

(18) $$cross - entropy = -\sum_{k=1}^{c} \hat{p}_{jk}\, log\left(\hat{p}_{jk}\right)$$

Here $\hat{p}_{jk}$ is the proportion of observation in region *j* from *kth* class and *0log0=1*. (Hastie and Tibshirani 2013) mention that for evaluating the quality of a split, Gini Index or cross-entropy are better than classification error rate because they are more sensitive to node purity. Figure **9** shows the behavior of these two measures for a two-class case as the probability varies. Both index reach their peak i.e. highest impurity when *p=0.5*.

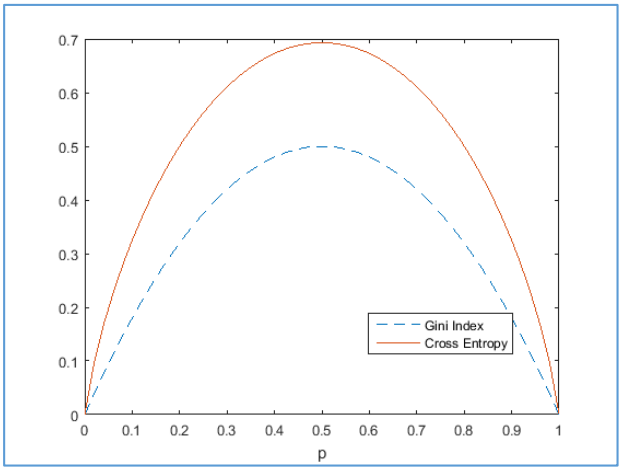


Figure 9: Behavior of Gini index and cross entropy.

- BKS classifier

This classifier uses discrete data and is a multinomial classifier which estimates the posterior probability of a class given data for all combination of the features. BKS stands for "Behavior Knowledge Space". BKS is implemented by forming a lookup table which contains all combination. Then, for a new observation, it searches the table to find the record that matches the new observation. Finally, it assigns the label with highest probability. In case the look up table does not contain the observation, label can be assigned according to the prior probability of the class. To illustrate, suppose *X1* and *X2* are features with 2 and 3 discrete level values. Table **2** shows a possible combination of these features and their label probabilities. The last column is the label assigned to a new observation. For instance, an observation with *X1=1*, *X2=2*, will be labeled *0*.

Table 2: BKS lookup table.

X1	X2	P(Y=1)	P(Y=0)	Assigned label
1	1	0.6	0.4	1
1	2	0.3	0.7	0
1	3	0.4	0.6	0
2	1	0.2	0.8	0
2	2	0.8	0.2	1
2	3	0.3	0.7	0

Two drawbacks of BKS are that, first, data set should be large to get reliable result. Second, it can get highly overtrained. On the other hand, the BKS is optimal for any dependencies.

Since this classifier uses discrete values, it can be considered for combining label outputs of different classifiers (Kuncheva 2014).

3.2. Estimation

When the purpose of estimating the model is to use it in practice especially for forecasting, two common schemes are (Rossi 2013):

- Recursive
- Rolling

Recursive and rolling scheme are used when data become available in a stream form such as time series data. Whenever a new data becomes available a new model based on the recent data is estimated and used to predict the new observation. Recursive estimation uses data from the first available observation up to the most recent one.

Rolling estimation uses last *R* data point to estimate the model. Figure **10** shows this two scheme schematically.

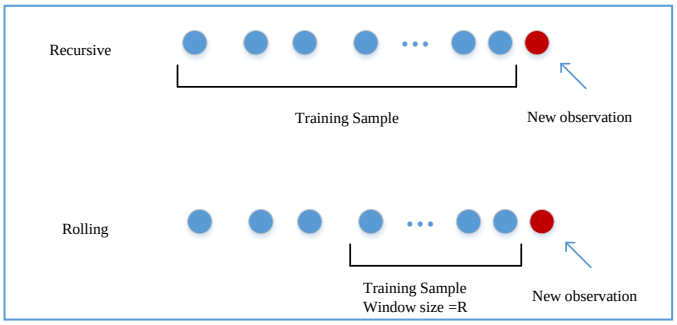


Figure 10: Recursive and rolling estimation scheme.

On the other hand, when the purpose of estimating is to evaluate the model which does not use data in a streaming form, there are more alternatives in addition to the two previously mentioned. These alternatives among others include various cross-validation methods, leave one out, and hold-out. Hold-out is simply putting aside a subsample called holdout sample, and estimate the model without using it, and finally evaluate the model on the holdout sample. In time series context, the holdout sample is from the recent time, but in other context it is usually chosen randomly.

3.3. Two Examples

In this section two examples are presented. First one uses an artificial data set in (Kuncheva 2014) called “Fish data” .This data is generated according to the following formula

$$label(x, y) = XOR(z_1, z_2)$$
$$z_1 = I(x^3 - 2xy + 1.6y^2 < 0.4) \quad (19)$$
$$z_2 = I(-x^3 + 2ysin(x) + y < 0.7)$$

XOR is the exclusive or logical operator, which takes true if only one of its input is true (for both true it yields false) and I(x) is an indicator function yielding 1 if its input is true and 0 otherwise. Figure **11** displays this data where black dots represent true (class label 1) .Figure **12** shows another version of this data where 10% of the labels are flipped randomly to create an imperfect data.

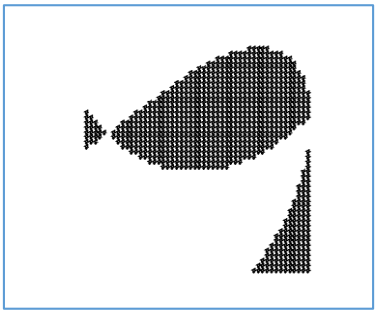


Figure 11: Fish data without noise.

Figure 12: Fish data with 10% noise.

Second data set used to demonstrate the performance of classifiers is taken from (Hastie and Tibshirani 2013) and called "Smarket". It consists of percentage return for the S&P 500 stock index from beginning of 2001 until the end of 2005 with the total of 1250

days. In this section we only use the first two lags in data to perform analysis. Figure **13** shows this data in the first two lags dimensions.

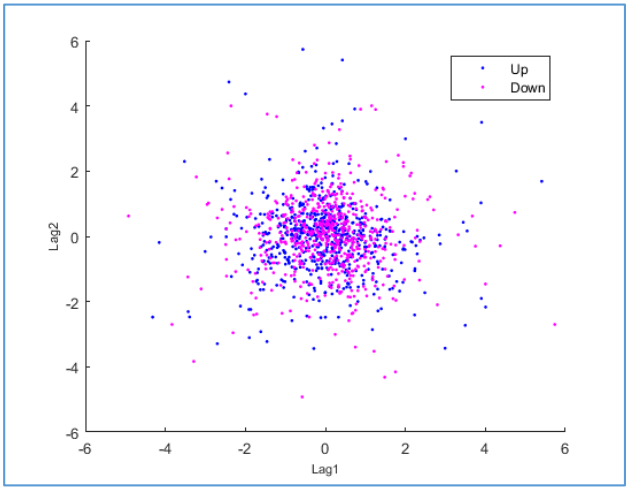


Figure 13: Smarket data.

Table **3** shows the parameters and method used to build the model where there needs to set a parameter such as number of neighbors for Knn or there are several alternative such as various methods to estimate the density for naïve base. Class labels are assigned according to the maximum discriminant value for each class.

Table 3: Parameter setting of classifiers.

Classifier	Mnemonic	Paremeter Setting
Logistic Regression	'Lgr'	---
K-nearest neighbor	'Knn'	K=3
Naïve Bayes	'Nb'	Kernel density
Tree Classifier	'Tre'	Tree is first grown with at least 20 observation in final node, then pruned back
Multinomial Classifier	'Mn'	---
Linear Regression	'Lr'	---
Naïve Classifier	'Nv'	---

For each data set the accuracy i.e. the percent of correct classification of the classifier is reported in table **4**.

Table 4: Accuracy of classifiers.

Classifier	Mnemonic	Fish data	Smarket
Logistic Regression	'Lgr'	63.52	55.95
K-nearest neighbor	'Knn'	86.40	53.17
Naïve Bayes	'Nb'	76.64	54.37
Tree Classifier	'Tre'	84.48	55.95
Multinomial Classifier	'Mn'	78.56	52.38
Linear Regression	'Lr'	63.36	57.54
Naïve Classifier	'Nv'	62.08	55.95

In figure **14**, the region for "Fish Data" that each classifier assign to positive label (class label is 1) is shown in green color. The black color represents positive class and the white color within the dashed box is negative class (class label is 0). Considering the accuracy and the regions for each classifier, it can be seen that for this dataset the nonlinear methods such as Knn and Tree classifier outperform the linear models such as logistic regression. This result is expected due to the highly nonlinear nature of the data.

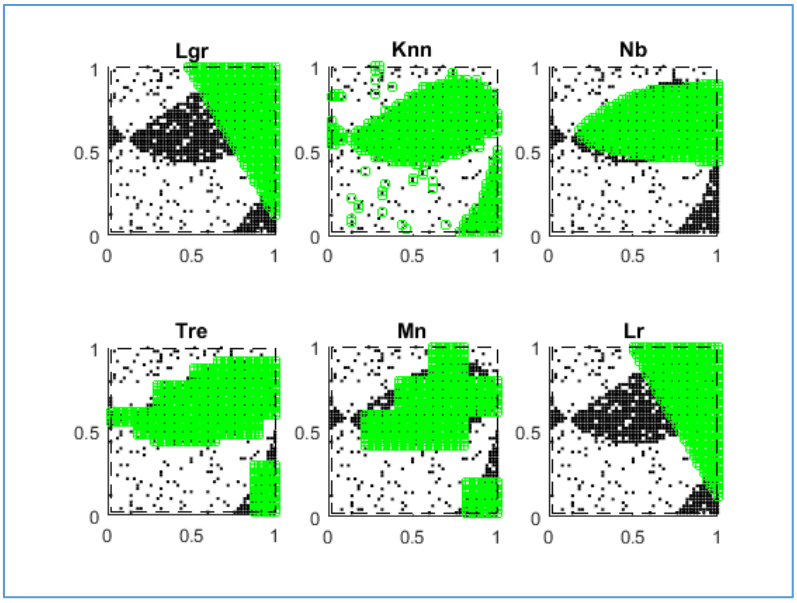


Figure 14: Region for Fish dataset.

In figure **15**, the region for “Smarket” that each classifier assign to positive label (class label is 1) is shown in green color. The white color within the dashed box is negative class (class label is 0). The actual data point is also added (blue up, magenta down). Considering the accuracy and the regions for each classifier, it seems none of the classifiers can outperform the “Naïve Classifier”. This is highly notable in case of “Tree Classifier” because it does not find any significant variable to add in to the model, and in this data set it just performs like a naïve classifier which assign labels based on prior probabilities.

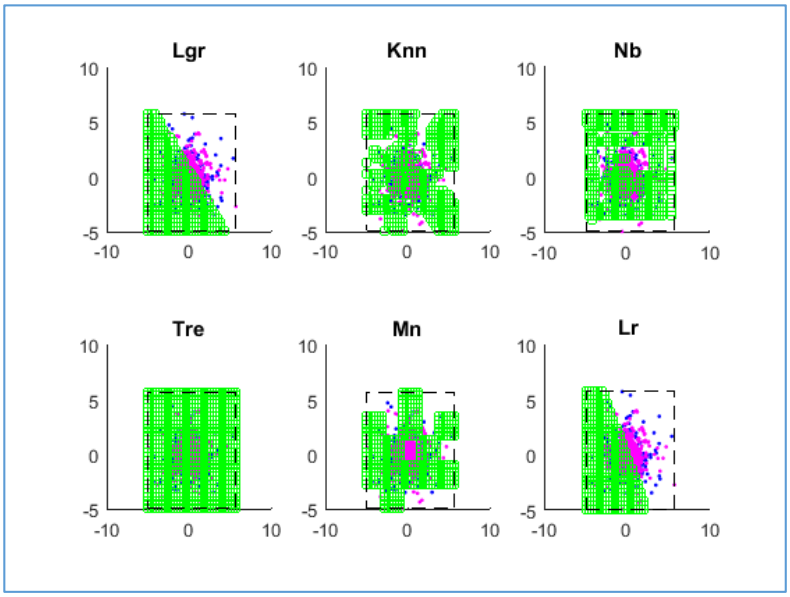


Figure 15: Region for Smarket dataset.

4. Classifier performance

In this chapter, various performance measures are introduced and statistical test to perform formal tests are presented. To organize the material we make use of contingency table and discuss in that framework.

