Precision and Recall for Regression

Luis Torgo[1] and Rita Ribeiro[2]

[1] FC / LIAAD-Inesc Porto LA, University of Porto, R. de Ceuta, 118, 6., 4050-190 Porto, Portugal
ltorgo@liaad.up.pt
[2] LIAAD-Inesc Porto LA, University of Porto, R. de Ceuta, 118, 6., 4050-190 Porto, Portugal
rpribeiro@liaad.up.pt

Abstract. Cost sensitive prediction is a key task in many real world applications. Most existing research in this area deals with classification problems. This paper addresses a related regression problem: the prediction of rare extreme values of a continuous variable. These values are often regarded as outliers and removed from posterior analysis. However, for many applications (e.g. in finance, meteorology, biology, etc.) these are the key values that we want to accurately predict. Any learning method obtains models by optimizing some preference criteria. In this paper we propose new evaluation criteria that are more adequate for these applications. We describe a generalization for regression of the concepts of precision and recall often used in classification. Using these new evaluation metrics we are able to focus the evaluation of predictive models on the cases that really matter for these applications. Our experiments indicate the advantages of the use of these new measures when comparing predictive models in the context of our target applications.

1 Introduction

Several important predictive data mining applications involve handling non-uniform costs and benefits of the predictions. This is almost always the case in event-based applications like prediction of ecological or meteorological catastrophes, fraud detection, network intrusions, financial forecasting, etc.. Many of these tasks are particular cases of regression problems where the continuous target variable values have differentiated importance. Often these prediction tasks are related to the anticipation of a critical phenomenon that is inherently continuous and for which an alarm may be triggered by a specific range of values of a continuous target variable. This type of applications requires techniques that are able to cope with differentiated costs and benefits of predictions.

In this paper we have as main goal to address a particular and highly relevant sub-class of non-uniform cost/benefit prediction tasks. These applications associate higher cost or benefit with rarity. For these applications the most (and often solely) important cases are the ones associated with unusual values of the target variable. We are thus facing a task of predicting outlier values of a continuous target variable.

Handling applications with differentiated costs and benefits of predictions is not new and many cost-sensitive techniques have been proposed in the literature (e.g. [6, 7]). Still, most of these works focus on predictive classification tasks. For regression the most common setup considers that the cost of predictions is uniform across the domain of the target variable and solely dependent on the magnitude of the prediction errors themselves.

Addressing cost-sensitive applications involves two major issues: i) defining proper evaluation metrics to correctly assert the merits of alternative models given the application preference biases; and ii) defining learning strategies to better tune the models towards these biases. These two issues have been throughly addressed within classification problems. However, they have been essentially ignored in research on regression. The goal of this paper is to address the first of these issues: the selection of proper evaluation metrics. The main contributions of the paper are: i) increasing the awareness of the research community for these important tasks and in general to cost-sensitive regression; ii) exposing the risks of using standard regression evaluation metrics on cost sensitive applications; iii) proposing a new evaluation framework for the prediction of rare extreme values of a continuous variable.

2 Problem Statement

In predictive data mining the goal is to learn a model of an unknown function that maps a set of predictor variables into a target variable. This model is to be obtained using a training set containing examples of this mapping. The training data is used to obtain the model parameters that minimise some preference criterion. The preference criteria that are commonly used in regression are the mean squared error, $MSE = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2$, and the mean absolute deviation, $MAD = \frac{1}{n}\sum_{i=1}^{n}|y_i - \hat{y}_i|$. These are average estimators of the true mean squared and absolute error of the model, respectively.

In this paper we are interested in a particular sub-class of regression problems. The main particularity of this sub-class of problems lies on their focus on the predictive performance at rare extreme values of the continuous target variable, i.e. extreme low and/or high values. Performance on the other more frequent values is basically irrelevant for the end user of these applications.

We claim that standard error measures, such as MSE and MAD, are not suitable for these tasks. They take all the prediction errors equally across the domain of the target variable, assuming that the magnitude of the error is the decisive factor for the "cost" of a prediction. We argue that while this magnitude is important it should be weighed by the "relevance" of the values involved in the prediction.

Let us illustrate our claim by a small example. In Table 1 we present the predictions of two artificial models (M_1 and M_2) for a set of 10 hypothetical returns of some financial asset given in percentage daily variation. For this prediction problem it is very clear that we want to be particularly accurate at predicting the large variations (positive or negative) as these are the ones on which we can

Table 1. The predictions of two artificial models.