Contingency table is way to show the joint frequency of two discrete variables in a matrix format. Each cell represents the count of observations. Table **5** shows a contingency table for analyzing the distribution of realized (Y) and predicted value ($\hat{Y}$).

Table 5: Contingency table structure.

	Y=1	*Y=0*	*Total*
$\hat{Y}=1$	n_{11}	n_{10}	$n_{1.}$
$\hat{Y}=0$	n_{01}	n_{00}	$n_{0.}$
Total	$n_{.1}$	$n_{.0}$	N

Here n_{ij} is the number of observation which are predicted to be i and the realized value j, *i,j= 0 ,1* . The notation "." stands for total of row or column. For instance, $n_{.1}$ is the total number of observation with *Y=1*.

If the table represents the population, we have the following:

$$
\begin{aligned}
P(Y=1) &= \frac{n_{.1}}{N}\,, univariate\ distribution \\
P\left(Y=1\middle|\hat{Y}=1\right) &= \frac{n_{11}}{n_{.1}}\,, conditional\ distribution \\
P\left(Y=1,\hat{Y}=1\right) &= \frac{n_{11}}{N}\ joint\ distirbution
\end{aligned}
\tag{20}
$$

If the table represents a sample from population, then all the above mentioned quantity are estimate of the population counterparts.

4.1. Performance measure

There are various performance measures, but in the following, the ones that are used in this study are introduced. The probability formula are used to make the link between the introduced measure and contingency table clearer. The presented descriptions are adapted from (Powers 2011) .

- Accuracy: The proportion of correctly classified observation.

(21) $$ACC = P_{correct} = P(Y = 1, \hat{Y} = 1) + P(Y = 0, \hat{Y} = 0)$$

- Error rate: The proportion of observations that are misclassified.

(22) $$Error\ rate = P_e = 1 - P_{correct}$$

- TPR: stands for True Positive Rate and shows the proportion of positive (Y=1) instances correctly recognized. Other names for *TPR* are "*Hit rate"* and "*Sensitivity*".

(23) $$TPR = P(\hat{Y} = 1 \,|Y = 1) = \frac{P(\hat{Y} = 1, Y = 1)}{P(Y = 1)}$$

- FPR: stands for “False Positive Rate” and shows the proportion of negative (Y=0) instances that are misclassified as positive. Other names for *FPR* are “False alarm rate” and “1-Specifity”; Specifity is the true negative rate.

(24) $$FPR = P(\hat{Y} = 1 | Y = 0) = \frac{P(\hat{Y} = 1, Y = 0)}{P(Y = 0)}$$

- PPV: stands for “Positive Predictive Value” is the proportion of the positive signals ($\hat{Y} = 1$) that are correct.

(25) $$PPV = P(Y = 1 | \hat{Y} = 1) = \frac{P(Y = 1, \hat{Y} = 1)}{P(\hat{Y} = 1)}$$

- NPV: stands for “Negative Predictive Value” and is the proportion of correct negative signals.

(26) $$NPV = P(Y = 0 | \hat{Y} = 0) = \frac{P(Y = 0, \hat{Y} = 0)}{P(\hat{Y} = 0)}$$

4.2. Classifier predictive power

It is important to note the measures introduced do not fully shown if a classifier has predictive power or not. A classic example is the case of an imbalanced population and a naïve classifier that only predicts the class with higher prior probability. Suppose the proportion of observation with positive label is 90% i.e. *P(Y=1) =0.9*. In this case, a naïve classifier that always predict 1 will have an accuracy of 90% which misleadingly

seems impressive. Therefore it is important to examine the dependency of predicted values and realized values.

Among several alternatives to formally test the predictive power of a classifier, in this study, the test proposed by (Pesaran and Timmermann, A simple nonparametric test of predictive performance 1992) is used. One feature of this test is that its statistic can be viewed as a market timing measure as shown by (Granger and Pesaran 2000).In the literature this test is usually referred to as PT, which is also used in this study. The original PT-statistic as introduced in (Pesaran and Timmermann, A simple nonparametric test of predictive performance 1992) is:

(27)
$$PT = \frac{\hat{P} - \hat{P}_*}{\left(v\hat{a}r(\hat{P}) - v\hat{a}r(\hat{P}_*)\right)^{\frac{1}{2}}} \; asym\; N(0,1)$$

Where $\hat{P}$ is the proportion of correct forecast. And $\hat{P}_*$ is calculated as:

$$\hat{P}_* = \hat{P}_Y\hat{P}_{\hat{Y}} + \left(1 - \hat{P}_Y\right)\left(1 - \hat{P}_{\hat{Y}}\right)$$
$$\hat{P}_Y: proportion\ of\ positive\ observation$$
$$\hat{P}_{\hat{Y}}: Proportion\ of\ positive\ signal$$

The sample estimate of variance terms are:

$$v\hat{a}r(\hat{P}) = n^{-1}\hat{P}_*(1 - \hat{P}_*)$$
$$v\hat{a}r(\hat{P}_*) = n^{-1}(2\hat{P}_Y - 1)^2\hat{P}_{\hat{Y}}(1 - \hat{P}_{\hat{Y}}) + n^{-1}(2\hat{P}_{\hat{Y}} - 1)^2\hat{P}_Y(1 - \hat{P}_Y) + 4n^{-2}\hat{P}_Y\hat{P}_{\hat{Y}} + (1 - \hat{P}_Y)(1 - \hat{P}_{\hat{Y}})$$

In this study we use this original form knowing that it is also a measure of market timing, therefore higher PT statistic is preferable which makes it a one sided test. The null hypothesis is as follows:

$$\begin{cases} H_0: forecast\ and\ relized\ value\ are\ independet \\ H_a: Predictive\ power\ exist \end{cases}$$

4.3. Comparing two classifiers

To compare the performance of two classifier on with the same test observation, (Kuncheva 2014) and (Gama, et al. 2014) suggest using McNemar test. To conduct the test, first output of the two classifiers say C1 and C2 are arranged in 2 by 2 table as in table **6**.

Table 6: Contingency table of two classifier performance.

	C2 correct(1)	*C2 wrong(0)*	*Total*
C1 correct(1)	n_{11}	n_{10}	$n_{1.}$
C1 wrong(0)	n_{01}	n_{00}	$n_{0.}$
Total	$n_{.1}$	$n_{.0}$	N

The null hypothesis of the test is:

$$H_0: No\ difference\ between\ accuracy\ of\ the\ two\ classifiers$$

There are three ways to conduct the test:

- Asymptotic

  The test statistic for the two-sided test is:

(28)
$$S = \frac{(n_{10} - n_{01})^2}{n_{10} + n_{01}}$$

If $= 1 - F_{\chi^2}(S, 1) < \alpha$, where $F_{\chi^2}(S, 1)$ is the C.D.F of χ^2 then reject H_0.

The asymptotic Mcnemar has acceptable statistical power, though it does not guarantee nominal coverage i.e. falsely rejecting H_0 can exceed α.Besides, since it is an asymptotic test, (Agresti 2002) suggest $n_{12} + n_{21}$ should be greater than 10.

- Exact conditional

The two-sided test statistic (McNemar 1947) and (Mosteller 1952) is:

(29) $$S = \min(n_{10}, n_{01})$$

If $Pvalue = 1 - F_{Bin}(S, n_{10} + n_{01}, 0.5) < \alpha/2$ then reject H_0. $F_{Bin}(S, n_{10} + n_{01}, 0.5)$ is the C.D.F of the binomial distribution.

Based on the simulation studies in (Fagerland, Lydersen and Laake 2013) the "exact conditional test" always achieve nominal coverage, but it lacks the statistical power relative to other variant.

- Mid-p-value test

The test statistic is the same as "exact conditional test" ,but the rejection is as follows :

$$Pvalue = 1 - F_{Bin}(S - 1, n_{10} + n_{01} - 1,0.5) + 0.5 * f_{Bin}(S, n_{10} + n_{01}, 0.5) < \alpha/2$$

Where $F_{Bin}\ and f_{Bin}$ are C.D.F and P.D.F of binomial distribution, respectively.

Simulation studies in (Fagerland, Lydersen and Laake 2013) suggest attains nominal coverage and has good statistical power.

4.4. Two Examples

In this section, the examples introduced in previous chapter are expanded in two ways. First, the performance measures introduced in this chapter are calculated. Then, each classifier is compared with a) naïve classifier, and b) linear regression model.

Table **7** displays the result for "Fish data". The last two columns are MaNemar test of no difference in accuracy of the classifier versus "Naïve Classifier" and Linear regression respectively. It is clear that nonlinear models outperform both benchmarks. Regarding logistic regression, although its accuracy is greater than both benchmark, it is not statistically significant. PT-statistic is shows all models are significant, i.e. the forecast value is not independent of actual value or in other words, it helps predicting it. This may sound conflicting with the result of McNemar test for logistic regression, but it is important to note that McNemar test only compares the accuracies.

Table 7: Performance measures and statistical test for Fish dataset.

Classifier	ACC	TPR	FPR	PPV	NPV	PT-Pval	Pvalue McNemar1	Pvalue McNemar2
Logistic Regression	63.52	38.82	21.39	52.57	67.78	0.0000	0.4976	0.6250
K-nearest neigbor	86.40	82.70	11.34	81.67	89.35	0.0000	0.0000	0.0000
Naïve Bayes	76.64	59.92	13.14	73.58	78.01	0.0000	0.0000	0.0000
Tree Classifier	84.48	81.01	13.40	78.69	88.19	0.0000	0.0000	0.0000
Multinomial Classifier	78.56	64.56	12.89	75.37	80.09	0.0000	0.0000	0.0000
Linear Regression	63.36	37.97	21.13	52.33	67.55	0.0000	0.5432	1.0000
Naïve Classifier	62.08	0.00	0.00		62.08	0.5000	1.0000	0.5432

Table **8** reports the result for "Smarket". The last two column indicates that none of the models can outperform the benchmarks. On the other hand, considering the PPV values, it can be seen linear models seems to perform relatively better in predicting the up movement of the market. Therefore, as mentioned in (Hastie and Tibshirani 2013), one trading strategy can be only using the up-movement signals of the model. The PT-Pvalue indicates none of these models have predictive powers except the linear regression model which is significant at any conventional level.

Table 8: Performance measures and statistical test for Smarket.

Classifier	ACC	TPR	FPR	PPV	NPV	PT-Pval	Pvalue McNemar1	Pvalue McNemar2
Logistic Regression	55.95	75.18	68.47	58.24	50.00	0.1185	1	0.5966
K-nearest neighbor	53.17	60.99	56.76	57.72	46.60	0.2481	0.4926	0.3028
Naïve Bayes	54.37	68.79	63.96	57.74	47.62	0.2092	0.6646	0.3135
Tree Classifier	55.95	100.00	100.00	55.95	---	0.5000	1	0.7227
Multinomial Classifier	52.38	53.19	48.65	58.14	46.34	0.2365	0.4190	0.0924
Linear Regression	57.54	56.74	41.44	63.49	51.59	0.0079	0.7227	1
Naïve Classifier	55.95	100.00	100.00	55.95	---	0.5000	1	0.7227

5. Combining Classifiers

(Kuncheva 2014) Defines 4 types of classifiers output:

- Class labels: Each classifier D_i produces a class label output $s_i \in \Omega, i = 1, \dots, L$. Therefore, for any observation there is a vector $s = [s_1, .. s_L]^T \in \Omega^L$.
- Ranked class labels: The outputs are ranked or ordered subset of Ω. This type is suitable for problems with a large number of classes.
- Numerical support for the classes: For a problem with c classes each classifier produce a c-dimensional vector where each value shows the degree of belief that the classifier associate to that class label.
- Oracle: This is an artificial output and cannot be defined for unlabeled data. It is the output of each classifier for a given observation which is either correct (1) or *wrong (0).*

In this study, we only consider the methods that combine the "Class label" outputs.

There are mainly two ways to combine class labels generated by each classifier:

- Untrainable combiner: In this method the outputs are combined based on some predefined method such as simple majority vote where the class label produced by most of the classifiers will be the result of the combiner.
- Trainable: This requires that the combiner is trained on some dataset preferably different from the one used to train the base classifiers. For example using the Naïve Bayes classifier as a combiner.

The implicit assumption of the above mentioned description is that the base classifier that are going to be combined have already been selected based on some criteria such as accuracy. But in practice, these two steps maybe performed in any order. For example, when using genetic algorithm to both select classifiers and maximize accuracy of the combined result, the two steps of selection and combining are performed interactively.

5.1. Simple majority vote

If the label outputs of a classifier i denoted as a *c-dimensional* binary vector $[d_{i1}, \dots, d_{ic}]^T, i = 1, \dots, L, d_{ij} = 1$ if the output label is ω_j and 0, otherwise. Then the "majority vote" will return ω_k if

$$\sum_{i=1}^{L} d_{ik} = \max_{j=1}^{c} \sum_{i=1}^{L} d_{ij} \tag{30}$$

In two-class case (c=2) it is the same as 50% of votes plus 1 or

$$label = 1 \text{ if at least } \left\lfloor \frac{L}{2} \right\rfloor + 1 \ outputs \ are \ 1. otherwise \ 0.$$

It should be noted that it is not guaranteed that majority votes outperform any individual classifier. (Kuncheva 2014) States the upper and lower bounds of majority vote as follows:

Assume classifier i with accuracy P_i are arranged such that $P_1 \leq P_2 \leq \cdots \leq P_L$. Let $k = \frac{L+1}{2}$. The Upper bound of the majority vote accuracy is:

$$\max P_{maj} = \min\{1, f(k), f(k-1), \dots, f(1)\}, \tag{31}$$

Where

$$f(m) = \frac{1}{m} \sum_{i=1}^{L-k+m} P_i, \qquad m = 1, \dots, k.$$

And the lower bound of the majority vote accuracy is:

(32) $$\max P_{min} = \min\{1, g(k), g(k-1), \dots, g(1)\},$$

Where

$$g(m) = \frac{1}{m} \sum_{i=k-m+1}^{L} P_i - \frac{L-k}{m}, \qquad m = 1, \dots, k.$$

The following example adapted from (Kuncheva 2014) elaborate on the above formulas:

- Example: There are 5 classifiers with accuracies (0.56, 0.58, 0.60, 0.60, 0.62) arranged from smallest to largest. The upper and lower bounds of the combining these classifiers with majority votes are:

$$L = 5, k = 3,$$

$$m = 1,\ f(1) = \frac{1}{1}(0.56 + 0.58 + 0.60) = 1.74$$

$$m = 2,\ f(2) = \frac{1}{2}(0.56 + 0.58 + 0.60 + 0.60) = 1.17$$

$$m = 3,\ f(3) = \frac{1}{3}(0.56 + 0.58 + 0.60 + 0.60 + 0.62) = 0.99$$

$$\max P_{maj} = \min\{1, 1.74, 1.17, 0.99\} = 0.99$$

And the lower bound

$$m = 1,\ g(1) = \frac{1}{1}(0.60 + 0.60 + 0.62 - (5-3)) = -0.18$$

$$m = 2,\ g(2) = \frac{1}{2}\left(0.58 + 0.60 + 0.60 + 0.62 - \frac{5-3}{2}\right) = 0.20$$

$$m = 3,\ g(3) = \frac{1}{3}\left(0.56 + 0.58 + 0.60 + 0.60 + 0.62 - \frac{5-3}{3}\right) = 0.32$$

$$\min P_{maj} = \max\{0, -0.18, 0.20, 0.32\} = 0.32$$

This example shows the potential of improvement and the possibility of obtaining a worse result than just simply using the best classifier.

In general, majority vote is optimal when:

1. The individual classifier accuracies are equal. And
2. The prior probabilities for the classes are the same, and the classifier give their decision independently conditioned on class label.
3. For each classifier, the probability of incorrect classification is equally distributed among the remaining classes.

5.2. Weighted majority vote

If the classifiers have different accuracy, it is reasonable to give the better ones more power in final decision. To put this idea into practice, we introduce weights of coefficient of importance so that the class label ω_k is chosen if

$$\sum_{i=1}^{L} w_i d_{ik} = \max_{j=1}^{c} \sum_{i=1}^{L} w_i d_{ij} \tag{33}$$

The optimality condition for "weighted majority vote" is the same as "majority vote" and under the optimality condition, the optimal weight is:

$$w_i = \log\left(\frac{P_i}{1-P_i}\right), \qquad 0 < P_i < 1\,, \qquad i = 1, \dots, L \tag{34}$$

5.3. Naive bayes and BKS combiner

Naïve bayes combiner belongs to the trainable category. That is, it needs to be trained before using. There are two common ways to train such combiners. First, to use the training set that was already used to train the classifiers. Second method is, to train the combined on a different dataset.

Naïve Bayse combiner is optimal, if the classifier give their decision independent of class labels. In this study we use a Naïve Bayes classifier to combine label outputs.

BKS is also a trainable combiner. It is optimal for any sort of dependencies between the classifier outputs (Kuncheva 2014).The major drawback of BKS is that it can become easily overstrained.

5.4. Two examples

In this section, we continue with the two example from previous chapters and implement the materials discussed in this chapter regarding combining the classifiers. Briefly, in previous chapters, we use two dataset and build several classifiers and measured their performance on two toy data. Now, we continue by combining their output and measure their performance using the combination methods discussed in this chapter. We use the following strategies to select the classifiers:

1. Select all classifiers regardless of any performance measures and combine them with “Majority Vote”
2. First, 100 observation from the train dataset are set aside. We call this dataset “dsTrain2”, then the classifier are trained on the training set (not including the 100 observations). We call it dsTrain1. Then predict the labels in dsTrain2 (100 observations already set aside). The classifiers are selected based on the performance on this dataset with accuracy greater than 50%. And combined using the method discussed in this chapter.

Last row of table **9** shows the result of using strategy 1 to combine the outputs of classifier for “Fish Data”. It is colored in green and italic format for visual ease.

Although the combiner give significant result, it does not outperform the single best classifier (Knn) in terms of accuracy.

Table 9: Performance of combining classifier based on strategy 1 for Fish dataset.

Classifier	ACC	TPR	FPR	PPV	NPV	PT-Pval	Pval Vs Naïve Classifier	Pval Vs Linear Regression
Logistic Regression	63.52	38.82	21.39	52.57	67.78	0.0000	0.4976	0.6250
K-nearest neighbor	86.40	82.70	11.34	81.67	89.35	0.0000	0.0000	0.0000
Naïve Bayes	76.64	59.92	13.14	73.58	78.01	0.0000	0.0000	0.0000
Tree Classifier	84.48	81.01	13.40	78.69	88.19	0.0000	0.0000	0.0000
Multinomial Classifier	78.56	64.56	12.89	75.37	80.09	0.0000	0.0000	0.0000
Linear Regression	63.36	37.97	21.13	52.33	67.55	0.0000	0.5432	1.0000
Naïve Classifier	62.08	0.00	0.00		62.08	0.5000	1.0000	0.5432
Majority Vote	*79.52*	*60.76*	*9.02*	*80.45*	*79.15*	*0.0000*	*0.0000*	*0.0000*

The result of strategy 1 for "Smarket" dataset is presented in table **10** .The result does not show considerable improvement over the single best classifier. Besides, it is not significant based on PT test.

Table 10: Performance of combining classifier based on strategy 1 for Smarket dataset.

Classifier	ACC	TPR	FPR	PPV	NPV	PT-Pval	Pval Vs Naïve Classifier	Pval Vs Linear Regression
Logistic Regression	55.95	75.18	68.47	58.24	50.00	0.1185	1.0000	0.5966
K-nearest neighbor	53.17	60.99	56.76	57.72	46.60	0.2481	0.4926	0.3029
Naïve Bayes	54.37	68.79	63.96	57.74	47.62	0.2092	0.6646	0.3135
Tree Classifier	55.95	100.00	100.00	55.95		0.5000	1.0000	0.7228
Multinomial Classifier	52.38	53.19	48.65	58.14	46.34	0.2365	0.4191	0.0925
Linear Regression	57.54	56.74	41.44	63.49	51.59	0.0079	0.7228	1.0000
Naïve Classifier	55.95	100.00	100.00	55.95		0.5000	1.0000	0.7228
Majority Vote	*56.75*	*78.72*	*71.17*	*58.42*	*51.61*	*0.0831*	*0.8013*	*0.8043*

Following strategy 2 for "Fish Data", first we choose the classifiers that outperform the naïve classifier in the dsTrain2, the classifier chosen are K-nearest neighbor, Naïve Bayes, Tree classifier and Multinomial classifier. Table **11** shows the result of combining the output of these classifiers using the combiner discussed in this chapter. The "Weighted Majority Vote" slightly outperforms the single best classifier (Knn) in

terms of accuracy. Although the improvement is not noteworthy, considering Positive Predictive Value and Negative Predictive Value, it can be seen it is a more balanced in terms of these two performance measure.

Table 11: Performance of combining classifier based on strategy 2 for Fish dataset.

Classifier	ACC	TPR	FPR	PPV	NPV	PT-Pval	Pval Vs Naïve Classifier	Pval Vs Linear Regression
Logistic Regression	63.52	38.82	21.39	52.57	67.78	0.0000	0.4976	0.6250
K-nearest neighbor	86.40	82.70	11.34	81.67	89.35	0.0000	0.0000	0.0000
Naïve Bayes	76.64	59.92	13.14	73.58	78.01	0.0000	0.0000	0.0000
Tree Classifier	84.48	81.01	13.40	78.69	88.19	0.0000	0.0000	0.0000
Multinomial Classifier	78.56	64.56	12.89	75.37	80.09	0.0000	0.0000	0.0000
Linear Regression	63.36	37.97	21.13	52.33	67.55	0.0000	0.5432	1.0000
Naïve Classifier	62.08	0.00	0.00		62.08	0.5000	1.0000	0.5432
Majority Vote	*85.60*	*83.54*	*13.14*	*79.52*	*89.63*	*0.0000*	*0.0000*	*0.0000*
Weighted MV	***86.88***	***81.43***	***9.79***	***83.55***	***88.83***	***0.0000***	***0.0000***	***0.0000***
Naïve Bayes Combiner	*85.60*	*83.54*	*13.14*	*79.52*	*89.63*	*0.0000*	*0.0000*	*0.0000*
BKS Combiner	*84.84*	*85.34*	*15.47*	*77.34*	*90.31*	*0.0000*	*0.0000*	*0.0000*

Regarding the "Smarket" data, following the strategy 2, the selected classifiers are Logistic Regression, K-nearest neighbor, Naïve Bayes, Tree classifier and Multinomial classifier. Table **12** displays the result. None of the classifier outperforms the single best classifier which in this dataset is the linear regression in terms of accuracy. The McNemar test also implies they are not different from the benchmarks either.

Table 12: Performance of combining classifier based on strategy 2 for Smarket dataset.

Classifier	ACC	TPR	FPR	PPV	NPV	PT-Pval	Pval Vs Naïve Classifier	Pval Vs Linear Regression
Logistic Regression	55.95	75.18	68.47	58.24	50.00	0.1185	1.0000	0.5966
K-nearest neighbor	53.17	60.99	56.76	57.72	46.60	0.2481	0.4926	0.3029
Naïve Bayes	54.37	68.79	63.96	57.74	47.62	0.2092	0.6646	0.3135
Tree Classifier	55.95	100.00	100.00	55.95		0.5000	1.0000	0.7228
Multinomial Classifier	52.38	53.19	48.65	58.14	46.34	0.2365	0.4191	0.0925
Linear Regression	57.54	56.74	41.44	63.49	51.59	0.0079	0.7228	1.0000
Naïve Classifier	55.95	100.00	100.00			0.5000	1.0000	0.7228
Majority Vote	*56.35*	*78.01*	*71.17*	*58.20*	*50.79*	*0.1060*	*0.9007*	*0.7122*
Weighted MV	*57.54*	*81.56*	*72.97*	*58.67*	*53.57*	*0.0514*	*0.5966*	*1.0000*
Naïve Bayes Combiner	*55.95*	*100.00*	*100.00*	*55.95*		*0.5000*	*1.0000*	*0.7228*
BKS Combiner	*57.38*	*95.71*	*94.23*	*57.76*	*50.00*	*0.3534*	*1.0000*	*0.4244*

6. Variable selection

Variable selection is an active research area and its importance becomes clearer in situation where there is a large number of variables to be selected. The methods used in this area is myriad. (May, Dandy and Maier 2011) Has reviewed these methods in the literature of Artificial Neural Network. Figure **16** is extracted from their study which elegantly depicts the variety of methods that can be used. Based on this frame work, this thesis uses the methods from information theoretic which is marked by a dashed rectangular. It should be mentioned that we do not use some of the methods mentioned in the figure. The purpose of reporting this figure from (May, Dandy and Maier 2011) is just to show are location in the literature.