True	−5.29	−2.65	−2.43	−0.20	−0.03	0.03	0.51	1.46	2.53	2.94
M_1	−4.40	−2.06	−2.20	0.10	−0.23	−0.27	0.97	2.00	1.86	2.15
M_2	−5.09	−2.95	−2.89	0.69	−0.82	0.70	−0.08	0.92	2.83	3.17

earn some money if they are correct. Smaller variations, even if correctly predicted are most of the times not tradable given the transaction costs. From the observation of this table, we can say that M_1 has more accurate predictions at smaller returns (in absolute terms), while M_2 achieves more accurate predictions at the larger variations. However, if we calculate the values of both MAD and MSE of these two models we observe that they are exactly the same, 0.497 and 0.29893, respectively, meaning that these two metrics tag these two models as having the same performance. The reason for this is that both models obtain the same total error magnitude value and thus both have same average error. This is a clearly misleading "conclusion" for this type of applications, as model M_2 is obviously more useful. This small example provides a simple illustration of the problem of assuming that the error amplitudes cost the same across all the domain of the target variable (as it is the case of all standard error metrics). For our target applications this is clearly not the case and, therefore, it is necessary to have an error metric that is sensitive to where the errors occur within the range of the target variable, i.e. that copes with differentiated relevance across the domain of this variable.

Another further problem with standard error metrics, not illustrated in the above example, is the fact that even though some model may have a clear advantage on extreme values, given their rarity, this advantage may well be diluted by its poorer performance on the "irrelevant" (but very frequent) normal values.

3 Existing Approaches to the Problem

3.1 Case Weights

Within the regression learning setup described in Section 2, there are a few alternatives to the standard error measures that could be considered more adequated to our applications. One such alternative is to use case weights. Some learning algorithms allow the user to attach a weight to each observation of the training sample. Model parameters can then be obtained by minimizing a criterion that takes into account these weights. Training cases with a target variable value that is more "relevant" should have higher weights. In the case of rare extreme values prediction this would mean to give more weight to the extreme values.

Assuming we can easily obtain the values of these weights this would apparently lead to a proper evaluation of the models' performance. However, the main drawback of this approach is that it only sees one side of the problem, the true values. In effect, this method does not try to avoid (or penalize) the situation where a "relevant" value is predicted by the model, but the true value is "normal". This is a kind of false alarm and would correspond, for instance, to predict

a high return for some stock that then turns out to have a really irrelevant (very small) return. This drawback stems from the fact that the weights are dependent solely on the true value of the cases, y_i, instead of being dependent on both y_i and $\hat{y}_i$. Because of this, the above example would have a low penalization (as the true value is irrelevant), which is contradictory to the application objectives where we clearly want to avoid these costly mistakes.

3.2 Special-purpose Loss Functions

Some authors (e.g. [4]) have addressed the issue of differentiated prediction costs by the use of so-called asymmetric loss functions. Their main goal was to be able to distinguish two types of errors, and assign costs accordingly, namely, the cost of under-predictions ($\hat{y} < y$) and the cost of over-predictions ($\hat{y} > y$). That is the case of the $LINLIN$ loss function, presented in Equation 1.

$$LINLIN = \begin{cases} c_o|y - \hat{y}|, \text{ if } \hat{y} > y; \\ 0, \qquad\quad \text{if } \hat{y} = y; \\ c_u|y - \hat{y}|, \text{ if } \hat{y} < y. \end{cases} \tag{1}$$

where c_o and c_u are constants for penalizing over- and under-predictions.

In spite of its use for some type of applications, the $LINLIN$ loss function is far from being a general cost-sensitive approach for any regression task as it only distinguishes between two types of differentiated costs: under- and over-predictions. Moreover, even on these situations it considers all under-(over-) predictions as equally serious, only looking at the error amplitude as "standard" error metrics. For instance, in stock market forecasting, predicting a future price change of -1% for a true value of 1%, has the same error amplitude as predicting 6% for a true value of 8%, and both are under-predictions. Nonetheless, they may lead to very different trading actions, and thus different costs/benefits.

4 Precision and Recall for Regression

Our target applications are driven by rare events - the occurrence of rare extreme values of a continuous variable. Within research on classification, this type of event-driven prediction tasks are usually evaluated using the notions of *precision* and *recall*, which are preferred over other alternatives when in presence of large skew in the class distribution [5]. The main advantage of these statistics is that they are focused on the performance of the models on the events, completely ignoring their accurate predictions for the non-event classes. Informally, precision measures the proportion of events signalled by the model that are real events. Recall measures the proportion of events occurring in the domain that are "captured" by the models. There is usually a trade-off between these two statistics (always outputting an event signal will get you 100% recall but with a very poor precision as most signals will be wrong), and often the two are put together in a single weighted score like for instance the F-measure [11]. Conceptually, our proposal in this paper is to provide the equivalents of these two

statistics for regression problems in order to properly evaluate the performance of the models on the values that really matter.