Before proceeding to next section, a few terminology in variable selection literature is explained in the context of classification:

- Wrapper: Includes methods where some performance measure such as accuracy of the classifier is used to select the variables. In other words, the variables are selected so that they optimize some objective value (function), which in case of wrapper methods this objective is some performance measure (can be combination of several performance measures) of the classifier such as accuracy.
- Filter: Includes methods where general characteristics of the data are used to select variables without involving the specific classifier. For example, using correlation to rank the variables, and then select the top 5 variables belong to the "filter method" groups.

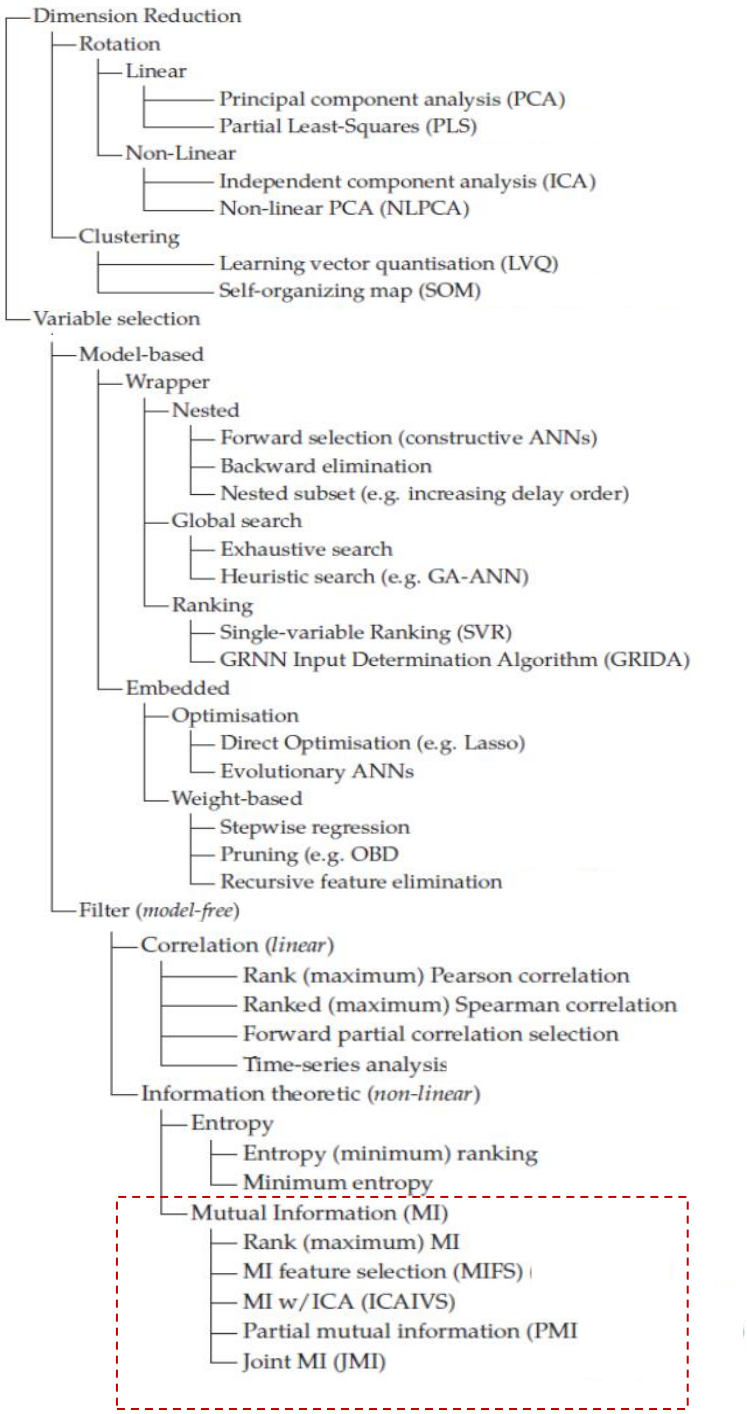


Figure 16: Categories of input variable selectoin.

6.1. Entropy and Mutual information

The materials in this section are based on (Cover and Thomas 1991)

- Probability theory background :

(35) $$P_{Y|X}(y|x) = \frac{P_{XY}(x,y)}{P_X(x)}$$

Which can be written in the chain rule form

(36) $$P_{XY}(x,y) = P_X(x)P_{Y|X}(y|x)$$

Two random variables are independent if

$$X \; indpendent \; of \; Y \Leftrightarrow \; P_{XY}(x,y) = P_X(x)P_Y(y)$$

If U=g(X) then the expected value of U is

(37) $$E(U) = E\big(g(X)\big) = \sum_{x \in X} g(x)P_X(x)$$

That is, the expected value of a function of X can be calculated using the distribution of X, and it is not necessary to know the distribution of U.

- Entropy: Entropy is a measure of uncertainty. For a discrete random variable X is defined by

(38) $$H(X) = -\sum_{x \in X} P_X(x) log P_X(x), \quad 0log0 = 0$$

Entropy is function of distribution of X i.e. it does not depend on the actual values taken by the random variable X, but only on the probabilities. Using the relation for expected value of a function of X, the expression for entropy can be written as

(39) $$H(X) = E\left(\log \frac{1}{P_X(X)}\right) = -E(log P_X(X))$$

- Joint entropy: The joint entropy H(X,Y) of a pair of discrete random variables is defined as

(40) $$H(X,Y) = -\sum_{x \in X} \sum_{y \in Y} P_{XY}(x,y) log P_{XY}(x,y) = -E(log P_{XY}(x,y))$$

- Conditional entropy: The conditional entropy is defined as

(41) $$H(Y|X) = -\sum_{x\in X}\sum_{y\in Y} P_{XY}(x,y) log P_{Y|X}(y|x)$$

- Chain rule: The following relationship holds

(42) $$H(X,Y) = H(X) + H(Y|X)$$

- Mutual information: The definition of mutual information (I) for two random variable *X* and *Y* is

(43) $$I(X,Y) = \sum_{x\in X}\sum_{y\in Y} P_{XY}(x,y)\ log\frac{P_{XY}(x,y)}{P_X(x)P_Y(y)} = E(log\frac{P_{XY}(x,y)}{P_X(x)P_Y(y)})$$

As mentioned in earlier entropy is a measure of uncertainty. It can be shown that mutual information can be written as

(44) $$I(X,Y) = H(Y) - H(Y|X)$$

This formula shows an interpretation of mutual information. It indicates the amount of reduction in entropy of Y (or uncertainty of Y) if the X is known. Another point worth noting is

$$X \; indpendent \; of \; Y \Leftrightarrow I(X,Y) = 0$$

Zero correlation between two variables does not imply independence, but unlike correlation, zero mutual information implies independence.

- Conditional mutual information : The conditional mutual information of random variables X and Y *given* Z is defined by

(45) $$I(X,Y|Z) = H(X|Z) - H(X|Y,Z)$$

It can be thought of as the reduction of uncertainty of X when Y is known in the world that Z is already given.

- Example: Suppose *Y* is a function of *X* such that $Y=X^2$ .Given the information about X, we calculate *E(X), E(Y), Cor(X,Y), H(X), H(Y) and I(X,Y)* .

X	**-1**	**0**	**1**
Px	1/3	1/3	1/3

Information about X result in the following probability distribution

$\boldsymbol{Y = X^2}$	**0**	**1**
Py	1/3	2/3

Pxy		Y	
		0	***1***
X	***-1***	0	1/3
	0	1/3	0
	1	0	1/3

$$E(X) = \sum x P_X(x) = 0\,, \boldsymbol{and} \quad E(Y) = \frac{2}{3}, \boldsymbol{and} \quad Cor(X,Y) = \frac{Cov(X,Y)}{\sigma_X \sigma_Y} = 0$$

$$H(X) = -\sum_{x \in X} P_X(x) log P_X(x) = \log(3) = 1.0986, \ \boldsymbol{and} \ H(Y) = 0.6365$$

$$I(X,Y) = \sum_{x \in X} \sum_{y \in Y} P_{XY}(x,y) \ log \frac{P_{XY}(x,y)}{P_X(x) P_Y(y)} = 0.6365 \quad \Rightarrow \quad H(Y|X) = 0$$

Here, although there exist a relationship between *X* and *Y*, correlation is zero. In contrast, mutual information is not. Besides, it indicates uncertainty about *Y* will be zero given *X*.

6.2. Variable selection using filter method

In this method, a criteria *J* is defined with the intention to measure how potentially useful a feature (variable) or feature subset can be to be used in classification (Brown, et al. 2012). For instance, correlation between feature and class label is an intuitive example i.e. it is expected that stronger correlation should result in better predictive ability. Numerous criteria have been proposed in the literature. This thesis study only considers a few criteria which are from information theory literature. Specifically, the selected criteria are *MIM*, *MRMR*, *JMI* and *DISR*. In the following these criteria are introduced. Almost all the material which follows is from (Brown, et al. 2012).

MIM: It stands for "*Mutual Information Maximization*". To use this criteria, features are ranked in order of their MIM score. Then, the first K features are selected. The MIM score is

(46) $$J_{MIM}(X_j) = I(X_j, Y)$$

An important limitation is that it assumes each feature is independent of all others. In other words, it does not take into account the dependency between features and basically ranks them in descending order based on their individual mutual information. Therefore it is suboptimal.

MRMR: It stands for "*Minimum-Redundancy Maximum-Relevance*". MRMR criteria (score) is

(47) $$J_{MRMR}(X_j) = I(X_j, Y) - \frac{1}{|S|} \sum_{k \in S} I(X_j, X_k)$$

MRMR tries to include variable with high mutual information with the target Y (Relevance, first term in expression) and low mutual information with the variables already selected (Redundancy, second term)

JMI: It stands for "Joint Mutual Information". It tries to increase complementary information between features. The JMI score is

(48) $$J_{JMI}(X_j) = \sum_{X_j \in S} I(X_j, X_k, Y)$$

This is the mutual information between the joint random variables $X = [X_j, X_K]$ and the target variable *Y*. The idea is to include a variable if it is complementary with current selected features.

DISR: It stands for "*Double Input Symmetrical Relevance*". It belongs to the group of criteria that use a normalization term on the mutual information. The DISR score is

$$J_{DISR}(X_j) = \sum_{X_j \in S} \frac{I(X_j, X_k, Y)}{H(X_j, X_k, Y)} \tag{49}$$

6.3. Estimation of mutual information

Similar to mean, variance and other statistical measures, mutual information needs to be estimated using the sample. There are various methods to obtain an estimation of mutual information. In this thesis study we follow (Brown, et al. 2012) . For discrete variables, the required probability distribution are estimated using the histogram. Then, the mutual information is estimated as the expected value of the ratio $log \frac{P_{XY}(x,y)}{P_X(x)P_Y(y)}$ that is

$$I(X,Y) = E(log \frac{P_{XY}(x,y)}{P_X(x)P_Y(y)}) \tag{50}$$

$$\hat{I}(X,Y) = \frac{1}{N} \sum_{i=1}^{N} \frac{\hat{P}(x_i, y_i)}{\hat{P}(x_i)\hat{P}(y_i)} \tag{51}$$

Regarding continuous variables, first, the distribution is estimated by histogram. Then, the mutual information is estimated using the method used for discrete variables. It should be noted that abovementioned approximation requires large enough N, depending on the problem.

6.4. An experiment

An experiment is performed using the various method of variable selection mentioned in this chapter. We use the "Fish Data" since the true variables are known $X_{true} = [x_1, x_2]$, then 200 irrelevant (or noise) variables are added to the existing variables $X_{noise} = [Xnoise_1, \dots, X_{noise200}]$ the input space now is $Input = [X_{true}, X_{noise}]$.Then different random samples of size N_i are drawn from this dataset. The algorithm mentioned are used to select 2 variables from the 202 input variables. This experiment is replicated for 1000 times for each sample size. Finally, the number of times at least one variable is selected correctly, and the number of times both variables are selected correctly are calculated.

Figure **17** depicts the result. The top chart shows that all algorithm were able to choose at least one variable correctly. The bottom row chart indicates JMI and MIM performs well for all sample size. But MRMR and DISR require more sample to give reliable results. DISR results improves as the sample size increases, but the behavior of MRMR is unreliable for sample size but it improves after as sample size becomes more than 200.