4.1 Our Proposal

The standard setup for event-driven classification is to have a so-called "positive" class that represents the target events while the "negative" class represents all non-events. Confusion matrices provide a good characterization of the performance of a model. The numbers in this matrix can be used to calculate several statistics among which are precision and recall [8]. Table 2 shows a general confusion matrix for these type of applications. Recall is defined as the ratio TP/POS, while precision as the ratio $TP/PPOS$.

Table 2. The 2-classes confusion matrix.

	Predicted Pos	Predicted Neg	
Pos	TP	FN	POS
Neg	FP	TN	NEG
	PPOS	PNEG	

In these classification problems, relevance (importance) is established by declaring the "target" class. This enumeration strategy is not possible in regression given the infinite domain of the target variable. We propose the use of a relevance function, $\phi()$, that maps the original domain of the target variable into a continuous scale of relevance[3],

$$\phi(Y): \quad]-\infty,\infty[\rightarrow [0,1] \tag{2}$$

This function allows the specification of different degrees of relevance with the obvious advantages in terms of sensibility of the method with respect to the different values of the target variable.

We can also describe the strategy followed in classification using this notion of relevance. In effect, from this perspective it corresponds to specifying the following relevance function,

$$\phi(Y) = I(Y = C_E) \tag{3}$$

where $I()$ is the indicator function given 1 if its argument is true and 0 otherwise, and C_E is the label of the class describing the events (i.e. the positive class).

The information on the relevance function is obviously domain-dependent. In classification this information consists of choosing the positive class. In regression, given the infinite nature of the domain of the target variable, a real

[3] We use the value of zero for completely irrelevant values, and one for maximally relevant values.

valued function makes more sense. Specifying such function in an analytical way may not be always easy for a user. Still, for some applications we can come up with a reasonable automatically generated relevance function. That is the case of our target applications. In effect, in these domains relevance is associated with rarity and extremeness of the values. In this context we may say that the relevance function is the complement of the probability distribution function (*pdf*) of the target variable. Box plots provide key information on this *pdf* in particular regards extreme values. In effect, they are at the basis of a parametric test for outliers, the box-plot rule. This test assumes a Gaussian distribution of the variable and tags as outliers all values above the high adjacent value given by $adj_H = Q_3 + 1.5 \cdot IQR$, where Q_3 is the third quartile and $IQR = Q_3 - Q_1$. Equivalently, all values below the low adjacent value, $adj_L = Q_1 - 1.5 \cdot IQR$, are also tagged as outliers. These values correspond to rare high (low) extreme values. For our target applications we may have both types of outliers or only high (low) outliers. Our proposal consists of using a sigmoid-like relevance function whose shape is a function of these adjacent values for each of these two "sides" of extremeness. Let us see how we can derive this function from the training sample we have available for each application.

The relevance function is based in the following sigmoid,

$$f(Y) = \frac{1}{1 + \exp^{-s \cdot (Y - c)}} \tag{4}$$

where c is the center of the sigmoid and s is the shape of the sigmoid. The values of these parameters are also dependent on the type of extremes the variable has (low, high or both types of extremes). For applications with only low or high extremes the relevance function is defined by a single sigmoid, while for applications with both types of extremes (like stock market prediction tasks) we will have two of these sigmoids defining $\phi(Y)$.

The parameter c, the center of the sigmoid, represents the value where $\phi(Y) = 0.5$. The meaning of c is that of a threshold above which the values of target variable start to be more relevant. We set the c values of the sigmoids to the values of the respective adjacent values, i.e. $c_L = adj_L$ and $c_H = adj_H$.

With respect to the parameter s we want to set it in such a way that for the high extreme values $\phi(c - c \cdot k) \simeq 0$, and for low extremes $\phi(c + c \cdot k) \simeq 0$, where k is a kind of decay factor that determines how fast the sigmoid decays to 0. By selecting a certain precision value Δ (e.g. $1e - 04$) and solving the equation in order to s we get,

$$s = \pm \frac{\ln(\Delta^{-1} - 1)}{|c \cdot k|} \tag{5}$$

where the $+$ signal is used for high extremes, while the $-$ signal is for low extremes.

In the case of applications with both extremes, each sigmoid is obtained using the parameter values described above. Figure 1 shows two relevance functions

generated using this method for two types of applications: only with high extremes ; and with both types of extremes. We provide R code[4] that implements this type of relevance functions that are adequate for applications of predicting rare extreme values.