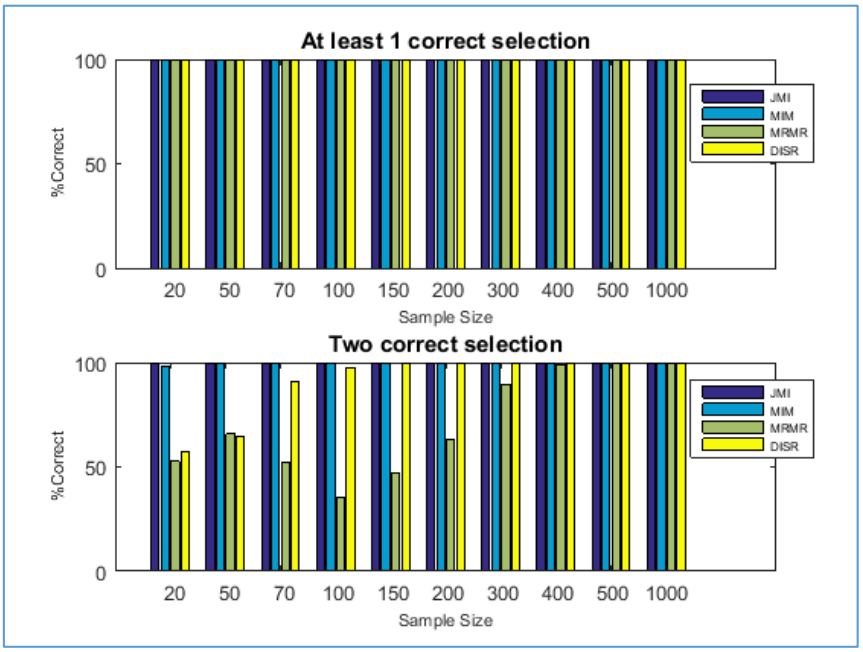


Figure 17: Performance of input selection algorithms.

7. Previous studies

Regarding forecasting exchange rates, the literature abounds with studies originating from different fields such as economics, statistics, signal processing and machine learning. Maybe it is not exaggerating to state that for any method of prediction there is a study using that method in forecasting currency movements.

(Rossi 2013) Conducted a comprehensive literature review regarding out-of-sample performance of previous studies mainly using economic models. She concluded predictability of exchange rates depends on the following:

- Choice of predictors.
- Forecast horizon.
- Sample period for evaluation.
- Model specification.
- Forecast evaluation method.

Other findings of the study was instabilities in the models' forecasting performance across models, predictors and time. These finding leads to two questions:

3. What are the reasons predictability change over time?
4. Is it possible to find a way to exploit instabilities to improve forecast?

(Yu, Wang and Lai 2007) Analyzed around 45 studies that just used different types of Neural Network for predicting currencies. They tried to find out why literature gives mixed result regarding performance of the models. They found that the difference in the reported results can be due to

- Prediction horizon.
- Data frequency (daily, weekly, monthly and quarterly).
- Training sample size.
- Performance measure.

Actually, the authors mentioned several other factors specific to training neural network.

Although the classification methods have been known to researchers, regardless of different method used for model building, the majority of the previous studies seek to find the expected value of the currency (level $E_t(S_{t+h})$ or change $E_t(\Delta S_{t+h})$) in next h period, and compare it with some benchmark which is usually random walk or some AR process. Here for the sake of completeness, a few of such studies are mentioned. (Li, Tsiakas and Wang, Predicting exchange rates out of sample: Can economic fundamentals beat the random walk? 2015) used elastic-net shrinkage method to forecast foreign exchange rates, based on "kitchen-sink" predictors and the reported this method of estimation outperforms both random walk and several other methods such as a kitchen-sink regression estimated with ordinary regression. (Dal Bianco, Camacho and Perez-Quiros 2012) Forecast EURUSD with the distinctive characteristic of using variables with different frequency in a state-space representation of the dynamic relationship between the exchange rate and its economic fundamental.

Unlike other methods, there are not too many studies using classification approach to forecast the direction of exchange rate. (Ullrich 2009) used SVM (support vector machine) to forecast several currencies including EURUSD for daily data from 1.1.1997 to 31.12.2004 totaling 2349 trading day which is divide into 1738 observation for model building and 350 observation for testing the model. Although the results were not impressive in absolute term (almost 53% accuracy), they all outperformed the benchmarks. (Qian and Rasheed 2010) Used a multiple classifier approach to forecast direction of GBPUSD for daily data from 4.1.1971 to 5.7.2005 with a hold-out of 20% for evaluating the final model. The base classifier includes neural network, k-nearest neighbor, classification tree, and naïve bayes. The best result on the test set has almost 60% accuracy achieved by consistent voting (when all classifiers have the same output). (Zhang and Zhao 2009) Used SVM to forecast the direction of EURUSD using technical indicators as inputs. The data is daily from 10.7.2007 to 9.7.2009 totaling 523 days. The test sample size is 50 and they obtained 60% accuracy on test dataset. (Plakandaras, Gogas and Papadimitriou, Directional Forecasting in financial time series using Support Vector machines: the USD/Euro exchange rate. n.d.) Used SVM to forecast EURUSD for daily data collected from 1.1.1999 to 30.10.2011 a total sample size of 3280 observations. The models are evaluated on 163 test dataset. The accuracy achieved on test dataset is about 60%.

There are some studies that deal with the nature of directional forecasting. (Christoffersen and Diebold 2006) investigate the directional forecasting from a theoretic aspect and concluded " a) Volatility dependence produces sign dependence, so long as expected returns are nonzero, so that one should expect sign dependence without conditional mean independence b) It is possible to have sign dependence without conditional mean dependence c) Sign dependence is not likely to be found via analysis of sign autocorrelations, runs tests, or traditional market timing test, because of the special nonlinear nature of sign dependence, so that traditional market timing tests are best viewed as tests for sing dependence arising from variation in expected returns rather than from variation in volatility or higher moments d) Sign dependence is not likely to be found in very high-frequency (e.g., daily) or very low-frequency (e.g., annual) returns; instead, it is more likely to be found at intermediate horizons e) The link between volatility dependence and sign dependence remains intact in conditionally non-Gaussian environments, as for example with time-varying conditional skewness and/or kurtosis" (To avoid my own misinterpretation , the exact wording of authors are used). (Chung and Hong 2007) Used a model free approach to investigate the nature of directional predictability and concluded that "Despite the weak conditional mean dynamics of foreign exchange returns, directional predictability can be explained by strong dependence derived from higher-order conditional moments such as the volatility, skewness and kurtosis of past exchange returns".

It is worth noting there are studies using directional forecasting in equity market. (Pesaran and Timmermann, Predictability of stock returns: Robustness and economic significance. 1995) Find that in stock market the predictive power of factors change over time. (Leung, Daouk and Chen 2000) Compared the performance of classification and level estimation in forecasting stock indices and concluded classification models outperform level forecasting models in terms of directional forecasting.

8. Data and methodology

8.1. Data

Table **13** contains the major data series used in this study. All series are from 1.1.2004 to 8.2.2016. The interest rates are obtained from FRED (https://research.stlouisfed.org/fred2/). Currency pairs are obtained from Google Finance (https://www.google.com/finance). Market index data before 2011 was obtained from DataStream, and from 2011 was obtained from Google Finance.

Table 13: Major data series.

Category	Name	Description	Frequency
Interest rates	DBAA	Moody's Seasoned Baa Corporate Bond Yield	Daily
	DAAA	Moody's Seasoned Aaa Corporate Bond Yield	Daily
	DGS10	10-Year Treasury Constant Maturity Rate	Daily
	DGS1	1-Year Treasury Constant Maturity Rate	Daily
	DTB3	3-Month Treasury Bill: Secondary Market Rate	Daily
	DTB4WK	4-Week Treasury Bill: Secondary Market Rate	Daily
	DFF	Effective Federal Funds Rate	Daily
	USDLib3M	3-Month London Interbank Offered Rate (LIBOR), based on U.S. Dollar	Daily
	EURLib3M	3-Month London Interbank Offered Rate (LIBOR), based on Euro	Daily
	GBPLib3M	3-Month London Interbank Offered Rate (LIBOR), based on British Pound	Daily
	JPYLib3M	3-Month London Interbank Offered Rate (LIBOR), based on Japanese Yen	Daily
	CHFLib3M	3-Month London Interbank Offered Rate (LIBOR), based on Swiss Franc	Daily
	USDLib1W	1-Week London Interbank Offered Rate (LIBOR), based on U.S. Dollar	Daily
	EURLib1W	1-Week London Interbank Offered Rate (LIBOR), based on Euro	Daily
	GBPLib1W	1-Week London Interbank Offered Rate (LIBOR), based on British Pound	Daily
	JPYLib1W	1-Week London Interbank Offered Rate (LIBOR), based on Japanese Yen	Daily
	CHFLib1W	1-Week London Interbank Offered Rate (LIBOR), based on Swiss Franc	Daily
Market Index	SP500	Standard & Poor's 500 Index	Daily
	EUR STOXX 50	Euro 50 blue chip stocks	
	VIX	CBOE volatility index	Daily
	GSCI	Benchmark for investment in commodity market	Daily
	GSCIPM	GSCI precious metal	Daily
	GSCIIM	GSCI industrial metal	Daily
	GSCIE	GSCI energy	Daily
Currency Pair	EURUSD,JPYUSD CADUSD,GBPUSD AUDUSD,CHFUSD NZDUSD	USD per Currency. EURUSD=1.3 means 1.3 dollar per 1 euro.	Daily

8.2. Overall description of methodology

In this thesis study, various models are run under different scheme, then their significance are tested using the test suggested in (Pesaran and Timmermann, A simple nonparametric test of predictive performance 1992) . This test (referred to as PT-test) is usually used in the literature to test the significance of generated forecast in predicting the direction of the realized value. In addition to PT-test, to compare the model accuracy with a benchmark, the McNemar test suggested in (Gama, et al. 2014) is used. The benchmark in this study is the accuracy performance of the linear regression counterpart of the model which its dependent variable is not coded into binary i.e. the return itself. For example, if a model is estimated in a rolling fashion with a window size of 100 using Knn classifier, then it is compared to a linear regression counterpart estimated exactly the same way. The test period is from 7.3.2014 to 2.5.2016 totaling 100 observation. These observation are not used at all during model building steps.

Each model regardless of its input variable is estimated both in recursive and rolling fashion. Regarding rolling scheme three window size are used 100,200 and 358.

Each model is estimated using two types of input “Original” and “Discretized”. The reason for using the discretized form is to mitigate the adverse effect of noise.

Each model is estimated using seven classification method i.e. logistic regression, k-nearest neighbor (Knn), naïve Bayes, classification tree, multinomial classifier, Naïve classifier, and linear regression. All classifiers except linear regression use the coded binary of return as dependent variable. In the linear regression case, the return magnitude is used as the dependent variable. All classifiers except multinomial classifier are estimated using the two types of input data (original, discretized). Multinomial classifier can only be estimated using discretized data.

Regarding the selecting the input variables two overall models are constructed. First are models that use a fixed set of inputs throughout the estimation and evaluation. These inputs are chosen from ideas suggested or inferred from the literature and theory. Second are models whose inputs are selected at each estimation steps using various input selection methods. That is the input variables can potentially be different for each forecasted realize value. The input selection methods used in this study are JMI, MIM, MRMR, and DISR.

Finally, the outputs of constructed models can be combined using various strategies. The combining methods used in this study are: simple majority vote, naïve Bayes combiner, and multinomial combiner.

Table **14** depicts the combination of different elements in model building and figure **18** shows the overall practical steps required in model building.

Table 14: Combination of elements in model building.

Modle_Id	Classifier	Estimation	Rolling Window Size	Predictors Selector	Predictor Type	Number of Predictors	Forecast horizon	Model Evaluation	Forecast combiner
Model$_i$	-Logistic Regression -Knn -Naïve Bayes -Decision Tree -Linear Regression -Multinomial -Naïve	Recursive Rolling	100 200 358	-JMI -MIM -MRMR -DISR	-Original -Discretized	5 7 10 15	1 week	-PT-test -McNemar	-Majority Vote -Naïve Bayes -Multinomial

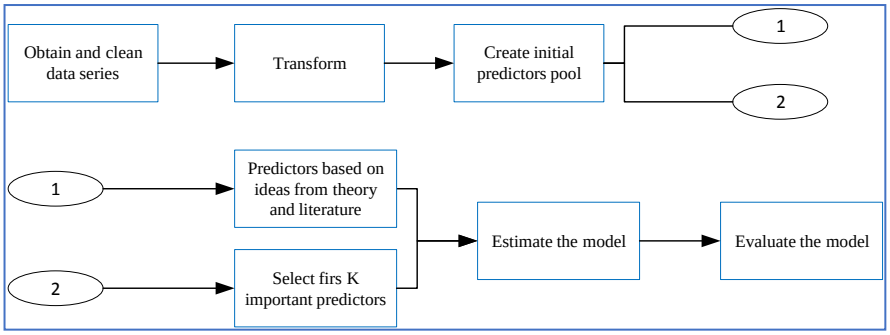


Figure 18: Practical steps in model building.