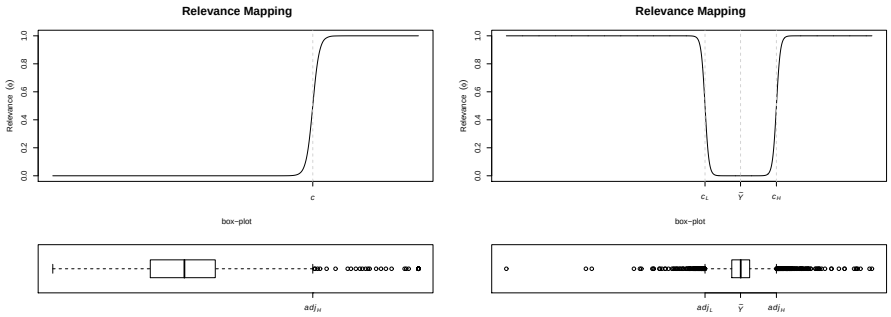

Fig. 1. Two examples of relevance functions generated from box plots.

Please note that our proposal in no way depends on this illustrative and heuristic definition of a relevance function. This definition is only of use in cases where the user does not have a precise notion of relevance/importance of his application, simply having the "intuition" that relevance is associated with extreme and rare values. In these cases, this heuristic function we have described may help in defining a relevance function that is required for the application of our evaluation framework that we now describe.

Recall is informally defined as the proportion of relevant events that are retrieved by a model. Having defined events as a function of relevance we can say that relevant events are those for which $\phi(Y) \geq t_E$, where t_E is a domain-dependent threshold on relevance. In classification, as relevance is usually a 0/1 function (c.f. Equation 3), this threshold is 1. For regression, this will most probably be a value near 1, depending on the values we want to consider as the targets of our prediction task.

We now need to clarify the notion of "events that are retrieved by a model" in order to fully define recall. In classification this consists of achieving a correct prediction that is asserted by the usual 0/1 loss function, i.e. having $L_{0/1}(\hat{y}_i, y_i) = 0 \Leftrightarrow \hat{y}_i = y_i$. In regression loss functions are usually metric with domain $[0, \infty[$. Imposing that $\hat{y}_i = y_i$ will be, in most cases, too strict. Generally, we can say that a prediction is "correct" in regression if $L(\hat{y}_i, y_i) \leq t_L$, where t_L is a threshold on the range of the loss function. We may generalize even further this notion by allowing different degrees of "accuracy" within the interval of "admissible" errors, i.e. errors that are less than t_L. The value of t_L is again domain-dependent.

[4] Available in http://www.liaad.up.pt/∼ltorgo/DS09 .

Having defined the two general concepts involved in the notion of recall we can now propose a general definition for this statistic that can cope with both classification and regression tasks,

$$Recall = \frac{\sum_{\phi(y_i) \geq t_E} \alpha(\hat{y}_i, y_i) \cdot \phi(y_i)}{\sum_{\phi(y_i) \geq t_E} \phi(y_i)} \tag{6}$$

where $\alpha()$ is a function that defines the accuracy of a prediction.

In classification the $\alpha()$ function is defined as follows,

$$\alpha(\hat{y}_i, y_i) = I(L_{0/1}(\hat{y}_i, y_i) = 0) \tag{7}$$

where $L_{0/1}()$ is a standard $0/1$ loss function.

Given the definition of relevance for classification problems we have described above, and this definition of what is an accurate prediction in classification, it is easy to see that our proposed definition of Recall reduces to the standard proportion TP/POS.

For regression we may define $\alpha()$ using a similar indicator function,

$$\alpha(\hat{y}_i, y_i) = I(L(\hat{y}_i, y_i) \leq t_L) \tag{8}$$

where t_L is the above mentioned threshold defining an admissible error within the domain a metric loss function $L()$ (e.g. the absolute deviation).

Alternatively, we may use a smoother notion of accuracy by using a continuous function in the interval $[0, 1]$, instead of the $0/1$ function of Equation 8. This allows a more accurate assessment of the quality of the signals of a regression model. There are many ways of mapping the loss function values in the interval $[0, t_L]$ into a $[1, 0]$ scale. Examples include variations of linear interpolation or the ramping function. Another alternative is to use a variant of the complementary error function [1], that has a Gaussian-type shape that we think is more adequate for our goals,

$$\alpha(\hat{y}_i, y_i) = I(L(\hat{y}_i, y_i) \leq t_L) \cdot \left(1 - \exp^{-k \cdot \frac{(L(\hat{y}_i, y_i) - t_L)^2}{t_L^2}}\right) \tag{9}$$

where k is a positive integer that determines the shape of the function. Larger values lead to steeper decreases.

Precision is the proportion of the events retrieved by a model that are effective events. We have already seen what is an event in both classification and regression. The only difference here is that we are talking about "retrieved" events and not the "real" events (i.e. predictions and not true values). Some of these correspond to "real" events but others not, and the goal of precision is to assert this proportion. In classification a retrieved event is a prediction of the "positive" class. In regression this is a prediction of a value whose relevance is greater than the user-defined relevance threshold t_E. As we have seen, both can be described by the same condition using the relevance function. In this context, we propose the following generalized definition of precision,

$$Precision = \frac{\sum_{\phi(\hat{y}_i) \geq t_E} \alpha(\hat{y}_i, y_i) \cdot \phi(\hat{y}_i)}{\sum_{\phi(\hat{y}_i) \geq t_E} \phi(\hat{y}_i)} \tag{10}$$

You may have noticed that the numerators of definitions of Precision and Recall we are proposing are different (c.f. Equations 6 and 10), which is not in aggrement with the standard definitions of recall and precision that have in the numerator the number of true positives (TP). However, for the settings used in classification the numerators of these equations we propose are in effect equal. The $\alpha()$ function used for classification is a 0/1 function that is 1 if the classification is accurate, which implies that $\hat{y}_i = y_i$. This in turn implies that $\phi(\hat{y}_i) = \phi(y_i)$ and thus the numerators are equal. However, we should remark that this may not be the case for regression setups where an accurate prediction may not mean that $\hat{y}_i = y_i$, namely if $t_L > 0$.

Precision and recall may be aggregated into composite measures, like for instance the F-measure [11],

$$F = \frac{\left(\beta^2 + 1\right) \cdot Precision \cdot Recall}{\beta^2 \cdot Precision + Recall} \tag{11}$$

where $0 \leq \beta \leq 1$, controls the relative importance of recall to precision.

These composite measures have the advantage of facilitating comparisons among models as they provide a single score.

5 Experimental Analysis

5.1 Artificial Data

On Table 1 we have presented an artificial example on stock returns prediction with the predictions of two models that, in spite of their clearly different approach to rare extreme values, had exactly the same score in terms of standard error metrics like MSE and MAD. Let us examine this example with our new proposed measures of recall and precision. Let us suppose that we use as threshold for events (t_E) a value of relevance greater than 0.75 ,i.e. $\phi(Y) \geq 0.75$. We will use an automatically generated relevance function for extremes (c.f. Equation 4). The generated function uses a larger sample of values than those shown on Table 1. Using this sample we estimate $adj_L = -1.5$ and $adj_H = 1.5$. These values setup the value of the c parameter of the function and together with a value of $k = 0.5$ we define our relevance function (c.f. Equation 4). We will also use the smooth $\alpha()$ function defined in Equation 9 with a threshold for accurate predictions of half percent return, i.e. $t_L = 0.5$. In this context, we come up with the results show in Table 3.

These values correspond to a recall of 0.178 for model M_1 and of 0.670 for M_2. Precision is of 0.292 for model M_1 and of 0.668 for M_2. These scores provide a completely different (and more correct with respect to the preference bias of this application) perspective on the performance of the models, which according to both MSE and MAD are equal.

Table 3. Evaluating the two artificial models with the new metrics.

True	−5.29	−2.65	−2.43	−0.20	−0.03	0.03	0.51	1.46	2.53	2.94
$\phi(Y)$	1.00	1.00	0.98	0.00	0.00	0.00	0.00	0.01	0.99	1.00
M_1	−4.40	−2.06	−2.20	0.10	−0.23	−0.27	0.97	2.00	1.86	2.15
$\phi(\hat{Y}_1)$	1.00	0.63	0.86	0.00	0.00	0.00	0.00	0.50	0.22	0.80
$L(\hat{Y}_1, Y)$	0.89	0.59	0.23	0.30	0.20	0.30	0.46	0.54	0.67	0.79
$\alpha(\hat{Y}_1, Y)$	0.00	0.00	0.90	0.72	0.94	0.72	0.05	0.00	0.00	0.00
M_2	−5.09	−2.95	−2.89	0.69	−0.82	0.70	−0.08	0.92	2.83	3.17
$\phi(\hat{Y}_2)$	1.00	1.00	1.00	0.00	0.00	0.00	0.00	0.00	1.00	1.00
$L(\hat{Y}_2, Y)$	0.20	0.30	0.46	0.89	0.79	0.67	0.59	0.54	0.30	0.23
$\alpha(\hat{Y}_2, Y)$	0.94	0.72	0.05	0.00	0.00	0.00	0.00	0.00	0.72	0.90

5.2 Predicting Stock Market Returns

In this section we illustrate the use of the proposed precision and recall statistics in the context of the prediction of rare extreme returns of a set of stocks. The purpose of this study is to illustrate both the "danger" of using standard regression evaluation statistics in this type of problems, as well as presenting and measuring the advantages of our proposals.