Briefly, the null hypotheses considered in this study is actually the null hypothesis of the PT test which is forecasted value (f_t) does not have any power to forecast the actual value (y_t). In other words they are independent. The rejection of the null means the model has predictive power.

H_{0i}: $Model_i$ forecast is independent of actual value. (For i=1 to number of models).

8.3. Ideas from literature and theory- fixed input

In this section, the rational for constructing a few models based on ideas from literature and theory is described.

(Chung and Hong 2007) Investigated the directional predictability of foreign exchange rates using a model free approach and concluded "the level, volatility, skewness, kurtosis and direction of past returns and interest rate differentials are more or less useful in predicting the currency returns". (Chung and Hong 2007) Also remind a simple model does not give necessarily a good out-of-sample performance.

One of conclusions from (Christoffersen and Diebold 2006) was that volatility dependence will produce sign dependence and in a case study demonstrated using RiskMetrics as an estimation of future volatility in a logistic regression framework to forecast S&P100 index direction.

As mentioned in the chapter regarding the theory of exchange rates, interest rates differential is a plausible candidate. Besides, as mentioned earlier, (Chung and Hong 2007) also suggest interest rate differential predictability power.

Table **15** presents the model built in this thesis study based on the abovementioned ideas.

Table 15: Models based on ideas from literature.

Model	Predictors	Number of Predictors
1	2 lags of EURUSD past log return	2
2	Inverse of RiskMetrics	1
3	Interest rate differential (libor differential)	1
4	2 lags return, Riskmetrics, Skewness	4

This list can be extended in further researches.

8.4. Adaptive variable selection

Findings of two specific studies (among other studies) that encouraged using an adaptive variable selection scheme in this thesis study. First is (Rossi 2013) which mentions the factors affecting the performance of a forecasting model which was mentioned earlier in the chapter regarding previous studies. Second is (Li and Chen, Forecasting macroeconomic time series: LASSO-based approaches and their forecast combinations with dynamic factor models. 2014)that finds a) LASSO based models and b) combination of forecast are promising approach in forecasting models. It is worth noting that LASSO can be thought of as a method that has internal variable selector, therefore when it is used in rolling or recursive regression, in essence, it can potentially choose different predictors which can be considered as a method of dealing with uncertainty in model's predictors. More information about LASSO based models can be found in (Hastie and Tibshirani 2013).

Figure **19** shows the overall steps of the adaptive variable selection scheme.

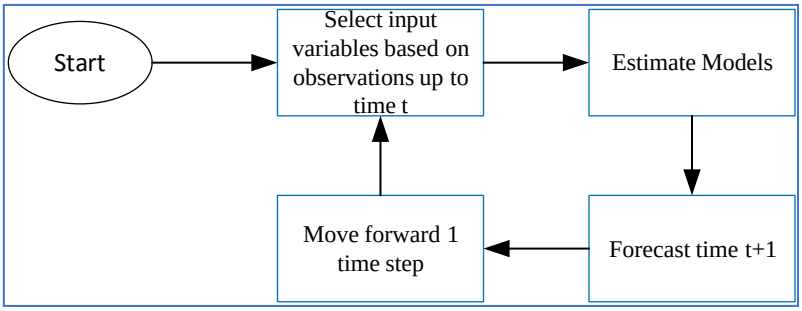


Figure 19: Adaptive input variable selection.

8.5. Combining classifiers

The models generated using the abovementioned procedures can be combined in a variety of ways. (Kuncheva 2014) Introduces a comprehensive ways of combining outputs of classifiers. One possible way to categorize combiners is into non-trainable and trainable. Trainable combiners need to be trained or estimated on dataset. In this study, simple majority vote is used as a non-trainable combiner. Two trainable combiners are naïve Bayes and multinomial combiners. The strategies used to combine the outputs are as follows:

- Combining strategy 1: Models estimated in fixed variables scheme are combined using both type of combiners and their significance are tested using the PT-test.
- Combining strategy 2: First a number of classifiers are selected based on their accuracy performance on the previous 100 weeks out-of-sample performance then both type of combiners are used to combine their outputs and finally their significance is tested.

These two strategies can be extended in a variety of ways, such as using different performance measure other than accuracy, or building combiners for each class label i.e. one combining scheme to predict the upward movement, and another to predict the downward movement.

9. Result

This chapter presents the result of the model building procedure described in previous chapter.

9.1. Descriptive statistics and preliminary analysis

This section presents some statistical properties of the EURUSD series both as a whole and through time.

Table **16** shows the descriptive statistics of the *annualized* weekly log-return of EURUSD from 2.1.2004 to 5.2.2016. The weekly returns are annualized by multiplying the log-return by 52. The "Mode" is calculated using the kernel density of the distribution.

Table 16: Descriptive statistics of EURUSD anualized weekly log-retrun.

Series	Mean	Trimmed Mean 5%	Median	Standard Deviation	Quantile 25%	Quantile 75%	Skewness	%Up	Mode
Log Return	-0.0099	-0.0033	0.0121	0.7287	-0.4447	0.4563	-0.3052	0.5055	0.0969

It can be seen from table **16** the skewness of EURUSD return is negative. In addition to the observed negative skewness, from statistics theory we know the relationship between, mean, median and mode of a negatively skewed distribution is

$$X_{mean} < X_{median} < X_{Mode}$$

This relationship also suggests the returns have slightly negative skewness. Although it may or may not continue this behavior in future time, it is useful to know this currency can have risk of crash. As in the literature skewness is associated with crash risk and is not diversifiable. The percent of up movement (%Up in table **16**) is very close to 50%

suggesting the unconditional percentage of up or down movement does not have too much predictive power.

Figure **20** depicts the histogram and kernel density of the *annualized* weekly log-return of EURUSD

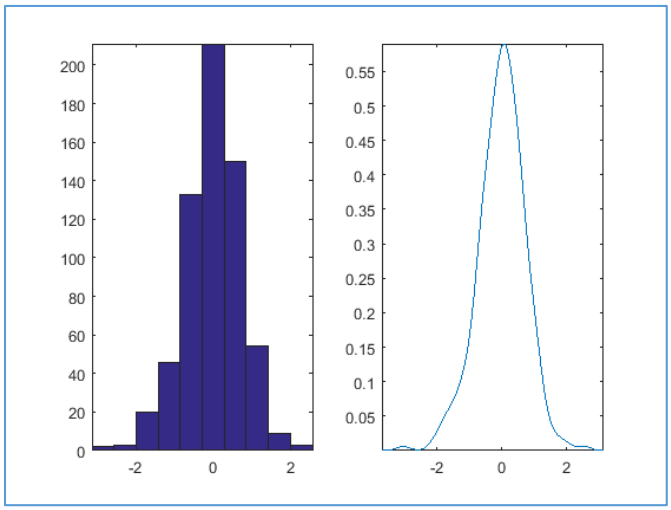


Figure 20: Histogram and kernel density of EURUSD annualized weekly log-return.

In what follows, a few characteristics of this series is presented through time. Figure **21** shows the weekly EURUSD series in level. The shaded area is recession period in the United States.

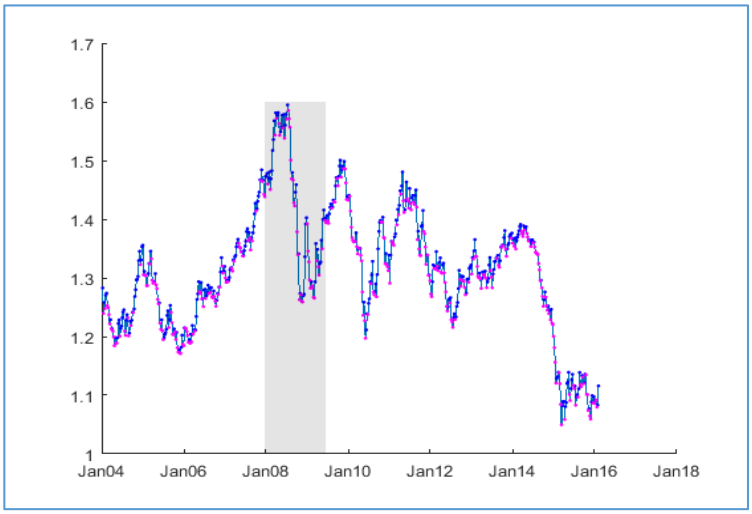


Figure 21: Weekly EURUSD in level.

Figure **22** and **23** depicts two measure of volatility for the EURUSD series "moving standard deviation" and "RiskMetrics". Both measures are calculated using "daily" data and are annualize. The period used for using "moving standard deviation" is 60 past days. The "RiskMetrics" metric is calculated using the following formula

$$\sigma_t^2 = 0.94\sigma_{t-1}^2 + 0.06r_t^2 \quad (52)$$

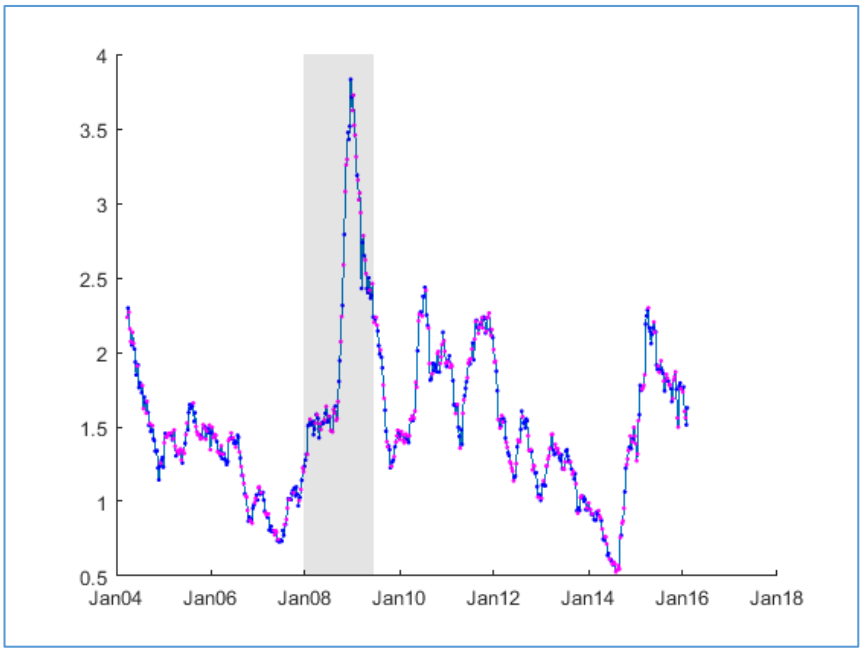


Figure 22: Moving standard deviation

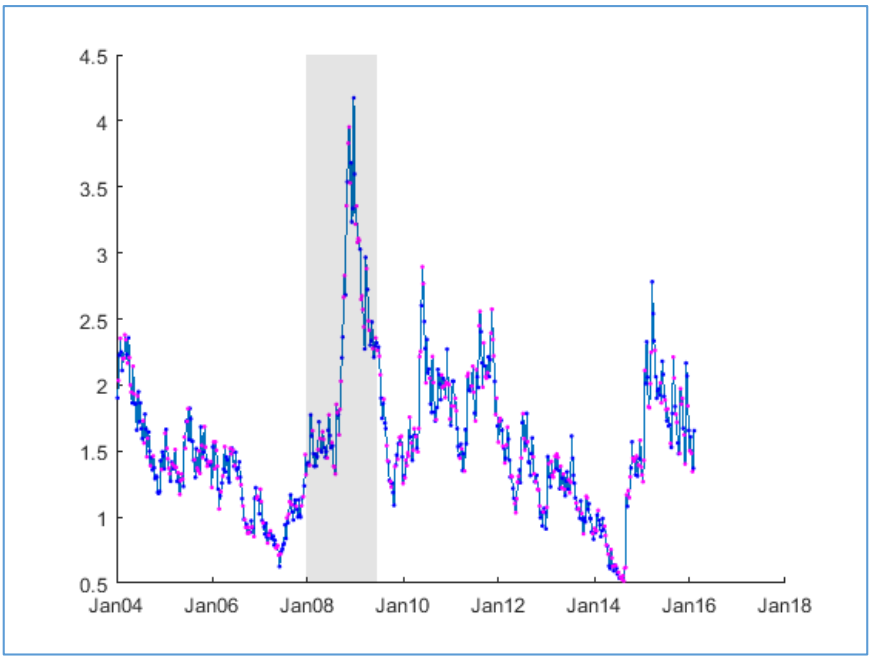


Figure 23: RisckMetrics.