The Data The base data we will use in our study are the standard daily quotes of four companies: International Business Machines (IBM), Coca-Cola (KO), Boeing (BA) and General Motors (GM). This daily data was obtained from Yahoo finance[5] and it contains the usual quotes and volume information.

Most applications of this type based on daily data focus on predicting the Adjusted Close prices of the stocks. Namely, a common procedure consists predicting the h-days returns defined as,

$$R_h(t) = \frac{Close(t) - Close(t-h)}{Close(t-h)} \tag{12}$$

Using this time series of returns we have defined a prediction task consisting of trying to predicted the future value of these returns, $R_h(t+h)$, using a set of p previous values of the time series (usually known as an embed of the time series). In our experiments we used an embed of 24 days back of the $R_h(t)$ variable. This modelling task was selected without any particular concern on whether this was the best setup for predicting future returns. That is not our main goal here. Our objective is to compare alternative modelling techniques on the same stock market prediction problems and check the model rankings we obtain when using both the standard evaluation metrics and our new proposals. Our hypothesis is that the model rankings obtained with our metrics are "better" from the perspective of the application objectives, which are being accurate at the rare extreme returns where profitable trading can take place.

[5] http://finance.yahoo.com.

Using this approach we have obtained datasets for the 1-, 3- and 5-days returns of the four companies used in our study, i.e. 12 regression tasks.

The Experimental Methodology The used quotes data covers the period from 1970-01-02 till 2008-07-11, in a total of 9725 daily sessions.

In order to provide an accurate estimate of the statistics that we will use to compare our alternative models we have divided the period mentioned above in two main consecutive time windows. The first spans from the first date till 1990-01-01. The second time window goes from this latter date till 2008-07-11. The first time window (first 20 years) will be used for obtaining the prediction models, while the second window (around 18 and a half years) will be used to evaluate and compare the models.

The Modelling Tools All tools we have used are available in the (free) R statistical environment[6], which allows easy replication of our results. We have considered 4 different regression techniques, each with several parameter variants, in a total of 57 different models being compared for each data set.

Artificial Neural Networks We have used the neural networks provided by the `nnet` package of R. This package has a function to obtain feed-forward neural networks with one hidden layer using the back-propagation learning algorithm.

Regarding model tuning we have considered 15 alternatives varying the number of inner nodes (parameter `Size`) of the hidden layer between 5, 10, 15, 20 and 30, and also the learning rate (parameter `Decay`) between 0.01, 0.05 and 0.1.

Multivariate adaptive regression splines The package `mda` of R has a re-implementation of MARS [9] done by Trevor Hastie and Robert Tibshirani. We have used this system in our experiments.

Regarding model tuning we have considered 16 variants formed by different combinations of the parameter setting the penalty for extra degrees of freedom (parameter `Pen` which was used with values 1,2,3,4), and of the parameter specifying the forward stepwise stopping threshold (parameter `Thr` that was tried with values 0.01, 0.005, 0.001 and 0.0005).

Support Vector Machines Package `e1071` of R includes a function implementing SVMs [10]. This implementation provides an interface to the award-winning `libsvm` library by Chang and Lin [3].

We have considered 16 variants of SVMs during our model tuning experiments. These variants were chosen according to the suggestions given in [10]. They include different values for the parameter `Cost` (tried values 400, 500, 600 and 700) and `Gamma` (tried values 0.01, 0.005, 0.001 and 0.0005). The former is a constraints violation parameter, while the latter is the radial basis function kernel parameter.

[6] http://www.R-project.org.

Random Forests Package `randomForest` of R includes a function that implements random forests [2] based on original Fortran code by L. Breiman and A. Cutler.

We have considered 10 variants of these models by setting the parameter `ntree`, which controls the number of trees in the ensembles, to values from 50 to 500 in steps of 50.

The Results We have obtained the 57 model variants using the experimental methodology described before on the returns data sets. The main hypothesis that we are trying to check is that the model rankings obtained by using our proposed metrics are significantly different from the rankings obtained with standard regression statistics. Moreover, that these rankings obtained with our metrics are clearly advantageous in terms of the application preference bias, that in this case is related to having good "signals" of rare and extreme movements of the markets.

In terms of our evaluation framework we have used the following settings. We have assumed that, giving the transaction costs, users of these applications are not willing to trade on returns smaller than 2% (-2%) for buying (selling) actions. In this context, we have setup the notion of rare extreme values around these two thresholds. Namely, with respect to the relevance function we have used as centers of the two sigmoids the values $c_L = -0.02$ and $c_H = 0.02$, while for the shapes of the sigmoids we have calculated them using Equation 5 with $k = 0.5$ and $\Delta = 1e - 04$. In the context of precision and recall we have used $t_E = 0.5$ (thus any return above 2% or below -2% will be considered an event, given the definition of the $\phi()$ function), and $t_L = 0.005$ (i.e. errors above 0.5% are not considered, c.f. Equation 9).