Figure **24** displays the behavior of the "moving skewness". Similar to volatility measures, the skewness is calculated using daily data of the past 60 days.

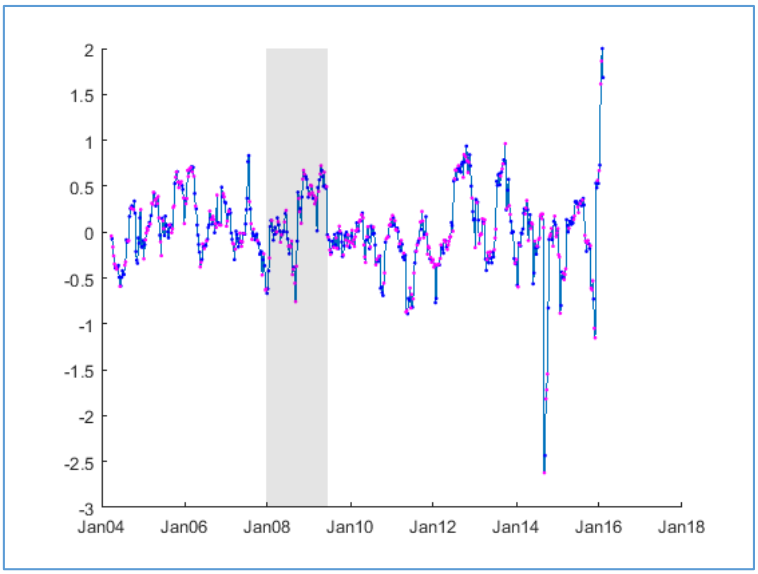


Figure 24: Moving skewness.

Figure **25** shows the moving weekly z-score of the EURUSD (level) based on the past 12 week data. The z-score gives a measure of the position of the series relative to its past behavior. The formula used for calculating the moving z-score is

(53) $$Zscore_t = \frac{X_t - MA_t(n)}{SD_t(n)}$$

MA(n) and SD(n) are the moving average ,and moving standard deviation, calculated using past n data. Z-score is between -3 and 3 most of the time.

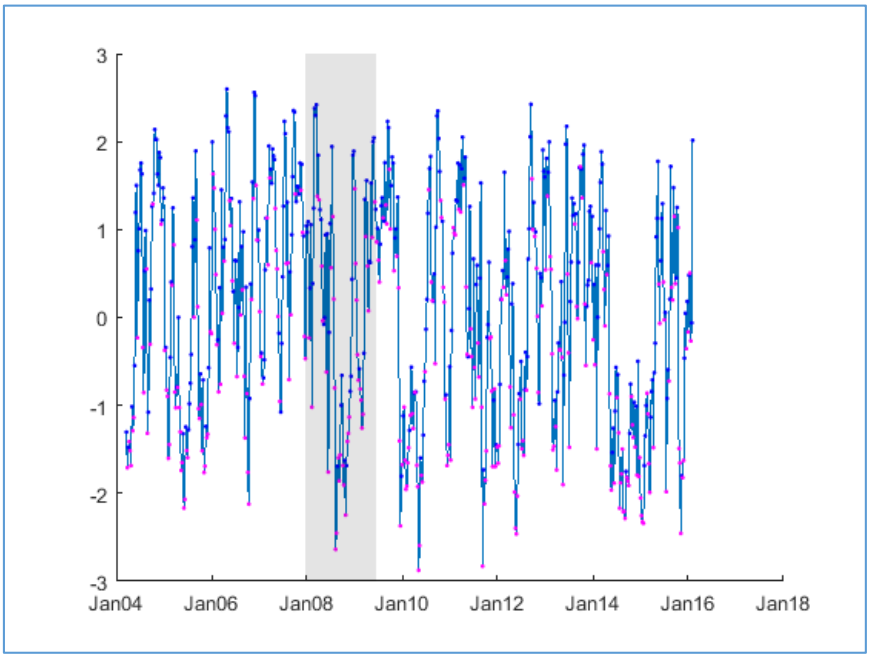


Figure 25: Moving Z-score of EURUSD weekly level.

9.2. Models with fixed variables

The models in this section are built based on the ideas from literature. Table **17**shows the variables used in the models.

Table 17: Models predictors.

Model	Predictors	Number of Predictors
1	2 lags of EURUSD past log return	2
2	Inverse of RiskMetrics	1
3	Interest rate differential (libor differential)	1
4	2 lags return, Riskmetrics, Skewness	4

Each of the aforementioned predictors are used in a model built according to the combination in table **18** .

Table 18: Combination of model elements for fixed variable models.

Modle_Id	Classifier	Estimation	Rolling Window Size	Predictor Type	Forecast horizon	Model Evaluation	Forecast combiner
Model$_i$	-Logistic Regression -Knn -Naïve Bayes -Decision Tree -Linear Regression -Multinomial -Naïve	Recursive Rolling	100 200 358	-Original -Discretized	1 week	-PT-test -McNemar	-Majority Vote -Naïve Bayes -Multinomial

Therefore in total, ((7*1*2)+(7*1*3*2))*4 =224 models are estimated. (7*1*2 is the number of models estimated in recursive scheme, the other part is for rolling scheme). This number of model is estimated for all the test set data.

One important note regarding "Multinomial classifier" in this study needs to be mentioned at this point. As pointed out in the classification chapter, this classifier uses a lookup table to classify objects. If the number of inputs are moderately high, due to a large number of combination, it happens that the exact match of the new observation is not in that lookup table. In that case, one common way is to report the class with the highest prior probability. But, in this study, we do not force this classifier to generate a label which is not in its lookup table, and it simply generates no value. The implication is that, cares must be taken when it is compared to other classifiers when its classification rate is less than 100%.

Table **19** presents the result of the top 20 models sorted based on PT p-value.

Table 19: Top 20 models with fixed variables.

Row	Model	Estimation	Rolling Window	Predictor Type	Classifier	ACC	PPV	NPV	PT-Pvalue	McNemar Pvalue
1	1	Rolling	358	'Orginal'	'Knn'	59.00 %	52.83 %	65.96 %	0.0288	0.4799
2	1	Recursive	---	'Discrete'	'Knn'	59.00 %	52.94 %	65.31 %	0.0324	0.3915
3	2	Recursive	---	'Orginal'	'Mn'	56.00 %	50.00 %	64.29 %	0.0767	0.1112
4	2	Recursive	---	'Discrete'	'Lgr'	56.00 %	50.00 %	64.29 %	0.0767	1.0000
5	2	Recursive	---	'Discrete'	'Nb'	56.00 %	50.00 %	64.29 %	0.0767	1.0000
6	2	Recursive	---	'Discrete'	'Mn'	56.00 %	50.00 %	64.29 %	0.0767	1.0000
7	2	Recursive	---	'Discrete'	'Lr'	56.00 %	50.00 %	62.00 %	0.1122	1.0000
8	4	Rolling	358	'Discrete'	'Lr'	56.00 %	50.00 %	62.00 %	0.1122	1.0000
9	2	Recursive	---	'Discrete'	'Knn'	55.00 %	49.12 %	62.79 %	0.1162	0.7266
10	1	Recursive	---	'Orginal'	'Mn'	55.00 %	49.12 %	62.79 %	0.1162	0.4177
11	1	Recursive	---	'Discrete'	'Mn'	55.00 %	49.12 %	62.79 %	0.1162	0.8679
12	4	Recursive	---	'Discrete'	'Lgr'	55.00 %	49.06 %	61.70 %	0.1385	0.8036
13	4	Recursive	---	'Discrete'	'Lr'	56.00 %	50.00 %	60.71 %	0.1408	1.0000
14	2	Recursive	---	'Orginal'	'Knn'	54.00 %	48.28 %	61.90 %	0.1545	0.1524
15	1	Recursive	---	'Discrete'	'Lr'	54.00 %	48.28 %	61.90 %	0.1545	1.0000
16	4	Recursive	---	'Orginal'	'Tre'	45.00 %	44.44 %	100.00 %	0.1853	0.6516
17	1	Recursive	---	'Orginal'	'Knn'	54.00 %	48.08 %	60.42 %	0.1951	0.5758
18	1	Rolling	200	'Orginal'	'Knn'	53.00 %	47.46 %	60.98 %	0.2005	0.3817
19	4	Recursive	---	'Discrete'	'Knn'	52.00 %	46.77 %	60.53 %	0.2365	0.4731
20	4	Rolling	358	'Discrete'	'Lgr'	53.00 %	47.17 %	59.57 %	0.2478	0.5034

It can be seen from the table that the top performer based on accuracy is “Model 1” which contains only the first two lags of past returns. The one-sided PT-test shows it is significant at 5% level, though not at 1% level. The McNemare test does not detect any difference between this model and its linear regression counterpart.

The top two models are significant at 5% and the models in row 3 and 4 are significant at 10%, the interesting point is that these two models are solely based on price information. The model 3 which is using the interest rate differential, does not appear in the top 20 models.

Another noteworthy feature, is the competing of models which use the discretized form of the predictors. Seven models of the top ten are using discretized data. This can be a topic worthy of deeper investigation.

From the classification model perspective, the knn classifier which can handle nonlinearity and does not assume any specific assumption regarding the distribution of data seems to outperform other methods in this particular setting. In contrast, the tree classifier which is also designed to handle nonlinearity and is not based on any specific assumption regarding the distribution of data, does not appear in the first top 10 at all. This result indicates that, in spite of having similar characteristic, the classifiers' performance can be quite different.

Regarding the sample size, it can be seen with this particular predictors, larger training sample size results in a better performance. Actually, there is only one model with the training sample size of 200 in the top 20.

In order to examine if combining these models result in a reasonable improvement we follow two strategy:

- Strategy 1: Combing all models regardless of performance measures i.e. all the 224 models are combined in this strategy. Table **20** presents the result of this strategy. As expected, the multinomial could not produce any output. The performance of the two other combination method is not significant.

Table 20: Combining fixed variable models using strategy 1.

Row	Combiner	Accuracy	PPV	NPV	PT-Pvalue
1	Majority Vote	43.00%	35.56%	49.09%	93.90%
2	Naïve Bayes	51.00%	44.68%	56.60%	44.84%
3	Multinomial	---	---	---	---

- Strategy 2: First a number of models which have accuracy more than 55% in the last 100 week, based on out-of-sample performance are selected and their outputs are combined. In this thesis study, the classifiers are selected based on significance of PT test at 5% level and the accuracy greater than 55% over the past 100 weeks out-of-sample. Table **21** presents the result of this strategy. None of the combiner produce significant result, but using selection seems to result in more promising outputs than just simply combining all generated classifiers.

Table 21: Combining fixed variable models using strategy 2.

Row	Combiner	Accuracy	PPV	NPV	PT-Pvalue
1	Majority Vote	51.00%	45.76%	58.54%	0.3343
2	Naïve Bayes	55.00%	49.02%	61.22%	0.1499
3	Multinomial	41.18%	35.85%	50.00%	0.9018

9.3. Models with adaptive variables

In this section models are built adaptively as described in the methodology section. The classifier used are those of previous section except multinomial, linear regression and naïve classifiers. The McNemar test to compare the result with liner regression is not also performed. Models are run according to the combinations shown in table **22**. Similar to previous section, the forecast horizon is one week and the same combining strategies are used.

Table 22: Combination of model elements for adaptive variable models.

Modle_Id	Classifier	Estimation	Rolling Window Size	Predictors Selector	Predictor Type	Number of Predictors
Model$_i$	-Logistic Regression -Knn -Naïve Bayes -Decision Tree	Recursive Rolling	100 200 358	-JMI -MIM -MRMR -DISR	-Original -Discretized	5 7 10 15

Therefore in total, ((4*1*4*2*4) + (4*1*3*4*2*4)) =512 models are estimated for each observation in the test dataset. Table **23** presents the first top 20 models sorted based on the significance of the PT-test.

Table 23: Top 20 models with adaptive variables.