The first results we show are designed to test the hypothesis concerning the different rankings. We have used the MAD statistic as a representative of the "standard" approaches, and the composite F-measure (with $\beta = 0.5$ that gives twice importance to precision compared to recall, as inaccurate trading signals may be costly) as representing our proposals. For all 12 experimental setups (4 companies and 3 forecasting scenarios), we have obtained the two model rankings according to these two statistics. Due to lack of space we can not present all graphs illustrating these 12 experimental setups[7]. All setups follow a similar results trend. We have selected 1 setup that is shown in Figure 2. The figure has two graphs. The graph on the left shows the scores of the best five models according to the two statistics. We should remark that for MAD, lower values are better, contrary to what happens with the F measure. On the X-axis we have the identifiers (a number from 1 to 57) of the top 5 models according to each statistic (the 5 on the left according to MAD and the other 5 according to F). Ideally these two sets of numbers should be different indicating that the best 5 models according to the two statistics are also different. On the graph we plot the actual values of these 10 models for the two statistics: circles and left Y-scale for MAD; and triangles and right Y-scale for the F measure. The second graph

[7] All graphs may be obtained at http://www.liaad.up.pt/∼ltorgo/DS09 .

presented on the figures shows a global perspective (on all 57 models) of the two rankings produced by the statistics. On both axis we have the possible ranking positions (from 1 to 57). The coordinates of each of the 57 dots shown on the graphs are obtained using the rank position assigned by MAD (X coordinate), and the corresponding rank position assigned by the F measure (Y coordinate). If for any of the 57 models both statistics give it the same ranking position, the respective dot should lie in the dashed diagonal line. The vertical and horizontal dashed lines highlight the results for the top 10 rank positions (left of the vertical line, and below the horizontal line) according to the two statistics.

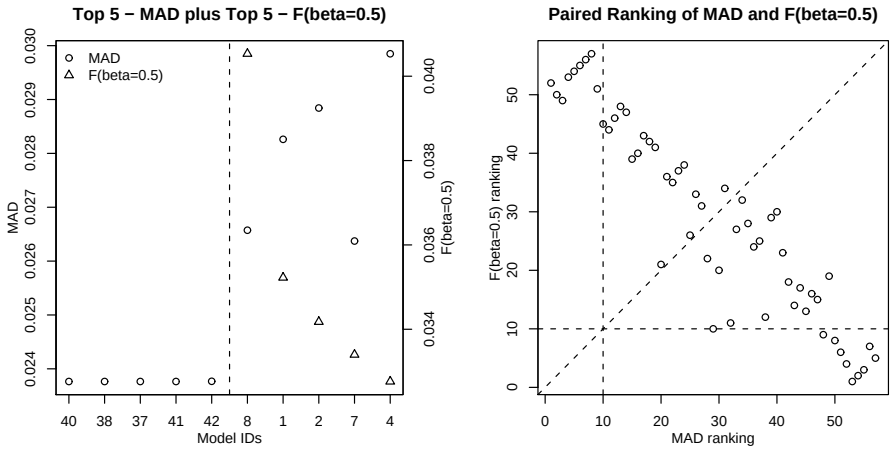

Fig. 2. The results for 3-days returns of Boeing (BA).

Analysing the results in Figure 2, namely the left graph, we observe that the top 5 models according to the two statistics are completely different. Moreover, we see that their concrete scores on the statistics are also very different. For instance the best models according to MAD achieve a much worse score in terms of F[8], when compared to the best 5 models according to this later measure. In terms of overall ranking we also observe a general tendency for all ranking positions to be different as most points in the right graph are far from the diagonal. In particular the top 10 models according to MAD are all below the 45th position in the F ranking. These results clearly indicate that the two metrics are evaluating very different aspects of the performance of the models.

We have also carried out a formal statistical test of the differences between the model rankings. On all 12 data sets we have observed some evidence of disagreement between the rankings, with only 4 lacking proper statistical significance. In summary, our experiments have confirmed the hypothesis that the

[8] Actually, no score at all because they do not produce any event signal, i.e. predictions with relevance higher than t_E, and thus they have no precision score.

two considered metrics (MAD and F-measure based on our proposed Recall and Precision statistics), often obtain significantly different model rankings on this type of applications. Moreover, we should remark that this experimental setup is not particularly favorable to our proposals. In effect, we are comparing 57 models that optimize some variant of the squared error. This means that these models are not particularly focused on predicting rare extreme values. Even on these conditions we have observed that our proposals are able to detect models that have some ability at predicting rare extreme values. We can expect that the differences would be even more marked if among the 57 models we had some that were particularly competent at predicting rare extremes (e.g. if they were optimizing our F measure instead of squared errors).