Row	Estimation	Rolling Window	number of Predictor	Predictor Type	Classifier	Feature Selector	ACC	PPV	NPV	PT-Pvalue
1	Recursive	---	5	'Orginal'	'Knn'	'mim'	59.00 %	53.06 %	64.71 %	0.0361
2	Recursive	---	5	'Orginal'	'Lgr'	'mim'	59.00 %	53.49 %	63.16 %	0.0476
3	Rolling	200	10	'Discrete'	'Nb'	'jmi'	59.00 %	53.85 %	62.30 %	0.0555
4	Rolling	358	15	'Orginal'	'Knn'	'mim'	58.00 %	52.17 %	62.96 %	0.0633
5	Rolling	200	10	'Discrete'	'Lgr'	'disr'	58.00 %	52.38 %	62.07 %	0.0744
6	Rolling	358	7	'Orginal'	'Knn'	'jmi'	57.00 %	51.02 %	62.75 %	0.0818
7	Recursive	---	7	'Orginal'	'Knn'	'mim'	57.00 %	51.11 %	61.82 %	0.0964
8	Recursive	---	10	'Discrete'	'Knn'	'disr'	57.00 %	51.11 %	61.82 %	0.0964
9	Rolling	200	7	'Discrete'	'Knn'	'disr'	58.00 %	52.94 %	60.61 %	0.0969
10	Rolling	200	15	'Discrete'	'Knn'	'disr'	58.00 %	52.94 %	60.61 %	0.0969
11	Rolling	358	10	'Orginal'	'Knn'	'disr'	57.00 %	51.16 %	61.40 %	0.1039
12	Recursive	---	15	'Orginal'	'Tre'	'mim'	53.00 %	47.89 %	65.52 %	0.1091
13	Rolling	358	7	'Discrete'	'Tre'	'mrmr'	57.00 %	51.22 %	61.02 %	0.1115
14	Rolling	200	10	'Orginal'	'Knn'	'mrmr'	55.00 %	49.09 %	62.22 %	0.1272
15	Recursive	---	15	'Orginal'	'Knn'	'mim'	56.00 %	50.00 %	60.71 %	0.1408
16	Rolling	200	15	'Orginal'	'Nb'	'jmi'	56.00 %	50.00 %	60.71 %	0.1408
17	Rolling	358	15	'Orginal'	'Lgr'	'mim'	56.00 %	50.00 %	60.71 %	0.1408
18	Rolling	200	5	'Discrete'	'Lgr'	'jmi'	56.00 %	50.00 %	60.71 %	0.1408
19	Recursive	---	15	'Orginal'	'Knn'	'jmi'	55.00 %	49.02 %	61.22 %	0.1499
20	Rolling	358	10	'Orginal'	'Lgr'	'mim'	56.00 %	50.00 %	60.34 %	0.1506

It can be seen from the table **23** there are several models with more or less similar performance. One interesting fact in this table is the competing of logistic regression

which is a linear classifier with nonlinear models. As it is apparent from the table, the logistic regression is ranked immediately after the knn classifier which is a nonlinear classifier.

The pattern observed regarding the discretized version of the predictors in previous section is almost repeated here, and indicates that discretizing the predictors does not have an adverse effect on the final results.

Regarding the sample size, the results indicate the plausible conclusion drawn in the previous section cannot be extended to the case where variables are not fixed and selected adaptively. But it can imply that in case a method is decided upon, it is good practice to estimate the model on various sample size to evaluate its influence.

In order to combine the outputs of the classifiers, the same two strategies that mentioned in previous section are used.

- Strategy 1: Combing all models regardless of performance measures. Table **24** presents the result of using this strategy. The results are not significant according to PT-test.

Table 24: Combining adaptive variable models using strategy 1.

Row	Combiner	Accuracy	PPV	NPV	PT-Pvalue
1	Majority Vote	49.00%	40.00%	53.85%	0.7238
2	Naïve Bayes	49.00%	44.26%	56.41%	0.4735
3	Multinomial	---	---	---	---

- Strategy 2: First a number of models which have accuracy more than 55% in the last 100 week, based on out-of-sample performance are selected and their outputs are combined. In this thesis study, the classifiers are selected based on significance of PT test at 5% level and the accuracy greater than 55% over the past 100 weeks out-of-sample. None of the combiners are significant based on PT-test. Table **25** presents the result of this strategy. None of the combiner

produce significant result, but using selection seems to result in more promising outputs than just simply combining all generated classifiers.

Table 25: Combining adaptive variable models using strategy 2.

Row	Combiner	Accuracy	PPV	NPV	PT-Pvalue
1	Majority Vote	47.47%	40.00%	52.54%	0.7693
2	Naïve Bayes	48.48%	40.00%	53.13%	0.7458
3	Multinomial	52.17%	48.00%	57.14%	0.3622

10. Conclusion

The aim of this thesis study is to explore the idea of using classification method in forecasting the EURUSD currency. At first, several theory regarding exchange rates were introduced. To lay a ground for empirical section, theories regarding a few common classification model and combining their outputs were described. One chapter was devoted to the important subject of variable selection using methods rooted in information theory. The empirical section examined a few ideas from literature and then extend the analysis to an adaptive variable selection scheme. The results were formally tested using a commonly used statistical test in the literature.

Although a few models looked promising and were significant at 5% significance level, they were not significant at 1% level. One plausible reason can be the asymptotic nature of the test. The models were tested on 100 observations, which is almost two years. Whether increasing the length of testing period is reasonable or not is a valid question. The same argument can be applied to failing to detect a significant difference with linear regression model.

The research can be extended in three major ways. First direction can be viewed as technical extension such as using other possible methods at each steps of empirical study. Because this study just examined a limited number of possibilities. For example, other classification models can be used, or other input variable selection methods or parameter setting, combining strategy and so on. Second way that this study can be enriched is to examine the methods that deal with the instability of models in forecasting exchange rates. This concept is known as conceptual drift in machine learning community. In econometrics literature it is called structural break. Finally, this study can be extended easily by adding a step that uses input reduction technique such as principle component analysis. It can be performed in a variety of ways such as simply using PCA on all inputs, combine PCA and variable selectors i.e. use PCA on a number of selected variables and so on.

One finding of the study is that, models with discretized variables yield comparable models. There can be a few reason for this behavior. Since the majority of models are not significant, it can be concluded that since all of them are generating random output,

they can be sorted randomly regardless of what variable they use. But if this is not at least partly the case, it seems the subject of using a better or more sophisticated method of discretizing variables deserves examination.

In the end, one question deserves attention. That is, as seen above, assuming 5% significance level is good enough, a few models generated significant results which means they can predict the market movement to some extent. Does this finding imply that market is not efficient? Or efficient market hypothesis is wrong?

As mentioned in the introduction section (Rossi 2013) argues "It is important to note that the efficient market hypothesis does not imply that exchange rate changes should be unpredictable". This argument implies if a model can predict exchange rates it does not mean efficient market hypothesis is rejected. On the other hand, (Plakandaras, Papadimitriou, et al. 2015) states "Employing econometric and machine learning methodologies we develop models that forecast in out-of-sample exercise the future direction of the four exchange rates. Our empirical findings reject the Efficient Market Hypothesis even in its weak form for all four exchange rates".

Apart from which point of view is correct in a theoretical and abstract sense, from practical perspective, both of the above mentioned line of thought imply that if we succeed in building a satisfactory model, it can be used. Because achieving a significant model either has no conflict with efficient market hypothesis (first opinion) or rejects efficient market hypothesis (second opinion).

References

Agresti, Alan. 2002. *Categorical Data Analysis.* 2nd. Wiely.

Bekaert, Geert, and Robert Hodric. 2011. *International Financial Management.* 2nd. Pearson.

Brown, Gavin, Adam Pocock, Ming-Jie Zhao, and Mikel Luján. 2012. "Conditional likelihood maximisation: a unifying framework for information theoretic feature selection." *The Journal of Machine Learning Research* 13 (1): 27-66.

Cheung, Yin-Wong, Menzie Chinn, and Antonio Garcia Pascual. 2005. "Empirical exchange rate models of the nineties: Are any fit to survive?" *Journal of international money and finance* 24 (7): 1150-1175.

Christoffersen, Peter , and Francis Diebold. 2006. "Financial asset returns, direction-of-change forecasting, and volatility dynamics." *Management Science* 52 (8): 1273-1287.

Chung, Jaehun, and Yongmiao Hong. 2007. "Model-free evaluation of directional predictability in foreign exchange markets." *Journal of Applied Econometrics* 22 (5): 855-889.

Cover, Thomas, and Joy Thomas. 1991. *Elements of Information Theory.* Wiely.

Dal Bianco, Marcos, Maximo Camacho, and Gabriel Perez-Quiros. 2012. "Short-run forecasting of the euro-dollar exchange rate with economic fundamentals." *Journal of International Money and Finance* 31 (2): 377-396.

Fagerland, Morten , Stian Lydersen, and Petter Laake. 2013. "The McNemar test for binary matched-pairs data: mid-p and asymptotic are better than exact conditional." *BMC medical research methodology* 13 (1).

Gama, João, Indrė Žliobaitė, Albert Bifet, Mykola Pechenizkiy, and Abdelhamid Bouchachia. 2014. "A survey on concept drift adaptation." *ACM Computing Surveys (CSUR)* 46 (4).

Granger, Clive , and Hashem Pesaran. 2000. "Economic and statistical measures of forecast accuracy." *Journal of Forecasting* 19 (7): 537-560.

Greene, William. 2008. *Econometric Analysis.* 6th. Prentice Hall.

Hastie, Trevor, and Robert Tibshirani. 2013. *An Introduction to Statistical Learning with Applications in R.* 1st. Springer.

James, Jessica, Ian Marsh, and Lucio Sarno. 2012. *Handbook of Exchange Rates.* Wiley.

Kuncheva, Ludmila I. 2014. *Combining pattern classifier: methods and algorithm.* 2nd. Wiley.

Leung, Mark, Hazem Daouk, and An-Sing Chen. 2000. "Forecasting stock indices: a comparison of classification and level estimation models." *International Journal of Forecasting* 16 (2): 173-190.

Li, Jiahan, and Weiye Chen. 2014. "Forecasting macroeconomic time series: LASSO-based approaches and their forecast combinations with dynamic factor models." *International Journal of Forecasting* 30 (4): 996-1015.

Li, Jiahan, Ilias Tsiakas, and Wei Wang. 2015. "Predicting exchange rates out of sample: Can economic fundamentals beat the random walk?" *Journal of Financial Econometrics* 13 (2): 293-341.

May, Robert, Graeme Dandy, and Holger Maier. 2011. "Review of input variable selection methods for artificial neural networks." *Methodological advances and biomedical applications* (INTECH Open Access Publisher).

McNemar, Quinn. 1947. "Note on the sampling error of the difference between correlated proportions or percentages." *Psychometrika* 12 (2): 153-157.

Meese, Richard, and Kenneth Rogoff. 1983. "The out-of-sample failure of empirical exchange rate models: sampling error or misspecification?" *Exchange rates and international macroeconomics.*

Mishkin, Fredric, Kent Matthews, and Massimo Giuliodori. 2013. *The Economics of Money, Banking & Financial Markets.* European Edition. Pearson.

Mosteller, Frederick. 1952. "Some statistical problems in measuring the subjective response to drugs." *Biometrics* 8 (3): 220-226.

Pesaran, Hashem, and Allan Timmermann. 1992. "A simple nonparametric test of predictive performance." *Journal of Business & Economic Statistics* 10 (4): 461-465.

Pesaran, Hashem, and Allan Timmermann. 1995. "Predictability of stock returns: Robustness and economic significance." *The Journal of Finance* 50 (4): 1201-1228.

Plakandaras, Vasilios, Periklis Gogas, and Theophilos Papadimitriou. n.d. "Directional Forecasting in financial time series using Support Vector machines: the USD/Euro exchange rate." *Journal of Computational Optimization in Economics and Finance* 5 (2): 125-139.

Plakandaras, Vasilios, Theophilos Papadimitriou, Periklis Gogas, and Konstantinos Diamantaras. 2015. "Market sentiment and exchange rate directional forecasting." *Algorithmic Finance* 4 (1-2): 69-79.

Powers, David Martin. 2011. "Evaluation: from precision, recall and F-measure to ROC, informedness, markedness and correlation." *Journal of Machine Learning Technologies* 2 (1): 37-63.

Qian, Bo, and Khaled Rasheed. 2010. "Foreign exchange market prediction with multiple classifiers." *Journal of Forecasting* 29 (3): 271-284.

Rossi, Barbara. 2013. "Exchange rate predictability."

Ullrich, Christian. 2009. *Forecasting and Hedging in the Foreign Exchange Markets.* . Springer.

Yu, Lean , Shouyang Wang, and Kin Keung Lai. 2007. *Foreign-Exchange-Rate Forecasting with Artificial Neural Networks.* Springer US.

Zhang, Zuoquan, and Qin Zhao. 2009. "The application of SVMs method on exchange rates fluctuation." *Discrete Dynamics in Nature and Society.*