What are the advantages of comparing a set of alternative models using our proposed metrics in alternative to standard statistics? Or in other words, what are the costs a user can expect if he uses a measure like MAD to select the model to apply for trading on stock markets? The results we have shown previously provided evidence that our metrics rank the 57 models we have considered, differently. However, do these ranking differences lead to better trading performance? In other words if a user uses the best model according to our F-measure, instead of the best model according to MAD, what does he have to gain or loose? For each of the 12 experimental setups we have selected the two best models according to MAD and F, respectively. We have used their respective predictions of the future returns for the 18 years and have calculated a set of trading-related statistics. We have assumed that we are going to trade with futures (thus allowing both short and long positions, i.e. trading when we predict the market goes down or up, respectively). Moreover, considering trading costs we only "trade" when a model predicts a future return above (below) 2% (-2%), i.e. we are going to take these situations as indicators for buying (selling). Under these conditions each model outputs a set of trading signals (predictions above 0.02 or below -0.02). The predicted signals were then compared to the "true" signals, i.e. did the prices go up (down) as predicted?

These experiments have confirmed the advantages of our metrics. In effect, the models "selected" by MAD almost never issue a single signal during the 18 testing years! On the contrary, the models selected using our metrics issue several trading signals during this period. Still, the accuracy of these signals is far from ideal as expected. This is expectable because: i) the candidate models are optimizing squared errors; ii) the information used to obtain the models (embed of 24 days) is clearly sub-optimal; and iii) predicting stock returns is a very difficult task!

6 Conclusions

This paper has presented a study on the prediction of rare extreme values of a continuous target variable that can be regarded as outliers. Our study is focused on the development of proper evaluation metrics for these tasks, which is a key step in addressing these problems.

We have described a generalization of the notions of precision and recall for regression tasks. These intuitive concepts are ideal for addressing our target problems as they focus the evaluation solely on the important events (the rare extreme values). Our proposals incorporate the standard definitions used in classification as particular cases.

We have illustrated the use of these metrics in the context of stock market forecasting applications. Namely, we have used our metrics to compare a large set of models in several experimental setups. Our experiments have confirmed that our evaluation metrics provide a significantly different perspective of the performance of the models, when compared to standard evaluation statistics. Moreover, this perspective is more adjusted to the preference biases of this type of applications. Our experimental results have also shown the danger of using standard evaluation metrics in this class of problems.

Acknowledgements

This work was partially supported by FCT projects oRANKI (PTDC/EIA/68322/2006) and MORWAQ (PTDC/EIA/68489/2006) , by a sabbatical scholarship of the Portuguese government (FCT/BSAB/388/2003) to L. Torgo and by a PhD scholarship of the Portuguese government (SFRH/BD/1711/2004) to R. Ribeiro.

References

1. M. Abramowitz and I. A. Stegun. *Handbook of Mathematical Functions*. New York: Dover, 1972.
2. L. Breiman. Random forests. *Machine Learning*, 1(45):5–32, 2001.
3. C. Chang and C. Lin. Libsvm: a library for support vector machines. Software available at http://www.csie.ntu.edu.tw/ cjlin/libsvm, detailed documentation at http://www.csie.ntu.edu.tw/ cjlin/papers/libsvm.ps.gz, 2001.
4. P. Christoffersen and F. Diebold. Further results on forecasting and model selection under asymmetric loss. *Journal of Applied Econometrics*, 11:561–571, 1996.
5. J. Davis and M. Goadrich. The relationship between precision-recall and roc curves. In *Proceedings of 23rd International Conference on Machine Learning*, 2006.
6. P. Domingos. Metacost: A general method for making classifiers cost-sensitive. In *Proceedings of the 5th International Conference on Knowledge Discovery and Data Mining (KDD-99)*, pages 155–164. ACM Press, 1999.
7. C. Elkan. The foundations of cost-sensitive learning. In *Proc. of 7th International Joint Conference of Artificial Intelligence (IJCAI'01)*, pages 973–978, 2001.
8. P. Flach. The geometry of roc space: understanding machine learning metrics through roc isometrics. In *Proceedings of the 20th International Conference on Machine Learning*, 2003.
9. J. Friedman. Multivariate adaptive regression splines. *The Annals of Statistics*, 19(1):1–141, 1991.
10. David Meyer. *Support Vector Machines, the interface to libsvm in package e1071*. Technische Universitat Wien, Austria, 2002.
11. C. Van Rijsbergen. *Information Retrieval*. Dept. of Computer Science, University of Glasgow, 2nd edition, 1979.